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 $LnW = b_0 + b_1Age + b_2Female*Age$

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.

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

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).

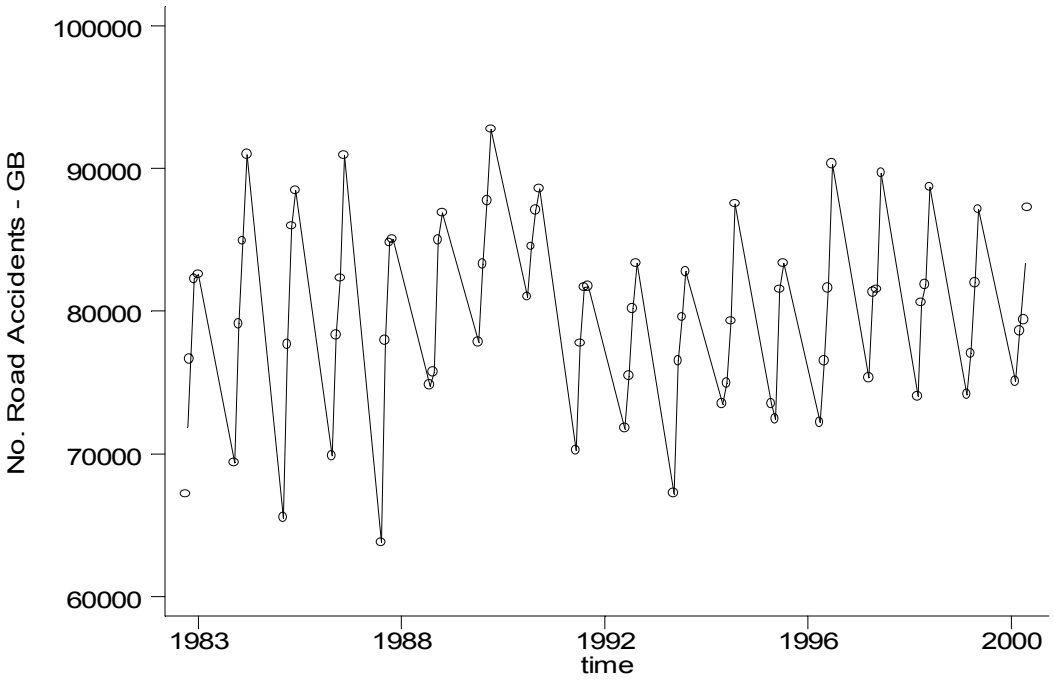
To obtain standard errors, pool the data and estimate the following

```
. 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
```

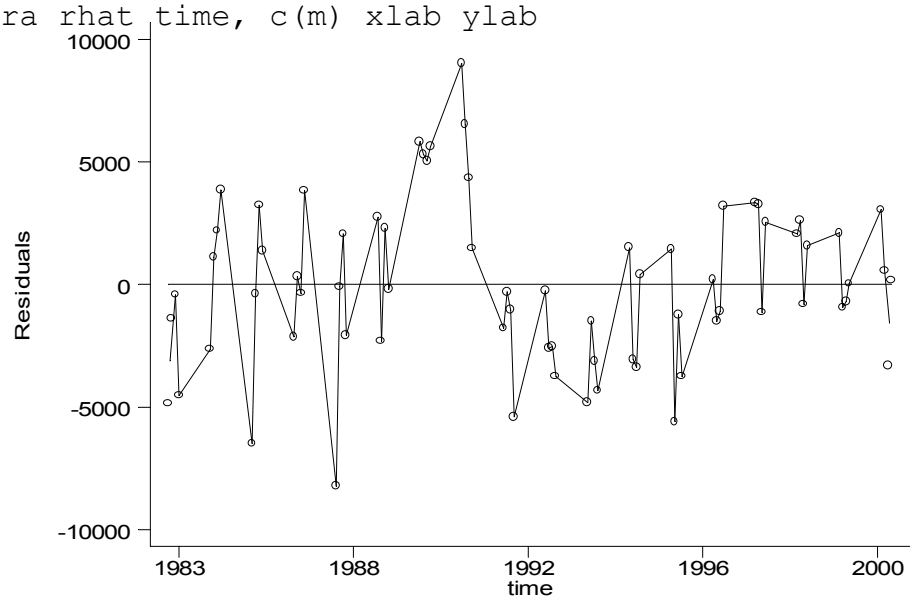
```
. reg acc q1 q2 q3
      Source |       SS       df       MS              Number of obs =      72
-------------+------------------------------           F(  3,    68) =   65.77
       Model |  2.2572e+09     3   752388623           Prob > F      =  0.0000
    Residual |   777899883    68  11439704.2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

This page shows an example of logistic regression regression analysis with footnotes explaining the output. These data were collected on 200 high schools students and are scores on various tests, including science, math, reading and social studies (**socst**). The variable **female** is a dichotomous variable coded 1 if the student was female and 0 if male.

Because we do not have a suitable dichotomous variable to use as our dependent variable, we will create one (which we will call **honcomp**, for honors composition) based on the continuous variable **write**. We do not advocate making dichotomous variables out of continuous variables; rather, we do this here only for purposes of this illustration.

```
use https://stats.idre.ucla.edu/stat/data/hsb2, clear

generate honcomp = (write >=60)
logit honcomp female read science
```

```
Iteration 0:   log likelihood = -115.64441
Iteration 1:   log likelihood = -84.558481
Iteration 2:   log likelihood = -80.491449
Iteration 3:   log likelihood = -80.123052
```

```
Iteration 4:   log likelihood = -80.118181
Iteration 5:   log likelihood =  -80.11818

Logit estimates                                   Number of obs   =        200
                                                  LR chi2(3)      =      71.05
                                                  Prob > chi2     =     0.0000
Log likelihood =  -80.11818                       Pseudo R2       =     0.3072

------------------------------------------------------------------------------
     honcomp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   1.482498   .4473993     3.31   0.001     .6056111    2.359384
        read |   .1035361   .0257662     4.02   0.000     .0530354    .1540369
     science |   .0947902   .0304537     3.11   0.002      .035102    .1544784
       _cons |   -12.7772    1.97586    -6.47   0.000    -16.64982   -8.904589
------------------------------------------------------------------------------
```

Iteration Log

```
Iteration 0:   log likelihood = -115.64441
Iteration 1:   log likelihood = -84.558481
Iteration 2:   log likelihood = -80.491449
Iteration 3:   log likelihood = -80.123052
Iteration 4:   log likelihood = -80.118181
Iteration 5:[a]  log likelihood = -80.11818
```

a. This is a listing of the log likelihoods at each iteration. (Remember that logistic regression uses maximum likelihood, which is an iterative procedure.) The first iteration (called iteration 0) is the log likelihood of the “null” or “empty” model; that is, a model with no predictors. At the next iteration, the predictor(s) are included in the model. At each iteration, the log likelihood increases because the goal is to maximize the log likelihood. When the difference between successive iterations is very small, the model is said to have “converged”, the iterating is stopped and the results are displayed. For more

information on this process, see Regression Models for Categorical and Limited Dependent Variables, Third Edition (https://www.stata.com/bookstore/regression-models-categorical-dependent-variables/) by J. Scott Long and Jeremy Freese.

Model Summary

```
Logit estimates                                   Number of obs[c]   =        200
                                                  LR chi2(3)[d]      =      71.05
                                                  Prob > chi2[e]     =     0.0000
Log likelihood =  -80.11818[b]                    Pseudo R2[f]       =     0.3072
```

b. **Log likelihood –** This is the log likelihood of the final model. The value -80.11818 has no meaning in and of itself; rather, this number can be used to help compare nested models.

c. **Number of obs –** This is the number of observations that were used in the analysis. This number may be smaller than the total number of observations in your data set if you have missing values for any of the variables used in the logistic regression. Stata uses a listwise deletion by default, which means that if there is a missing value for any variable in the logistic regression, the entire case will be excluded from the analysis.

d. **LR chi2(3) –** This is the likelihood ratio (LR) chi-square test. The likelihood chi-square test statistic can be calculated by hand as 2*(115.64441 – 80.11818) = 71.05. This is minus two (i.e., -2) times the difference between the starting and ending log likelihood. The number in the parenthesis indicates the number of degrees of freedom. In this model, there are three predictors, so there are three degrees of freedom.

e. **Prob > chi2 –** This is the probability of obtaining the chi-square statistic given that the null hypothesis is true. In other words, this is the probability of obtaining this chi-square statistic (71.05) if there is in fact no effect of the independent variables, taken together, on the dependent variable. This is, of course, the p-value, which is compared to a critical value, perhaps .05 or .01 to determine if the overall model is statistically significant. In this case, the model is statistically significant because the p-value is less than .000.

f. **Pseudo R2 –** This is the pseudo R-squared. Logistic regression does not have an equivalent to the R-squared that is found in OLS regression; however, many people have tried to come up with one. There are a wide variety of pseudo-R-square statistics. Because this statistic does not mean what R-square means in OLS regression (the proportion of variance explained by the predictors), we suggest interpreting this statistic with great caution.

Parameter Estimates

```
------------------------------------------------------------------------------
    honcomp[g]|     Coef.[h]  Std. Err.[i]     z[j]   P>|z|[j]    [95% Conf. Interval][k]
```

```
------------+----------------------------------------------------------------
     female |   1.482498   .4473993     3.31   0.001     .6056111    2.359384
       read |   .1035361   .0257662     4.02   0.000     .0530354    .1540369
    science |   .0947902   .0304537     3.11   0.002      .035102    .1544784
      _cons |   -12.7772    1.97586    -6.47   0.000    -16.64982   -8.904589
------------------------------------------------------------------------------
```

g. **honcomp –** This is the dependent variable in our logistic regression. The variables listed below it are the independent variables.

h. **Coef.** – These are the values for the logistic regression equation for predicting the dependent variable from the independent variable. They are in log-odds units. Similar to OLS regression, the prediction equation is

log(p/1-p) = **b0** + **b1***female** + **b2***read** + **b3***science**

where p is the probability of being in honors composition. Expressed in terms of the variables used in this example, the logistic regression equation is

log(p/1-p) = **-12.7772** + **1.482498***female** + **.1035361***read** + **0947902***science**

These estimates tell you about the relationship between the independent variables and the dependent variable, where the dependent variable is on the logit scale. These estimates tell the amount of increase in the predicted log odds of honcomp = 1 that would be predicted by a 1 unit increase in the predictor, holding all other predictors constant. Note: For the independent variables which are not significant, the coefficients are not significantly different from 0, which should be taken into account when interpreting the coefficients. (See the columns with the z-values and p-values regarding testing whether the coefficients are statistically significant). Because these coefficients are in log-odds units, they are often difficult to interpret, so they are often converted into odds ratios. You can do this by hand by exponentiating the coefficient, or by using the **or** option with **logit** command, or by using the **logistic** command.

female – The coefficient (or parameter estimate) for the variable **female** is 1.482498. This means that for a one-unit increase in **female** (in other words, going from male to female), we expect a 1.482498 increase in the log-odds of the dependent variable **honcomp**, holding all other independent variables constant. **read –** For every one-unit increase in reading score (so, for every additional point on the reading test), we expect a .1035361 increase in the log-odds of **honcomp**, holding all other independent variables constant. **science –** For every one-unit increase in science score, we expect a .0947902 increase in the log-odds of **honcomp**, holding all other independent variables constant. **constant –** This is the expected value of the log-odds of **honcomp** when all of the predictor variables equal zero. In

most cases, this is not interesting. Also, oftentimes zero is not a realistic value for a variable to take.

i. **Std. Err. –** These are the standard errors associated with the coefficients. The standard error is used for testing whether the parameter is significantly different from 0; by dividing the parameter estimate by the standard error you obtain a z-value (see the column with z-values and p-values). The standard errors can also be used to form a confidence interval for the parameter, as shown in the last two columns of this table.

j. **z** and **P>|z|** – These columns provide the z-value and 2-tailed p-value used in testing the null hypothesis that the coefficient (parameter) is 0. If you use a 2-tailed test, then you would compare each p-value to your preselected value of alpha. Coefficients having p-values less than alpha are statistically significant. For example, if you chose alpha to be 0.05, coefficients having a p-value of 0.05 or less would be statistically significant (i.e., you can reject the null hypothesis and say that the coefficient is significantly different from 0). If you use a 1-tailed test (i.e., you predict that the parameter will go in a particular direction), then you can divide the p-value by 2 before comparing it to your preselected alpha level. With a 2-tailed test and alpha of 0.05, you may reject the null hypothesis that the coefficient for **female** is equal to 0. The coefficient of 1.482498 is significantly greater than 0. The coefficient for **read** is .1035361 significantly different from 0 using alpha of 0.05 because its p-value is 0.000, which is smaller than 0.05. The coefficient for **science** is .0947902 significantly different from 0 using alpha of 0.05 because its p-value is 0.000, which is smaller than 0.05.

k. [**95% Conf. Interval**] – This shows a 95% confidence interval for the coefficient. This is very useful as it helps you understand how high and how low the actual population value of the parameter might be. The confidence intervals are related to the p-values such that the coefficient will not be statistically significant if the confidence interval includes 0.

Odds Ratios

In this next example, we will illustrate the interpretation of odds ratios. We will use the **logistic** command so that we see the odds ratios instead of the coefficients. In this example, we will simplify our model so that we have only one predictor, the binary variable **female**. Before we run the logistic regression, we will use the **tab** command to obtain a crosstab of the two variables.

```
tab female honcomp
```

```
           |        honcomp
    female |         0          1 |     Total
-----------+----------------------+----------
      male |        73         18 |        91
```

```
    female |        74          35 |       109
-----------+----------------------+----------
     Total |       147          53 |       200
```

If we divide the number of males who are in honors composition, 18, by the number of males who are not in honors composition, 73, we get the odds of being in honors composition for males, 18/73 = .24657534. If we do the same thing for females, we get 35/74 = .47297297. To get the odds ratio, which is the ratio of the two odds that we have just calculated, we get .47297297/.24657534 = 1.9181682. As we can see in the output below, this is exactly the odds ratio we obtain from the **logistic** command. The thing to remember here is that you want the group coded as 1 over the group coded as 0, so honcomp=1/honcomp=0 for both males and females, and then the odds for females/odds for males, because the females are coded as 1.

With regard to the 95% confidence interval, we do not want this to include the value of 1. When we were considering the coefficients, we did not want the confidence interval to include 0. If we exponentiate 0, we get 1 (exp(0) = 1). Hence, this is two ways of saying the same thing. As you can see, the 95% confidence interval includes 1; hence, the odds ratio is not statistically significant. Because the lower bound of the 95% confidence interval is so close to 1, the p-value is very close to .05.

There are a few other things to note about the output below. The first is that although we have only one predictor variable, the test for the odds ratio does not match with the overall test of the model. This is because the z statistic is actually the result of a Wald chi-square test, while the test of the overall model is a likelihood ratio chi-square. While these two types of chi-square tests are asymptotically equivalent, in small samples they can differ, as they do here. Also, we have the unfortunate situation in which the results of the two tests give different conclusions. This does not happen very often. In a situation like this, it is difficult to know what to conclude. One might consider the power, or one might decide if an odds ratio of this magnitude is important from a clinical or practical standpoint.

```
logistic honcomp female
```

```
Logistic regression                               Number of obs   =        200
                                                  LR chi2(1)      =       3.94
                                                  Prob > chi2     =     0.0473
Log likelihood =  -113.6769                       Pseudo R2       =     0.0170
```

This page shows an example of a multinomial logistic regression analysis with footnotes explaining the output. The data were collected on 200 high school students and are scores on various tests, including a video game and a puzzle. The outcome measure in this analysis is the preferred flavor of ice cream – vanilla, chocolate or strawberry- from which we are going to see what relationships exists with video game scores (**video**), puzzle scores (**puzzle**) and gender (**female**). Our response variable, **ice_cream**, is going to be treated as categorical under the assumption that the levels of **ice_cream** have *no* natural ordering, and we are going to allow Stata to choose the referent group. In out example, this will be vanilla. By default, Stata chooses the most frequently occurring group to be the referent group. The first half of this page interprets the coefficients in terms of multinomial log-odds (logits). These will be close to but not equal to the log-odds achieved in a logistic regression with two levels of the outcome variable. The second half interprets the coefficients in terms of relative risk ratios.

```
use https://stats.idre.ucla.edu/stat/stata/output/mlogit, clear
```

Before running the regression, obtaining a frequency of the ice cream flavors in the data can inform the selection of a reference group.

```
tab ice_cream

favorite flavor of ice cream
             | Freq. Percent Cum.
 ------------+-----------------------------------
   chocolate | 47 23.50 23.50
     vanilla | 95 47.50 71.00
  strawberry | 58 29.00 100.00
 ------------+-----------------------------------
       Total | 200 100.00
```

Vanilla is the most frequently occurring ice cream flavor and will be the reference group in this example.

```
mlogit ice_cream video puzzle female

Iteration 0:   log likelihood = -210.58254
Iteration 1:   log likelihood = -194.75041
```

```
Iteration 2:   log likelihood = -194.03782
Iteration 3:   log likelihood = -194.03485
Iteration 4:   log likelihood = -194.03485

Multinomial logistic regression                 Number of obs   =        200
                                                LR chi2(6)      =      33.10
                                                Prob > chi2     =     0.0000
Log likelihood = -194.03485                     Pseudo R2       =     0.0786

------------------------------------------------------------------------------
   ice_cream |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
chocolate    |
       video |  -.0235647   .0209747    -1.12   0.261    -.0646744     .017545
      puzzle |  -.0389243   .0195165    -1.99   0.046    -.0771759   -.0006726
      female |   .8166202   .3909813     2.09   0.037      .050311    1.582929
       _cons |   1.912256   1.127256     1.70   0.090    -.2971258    4.121638
-------------+----------------------------------------------------------------
strawberry   |
       video |    .022922   .0208718     1.10   0.272    -.0179861    .0638301
      puzzle |   .0430036   .0198894     2.16   0.031     .0040211     .081986
      female |   -.032862   .3500153    -0.09   0.925    -.7188793    .6531553
       _cons |  -4.057323   1.222939    -3.32   0.001     -6.45424   -1.660407
------------------------------------------------------------------------------
(ice_cream==vanilla is the base outcome)
```

```
Iteration Log[a]
```

```
Iteration 0:   log likelihood = -210.58254
Iteration 1:   log likelihood = -194.75041
Iteration 2:   log likelihood = -194.03782
Iteration 3:   log likelihood = -194.03485
Iteration 4:   log likelihood = -194.03485
```

a. **Iteration Log –** This is a listing of the log likelihoods at each iteration. Remember that multinomial logistic regression, like binary and ordered logistic regression, uses maximum likelihood estimation, which is an iterative procedure. The first iteration (called iteration 0) is the log likelihood of the "null" or "empty" model; that is, a model with no predictors. At the next iteration, the predictor(s) are included in

the model. At each iteration, the log likelihood increases because the goal is to maximize the log likelihood. When the difference between successive iterations is very small, the model is said to have "converged", the iterating stops, and the results are displayed. For more information on this process for binary outcomes, see Regression Models for Categorical and Limited Dependent Variables (/stata/examples/long/) by J. Scott Long (page 52-61).

Model Summary

Multinomial logistic regression	Number of obs[c]	=	200
	LR chi2(6)[d]	=	33.10
	Prob > chi2[e]	=	0.0000
Log likelihood = -194.03485[b]	Pseudo R2[f]	=	0.0786

b. **Log Likelihood –** This is the log likelihood of the fitted model. It is used in the Likelihood Ratio Chi-Square test of whether all predictors' regression coefficients in the model are simultaneously zero and in tests of nested models.

c. **Number of obs –** This is the number of observations used in the multinomial logistic regression. It may be less than the number of cases in the dataset if there are missing values for some variables in the equation. By default, Stata does a listwise deletion of incomplete cases.

d. **LR chi2(6) –** This is the Likelihood Ratio (LR) Chi-Square test that for both equations (chocolate relative to vanilla and strawberry relative to vanilla) that at least one of the predictors' regression coefficient is not equal to zero. The number in the parentheses indicates the degrees of freedom of the Chi-Square distribution used to test the LR Chi-Square statistic and is defined by the number of models estimated (2) times the number of predictors in the model (3). The LR Chi-Square statistic can be calculated by -2*(L(null model) – L(fitted model)) = -2*((-210.583) – (-194.035)) = 33.096, where L(null model) is from the log likelihood with just the response variable in the model (Iteration 0) and L(fitted model) is the log likelihood from the final iteration (assuming the model converged) with all the parameters.

e. **Prob > chi2 –** This is the probability of getting a LR test statistic as extreme as, or more so, than the observed statistic under the null hypothesis; the null hypothesis is that all of the regression coefficients across both models are simultaneously equal to zero. In other words, this is the probability of obtaining this chi-square statistic (33.10) or one more extreme if there is in fact no effect of the predictor variables. This p-value is compared to a specified alpha level, our willingness to accept a type I error, which is typically set at 0.05 or 0.01. The small p-value from the LR test, <0.00001, would lead us to conclude that at least one of the regression coefficients in the model is not equal to zero. The parameter of the chi-square distribution used to test the null hypothesis is defined by the degrees of freedom in the prior

line, **chi2(6)**.

f. **Pseudo R2** – This is McFadden's pseudo R-squared. Logistic regression does not have an equivalent to the R-squared that is found in OLS regression; however, many people have tried to come up with one. There are a wide variety of pseudo-R-square statistics. Because this statistic does not mean what R-square means in OLS regression (the proportion of variance of the response variable explained by the predictors), we suggest interpreting this statistic with great caution.

Parameter Estimates

ice_cream[g]	Coef.[h]	Std. Err.[j]	z[k]	P>|z|[k]	[95% Conf. Interval][l]
chocolate					
video	-.0235647	.0209747	-1.12	0.261	-.0646744 .017545
puzzle	-.0389243	.0195165	-1.99	0.046	-.0771759 -.0006726
female	.8166202	.3909813	2.09	0.037	.050311 1.582929
_cons	1.912256	1.127256	1.70	0.090	-.2971258 4.121638
strawberry					
video	.022922	.0208718	1.10	0.272	-.0179861 .0638301
puzzle	.0430036	.0198894	2.16	0.031	.0040211 .081986
female	-.032862	.3500153	-0.09	0.925	-.7188793 .6531553
_cons	-4.057323	1.222939	-3.32	0.001	-6.45424 -1.660407

(ice_cream==vanilla is the base outcome)[i]

g. **ice_cream** – This is the response variable in the multinomial logistic regression. Underneath **ice_cream** are two replicates of the predictor variables, representing the two models that are estimated: chocolate relative to vanilla and strawberry relative to vanilla.

h and i. **Coef.** and **referent group** – These are the estimated multinomial logistic regression coefficients and the referent level, respectively, for the model. An important feature of the multinomial logit model is that it estimates *k-1* models, where *k* is the number of levels of the outcome variable. In this instance, Stata, by default, set vanilla as the referent group, and therefore estimated a model for chocolate relative to vanilla and a model for strawberry relative to vanilla. Since the parameter estimates are relative to the referent group, the standard interpretation of the multinomial logit is that for a unit change in the predictor variable, the logit of outcome *m* relative to the referent group is expected to change by its respective parameter estimate (which is in log-odds units) given the variables in the model are held

constant.

chocolate relative to vanilla

video – This is the multinomial logit estimate for a one unit increase in **video** score for chocolate relative to vanilla, given the other variables in the model are held constant. If a subject were to increase his **video** score by one point, the multinomial log-odds for preferring chocolate to vanilla would be expected to decrease by 0.024 unit while holding all other variables in the model constant.

puzzle – This is the multinomial logit estimate for a one unit increase in **puzzle** score for chocolate relative to vanilla, given the other variables in the model are held constant. If a subject were to increase his **puzzle** score by one point, the multinomial log-odds for preferring chocolate to vanilla would be expected to decrease by 0.039 unit while holding all other variables in the model constant.

female – This is the multinomial logit estimate comparing females to males for chocolate relative to vanilla, given the other variables in the model are held constant. The multinomial logit for females relative to males is 0.817 unit higher for preferring chocolate to vanilla, given all other predictor variables in the model are held constant. In other words, females are more likely than males to prefer chocolate to vanilla.

_cons – This is the multinomial logit estimate for chocolate relative to vanilla when the predictor variables in the model are evaluated at zero. For males (the variable **female** evaluated at zero) with zero **video** and **puzzle** scores, the logit for preferring chocolate to vanilla is 1.912. Note that evaluating **video** and **puzzle** at zero is out of the range of plausible scores. If the scores were mean-centered, the intercept would have a natural interpretation: log odds of preferring chocolate to vanilla for a male with average **video** and **puzzle** scores.

strawberry relative to vanilla

video – This is the multinomial logit estimate for a one unit increase in **video** score for strawberry relative to vanilla, given the other variables in the model are held constant. If a subject were to increase his **video** score by one point, the multinomial log-odds for preferring strawberry to vanilla would be expected to increase by 0.023 unit while holding all other variables in the model constant.

puzzle – This is the multinomial logit estimate for a one unit increase in **puzzle** score for strawberry relative to vanilla, given the other variables in the model are held constant. If a subject were to increase his **puzzle** score by one point, the multinomial log-odds for preferring strawberry to vanilla would be expected to increase by 0.043 unit while holding all other variables in the model constant.

female – This is the multinomial logit estimate comparing females to males for strawberry relative to vanilla, given the other variables in the model are held constant. The multinomial logit for females relative to males is 0.033 unit lower for preferring strawberry to vanilla, given all other predictor variables in the model are held constant. In other words, males are more likely than females to prefer strawberry ice cream to vanilla ice cream.

_cons – This is the multinomial logit estimate for strawberry relative to vanilla when the predictor variables in the model are evaluated at zero. For males (the variable **female** evaluated at zero) with zero **video** and **puzzle** scores, the logit for preferring strawberry to vanilla is -4.057.

j. **Std. Err.** – These are the standard errors of the individual regression coefficients for the two respective models estimated. They are used in both the calculation of the **z** test statistic, superscript k, and the confidence interval of the regression coefficient, superscript l.

k. **z** and **P>|z|** – The test statistic **z** is the ratio of the **Coef.** to the **Std. Err.** of the respective predictor, and the p-value **P>|z|** is the probability the **z** test statistic (or a more extreme test statistic) would be observed under the null hypothesis. For a given alpha level, **z** and **P>|z|** determine whether or not the null hypothesis that a particular predictor's regression coefficient is zero, given that the rest of the predictors are in the model, can be rejected. If **P>|z|** is less than alpha, then the null hypothesis can be rejected and the parameter estimate is considered significant at that alpha level. The **z** value follows a standard normal distribution which is used to test against a two-sided alternative hypothesis that the **Coef.** is not equal to zero. In multinomial logistic regression, the interpretation of a parameter estimate's significance is limited to the model in which the parameter estimate was calculated. For example, the significance of a parameter estimate in the chocolate relative to vanilla model cannot be assumed to hold in the strawberry relative to vanilla model.

chocolate relative to vanilla

For chocolate relative to vanilla, the **z** test statistic for the predictor **video** (-0.024/0.021) is -1.12 with an associated p-value of 0.261. If we set our alpha level to 0.05, we would fail to reject the null hypothesis and conclude that for chocolate relative to vanilla, the regression coefficient for **video** has not been found to be statistically different from zero given **puzzle** and **female** are in the model.

For chocolate relative to vanilla, the **z** test statistic for the predictor **puzzle** (-0.039/0.020) is -1.99 with an associated p-value of 0.046. If we again set our alpha level to 0.05, we would reject the null hypothesis and conclude that the regression coefficient for **puzzle** has been found to be statistically different from zero for chocolate relative to vanilla given that **video** and **female** are in the model.

For chocolate relative to vanilla, the **z** test statistic for the predictor **female** (0.817/0.391) is 2.09 with an associated p-value of 0.037. If we again set our alpha level to 0.05, we would reject the null hypothesis

and conclude that the difference between males and females has been found to be statistically different for chocolate relative to vanilla given that **video** and **female** are in the model.

For chocolate relative to vanilla, the **z** test statistic for the intercept, **_cons** (1.912/1.127) is 1.70 with an associated p-value of 0.090. With an alpha level of 0.05, we would fail to reject the null hypothesis and conclude that a) the multinomial logit for males (the variable **female** evaluated at zero) and with zero **video** and **puzzle** scores in chocolate relative to vanilla are found not to be statistically different from zero; or b) for males with zero **video** and **puzzle** scores, you are statistically uncertain whether they are more likely to be classified as preferring chocolate or vanilla. We can make the second interpretation when we view the **_cons** as a specific covariate profile (males with zero **video** and **puzzle** scores). Based on the direction and significance of the coefficient, the **_cons** indicates whether the profile would have a greater propensity to be classified in one level of the outcome variable than the other **l**evel.

strawberry relative to vanilla

For strawberry relative to vanilla, the **z** test statistic for the predictor **video** (0.023/0.021) is 1.10 with an associated p-value of 0.272. If we set our alpha level to 0.05, we would fail to reject the null hypothesis and conclude that for strawberry relative to vanilla, the regression coefficient for **video** has not been found to be statistically different from zero given **puzzle** and **female** are in the model.

For strawberry relative to vanilla, the **z** test statistic for the predictor **puzzle** (0.043/0.020) is 2.16 with an associated p-value of 0.031. If we again set our alpha level to 0.05, we would reject the null hypothesis and conclude that the regression coefficient for **puzzle** has been found to be statistically different from zero for strawberry relative to vanilla given that **video** and **female** are in the model.

For strawberry relative to vanilla, the **z** test statistic for the predictor **female** (-0.033/0.350) is -0.09 with an associated p-value of 0.925. If we again set our alpha level to 0.05, we would fail to reject the null hypothesis and conclude that for strawberry relative to vanilla, the regression coefficient for **female** has not been found to be statistically different from zero given **puzzle** and **video** are in the model.

For strawberry relative to vanilla, the **z** test statistic for the intercept, **_cons** (-4.057/1.223) is -3.32 with an associated p-value of 0.001. With an alpha level of 0.05, we would reject the null hypothesis and conclude that a) the multinomial logit for males (the variable **female** evaluated at zero) and with zero **video** and **puzzle** scores in strawberry relative to vanilla are statistically different from zero; or b) for males with zero **video** and **puzzle** scores, there is a statistically significant difference between the likelihood of being classified as preferring strawberry or preferring vanilla**.** Such a male would be more likely to be classified as preferring vanilla to strawberry. We can make the second interpretation when

we view the **_cons** as a specific covariate profile (males with zero **video** and **puzzle** scores). Based on the direction and significance of the coefficient, the **_cons** indicates whether the profile would have a greater propensity to be classified in one level of the outcome variable than the other **l**evel.

[95% Conf. Interval] – This is the Confidence Interval (CI) for an individual multinomial logit regression

i. [**95% Conf. Interval**] – This is the Confidence Interval (CI) for an individual multinomial logit regression coefficient given the other predictors are in the model for outcome *m* relative to the referent group. For a given predictor with a level of 95% confidence, we'd say that we are 95% confident that the "true" population multinomial logit regression coefficient lies between the lower and upper limit of the interval for outcome *m* relative to the referent group. It is calculated as the **Coef.** $(z_{\alpha/2})$*(**Std.Err.**), where $z_{\alpha/2}$ is a critical value on the standard normal distribution. The CI is equivalent to the **z** test statistic: if the CI includes zero, we'd fail to reject the null hypothesis that a particular regression coefficient is zero given the other predictors are in the model. An advantage of a CI is that it is illustrative; it provides a range where the "true" parameter may lie.

Relative Risk Ratio Interpretation

The following is the interpretation of the multinomial logistic regression in terms of relative risk ratios and can be obtained by **mlogit, rrr** after running the multinomial logit model or by specifying the **rrr** option when the full model is specified. This part of the interpretation applies to the output below.

```
mlogit ice_cream video puzzle female, rrr
```

```
Iteration 0:   log likelihood = -210.58254
Iteration 1:   log likelihood = -194.75041
Iteration 2:   log likelihood = -194.03782
Iteration 3:   log likelihood = -194.03485
```

```
Iteration 4:   log likelihood = -194.03485

Multinomial logistic regression                 Number of obs   =        200
                                                LR chi2(6)      =      33.10
                                                Prob > chi2     =     0.0000
Log likelihood = -194.03485                     Pseudo R2       =     0.0786

------------------------------------------------------------------------------
   ice_cream |        RRR[a]   Std. Err.      z    P>|z|     [95% Conf. Interval][b]
-------------+----------------------------------------------------------------
chocolate    |
       video |   .9767108   .0204862    -1.12   0.261     .9373726      1.0177
      puzzle |   .9618236   .0187714    -1.99   0.046      .925727    .9993276
      female |   2.262839   .8847276     2.09   0.037     1.051598    4.869199
-------------+----------------------------------------------------------------
strawberry   |
       video |   1.023187   .0213558     1.10   0.272     .9821747    1.065911
      puzzle |   1.043942   .0207633     2.16   0.031     1.004029    1.085441
      female |   .9676721      .3387    -0.09   0.925     .4872981    1.921595
------------------------------------------------------------------------------
(ice_cream==vanilla is the base outcome)
```

a. **Relative Risk Ratio –** These are the relative risk ratios for the multinomial logit model shown earlier. They can be obtained by exponentiating the multinomial logit coefficients, e^{coef}, or by specifying the **rrr** option when the **mlogit** command is issued. Recall that the multinomial logit model estimates k-1 models, where the k^{th} equation is relative to the referent group. The RRR of a coefficient indicates how the risk of the outcome falling in the comparison group compared to the risk of the outcome falling in the referent group changes with the variable in question. An RRR > 1 indicates that the risk of the outcome falling in the comparison group relative to the risk of the outcome falling in the referent group increases as the variable increases. In other words, the comparison outcome is more likely. An RRR < 1 indicates that the risk of the outcome falling in the comparison group relative to the risk of the outcome falling in the referent group decreases as the variable increases. See the interpretations of the relative risk ratios below for examples. In general, if the RRR < 1, the outcome is more likely to be in the referent group.

chocolate relative to vanilla

video – This is the relative risk ratio for a one unit increase in **video** score for preferring chocolate to vanilla, given that the other variables in the model are held constant. If a subject were to increase her **video** score by one unit, the relative risk for preferring chocolate to vanilla would be expected to

video score by one unit, the relative risk for preferring chocolate to vanilla would be expected to decrease by a factor of 0.977 given the other variables in the model are held constant. So, given a one unit increase in **video**, the relative risk of being in the chocolate group would be 0.977 times more likely when the other variables in the model are held constant. More generally, we can say that if a subject were to increase her **video** score, we would expect her to be more likely to prefer vanilla ice cream over chocolate ice cream.

puzzle – This is the relative risk ratio for a one unit increase in **puzzle** score for preferring chocolate to vanilla, given that the other variables in the model are held constant. If a subject were to increase her **puzzle** score by one unit, the relative risk for preferring chocolate to vanilla would be expected to decrease by a factor of 0.962 given the other variables in the model are held constant. More generally, we can say that if two subjects have identical **video** scores and are both female (or both male), the subject with the higher **puzzle** score is more likely to prefer vanilla ice cream over chocolate ice cream than the subject with the lower **puzzle** score.

female – This is the relative risk ratio comparing females to males for preferring chocolate to vanilla, given that the other variables in the model are held constant. For females relative to males, the relative risk for preferring chocolate relative to vanilla would be expected to increase by a factor of 2.263 given the other variables in the model are held constant. In other words, females are more likely than males to prefer chocolate ice cream over vanilla ice cream.

strawberry relative to vanilla

video – This is the relative risk ratio for a one unit increase in **video** score for preferring strawberry to vanilla, given that the other variables in the model are held constant. If a subject were to increase her **video** score by one unit, the relative risk for strawberry relative to vanilla would be expected to increase by a factor of 1.023 given the other variables in the model are held constant. More generally, we can say that if a subject were to increase her **video** score, we would expect her to be more likely to prefer strawberry ice cream over vanilla ice cream.

puzzle – This is the relative risk ratio for a one unit increase in **puzzle** score for preferring strawberry to vanilla, given that the other variables in the model are held constant. If a subject were to increase her **puzzle** score by one unit, the relative risk for strawberry relative to vanilla would be expected to increase by a factor of 1.043 given the other variables in the model are held constant. More generally,

we can say that if two subjects have identical **video** scores and are both female (or both male), the subject with the higher **puzzle** score is more likely to prefer strawberry ice cream to vanilla ice cream than the subject with the lower **puzzle** score.

female – This is the relative risk ratio comparing females to males for strawberry relative to vanilla

female – This is the relative risk ratio comparing females to males for strawberry relative to vanilla, given that the other variables in the model are held constant. For females relative to males, the relative risk for preferring strawberry to vanilla would be expected to decrease by a factor of 0.968 given the other variables in the model are held constant. In other words, females are less likely than males to prefer strawberry ice cream to vanilla ice cream.

b. [**95% Conf. Interval**] – This is the CI for the relative risk ratio given the other predictors are in the model. For a given predictor with a level of 95% confidence, we'd say that we are 95% confident that the "true" population relative risk ratio comparing outcome *m* to the referent group lies between the lower and upper limit of the interval. An advantage of a CI is that it is illustrative; it provides a range where the "true" relative risk ratio may lie.

Click here to report an error on this page or leave a comment

How to cite this page (https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-cite-web-pages-and-programs-from-the-ucla-statistical-consulting-group/)


```
------------------------------------------------------------------------------
     honcomp | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |   1.918168   .6400451     1.95   0.051     .9973827    3.689024
------------------------------------------------------------------------------
```

For more information on interpreting odds ratios, please see How do I interpret odds ratios in logistic regression? (/stata/faq/how-do-i-interpret-odds-ratios-in-logistic-regression/) .

Click here to report an error on this page or leave a comment

How to cite this page (https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-cite-web-pages-and-programs-from-the-ucla-statistical-consulting-group/)

 HOME (/) CONTACT (/contact)

ANNOTATED STATA OUTPUT MULTIPLE REGRESSION ANALYSIS

This page shows an example multiple regression analysis with footnotes explaining the output. The analysis uses a data file about scores obtained by elementary schools, predicting **api00** from **ell**, **meals**, **yr_rnd**, **mobility**, **acs_k3**, **acs_46**, **full**, **emer** and **enroll** using the following Stata commands.

```
use https://stats.idre.ucla.edu/stat/stata/olc/reg/elemapi2
regress api00 ell meals yr_rnd mobility acs_k3 acs_46 full emer enroll
```

The output of this command is shown below, followed by explanations of the output.

Output

```
      Source[a] |       SS[b]         df[c]        MS[d]    Number of obs[e]   =    395
----------------+------------------------------------------  F(9, 385)[f]       =    232.41
          Model |  6740702.01          9   748966.89        Prob > F           =    0.0000
       Residual |  1240707.78        385  3222.61761        R-squared[g]       =    0.8446
```

```
-------------+----------------------------------   Adj R-squared[h]   =    0.8409
       Total |  7981409.79       394  20257.3852   Root MSE[i]        =    56.768

------------------------------------------------------------------------------
      api00[j] |    Coef.[k]   Std. Err.[l]   t[m]    P>|t|[m]   [95% Conf. Interval][n]
-------------+----------------------------------------------------------------
         ell |  -.8600707   .2106317    -4.08   0.000    -1.274203   -.4459382
       meals |  -2.948216   .1703452   -17.31   0.000     -3.28314   -2.613293
      yr_rnd |  -19.88875   9.258442    -2.15   0.032    -38.09219   -1.685309
    mobility |  -1.301352   .4362053    -2.98   0.003    -2.158995   -.4437088
      acs_k3 |     1.3187   2.252683     0.59   0.559    -3.110401    5.747801
      acs_46 |   2.032456   .7983213     2.55   0.011      .462841    3.602071
        full |    .609715   .4758205     1.28   0.201    -.3258169    1.545247
        emer |  -.7066192   .6054086    -1.17   0.244     -1.89694    .4837019
      enroll |   -.012164   .0167921    -0.72   0.469    -.0451798    .0208517
       _cons |   758.9418   62.28601    12.18   0.000     636.4785    881.4051
------------------------------------------------------------------------------
```

Footnotes

a. This is the source of variance, Model, Residual, and Total. The Total Variance is partitioned into the variance which can be explained by the independent variables (Model) and the variance which is not explained by the independent variables (Residual). Note that the Sums of Squares for the Model and Residual add up to the Total Variance, reflecting the fact that the Total Variance is partitioned into Model and Residual variance.

b. These are the Sum of Squares associated with the three sources of variance, Total, Model and Residual. These can be computed in many ways. Conceptually, these formulas can be expressed as:

SSTotal: The total variability around the mean. Σ (Y – Ybar)^2. SSResidual: The sum of squared errors in prediction. Σ (Y – Ypredicted)^2. SSModel: The improvement in prediction by using the predicted value of Y over just using the mean of Y. Hence, this would be the squared differences between the predicted value of Y and the mean of Y, S(Ypredicted – Ybar)^2. Another way to think of this is the SSModel is SSTotal –

SSResidual. Note that the SSTotal = SSModel + SSResidual. Note that SSModel / SSTotal is equal to .84, the value of R-Square. This is because R-Square is the proportion of the variance explained by the independent variables, hence can be computed by SSModel / SSTotal.

c. These are the degrees of freedom associated with the sources of variance. The total variance has N-1 degrees of freedom (DF). In this case, there were N=395 observations, so the DF for total is 394. The model degrees of freedom corresponds to the number of predictors minus 1 (K-1). You may think this would be 9-1 (since there were 9 independent variables in the model: **ell**, **meals**, **yr_rnd**, **mobility**, **acs_k3**, **acs_46**, **full**, **emer** and **enroll**). But, the intercept is automatically included in the model (unless you explicitly omit the intercept). Including the intercept, there are 10 predictors, so the model has 10-1=9 degrees of freedom. The Residual degrees of freedom is the DF total minus the DF model, 394 – 9 is 385.

d. These are the Mean Squares, the Sum of Squares divided by their respective DF. For the Model, 6740702.01 / 9 is equal to 748966.89. For the Residual, 124070.79 / 385 equals 3222.6176. These are computed so you can compute the F ratio, dividing the Mean Square Model by the Mean Square Residual (or Error) to test the significance of the predictors in the model.

e. This is the number of observations used in the regression analysis.

f. The F Value is the Mean Square Model (748966.89) divided by the Mean Square Residual (3222.61761), yielding F=232.41. The p-value associated with this F value is very small (0.0000). These values are used to answer the question “Do the independent variables reliably predict the dependent variable?”. The p-value is compared to your alpha level (typically 0.05) and, if smaller, you can conclude “Yes, the independent variables reliably predict the dependent variable”. You could say that the group of variables **ell**, **meals**, **yr_rnd**, **mobility**, **acs_k3**, **acs_46**, **full** , and **enroll** can be used to reliably predict **api00** (the dependent variable). If the p-value were greater than 0.05, you would say that the group of independent variables do not show a significant relationship with the dependent variable, or that the group of independent variables do not reliably predict the dependent variable. Note that this is an overall significance test assessing whether the group of independent variables when used together reliably predict the dependent variable, and does not address the ability of any of the particular independent variables to predict the dependent variables. The ability of each individual independent variable to predict the dependent variable is addressed in the table below where each of the individual variables are listed.

g. R-Square is the proportion of variance in the dependent variable (**api00**) which can be predicted from the independent variables (**ell**, **meals**, **yr_rnd**, **mobility**, **acs_k3**, **acs_46**, **full emer**, and **enroll**). This value indicates that 84% of the variance in **api00** can be predicted from the variables **ell**, **meals**, **yr_rnd**, **mobility**, **acs_k3**, **acs_46**, **full**, **emer** and **enroll**. Note that this is an overall measure of the strength of association, and does not reflect the extent to which any particular independent variable is associated with the dependent variable.

h. Adjusted R-square. As predictors are added to the model, each predictor will explain some of the

h. Adjusted R-square. As predictors are added to the model, each predictor will explain some of the variance in the dependent variable simply due to chance. One could continue to add predictors to the model which would continue to improve the ability of the predictors to explain the dependent variable, although some of this increase in R-square would be simply due to chance variation in that particular sample. The adjusted R-square attempts to yield a more honest value to estimate the R-squared for the population. The value of R-square was .8446, while the value of Adjusted R-square was .8409. Adjusted R-squared is computed using the formula 1 – ((1-R-sq)(N-1 / N – k – 1)). From this formula, you can see that when the number of observations is small and the number of predictors is large, there will be a much greater difference between R-square and adjusted R-square (because the ratio of (N-1 / N – k – 1) will be much less than 1. By contrast, when the number of observations is very large compared to the number of predictors, the value of R-square and adjusted R-square will be much closer because the ratio of (N-1)/(N-k-1) will approach 1.

i. Root MSE is the standard deviation of the error term, and is the square root of the Mean Square Residual (or Error)

j. This column shows the dependent variable at the top (**api00**) with the predictor variables below it (**ell**, **meals**, **yr_rnd**, **mobility**, **acs_k3**, **acs_46**, **full emer** and **enroll**). The last variable (**_cons**) represents the constant, also referred to in textbooks as the Y intercept, the height of the regression line when it crosses the Y axis.

k. These are the values for the regression equation for predicting the dependent variable from the independent variable. The regression equation is presented in many different ways, for example:

Ypredicted = **b0** + **b1*x1** +**b2*x2** + **b3*x3 . . .**

The column of estimates (coefficients or parameter estimates, from here on labeled coefficients) provides the values for b0, b1, b2, b3, b4, b5, b6, b7, b8 and b9 for this equation. Expressed in terms of the variables used in this example, the regression equation is

api00Predicted =778.83-.86*ell-2.95*meals-19.89*yr_rnd-1.30*mobility+1.32*acs_k3+2.03*acs_46+.61*full-.71*emer-.01*enroll

These estimates tell you about the relationship between the independent variables and the dependent variable. These estimates tell the amount of increase in api00 that would be predicted by a 1 unit increase in the predictor. Note: For the independent variables which are not significant, the coefficients are not significantly different from 0, which should be taken into account when interpreting the

coefficients. (See the columns with the t-value and p-value about testing whether the coefficients are significant.) **ell** – The coefficient (parameter estimate) is -.86. So, for every unit increase in ell, a .86 unit decrease in **api00** is predicted. Or, for every increase of one percentage point of **api00**, **ell** is predicted to be lower by .86. This is significantly different from 0. **meals** – For every unit increase in **meals**, there is a 2.95 unit decrease in the predicted **api00**.

yr_rnd – For every unit increase of **yr_rnd**, the predicted value of **api00** would be 19.89 units lower. **mobility** – For every unit increase in **mobility**, **api00** is predicted to be 1.30 units lower. **acs_k3** – For every unit increase in **acs_k3**, **api00** is predicted to be 1.32 units higher. **acs_46** – For every unit increase in **acs_46**, **api00** is predicted to be 2.03 units higher. **full** – For every unit increase in **full**, **api00** is predicted to be .61 units higher. **emer** – For every unit increase in **emer**, **api00** is predicted to be .71 units lower. **enroll** – For every unit increase in **enroll**, **api00** is predicted to be .01 units lower.

l. These are the standard errors associated with the coefficients. The standard error is used for testing whether the parameter is significantly different from 0 by dividing the parameter estimate by the standard error to obtain a t value (see the column with t values and p-values). The standard errors can also be used to form a confidence interval for the parameter, as shown in the last 2 columns of this table.

m. These columns provide the t value and 2 tailed p-value used in testing the null hypothesis that the coefficient/parameter is 0. If you use a 2-tailed test, then you would compare each p-value to your preselected value of alpha. Coefficients having p-values less than alpha are significant. For example, if you chose alpha to be 0.05, coefficients having a p-value of 0.05 or less would be statistically significant (i.e., you can reject the null hypothesis and say that the coefficient is significantly different from 0). If you use a 1-tailed test (i.e., you predict that the parameter will go in a particular direction), then you can divide the p-value by 2 before comparing it to your preselected alpha level. With a 2-tailed test and alpha of 0.05, you can reject the null hypothesis that the coefficient for **ell** is equal to 0. The coefficient of -.86 is significantly different from 0. Using a 2-tailed test and alpha of 0.01, the p-value of 0.000 is smaller than 0.01 and the coefficient for **ell** would still be significant at the 0.01 level. Had you predicted that this coefficient would be positive (i.e., a 1-tailed test), you would be able to divide the p-value by 2 before comparing it to alpha. This would yield a 1-tailed p-value of 0.000, which is less than 0.01, and then you could conclude that this coefficient is greater than 0 with a 1-tailed alpha of 0.01. The coefficient for **meals** is significantly different from 0 using alpha of 0.05 because its p-value of 0.000 is smaller than 0.05. The coefficient for **yr_rnd** (-19.89) is significantly different from 0 because its p-value is definitely smaller than 0.05 and even 0.01. The coefficient for **mobility** is significantly different from 0

using alpha of 0.05 because its p-value of 0.003 is smaller than 0.05. The coefficient for **acs_k3** is not significantly different from 0 using alpha of 0.05 because its p-value of .559 is greater than 0.05. The coefficient for **acs_46** is significantly different from 0 using alpha of 0.05 because its p-value of 0.011 is smaller than 0.05. The coefficient for **full** is not significantly different from 0 using alpha of 0.05 because

its p-value of .201 is greater than 0.05. The coefficient for **emer** is not significantly different from 0 using alpha of 0.05 because its p-value of .244 is greater than 0.05. The coefficient for **enroll** is not significantly different from 0 using alpha of 0.05 because its p-value of .469 is greater than 0.05. The constant (**_cons**) is significantly different from 0 at the 0.05 alpha level. However, having a significant intercept is seldom interesting.

n. This shows a 95% confidence interval for the coefficient. This is very useful as it helps you understand how high and how low the actual population value of the parameter might be. Consider the coefficients for **ell** (-.86) and **meals** (-2.95). Immediately you see that the estimate for **meals** is so much bigger, but examine the confidence interval for it (-3.28 to -2.61). Now examine the confidence interval for **ell** (-1.27 to -.45). Even though **meals** has a larger coefficient, it could be as small as -3.28. By contrast, the lower confidence level for **ell** is -1.27.

