Dummy variables

1

Treatment	*Control*
12	6
22	7
18	8
14	3
13	2
$\bar{Y}_1$	$\bar{Y}_0$

2

Y	*X*
18	1
14	1
13	1
10	1
6	0
7	0
8	0

X_i identifies treatment

X_i=1 if in treatment group
X_i=0 if in control

3

Are wages different across union/nonunion jobs

- H_o: $u_n = u_u$
  - Or alternatively
- H_o: $d = u_n - u_u = 0$
- H_o: $d \neq 0$

4

cps87.dta

```
. gen ln_weekly_earn=ln(weekly_earn)
. gen union=union_status==1
. gen nonwhite=((race==2)|(race==3))
```

5

```
. * test whether means are the same across two subsamples
. ttest weekly_earn, by(union)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |   15309    480.1503    2.017734    249.6532    476.1953    484.1053
       1 |    4597    515.2845    2.705061    183.4063    509.9813    520.5878
---------+--------------------------------------------------------------------
combined |   19906     488.264    1.676048    236.4713    484.9788    491.5492
---------+--------------------------------------------------------------------
    diff |          -35.13423    3.969334               -42.91446     -27.354
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -8.8514
Ho: diff = 0                                     degrees of freedom =    19904

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

$$\hat{t} = \frac{-35.13}{3.969} = -8.85$$

$$|\hat{t}| > 1.96 \quad \therefore \textit{reject null}$$

6

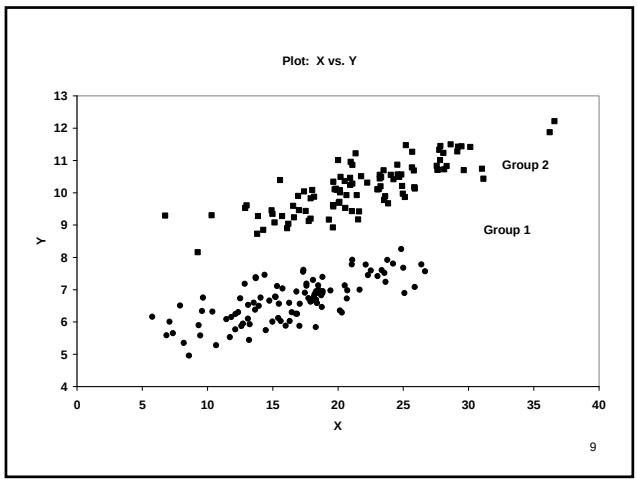
```
. reg weekly_earn union

      Source |       SS       df       MS              Number of obs =   19906
-------------+------------------------------           F(  1, 19904) =   78.35
       Model |  4364135.22     1  4364135.22           Prob > F      =  0.0000
    Residual |  1.1087e+09 19904  55702.2458           R-squared     =  0.0039
-------------+------------------------------           Adj R-squared =  0.0039
       Total |  1.1131e+09 19905  55918.6956           Root MSE      =  236.01

------------------------------------------------------------------------------
weekly_earn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
      union |   35.13423   3.969334     8.85   0.000       27.354    42.91446
      _cons |   480.1503   1.907493   251.72   0.000     476.4115    483.8891
------------------------------------------------------------------------------

.
```

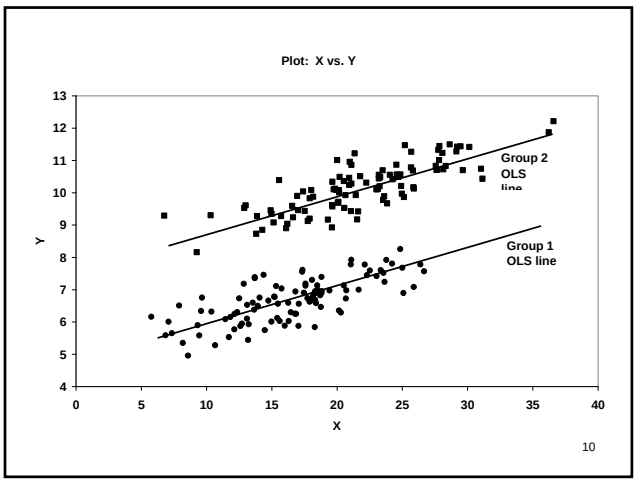
7

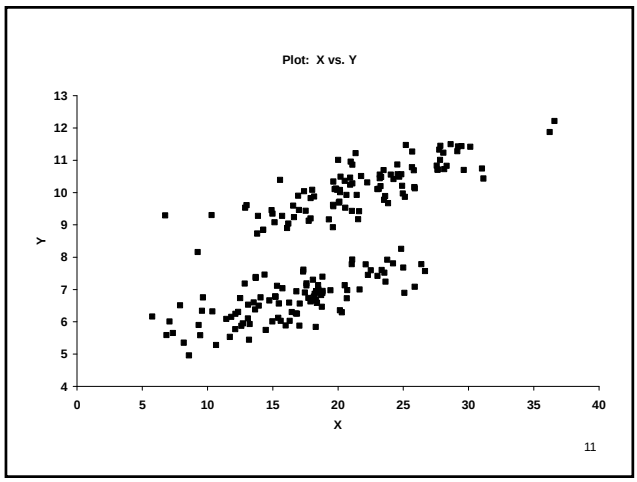
Synthetic problem

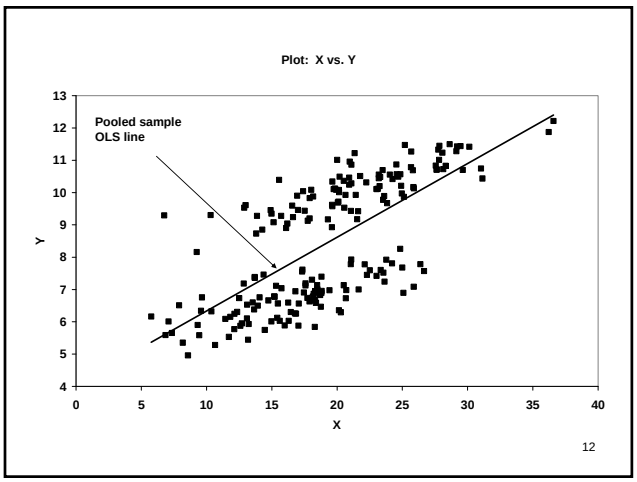
- X impacts Y
- But there are two groups of people in the population: 1 and 2
- Average of X and Y is higher for group 2 than 1
- Should you add a dummy for group 2?

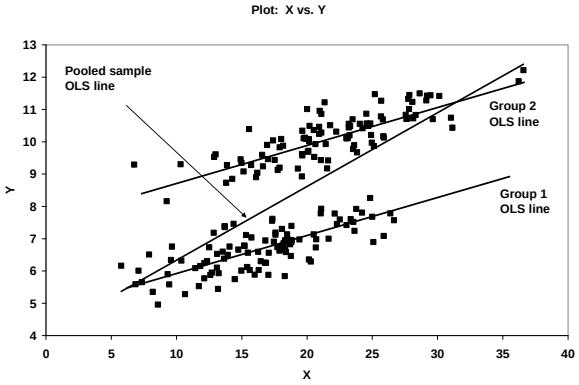
8


Plot: X vs. Y
Group 2
Group 1
Y
X
9


Plot: X vs. Y
Group 2
OLS
Group 1
OLS line
Y
X
10


Plot: X vs. Y
Y
X
11


Plot: X vs. Y
Pooled sample
OLS line
Y
X
12



13

Sort the data by groups

```
sort group

by group:  reg y x
```

Run a regression for each of the separate groups

14

```
-------------------------------------------------------------------------------
-> group = 1

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  109.99
       Model |  24.6462946     1  24.6462946           Prob > F      =  0.0000
    Residual |  21.9591909    98  .224073377           R-squared     =  0.5288
-------------+------------------------------           Adj R-squared =  0.5240
       Total |  46.6054855    99   .47076248           Root MSE      =  .47336

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .1038158   .0098988    10.49   0.000     .0841719    .1234596
       _cons |    4.96899   .1700462    29.22   0.000     4.630638    5.305541
------------------------------------------------------------------------------
-------------------------------------------------------------------------------
-> group = 2

      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  1,    98) =  183.45
       Model |  38.5560428     1  38.5560428           Prob > F      =  0.0000
    Residual |  20.5969602    98  .210173063           R-squared     =  0.6518
-------------+------------------------------           Adj R-squared =  0.6482
       Total |   59.153003    99   .59750508           Root MSE      =  .45845

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .1157525   .0085462    13.54   0.000     .0987929    .1327122
       _cons |   7.647898   .1922749    39.78   0.000     7.266335    8.029462
------------------------------------------------------------------------------
```

The coefficients on X in both models are pretty similar

15

```
. reg y x

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  1,   198) =  182.39
       Model |  343.730319     1  343.730319           Prob > F      =  0.0000
    Residual |  373.140785   198  1.88454942           R-squared     =  0.4795
-------------+------------------------------           Adj R-squared =  0.4769
       Total |  716.871104   199  3.60236736           Root MSE      =  1.3728

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .2282889   .0169036    13.51   0.000     .1949547    .2616231
       _cons |   4.051698   .3383412    11.98   0.000     3.384483    4.718912
------------------------------------------------------------------------------
```

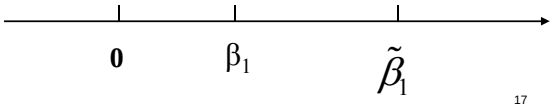
When you ignore the fact that group 2 has higher outcomes And higher x's, this overstates the impact on x

16

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2 \hat{\delta}_1$$

$$x_{2i} = \delta_0 + \delta_1 x_{1i} + \phi_i$$

$$\hat{\beta}_1 > 0, \hat{\delta}_1 > 0 \quad and \quad \beta_2 > 0$$



17

Generate dummy variable for One of the groups using logical operators

- gen group2=group==2
- reg y x group2

Run a regression with x and the variable

18

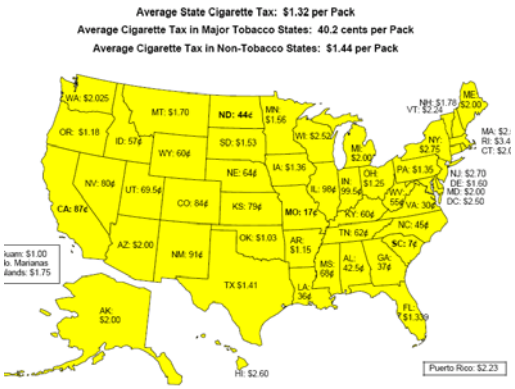
```
      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  2,   197) = 1553.71
       Model |  674.133396     2  337.066698           Prob > F      =  0.0000
    Residual |  42.7377077   197  .216942678           R-squared     =  0.9404
-------------+------------------------------           Adj R-squared =  0.9398
       Total |  716.871104   199  3.60236736           Root MSE      =  .46577

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |    .110467   .0064813    17.04   0.000     .0976853    .1232486
      group2 |   2.905037   .0744393    39.03   0.000     2.758237    3.051837
       _cons |   4.850349   .1166412    41.65   0.000     4.628323    5.088374
------------------------------------------------------------------------------
```

19

Return to tobacco model

- Regress ln(per capita consumption) on taxes and a time trend
- Concern: who are the lowest taxing states?
- Model subject to an omitted variables bias?

20



State rank per capita consumption - 2004

State	Rank	Per capita packs/year
KY	2	174.4
VA	6	97.4
TN	7	96.5
NC	8	95.6
SC	9	92.2
MD	38	48.9
US		74.2


```
. * time trend
. gen trend=year-1981

. label var trend "=1 in 1st year, 2 in second, etc"

..
. * tobacco producing state
. gen tob_state=(state=="NC"|state=="VA"|state=="SC"|state=="KY"|state=="MD"|st
> ate=="TN")
```

Two new variables:

A time trend, =1 in 1st year, 2 in second, etc

A dummy if the state produces tobacco


```
. * run regression with tax and trend
. reg packs_pc real_tax trend

      Source |       SS       df       MS              Number of obs =    1020
-------------+------------------------------           F(  2,  1017) =  418.36
       Model |   368205.64     2   184102.82           Prob > F      =  0.0000
    Residual |  447541.752  1017   440.06072           R-squared     =  0.4514
-------------+------------------------------           Adj R-squared =  0.4503
       Total |  815747.392  1019  800.537186           Root MSE      =  20.978

------------------------------------------------------------------------------
   packs_pc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   real_tax |  -.7349566   .0394065   -18.65   0.000     -.812284   -.6576292
      trend |  -1.510832   .1260754   -11.98   0.000    -1.758229   -1.263434
      _cons |   162.1518   2.115978    76.63   0.000     157.9996    166.304
------------------------------------------------------------------------------
```

Each year, tobacco consumption falls 1.5 packs/person

Every cent increase in the tax reduces consumption by .74 packs

```
. ttest packs_pc, by(tob_state)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     900    103.4318    .8854382    26.56315     101.694    105.1695
       1 |     120    130.3792    2.705486    29.63711     125.022    135.7363
---------+--------------------------------------------------------------------
combined |    1020    106.6021    .8859122    28.29377    104.8636    108.3405
---------+--------------------------------------------------------------------
    diff |           -26.94739    2.618155               -32.08499   -21.80979
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t = -10.2925
Ho: diff = 0                                     degrees of freedom =     1018

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Tobacco producing states have substantially higher consumption Than non-tobacco states


```
. * correlation between tax and other variables
. reg real_tax trend tob_state

      Source |       SS       df       MS              Number of obs =    1020
-------------+------------------------------           F(  2,  1017) =  257.06
       Model |  116564.372     2  58282.1858           Prob > F      =  0.0000
    Residual |  230584.642  1017  226.730228           R-squared     =  0.3358
-------------+------------------------------           Adj R-squared =  0.3345
       Total |  347149.013  1019  340.676166           Root MSE      =  15.058

------------------------------------------------------------------------------
    real_tax |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       trend |   1.371175   .0817634    16.77   0.000     1.210731     1.53162
  tob_state |  -22.33085   1.463332   -15.26   0.000    -25.20235   -19.45936
       _cons |   45.65438    .924806    49.37   0.000     43.83963    47.46913
------------------------------------------------------------------------------
```

Tobacco producing states have substantially lower taxes that Other states

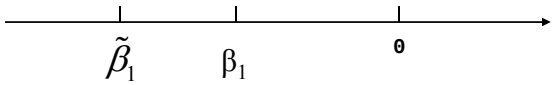
Know two facts

- Consumption is higher in tobacco producing states
- Taxes are lower in tobacco producing states
- What should happen to the tax coefficient when the tob_state dummy is added to the model?

$$E[\tilde{\beta}_1] = \beta_1 + \beta_2\hat{\delta}_1$$

$$x_{1i} = \delta_0 + \delta_1 x_{2i} + \phi_i$$

$$\hat{\beta}_1 < 0, \hat{\delta}_1 < 0 \quad and \quad \beta_2 > 0$$



$\tilde{\beta}_1$ β_1

```
. * add tobacco producing state dummy
. reg packs_pc real_tax trend tob_state

      Source |       SS       df       MS              Number of obs =    1020
-------------+------------------------------           F(  3,  1016) =  299.22
       Model |  382648.514     3  127549.505            Prob > F      =  0.0000
    Residual |  433098.878  1016  426.278423            R-squared     =  0.4691
-------------+------------------------------           Adj R-squared =  0.4675
       Total |  815747.392  1019  800.537186            Root MSE      =  20.647

------------------------------------------------------------------------------
    packs_pc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    real_tax |  -.6269276   .0429963   -14.58   0.000    -.7112994   -.5425558
       trend |  -1.658958    .126668   -13.10   0.000    -1.907519   -1.410398
  tob_state |   12.94756   2.224374     5.82   0.000     8.582669    17.31246
       _cons |   155.9803   2.336932    66.75   0.000     151.3946    160.5661
------------------------------------------------------------------------------
```

29

```
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
male            float  %9.0g                  dummy variable, =1 of male
business        float  %9.0g                  dummy variable, =1 if business
                                                major
engineer        float  %9.0g                  dummy variable, =1 if engineer
greek           float  %9.0g                  dummy variable, =1 if in
                                                sor/fraternity
college_gpa     float  %9.0g                  college GPA, 4.0 scale
hs_gpa          float  %9.0g                  high school GPA, 4.0 scale
act             float  %9.0g                  act score, 1-36
pc              float  %9.0g                  dummy variable, =1 if own a PC
-------------------------------------------------------------------------------
Sorted by:
```

30

```
. * run regression
. reg college_gpa hs_gpa act male greek business engineer pc

      Source |       SS       df       MS              Number of obs =     141
-------------+------------------------------           F(  7,   133) =    6.41
       Model |  4.89514299     7  .699306142            Prob > F      =  0.0000
    Residual |  14.5109565   133  .109104936            R-squared     =  0.2522
-------------+------------------------------           Adj R-squared =  0.2129
       Total |  19.4060994   140  .138614996            Root MSE      =  .33031

------------------------------------------------------------------------------
college_gpa |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hs_gpa |    .468164   .0977183     4.79   0.000      .274881     .661447
        act |   .0086761   .0108872     0.80   0.427    -.0128583    .0302105
       male |   .0198939   .0598948     0.33   0.740    -.0985757    .1383635
      greek |   .0322331   .0601106     0.54   0.593    -.0866633    .1511295
   business |   .0555991   .0745173     0.75   0.457    -.0917932    .2029914
   engineer |  -.2915697   .1661098    -1.76   0.082    -.6201284    .0369889
         pc |   .1779455   .0578525     3.08   0.003     .0635155    .2923756
      _cons |   1.129177   .3443683     3.28   0.001     .4480297    1.810324
------------------------------------------------------------------------------
```

31

cps87.dta

```
. gen ln_weekly_earn=ln(weekly_earn)

. gen union=union_status==1

. gen nonwhite=((race==2)|(race==3))
```

32

```
. * run basic regression
. * ln(weekly earnings) on age, educ, union nonwhite
. reg ln_weekly age years_educ union nonwhite

      Source |       SS       df       MS              Number of obs =   19906
-------------+------------------------------           F(  4, 19901) = 1952.80
       Model |  1476.80313     4  369.200784            Prob > F      =  0.0000
    Residual |  3762.53555 19901  .189062638           R-squared     =  0.2819
-------------+------------------------------           Adj R-squared =  0.2817
       Total |  5239.33869 19905  .263217216           Root MSE      =  .43481

------------------------------------------------------------------------------
ln_weekly_~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0130299   .0002794    46.64   0.000     .0124823    .0135775
  years_educ |   .0740252   .0011333    65.32   0.000     .0718038    .0762466
       union |   .1470411   .0074052    19.86   0.000     .1325263    .1615559
    nonwhite |  -.1680299   .0090807   -18.50   0.000    -.1858289   -.1502309
       _cons |   4.588052     .01929   237.85   0.000     4.550242    4.625861
------------------------------------------------------------------------------
```


Now change the reference group

```
. gen non_union=union_status==2
. gen white=race==1

. * no change the reference groups for the
. * dummy variables, adding non_union and white
. * to the model

. * ln(weekly earnings) on age, educ, nonunion white
. reg ln_weekly age years_educ non_union white
```


Notice that changing the reference groups on the DVs does not change R2 or the coef's on other parameters

```
. * ln(weekly earnings) on age, educ, nonunion white
. reg ln_weekly age years_educ non_union white

      Source |       SS       df       MS              Number of obs =   19906
-------------+------------------------------           F(  4, 19901) = 1952.80
       Model |  1476.80313     4  369.200784            Prob > F      =  0.0000
    Residual |  3762.53555 19901  .189062638           R-squared     =  0.2819
-------------+------------------------------           Adj R-squared =  0.2817
       Total |  5239.33869 19905  .263217216           Root MSE      =  .43481

------------------------------------------------------------------------------
ln_weekly_~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0130299   .0002794    46.64   0.000     .0124823    .0135775
  years_educ |   .0740252   .0011333    65.32   0.000     .0718038    .0762466
  non_union  |  -.1470411   .0074052   -19.86   0.000    -.1615559   -.1325263
       white |   .1680299   .0090807    18.50   0.000     .1502309    .1858289
       _cons |   4.567063   .0198475   230.11   0.000      4.52816    4.605966
------------------------------------------------------------------------------
```

Notice that the only thing that has changed is that the sign on the DVs has flipped


```
. * generate regional dummy variables
. gen region1=region==1

. gen region2=region==2

. gen region3=region==3

. gen region4=region==4
```

Generate dummies for each region of the country

Do something silly – include all four dummy variables In the model --

```
. * do something dumb -- include all dummy variables
. reg ln_weekly age years_educ union nonwhite region1-region4

      Source |       SS       df       MS              Number of obs =   19906
-------------+------------------------------           F(  7, 19898) = 1138.38
       Model |  1498.22885     7  214.032692            Prob > F      =  0.0000
    Residual |  3741.10984 19898  .188014365            R-squared     =  0.2860
-------------+------------------------------           Adj R-squared =  0.2857
       Total |  5239.33869 19905  .263217216            Root MSE      =  .43361

------------------------------------------------------------------------------
ln_weekly_~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0130034   .0002787    46.66   0.000     .0124572    .0135496
 years_educ |     .07325   .0011326    64.68   0.000     .0710301    .0754699
       union |   .1386003   .0074886    18.51   0.000      .123922    .1532787
   nonwhite |  -.1633235   .0091623   -17.83   0.000    -.1812823   -.1453647
    region1 |  -.0082481   .0091485    -0.90   0.367      -.02618    .0096837
    region2 |  -.0538852   .0092398    -5.83   0.000    -.0719961   -.0357743
    region3 |  -.0794961   .0088407    -8.99   0.000    -.0968246   -.0621676
    region4 |  (dropped)
      _cons |   4.639453   .0203976   227.45   0.000     4.599473    4.679434
------------------------------------------------------------------------------
```

STATA will remind you cannot run a model with all the Dummies included


```
. * run model with regional dummmy variables
. reg ln_weekly age years_educ union nonwhite region2-region4

      Source |       SS       df       MS              Number of obs =   19906
-------------+------------------------------           F(  7, 19898) = 1138.38
       Model |  1498.22885     7  214.032692            Prob > F      =  0.0000
    Residual |  3741.10984 19898  .188014365            R-squared     =  0.2860
-------------+------------------------------           Adj R-squared =  0.2857
       Total |  5239.33869 19905  .263217216            Root MSE      =  .43361

------------------------------------------------------------------------------
ln_weekly_~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0130034   .0002787    46.66   0.000     .0124572    .0135496
 years_educ |     .07325   .0011326    64.68   0.000     .0710301    .0754699
       union |   .1386003   .0074886    18.51   0.000      .123922    .1532787
   nonwhite |  -.1633235   .0091623   -17.83   0.000    -.1812823   -.1453647
    region2 |  -.0456371    .008755    -5.21   0.000    -.0627975   -.0284766
    region3 |  -.0712479   .0084693    -8.41   0.000    -.0878486   -.0546473
    region4 |   .0082481   .0091485     0.90   0.367    -.0096837       .02618
      _cons |   4.631205   .0201706   229.60   0.000     4.591669    4.670741
------------------------------------------------------------------------------
```

Difference between region 3 and region 4:
-0.0712 - 0.0082 = -0.0794

Difference between region 2 and region 4:
-0.0456 – 0.0082 = -0.0538

5% Critical values of F-Distribution

degrees of freedom in denominator	Degrees of Freedom in numerator 1	2	3	4	5	6
10	4.96	4.10	3.71	3.48	3.33	3.22
11	4.84	3.98	3.59	3.36	3.20	3.09
12	4.75	3.89	3.49	3.26	3.11	3.00
13	4.67	3.81	3.41	3.18	3.03	2.92
14	4.60	3.74	3.34	3.11	2.96	2.85
15	4.54	3.68	3.29	3.06	2.90	2.79
16	4.49	3.63	3.24	3.01	2.85	2.74
17	4.45	3.59	3.20	2.96	2.81	2.70
18	4.41	3.55	3.16	2.93	2.77	2.66
19	4.38	3.52	3.13	2.90	2.74	2.63
20	4.35	3.49	3.10	2.87	2.71	2.60
21	4.32	3.47	3.07	2.84	2.68	2.57
22	4.30	3.44	3.05	2.82	2.66	2.55
23	4.28	3.42	3.03	2.80	2.64	2.53
24	4.26	3.40	3.01	2.78	2.62	2.51
30	4.17	3.32	2.92	2.69	2.53	2.42
40	4.08	3.23	2.84	2.61	2.45	2.34
60	4.00	3.15	2.76	2.53	2.37	2.25
90	3.95	3.10	2.71	2.47	2.32	2.20
120	3.92	3.07	2.68	2.45	2.29	2.18
infinity	3.84	3.00	2.61	2.37	2.21	2.10


```
. *test whether the regional effects are all zero
. test region2 region3 region4

 ( 1)  region2 = 0
 ( 2)  region3 = 0
 ( 3)  region4 = 0

       F(  3, 19898) =   37.99
            Prob > F =    0.0000
```

- Change the reference group from region 1 to region 4
- All the coefficients are now in relation to the omitted group #4
- E.g., The coefficient on region 3 is now the difference between region 3 and 4

41

The coef's on the other parameters stay the same. Notice The the SSE, SSM, R2 do not change at all

```
. *change the reference group from region1 to region4
. reg ln_weekly age years_educ union nonwhite region1-region3

     Source |       SS       df       MS              Number of obs =   19906
------------+------------------------------           F(  7, 19898) = 1138.38
      Model |  1498.22885     7  214.032692            Prob > F      =  0.0000
   Residual |  3741.10984 19898  .188014365           R-squared     =  0.2860
------------+------------------------------           Adj R-squared =  0.2857
      Total |  5239.33869 19905  .263217216           Root MSE      =  .43361

------------------------------------------------------------------------------
ln_weekly_~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0130034   .0002787    46.66   0.000     .0124572    .0135496
 years_educ |     .07325   .0011326    64.68   0.000     .0710301    .0754699
       union |   .1386003   .0074886    18.51   0.000      .123922    .1532787
    nonwhite |  -.1633235   .0091623   -17.83   0.000    -.1812823   -.1453647
     region1 |  -.0082481   .0091485    -0.90   0.367      -.02618    .0096837
     region2 |  -.0538852   .0092398    -5.83   0.000    -.0719961   -.0357743
     region3 |  -.0794961   .0088407    -8.99   0.000    -.0968246   -.0621676
       _cons |   4.639453   .0203976   227.45   0.000     4.599473    4.679434
------------------------------------------------------------------------------
```

Coef on Region 1 is negative of the coef on region 4 from previous model. Coef on regions 2 and 3 exactly as we would expect

42

Definitions

- Obesity based on Body Mass Index
- BMI = weight (kg)/(height in cm)^2
- = 703 x weight (pounds)/(height in inches)^2
- BMI < 20 Underweight
- 20 ≤ BMI < 25 Ideal
- 25 ≤ BMI < 30 overweight
- 30 ≤ BMI obese

43

Obesity Rates Over Time

	Obesity		Overweight	
Group	1971/74	1999/00	1971/74	1999/00
All	14.6	30.9	47.7	64.5
Males	12.2	27.7	54.7	67.0
Females	16.8	34.0	41.1	62.0
Black F.	29.7	50.8	60.5	78.0

44

```
Contains data from bmi1.dta
  obs:         1,259
 vars:             9                          26 Sep 2008 09:45
 size:        33,993 (98.6% of memory free)
------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
age             byte   %8.0g                  age in years
sex             byte   %8.0g                  =1 if male, =2 if female
income          int    %8.0g                  annual family income
educ            byte   %8.0g                  years of education
srhealth        byte   %8.0g                  self report
                                                health,1=excel,2=vgood,3=good,4
                                                =fair,5=poor
bmi             float  %9.0g                  body mass index
totalexp        long   %12.0g                 total annual expenditures on
                                                medical care
smoker          byte   %8.0g                  dummy variable, =1 if current
                                                smoker
race            float  %9.0g                  =1 if white non-hisp,=2 if
                                                black nonhisp,=3 other
                                                race,4=hispanic
------------------------------------------------------------------------------
```


- . * generate race dummy variables;
- . gen black=race==2
- . gen other_race=race==3
- . gen hispanic=race==4
- . label var black "=1 of black, non hispanic"
- . label var other_race "=1 if other race, non hispanic"
- . label var hispanic "=1 if hispanic"
- .
- .
- . * generate overweight dummy
- . gen overweight=bmi>=25
- . label var overweight "dummy, =1 if overweight"


```
. * get table of overweight
. tab overweight
    dummy, =1 |
           if |
 overweight |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        377       29.94       29.94
          1 |        882       70.06      100.00
------------+-----------------------------------
      Total |      1,259      100.00
```



```
. reg overweight age educ incomel male black hispanic other_race smoker

      Source |       SS       df       MS              Number of obs =    1259
-------------+------------------------------           F(  8,  1250) =   11.78
       Model |  18.5118546     8  2.31398183           Prob > F      =  0.0000
    Residual |  245.597756  1250  .196478205           R-squared     =  0.0701
-------------+------------------------------           Adj R-squared =  0.0641
       Total |  264.109611  1258  .209944047           Root MSE      =  .44326

------------------------------------------------------------------------------
  overweight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0075824   .0014394     5.27   0.000     .0047585    .0104063
        educ |  -.0129821   .0043338    -3.00   0.003    -.0214843   -.0044798
     incomel |  -.0151473   .0379148    -0.40   0.690     -.089531    .0592364
        male |    -.12196   .0252562    -4.83   0.000    -.1715093   -.0724108
       black |    .173329   .0364559     4.75   0.000     .1018075    .2448504
    hispanic |   .1104625   .0338091     3.27   0.001     .0441337    .1767913
  other_race |  -.0516011   .0513305    -1.01   0.315    -.1523045    .0491022
      smoker |  -.0178954   .0315521    -0.57   0.571    -.0797963    .0440056
       _cons |   .7245841   .3746124     1.93   0.053    -.0103544    1.459523
------------------------------------------------------------------------------
```