Solutions for Session 6

08/11/2022

```
. do solution.do

. global basedir http://personalpages.manchester.ac.uk/staff/mark.lunt

. global datadir $basedir/stats/6_LinearModels2/data

. sysuse auto, clear
(1978 Automobile Data)

. regress weight foreign

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  1,    72) =   39.02
       Model |  15496779.3     1  15496779.3           Prob > F      =  0.0000
    Residual |  28597399.1    72  397186.099           R-squared     =  0.3514
-------------+------------------------------           Adj R-squared =  0.3424
       Total |  44094178.4    73  604029.841           Root MSE      =  630.23

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |  -1001.206   160.2876    -6.25   0.000    -1320.734   -681.6788
       _cons |   3317.115   87.39676    37.95   0.000     3142.893    3491.338
------------------------------------------------------------------------------
```

1.1 foreign vehicles are, on average, 1000 lbs lighter than US vehicles
The difference is significant, p = 0.000

```
. regress weight i.foreign

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  1,    72) =   39.02
       Model |  15496779.3     1  15496779.3           Prob > F      =  0.0000
    Residual |  28597399.1    72  397186.099           R-squared     =  0.3514
-------------+------------------------------           Adj R-squared =  0.3424
       Total |  44094178.4    73  604029.841           Root MSE      =  630.23

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |
     Foreign |  -1001.206   160.2876    -6.25   0.000    -1320.734   -681.6788
       _cons |   3317.115   87.39676    37.95   0.000     3142.893    3491.338
------------------------------------------------------------------------------
```

1.2 This makes no difference at all

```
. ttest weight, by(foreign)
Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Domestic	52	3317.115	96.4296	695.3637	3123.525	3510.706
Foreign	22	2315.909	92.31665	433.0035	2123.926	2507.892
combined	74	3019.459	90.34692	777.1936	2839.398	3199.521
diff		1001.206	160.2876		681.6788	1320.734

```
    diff = mean(Domestic) - mean(Foreign)                         t =   6.2463
Ho: diff = 0                                     degrees of freedom =       72

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

1.3 the mean difference and standard error are exactly the same (except for the minus sign)

```
. graph box weight, over(foreign)


. graph export graph1.eps replace
(file graph1.eps written in EPS format)
```

1.4 There is a wider spread of weights for Domestic cars compared to Foreign cars, i.e. greater variance

```
. by foreign: summ weight
```

-> foreign = Domestic

Variable	Obs	Mean	Std. Dev.	Min	Max
weight	52	3317.115	695.3637	1800	4840

-> foreign = Foreign

Variable	Obs	Mean	Std. Dev.	Min	Max
weight	22	2315.909	433.0035	1760	3420

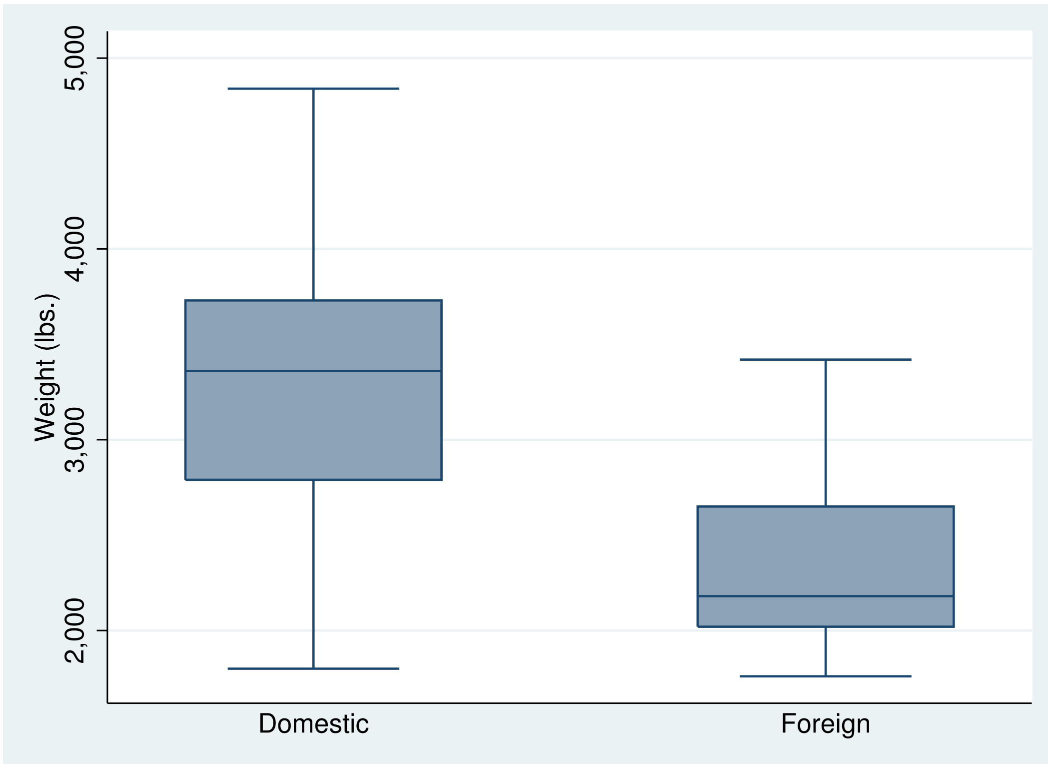

Figure 1: . graph box weight, over(foreign)

1.5 the SD is much higher for Domestic (~700) compared to Foreign (~430)

```
. hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of weight

         chi2(1)      =     4.51
         Prob > chi2  =   0.0337
```

1.6 The difference in variance is significant. Therefore, a linear model is inappropriate

```
. use $datadir/soap, clear

. graph box appearance, over(operator)
```

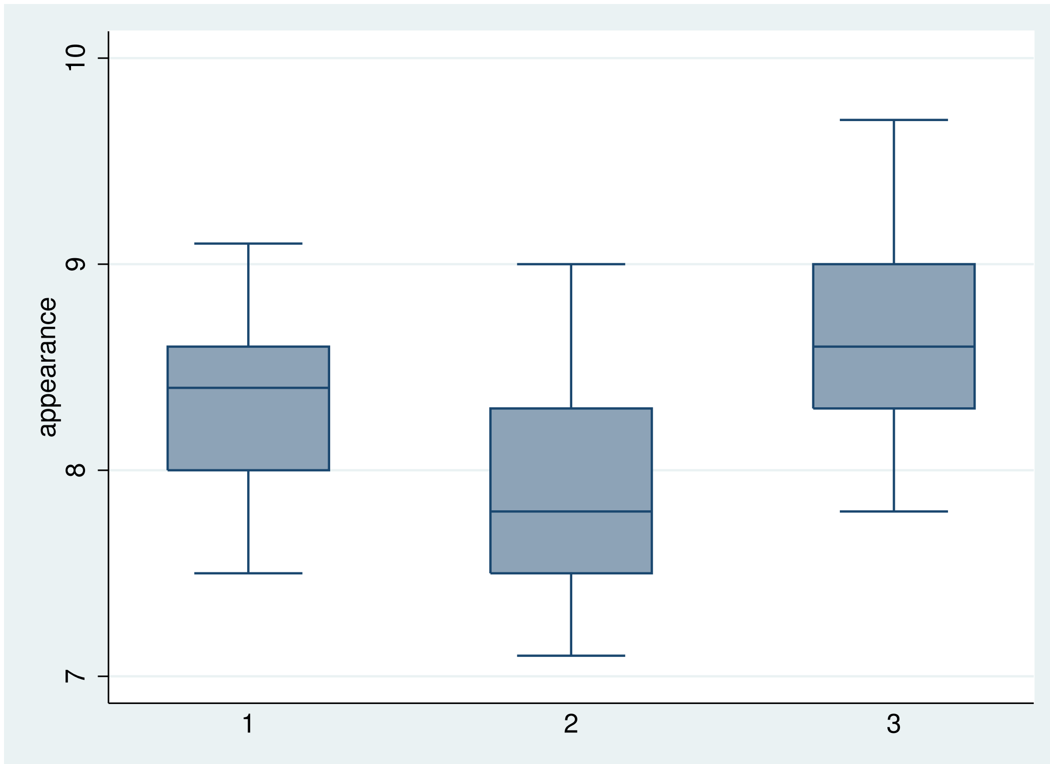

Figure 2: . graph box appearance, over(operator)

```
. graph export graph2.eps replace
(file graph2.eps written in EPS format)
```

1.7 Operator 3 has the highest scores: 25% of scores are above 9

```
. sort operator
```

```
. by operator: summm appearance

-------------------------------------------------------------------------------
-> operator = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  appearance |        30    8.306667    .4630732        7.5        9.1

-------------------------------------------------------------------------------
-> operator = 2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  appearance |        30    7.896667    .4766863        7.1          9

-------------------------------------------------------------------------------
-> operator = 3

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
  appearance |        30    8.626667    .4653018        7.8        9.7


. regress appearance i.operator

      Source |       SS       df       MS              Number of obs =      90
-------------+------------------------------           F(  2,    87) =   18.31
       Model |  8.03400033     2  4.01700016           Prob > F      =  0.0000
    Residual |  19.0869988    87  .219390791           R-squared     =  0.2962
-------------+------------------------------           Adj R-squared =  0.2800
       Total |  27.1209991    89  .304730327           Root MSE      =  .46839

------------------------------------------------------------------------------
  appearance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    operator |
          2  |       -.41   .1209382    -3.39   0.001    -.6503778   -.1696222
          3  |   .3200001   .1209382     2.65   0.010     .0796223    .5603779
             |
       _cons |   8.306667   .0855162    97.14   0.000     8.136694    8.476639
------------------------------------------------------------------------------
```

1.9 Yes: Prob > F = 0.0000 is testing the null hypothesis that all operators are the same.
1.10 p= 0.0000
1.11 Operator 1 is the baseline: there is no line for operator 1

```
. lincom _cons + 2.operator

 ( 1)  2.operator + _cons = 0

------------------------------------------------------------------------------
  appearance |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   7.896667   .0855162    92.34   0.000     7.726694    8.066639
------------------------------------------------------------------------------
```

1.12 This is the same as we have already seen

```
. lincom 2.operator - 3.operator
 ( 1)  2.operator - 3.operator = 0
```

appearance	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
(1)	-.73	.1209382	-6.04	0.000	-.9703778	-.4896222

1.13 Yes: t = -6.04, p= 0.000

```
. use $datadir/cadmium, clear

. scatter capacity age
```

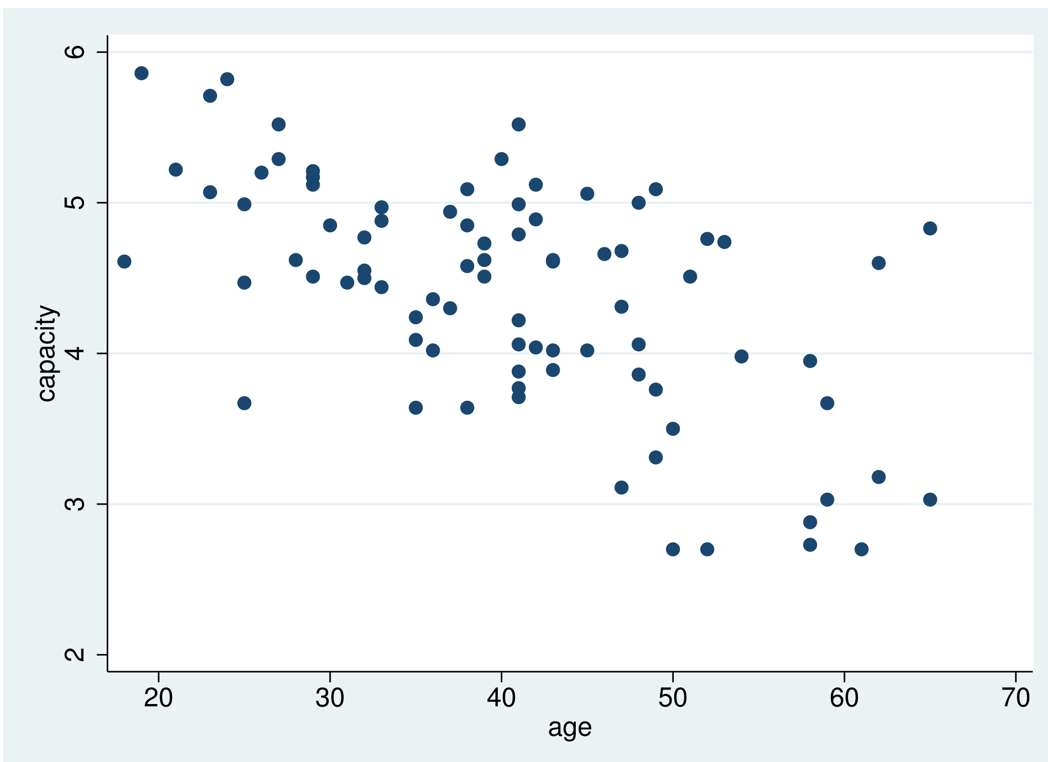

Figure 3: . scatter capacity age

```
. graph export graph3.eps replace
(file graph3.eps written in EPS format)
```

```
. regress capacity age
      Source |       SS       df       MS              Number of obs =      84
-------------+------------------------------           F(  1,    82) =   47.37
       Model |  17.4445864     1  17.4445864           Prob > F      =  0.0000
    Residual |  30.1963679    82  .368248388           R-squared     =  0.3662
-------------+------------------------------           Adj R-squared =  0.3584
       Total |  47.6409543    83  .573987401           Root MSE      =  .60683

------------------------------------------------------------------------------
    capacity |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0404781   .0058811    -6.88   0.000    -.0521776   -.0287787
       _cons |   6.033316    .247487    24.38   0.000     5.540986    6.525647
------------------------------------------------------------------------------
```

2.2 The regression coefficient for age is negative, showing that capacity decreases as age increases.

```
. gen cap1 = capacity if exposure == 1
(40 missing values generated)

. gen cap2 = capacity if exposure == 2
(56 missing values generated)

. gen cap3 = capacity if exposure == 3
(72 missing values generated)

. scatter cap1 cap2 cap3 age


. graph export graph4.eps replace
(file graph4.eps written in EPS format)

. regress capacity i.exposure
      Source |       SS       df       MS              Number of obs =      84
-------------+------------------------------           F(  2,    81) =    2.48
       Model |  2.74733751     2  1.37366875           Prob > F      =  0.0902
    Residual |  44.8936168    81  .554242182           R-squared     =  0.0577
-------------+------------------------------           Adj R-squared =  0.0344
       Total |  47.6409543    83  .573987401           Root MSE      =  .74447

------------------------------------------------------------------------------
    capacity |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exposure |
 < 10 years  |   .0097403   .1799744     0.05   0.957    -.3483523    .3678329
 > 10 years  |  -.5128788   .2424526    -2.12   0.037    -.9952834   -.0304741
             |
       _cons |   4.462045   .1122337    39.76   0.000     4.238735    4.685355
------------------------------------------------------------------------------
```

2.3 Its borderline, p = 0.09

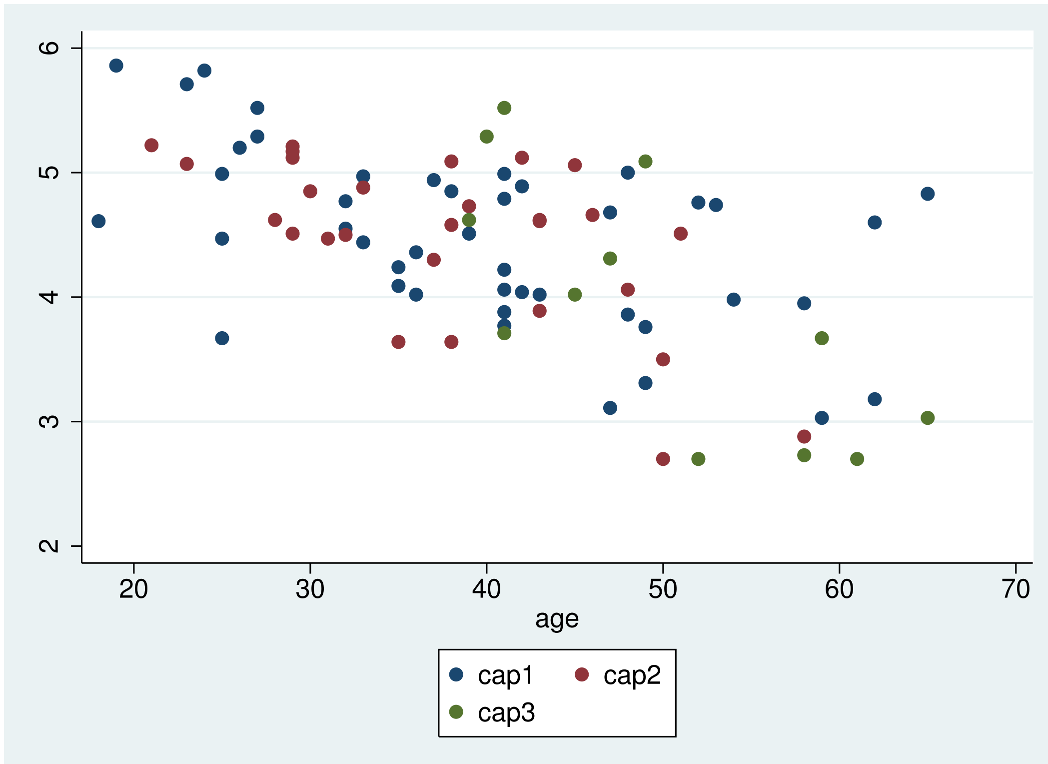

Figure 4: . scatter cap1 cap2 cap3 age

```
. regress capacity age i.exposure
      Source |       SS           df       MS      Number of obs   =        84
-------------+----------------------------------   F(  3,     80)  =     15.63
       Model |  17.6062849         3  5.86876164   Prob > F        =    0.0000
    Residual |  30.0346693        80  .375433367   R-squared       =    0.3696
-------------+----------------------------------   Adj R-squared   =    0.3459
       Total |  47.6409543        83  .573987401   Root MSE        =    .61273

------------------------------------------------------------------------------
    capacity |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0397752   .0063224    -6.29   0.000    -.0523572   -.0271931
             |
    exposure |
  < 10 years |  -.0701975   .1486686    -0.47   0.638    -.3660575    .2256626
  > 10 years |  -.1169349   .2092361    -0.56   0.578    -.5333281    .2994582
             |
       _cons |   6.044917   .2680248    22.55   0.000      5.51153    6.578303
------------------------------------------------------------------------------
```

```
. testparm i.exposure
 ( 1)  2.exposure = 0
 ( 2)  3.exposure = 0
       F(  2,    80) =    0.22
            Prob > F =    0.8067
```

2.4 There are now no significant differences between groups

```
. predict ppred, xb

. gen ppred1 = ppred if exposure == 1
(40 missing values generated)

. gen ppred2 = ppred if exposure == 2
(56 missing values generated)

. gen ppred3 = ppred if exposure == 3
(72 missing values generated)

. scatter cap1 cap2 cap3 age || line ppred1 age || line ppred2 age || /* */
 line ppred3 age


. graph export graph5.eps replace
(file graph5.eps written in EPS format)

. regress capacity i.exposure##c.age
      Source |       SS       df       MS              Number of obs =      84
-------------+------------------------------           F(  5,    78) =   11.39
       Model |  20.1057424     5  4.02114849           Prob > F      =  0.0000
    Residual |  27.5352118    78  .353015536           R-squared     =  0.4220
-------------+------------------------------           Adj R-squared =  0.3850
       Total |  47.6409543    83  .573987401           Root MSE      =  .59415

------------------------------------------------------------------------------
    capacity |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exposure |
  < 10 years |   .5497403   .5758844     0.95   0.343    -.5967574    1.696238
  > 10 years |   2.503148   1.041842     2.40   0.019     .4289997    4.577296
             |
         age |  -.0306127   .0075475    -4.06   0.000    -.0456385   -.0155868
             |
exposure#c.age |
  < 10 years |  -.0159193   .0145469    -1.09   0.277    -.0448799    .0130413
  > 10 years |  -.0544983   .0210698    -2.59   0.012    -.0964451   -.0125516
             |
       _cons |   5.680291    .313426    18.12   0.000     5.056307    6.304274
------------------------------------------------------------------------------

. testparm i.exposure#c.age
 ( 1)  2.exposure#c.age = 0
 ( 2)  3.exposure#c.age = 0
       F(  2,    78) =    3.54
            Prob > F =    0.0338
```

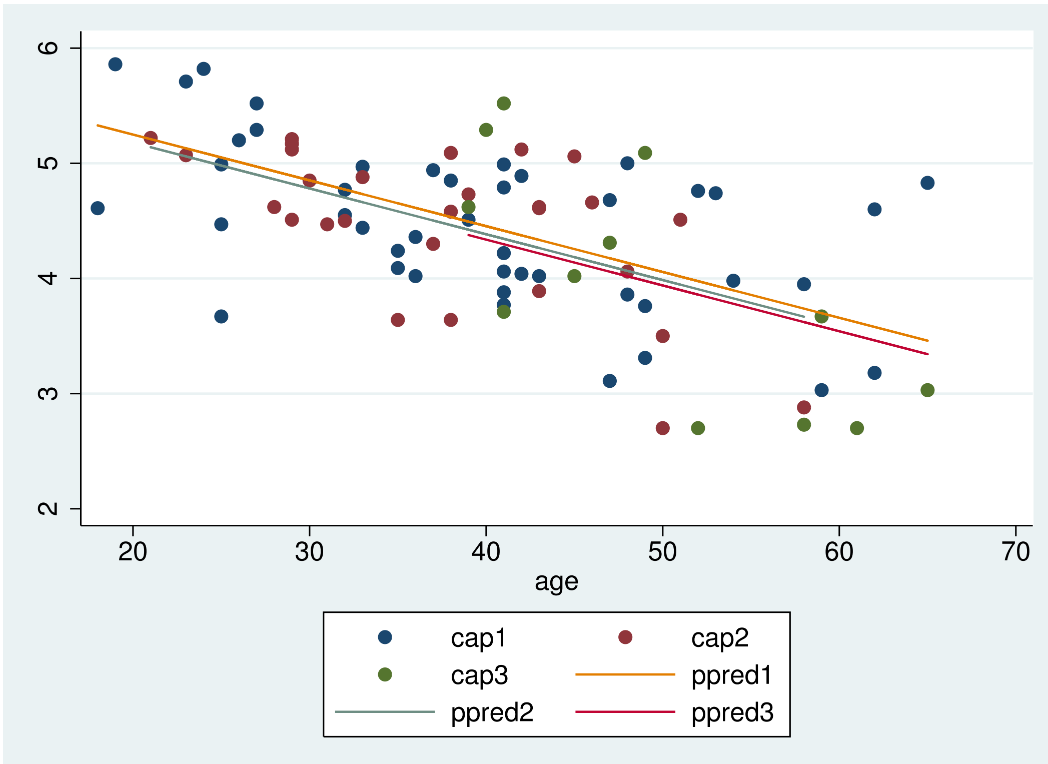

Figure 5: . scatter cap1 cap2 cap3 age —— line ppred1 age —— line ppred2 age —— /*

2.5 Yes, the slopes in the different exposure groups are different

```
. predict ipred, xb

. gen ipred1 = ipred if exposure == 1
(40 missing values generated)

. gen ipred2 = ipred if exposure == 2
(56 missing values generated)

. gen ipred3 = ipred if exposure == 3
(72 missing values generated)

. scatter cap1 cap2 cap3 age || line ipred1 age || line ipred2 age || /* */
 line ipred3 age
```

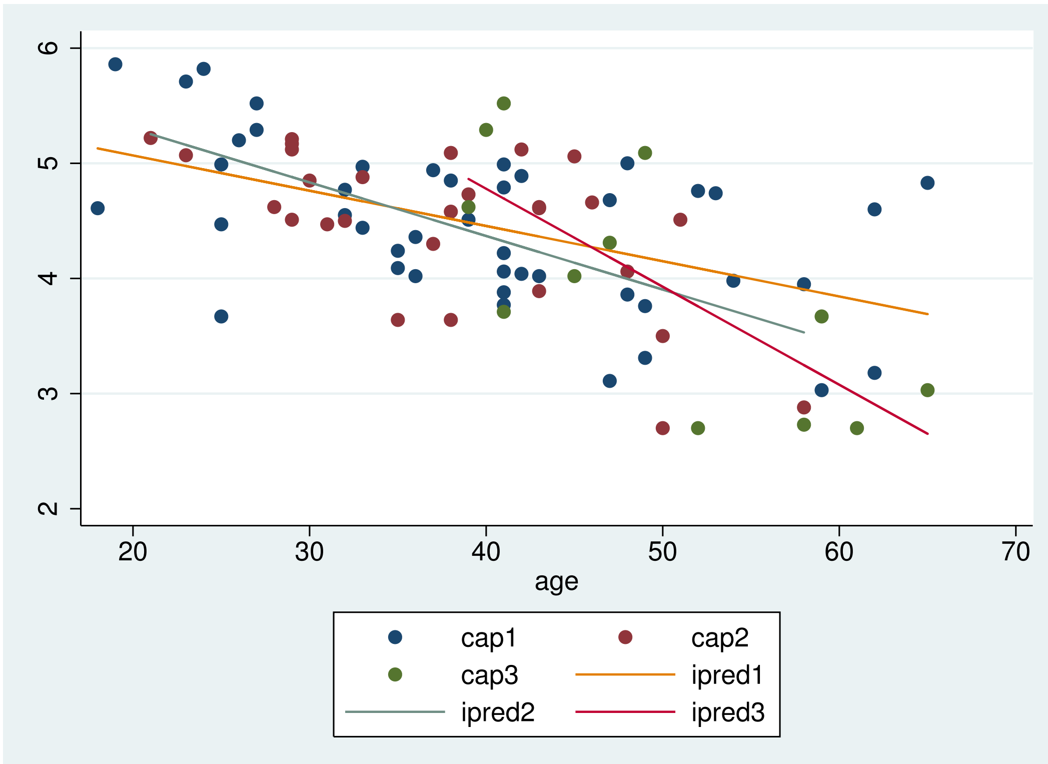

Figure 6: . scatter cap1 cap2 cap3 age —— line ipred1 age —— line ipred2 age —— /*

```
. graph export graph6.eps replace
(file graph6.eps written in EPS format)
```

2.6 The least steep is in the baseline (least exposed group)
The steepest is in the most exposed group

```
. lincom age + 3.exposure#c.age
 ( 1)  age + 3.exposure#c.age = 0
```

capacity	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
(1)	-.085111	.0196716	-4.33	0.000	-.1242742	-.0459478

```
. use $datadir/hald, clear
```

```
. sw regress y x1 x2 x3 x4, pe(0.05)
                      begin with empty model
p = 0.0006 <  0.0500  adding   x4
p = 0.0000 <  0.0500  adding   x1
      Source |       SS       df       MS              Number of obs =      13
-------------+------------------------------           F(  2,    10) =  176.63
       Model |  2641.00094     2  1320.50047           Prob > F      =  0.0000
    Residual |  74.7621108    10  7.47621108           R-squared     =  0.9725
-------------+------------------------------           Adj R-squared =  0.9670
       Total |  2715.76305    12  226.313587           Root MSE      =  2.7343

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x4 |  -.6139536   .0486446   -12.62   0.000    -.7223404   -.5055668
          x1 |   1.439958   .1384166    10.40   0.000     1.131547     1.74837
       _cons |   103.0974   2.123984    48.54   0.000     98.36485    107.8299
------------------------------------------------------------------------------
```

3.1 x1 & x4 are retained

```
. sw regress y x1 x2 x3 x4, pr(0.05)
                      begin with full model
p = 0.8959 >= 0.0500  removing x3
p = 0.2054 >= 0.0500  removing x4
      Source |       SS       df       MS              Number of obs =      13
-------------+------------------------------           F(  2,    10) =  229.50
       Model |  2657.85857     2  1328.92929           Prob > F      =  0.0000
    Residual |  57.9044793    10  5.79044793           R-squared     =  0.9787
-------------+------------------------------           Adj R-squared =  0.9744
       Total |  2715.76305    12  226.313587           Root MSE      =  2.4063

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.468306   .1213009    12.10   0.000      1.19803    1.738581
          x2 |   .6622505   .0458547    14.44   0.000     .5600798    .7644212
       _cons |   52.57735   2.286174    23.00   0.000     47.48344    57.67126
------------------------------------------------------------------------------
```

3.2 This time x1 & x2 are retained

```
. sw regress y x1 x2 x3 x4, pe(0.05) pr(0.0500005)
                      begin with full model
p = 0.8959 >= 0.0500  removing x3
p = 0.2054 >= 0.0500  removing x4
      Source |       SS       df       MS              Number of obs =      13
-------------+------------------------------           F(  2,    10) =  229.50
       Model |  2657.85857     2  1328.92929           Prob > F      =  0.0000
    Residual |  57.9044793    10  5.79044793           R-squared     =  0.9787
-------------+------------------------------           Adj R-squared =  0.9744
       Total |  2715.76305    12  226.313587           Root MSE      =  2.4063

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.468306   .1213009    12.10   0.000      1.19803    1.738581
          x2 |   .6622505   .0458547    14.44   0.000     .5600798    .7644212
       _cons |   52.57735   2.286174    23.00   0.000     47.48344    57.67126
------------------------------------------------------------------------------
```

3.3 This is the same as the backwards model

```
. corr x*
(obs=13)

             |       x1       x2       x3       x4
-------------+------------------------------------
          x1 |   1.0000
          x2 |   0.2286   1.0000
          x3 |  -0.8241  -0.1392   1.0000
          x4 |  -0.2454  -0.9730   0.0295   1.0000
```

3.4 Correlation between x2 & x4 is -0.97
3.5 x2 & x4 are very strongly correlated: they contain the same information, so they are largely interchangeable

```
. regress y x1 x2 x3 x4
      Source |       SS       df       MS              Number of obs =      13
-------------+------------------------------           F(  4,     8) =  111.48
       Model |  2667.89941     4  666.974853           Prob > F      =  0.0000
    Residual |   47.863637     8  5.98295463           R-squared     =  0.9824
-------------+------------------------------           Adj R-squared =  0.9736
       Total |  2715.76305    12  226.313587           Root MSE      =   2.446

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          x1 |   1.551103   .7447698     2.08   0.071    -.1663395    3.268545
          x2 |   .5101677    .723788     0.70   0.501     -1.15889    2.179226
          x3 |   .1019096    .754709     0.14   0.896    -1.638453    1.842272
          x4 |  -.1440609    .709052    -0.20   0.844    -1.779138    1.491016
       _cons |   62.40535   70.07096     0.89   0.399    -99.17856    223.9893
------------------------------------------------------------------------------
```

3.6 The F statistic says that the model is very highly significant: the null hypothesis that all coefficients are 0 could not have given rise to this data
3.7 98% of the variance is explained
3.8 None of the coefficients are significant, due to the strong correlations between them

```
. use $datadir/growth, clear

. scatter weight week
```

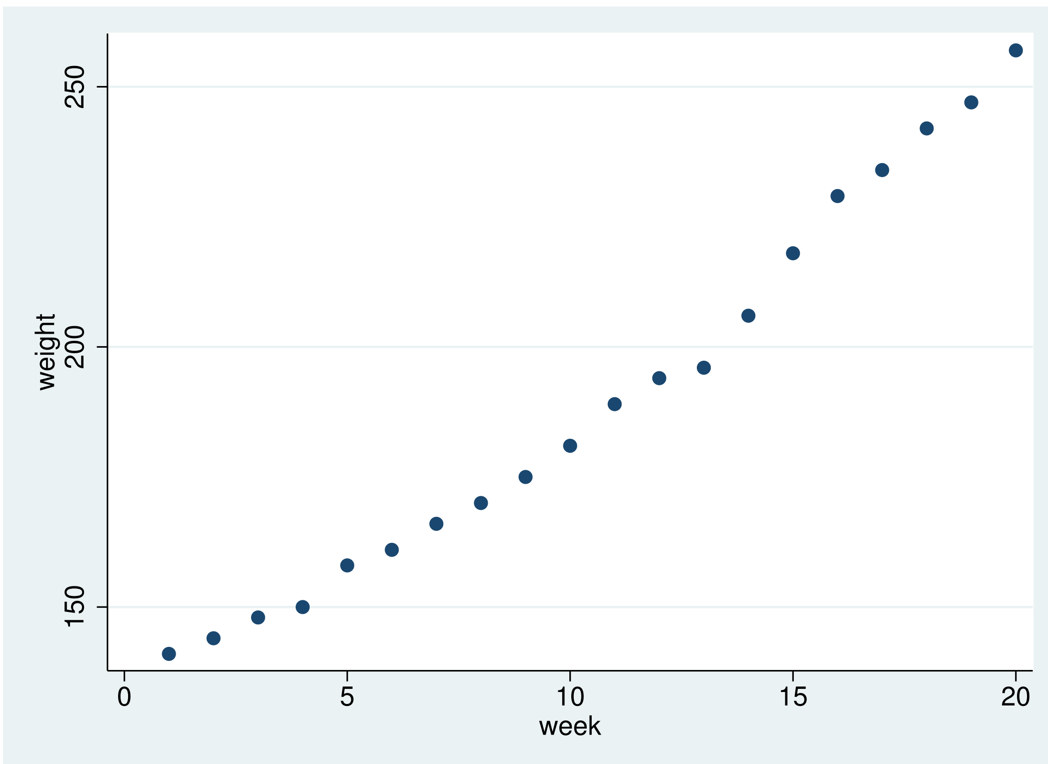

Figure 7: . scatter weight week

```
. graph export graph7.eps replace
(file graph7.eps written in EPS format)
```

4.1 The line does not look quite straight: there appears to be some curvature

```
. regress weight week
      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  1,    18) =  790.23
       Model |  25438.7504     1  25438.7504           Prob > F      =  0.0000
    Residual |  579.449624    18  32.1916458           R-squared     =  0.9777
-------------+------------------------------           Adj R-squared =  0.9765
       Total |     26018.2    19  1369.37895           Root MSE      =  5.6738

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        week |   6.184962   .2200193    28.11   0.000     5.722719    6.647206
       _cons |   125.3579   2.635644    47.56   0.000     119.8206    130.8952
------------------------------------------------------------------------------

. cprplot week
```

4.2 There is definitely curvature around the line

```
. gen week2 = week * week

. regress weight week week2
      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  2,    17) = 2436.58
       Model |  25927.7513     2  12963.8756           Prob > F      =  0.0000
    Residual |  90.4487127    17  5.32051251           R-squared     =  0.9965
-------------+------------------------------           Adj R-squared =  0.9961
       Total |     26018.2    19  1369.37895           Root MSE      =  2.3066

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        week |   2.680178   .3763642     7.12   0.000     1.886119    3.474237
       week2 |   .1668945   .0174086     9.59   0.000     .1301656    .2036235
       _cons |   138.2088   1.716086    80.54   0.000     134.5881    141.8294
------------------------------------------------------------------------------
```

4.3 week2 is very highly significant (p = 0.000)

```
. predict pred2, xb

. twoway scatter weight week || line pred2 week

. graph export graph8.eps replace
(file graph8.eps written in EPS format)
```

4.4 Curved predictor fits the data very well

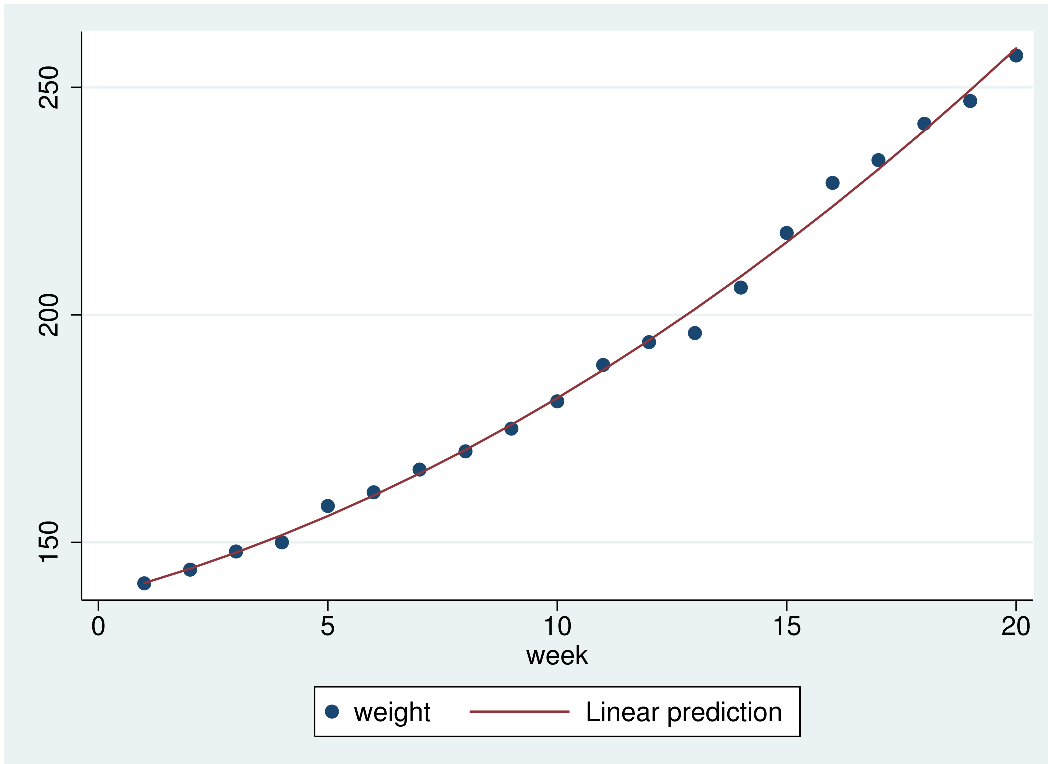

Figure 8: . twoway scatter weight week —— line pred2 week

```
. gen week3 = week2*week

. regress weight week week2 week3
      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  3,    16) = 1548.58
       Model |  25928.9007     3  8642.96691           Prob > F      =  0.0000
    Residual |  89.2992705    16  5.58120441           R-squared     =  0.9966
-------------+------------------------------           Adj R-squared =  0.9959
       Total |     26018.2    19  1369.37895           Root MSE      =  2.3625

------------------------------------------------------------------------------
      weight |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        week |   2.242641   1.038333     2.16   0.046     .0414737    4.443808
       week2 |   .2177334   .1134353     1.92   0.073    -.0227388    .4582055
       week3 |  -.0016139   .0035564    -0.45   0.656    -.0091531    .0059252
       _cons |   139.0663   2.580587    53.89   0.000     133.5957    144.5369
------------------------------------------------------------------------------
```

4.5 week3 is not significant

```
. corr week*
(obs=20)

             |     week    week2    week3
-------------+---------------------------
        week |   1.0000
       week2 |   0.9713   1.0000
       week3 |   0.9221   0.9865   1.0000
```

4.6 Correlation between week and week2 is 0.97
end of do-file