Unit 9 – STATA for Normal Theory Regression
Homework
Solutions

Before you Begin
Download from the course website
companies.dta
hersdata_small.dta

Description of companies.dta

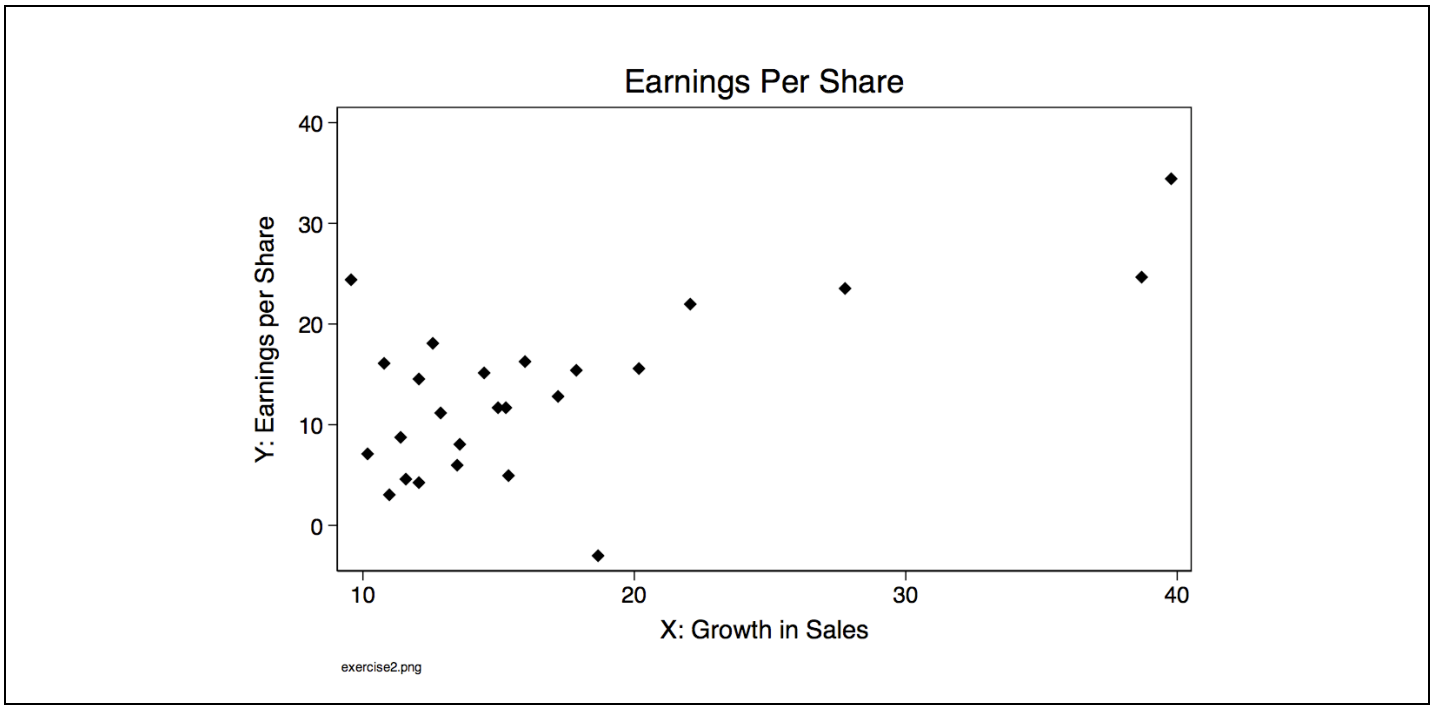
This data set contains 30 observations. It is from a study reported in the January 1981 issue of *Forbes* magazine of the characteristics of the 30 largest chemical companies. The two variables that we will use in a simple linear regression analysis are the following:
eps5 and **salesgr5**

Codebook

Variable Name	Variable Coding	Description
salesgr5	Continuous	**Independent (X)**: Per cent annual compound growth rate of sales, computed from the most recent five years compared with the previous five years.
eps5	Continuous	**Dependent (Y)**: Per cent annual compound growth in earnings per share, computed from the most recent five years compared with the preceding five years.

Simple Linear Regression Using companies.dta

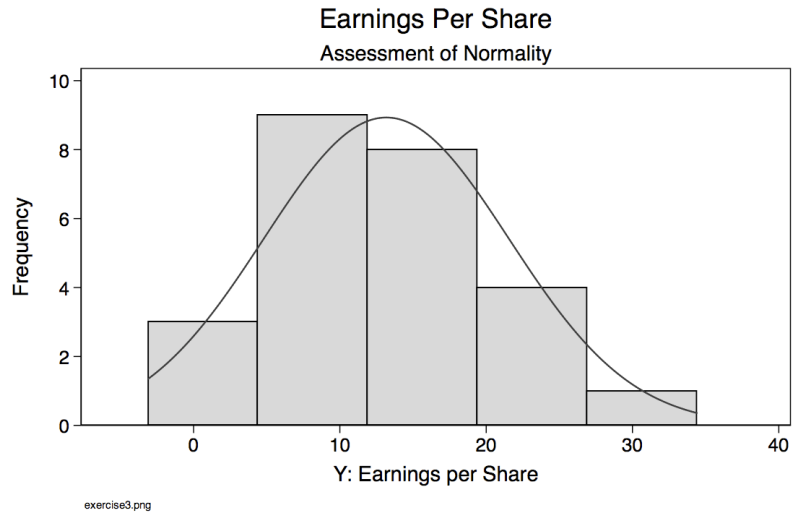
```
. ***** 1).  Label variables
. label variable eps5 "Y: Earnings per Share"
. label variable salesgr5 "X: Growth in Sales"


. *
. ***** 2)  Scatterplot
. graph twoway (scatter eps5 salesgr5, symbol(d)), title("Earnings Per Share")
caption("exercise2.png", size(vsmall))
```



Intepretation – Scatter plot suggests a linear relationship. It also suggests some sparseness of data for large values of X: Growth in Sales.

```
. *
```

```
. ***** 3)  Graphical Assessment of Normality of Y = eps5
. histogram eps5, normal frequency title("Earnings Per Share") subtitle("Assessment of
Normality") caption("exercise3.png", size(vsmall))
(bin=5, start=-3.0999999, width=7.5000003)
```



Interpretation: The histogram with overlay normal suggests reasonableness of the assumption of normality of Y.

```
. *
. ***** 4) Hypothesis Test of Normality of Y
. swilk eps5

                  Shapiro-Wilk W test for normal data

    Variable |    Obs       W           V         z       Prob>z
-------------+--------------------------------------------------
        eps5 |     25    0.97412      0.719    -0.674     0.74974


. sfrancia eps5
                  Shapiro-Francia W' test for normal data

    Variable |    Obs       W'          V'        z       Prob>z
-------------+--------------------------------------------------
        eps5 |     25    0.96853      0.971    -0.053     0.52129
```

Interpretation – These test results are consistent with the histogram and overlay normal. Both the Shapiro-Wilk and Francia tests of the null hypothesis of normality fail to reject (p-values: .75 and .52, respectively). Thus, we may assume that the assumption of normality of Y is reasonably satisfied for these data.

```
. *
```

```
. ***** 5) Estimate the straight line regression
. regress eps5 salesgr5

      Source |       SS       df       MS              Number of obs =      25
-------------+------------------------------           F(  1,    23) =   16.18
       Model |  695.264495     1  695.264495           Prob > F      =  0.0005
    Residual |  988.101157    23  42.9609199           R-squared     =  0.4130
-------------+------------------------------           Adj R-squared =  0.3875
       Total |  1683.36565    24  70.1402355           Root MSE      =  6.5545

------------------------------------------------------------------------------
        eps5 |      Coef.       Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    salesgr5 |    .6792279 = b1  .1688408     4.02   0.001      .3299542    1.028502
       _cons |    1.764971 = b0  3.124789     0.56   0.578    -4.699148    8.229091
------------------------------------------------------------------------------
```

Intepretation: The fitted simple linear regression model is:

eps5 = 1.76 + 0.679 * salesgr5

For each one percentage increase in annual compound growth of sales, it is estimated that there is a 0.679 percentage increase in annual compound growth in earnings per share

. *** 6a) How much of the variability in Y is explained by the model?**

This is shown at right, as R-squared = 0.4130 which says that 41.3% of the variability in Y is explained by the fitted model

. *** 6b) Overall F-Test**

This is shown at right, as Prob > F = 0.0005 which says that the p-value for the overall F test of the null hypothesis that the fitted simple linear regression has slope=0 is 5 chances in 10,000. The null hypothesis of zero slope has led to a very unlikely outcome prompting statistical rejection of the null hypothesis. Conclude that the fitted line explains statistically significantly more of the variability in Y=eps5 than is explained by the fit of no model.
Reminder - The "fit of no model" is really saying that the model fit is instead the average of Y.

. *** 7a) Test of Zero Slope**

This is shown in the coefficients table, as P>|t|= 0.001 which says that the p-value for the Student t-test of the null hypothesis of zero slope is 1 chance in 1,000. Again, the null hypothesis has led to a very unlikely outcome, prompting statistical rejection of the null hypothesis. Conclude that the fitted line explains statistically significantly more of the variability in Y=eps5 than is explained by the average of Y (no model).

NOTE - In theory, in simple linear regression, the F-test for the overall regression is equivalent to the Student t-test for zero slope, with [Student t]^2 = [Overall F]. The reason the p-values do not match exactly is the result of rounding.

. *** 7b) Confidence Interval Estimate of the Slope**

This is shown in the coefficients table, as (.3299542 , 1.028502)

Multiple Linear Regression Using hersdata_small.dta

Description of hersdata_small.dta

These data are a simple random sample of 1000 observations from the HERS study called *hersdata_small.dta*. The HERS study was a randomized clinical trial of hormone therapy (estrogen plus progestin) for the reduction of cardiovascular disease risk in post-menopausal women with established coronary disease. Study participants were n=2,763 women who were: (1) post-menopausal (2) with coronary disease; and (3) with an intact uterus. The data set for this homework is a simple random sample of n=1000.

The following variables are considered (*Tip* – Remember that Stata is case sensitive)

```
. summarize age BMI drinkany exercise glucose HT LDL physact statins
```

Variable	Obs	Mean	Std. Dev.	Min	Max
age	1000	66.701	6.575168	45	79
BMI	999	28.31502	5.45297	15.21	54.13
drinkany	999	.4134134	.4926924	0	1
exercise	1000	.393	.4886612	0	1
glucose	1000	111.214	35.46136	29	298
HT	1000	.508	.5001862	0	1
LDL	997	144.6558	36.84347	44.4	393.4
physact	1000	3.219	1.091893	1	5
statins	1000	.365	.4816711	0	1

Exercises # 8-13
Multiple Linear Regression Analysis of Y=Glucose, among non-diabetics only

```
. *
. ***** 8).  Single predictor model of Y=glucose to X=exercise among non-diabetics only
. * preliminary:  tabulation of diabetes
. fre diabetes

diabetes -- diabetes
```

		Freq.	Percent	Valid	Cum.
Valid	0 no	749	74.90	74.90	74.90
	1 yes	251	25.10	25.10	100.00
	Total	1000	100.00	100.00	

```
. regress glucose exercise if diabetes==0

      Source |       SS       df       MS              Number of obs =     749
-------------+------------------------------           F(  1,   747) =    6.74
       Model |  606.158226     1  606.158226            Prob > F      =  0.0096
    Residual |  67185.8204   747  89.9408573           R-squared     =  0.0089
-------------+------------------------------           Adj R-squared =  0.0076
       Total |  67791.9786   748  90.6309875           Root MSE      =  9.4837

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exercise |  -1.823973   .7025933    -2.60   0.010    -3.203266    -.444681
       _cons |   97.75688   .4541876   215.23   0.000     96.86524    98.64852
------------------------------------------------------------------------------
```

(1) The fitted line is Y = 97.76 – (1.82)*exercise for Y=glucose

(2) R-squared = .0089 tells us that less than 1% of the variation in glucose is explained.

(3) Even so, the p-value of .0096 for the overall F test (value=6.74) tells us that this fitted model performs better than the null model of using the average Y only.

```
. *
. *****  9) Multiple predictor model among non-diabetics only
. regress glucose exercise age drinkany BMI if diabetes==0

      Source |       SS       df       MS              Number of obs =     748
-------------+------------------------------           F(  4,   743) =   14.79
       Model |  4950.25806     4  1237.56452            Prob > F      =  0.0000
    Residual |  62164.5387   743  83.6669431           R-squared     =  0.0738
-------------+------------------------------           Adj R-squared =  0.0688
       Total |  67114.7968   747  89.8457788           Root MSE      =   9.147

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    exercise |  -1.090857   .6881622    -1.59   0.113    -2.441831     .260117
         age |   .0532612   .0508681     1.05   0.295    -.0466012    .1531236
    drinkany |  -.4022234   .6757454    -0.60   0.552    -1.728821    .9243744
         BMI |   .4660455   .0661306     7.05   0.000     .3362205    .5958705
       _cons |   81.20638   4.116258    19.73   0.000      73.1255    89.28726
------------------------------------------------------------------------------
```

(1) The fitted line is now:
Y = 81.21 – (1.09)*exercise + (0.05)*age – (0.40)*drinkany + (0.47)*BMI

(2) R-squared = .0738 says that 7.4% of the variation in glucose is explained by this model.

(3) The p-value of the overall F-test (value=14.79 is < .00001. So again, the fitted model performs better than the null model of using the average Y only.

```
. *
. *****  10) Partial F-test for exercise controlling for:  age, drinkany, BMI
. testparm exercise

 ( 1)  exercise = 0

       F(  1,   743) =    2.51
            Prob > F =    0.1134
```

The partial F-test (null hypothesis: exercise is not significant in the adjusted model) has p-value = .11. Do NOT reject. Conclude that, controlling for age, drink and BMI, exercise is not statistically significant for the prediction of glucose.

```
. *
. *****  11) Create 5 0/1 design variables for the 5 outcomes of physact.
. tab physact, gen(physact)

   comparative physical |
               activity |      Freq.     Percent        Cum.
------------------------+-----------------------------------
    1. much less active |         72        7.20        7.20
2. somewhat less active |        179       17.90       25.10
     3. about as active |        322       32.20       57.30
4. somewhat more active |        312       31.20       88.50
    5. much more active |        115       11.50      100.00
------------------------+-----------------------------------
                  Total |      1,000      100.00
```

TIP - There is more than one way to do this. One is to use the command tab with the option gen( ) to produce a 0/1 dummy variable for each level of a discrete variable. Stata names the 0/1 dummy variables same variable name with a suffix for each level. For physact with 5 levels, Stata produced physact1, physact2, physact3, physact4, and physact5. Don't see them? To confirm, look in the variables window. The new variables should appear at the bottom.

```
. ******   Check.
. tab2 physact1 physact
-> tabulation of physact1 by physact

physact==m |
  uch less |             comparative physical activity
    active | much less  somewhat   about as   somewhat   much more |     Total
-----------+-------------------------------------------------------+----------
         0 |         0        179        322        312        115 |       928
         1 |        72          0          0          0          0 |        72
-----------+-------------------------------------------------------+----------
     Total |        72        179        322        312        115 |     1,000
```

```
. tab2 physact2 physact
-> tabulation of physact2 by physact
```

physact==s omewhat less active	comparative physical activity: much less	somewhat	about as	somewhat	much more	Total
0	72	0	322	312	115	821
1	0	179	0	0	0	179
Total	72	179	322	312	115	1,000

```
. tab2 physact3 physact
-> tabulation of physact3 by physact
```

physact==a bout as active	comparative physical activity: much less	somewhat	about as	somewhat	much more	Total
0	72	179	0	312	115	678
1	0	0	322	0	0	322
Total	72	179	322	312	115	1,000

```
. tab2 physact4 physact
-> tabulation of physact4 by physact
```

physact==s omewhat more active	comparative physical activity: much less	somewhat	about as	somewhat	much more	Total
0	72	179	322	0	115	688
1	0	0	0	312	0	312
Total	72	179	322	312	115	1,000

```
. tab2 physact5 physact
-> tabulation of physact5 by physact
```

physact==m uch more active	comparative physical activity: much less	somewhat	about as	somewhat	much more	Total
0	72	179	322	312	0	885
1	0	0	0	0	115	115
Total	72	179	322	312	115	1,000

```
. *****   12) Multiple predictor model of Y=glucose using explicitly defined dummies.
regress glucose physact2 physact3 physact4 physact5 if diabetes==0

      Source |       SS       df       MS              Number of obs =     749
-------------+------------------------------           F(  4,   744) =    4.33
       Model |  1540.65437     4  385.163593            Prob > F      =  0.0018
    Residual |  66251.3243   744  89.0474788            R-squared     =  0.0227
-------------+------------------------------           Adj R-squared =  0.0175
       Total |  67791.9786   748  90.6309875            Root MSE      =  9.4365

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    physact2 |  -2.333766   1.694661    -1.38   0.169    -5.660654    .9931214
    physact3 |  -2.878975   1.546046    -1.86   0.063    -5.914106    .1561562
    physact4 |  -5.177782   1.538707    -3.37   0.001    -8.198506   -2.157057
    physact5 |  -4.162338   1.712441    -2.43   0.015    -7.524129   -.8005461
       _cons |   100.5909   1.422605    70.71   0.000     97.79811    103.3837
------------------------------------------------------------------------------
```

Don't forget!! When fitting a nominal predictor variable having k levels, include only (k-1) of the dummy variables that you have created. The omitted dummy variable is your 'referent'.

```
. *****  13) Multiple predictor model using the xi: prefix
. xi: regress glucose i.physact if diabetes==0

i.physact         _Iphysact_1-5       (naturally coded; _Iphysact_1 omitted)

      Source |       SS       df       MS              Number of obs =     749
-------------+------------------------------           F(  4,   744) =    4.33
       Model |  1540.65437     4  385.163593            Prob > F      =  0.0018
    Residual |  66251.3243   744  89.0474788            R-squared     =  0.0227
-------------+------------------------------           Adj R-squared =  0.0175
       Total |  67791.9786   748  90.6309875            Root MSE      =  9.4365

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 _Iphysact_2 |  -2.333766   1.694661    -1.38   0.169    -5.660654    .9931214
 _Iphysact_3 |  -2.878975   1.546046    -1.86   0.063    -5.914106    .1561562
 _Iphysact_4 |  -5.177782   1.538707    -3.37   0.001    -8.198506   -2.157057
 _Iphysact_5 |  -4.162338   1.712441    -2.43   0.015    -7.524129   -.8005461
       _cons |   100.5909   1.422605    70.71   0.000     97.79811    103.3837
------------------------------------------------------------------------------
```

(1) The output is the same, the only difference being that with the prefix xi:, stata has created dummy variables for you and they are named slightly differently.

(2) Stata, by default, omits the lowest level, hence using this as the referent.

(3) What if you want to use a different level as the referent? This is possible using the command **ib#**. For example, suppose we wanted to use the value of physact=5 as the referent. Our regression command would have then been

regress glucose ib5.physact if diabetes==0

Exercises # 14-19
Multiple Linear Regression Analysis of Y=LDL, entire sample.

```
.*
. *****  14) Two way contingency table analysis of HT and statins
. tab2 HT statins, row column chi2

-> tabulation of HT by statins
 random assignment |      statin use
to hormone therapy |     0. no     1. yes |     Total
-------------------+----------------------+----------
        0. placebo |       304        188 |       492
                   |     61.79      38.21 |    100.00
                   |     47.87      51.51 |     49.20
-------------------+----------------------+----------
1. hormone therapy |       331        177 |       508
                   |     65.16      34.84 |    100.00
                   |     52.13      48.49 |     50.80
-------------------+----------------------+----------
             Total |       635        365 |     1,000
                   |     63.50      36.50 |    100.00
                   |    100.00     100.00 |    100.00

          Pearson chi2(1) =   1.2239   Pr = 0.269
```

(1) The chi square test (null hypothesis: no association) is not statistically significant (p-value = .27). Do NOT reject. In this sample, there is no statistically significant evidence of an association of receipt of hormone treatment with statin use.

```
. *
. *****  15) Single predictor model of Y=LDL to X=HT
. regress LDL HT

      Source |       SS       df       MS              Number of obs =     997
-------------+------------------------------           F(  1,   995) =    0.06
       Model |  82.0397956     1  82.0397956           Prob > F      =  0.8059
    Residual |  1351929.67   995  1358.72328           R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.0009
       Total |  1352011.71   996  1357.44147           Root MSE      =  36.861

------------------------------------------------------------------------------
         LDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          HT |    .573778   2.335055     0.25   0.806    -4.008419    5.155975
       _cons |   144.3646   1.663507    86.78   0.000     141.1002    147.6289
------------------------------------------------------------------------------
```

(1) The fitted line is Y = 144.36 + (0.57)*HT for Y=LDL
(2) R-squared = .0001 tells us that less than .01% of the variation in LDL is explained.
(3) So, not surprisingly, the overall F test (value=0.06) is NOT statistically significant. Do NOT reject. Conclude that LDL is not associated with HT in this sample.

```
. *
. *****  16) Single predictor model of Y=LDL to X=statins
. regress LDL statins

      Source |       SS       df       MS              Number of obs =     997
-------------+------------------------------           F(  1,   995) =   55.48
       Model |  71403.2316     1  71403.2316           Prob > F      =  0.0000
    Residual |  1280608.47   995  1287.04369           R-squared     =  0.0528
-------------+------------------------------           Adj R-squared =  0.0519
       Total |  1352011.71   996  1357.44147           Root MSE      =  35.875

------------------------------------------------------------------------------
         LDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     statins |  -17.58768   2.361273    -7.45   0.000    -22.22133   -12.95403
       _cons |   151.0593   1.424794   106.02   0.000     148.2634    153.8553
------------------------------------------------------------------------------
```

(1) The fitted line is now Y = 151.06 - (17.59)*statins
(2) R-squared = .0528 says that 5% of the variation in LDL is explained.
(3) Here the p-value for the overall F test is < .0001. Reject. Conclude that this sample gives statistically significant evidence that LDL is associated with use of statins (just as you thought!).
(4) From the fitted line, we see that statin use is associated with lower LDL levels, because the regression coefficient is negative.

```
. *****  17) Create HTstatins that is the interaction of HT and statins
. generate HTstatins=HT*statins
. label variable HTstatins "Interaction HT*statins"

. *
. *****  18) Fit 3 predictor multiple linear regression
. regress LDL HT statins HTstatins

      Source |       SS       df       MS              Number of obs =     997
-------------+------------------------------           F(  3,   993) =   18.48
       Model |   71482.457     3  23827.4857           Prob > F      =  0.0000
    Residual |  1280529.25   993  1289.55614           R-squared     =  0.0529
-------------+------------------------------           Adj R-squared =  0.0500
       Total |  1352011.71   996  1357.44147           Root MSE      =   35.91

------------------------------------------------------------------------------
         LDL |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          HT |  -.5008528   2.855153    -0.18   0.861     -6.10368    5.101974
     statins |   -18.1676   3.333962    -5.45   0.000    -24.71002   -11.62518
   HTstatins |   1.161376   4.730765     0.25   0.806    -8.122067    10.44482
       _cons |   151.3208   2.062998    73.35   0.000     147.2725    155.3691
------------------------------------------------------------------------------
```

(1) We now have fitted Y = 151.32 – (0.5)*HT - (18.17)*statins + (1.16)*HTstatins

(2) Oh my goodness. R-squared = .0529 is essentially the same as the R-squared in the simpler model that did not include the interaction (there, R-squared = .0528).

```
. *
. *****  19) Perform a partial F-test for HTstatins controlling for HT, statins
. testparm HTstatins

 ( 1)  HTstatins = 0

       F(  1,    993) =    0.06
            Prob > F =    0.8061
```

Just as we expected. The partial F-test (null hypothesis: the addition of the interaction to the model containing the main effects) has p-value = .81. Do NOT reject. Conclude that, in this sample, there is no statistically significant evidence of an interaction of hormone therapy and statin use for prediction of LDL, after controlling for the main effects of hormone therapy and statin use. Note – in a real world analysis, we would choose as our final model the single predictor model containing statins.