Statistics/Data Analysis

```
      name:  <unnamed>
       log:  C:\Users\Conor\Documents\Conor\Grad School\TA Work\Econ 103 - Econometric
> s\STATA Work\Week 2\wk2_section_log.smcl
  log type:  smcl
 opened on:  16 Jan 2018, 14:08:05

. 
. // Demonstration STATA code for week 2
. // Principles of Econometrics 4th Edition
. 
. set more off

. clear all

. cd "C:\Users\Conor\Documents\Conor\Grad School\TA Work\Econ 103 - Econometrics\STATA
>  Work\Week 2"
C:\Users\Conor\Documents\Conor\Grad School\TA Work\Econ 103 - Econometrics\STATA Work\
> Week 2

. 
. /////////////////////////////////////\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
> ///////////////////////////// Question 3.8 \\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
> /////////////////////////////////////\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
> 
. * Analyze data on home sales in Baton Rouge, Lousiana in mid-2005
. use br2.dta, clear

. 
. // Notice that throughout this question, we are asked to use alpha = 0.01 as our
. // critical value. Rather than hard-code this value (i.e. type 0.01), we can
. // create a variable that we can reuse. If we need to change alpha later on, it
. // will be much easier to just adjust this variable instead of finding 0.01
. // (and making sure 0.01 means alpha in whatever context)
. // Question 3.12 will use a different alpha, so to be clear in this file, I
. // define my variable as alpha_38 (for 3.8). Other scalar variables for this
. // section will also have a 38.
. scalar alpha_38 = 0.01

. 
. ********************************************************************************
. *Part A: For the traditional-style houses, estimate the linear regression
. *model PRICE = beta1 + beta2*SQFT + e. Test the null hypothesis that the slope
. *is zero against the alternative that it is positive, using the alpha = 0.01
. *level of significance. Follow and show all the test steps described in
. *Chapter 3.4
. ********************************************************************************
. 
. // Note that all the exercises in this question ask that we look only at
. // traditional-style houses. We could put " if trad == 1" into all of our
. // reg commands, as we did last week with female-only or black-only regressions.
. reg price sqft if trad == 1

      Source |       SS           df       MS      Number of obs   =       582
-------------+----------------------------------   F(1, 580)       =   1027.92
       Model |  2.4362e+12         1  2.4362e+12   Prob > F        =    0.0000
    Residual |  1.3746e+12       580  2.3700e+09   R-squared       =    0.6393
-------------+----------------------------------   Adj R-squared   =    0.6387
       Total |  3.8108e+12       581  6.5591e+09   Root MSE        =     48683

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sqft |   73.77195    2.30097    32.06   0.000      69.2527     78.2912
       _cons |  -28407.56   5728.161    -4.96   0.000    -39658.02   -17157.09
------------------------------------------------------------------------------
```

```
35.
36. // Alternatively, we can remove all the non-traditional observations from the
37. // data set (since we don't save over br2.dta, we can always reload to recover
38. // those observations if we need to).
39. drop if trad != 1
  (498 observations deleted)

40. reg price sqft

      Source |       SS           df       MS      Number of obs   =       582
-------------+----------------------------------   F(1, 580)       =   1027.92
       Model |  2.4362e+12         1  2.4362e+12   Prob > F        =    0.0000
    Residual |  1.3746e+12       580  2.3700e+09   R-squared       =    0.6393
-------------+----------------------------------   Adj R-squared   =    0.6387
       Total |  3.8108e+12       581  6.5591e+09   Root MSE        =     48683

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        sqft |   73.77195    2.30097    32.06   0.000      69.2527     78.2912
       _cons |  -28407.56   5728.161    -4.96   0.000    -39658.02   -17157.09
------------------------------------------------------------------------------

41.
42. // Steps for setting up a hypothesis test:
43. // (1) Determine the null and alternative hypothesis
44. // --> Null (H0): beta2 = 0     Alternative (H1): beta2 > 0
45. //
46. // (2) Test statistic and its distribution
47. // --> t-stat: b2/se(b2)    Distribution: t with 580 degrees of freedom
48. //     for the degrees of freedom, see the residual df in the regression output,
49. //     also stored in e(df_r)
50. //
51. // (3) Select alpha and determine the rejection region
52. // --> Problem tells us to use alpha = 0.01.
53. //     Comment on STATA's invttail function:
54. // The function has inputs df and q, where df = degrees of freedom (i.e. which
55. // t distribution we're using) and q, with answer = invttail(df, q) matches t(df, an
  > swer) = 1-q/
56. // Alternatively, we could use the invt function, which would give us:
57. // answer = invt(df, p) matches t(df, answer) = p
58. // where t(df, x) is the CDF for the t distribution with df degrees of freedom
59. // --> Rejection region for a right-side (beta2 > x) test is invt(1-alpha,df)
60. //     Let's call this critical value tc
61.
62. scalar tc_38_1side = invttail(e(df_r),alpha_38) // = -1*invt(e(df_r), alpha_38)

63. disp tc_38_1side // should be positive since we're doing a right-sided (beta2>0) tes
  > t
  2.3327943

64.
65. // (4) Calculate the sample value of the test statistic
66.
67. scalar tstat_38a = _b[sqft]/_se[sqft]

68. disp tstat_38a
  32.061242

69.
```

```
// --> Note that this t-stat value matches what was reported in the regression outpu
> t
//     since the null is the same (beta2 = 0)
//
// (5) State your conclusion
// --> Given the t-statistic of about 32 compared to a critical value of about 2.3
//     we reject the null that beta2 = 0

*****************************************************************************
*Part B: Using the linear model in (a), test the null hypothesis (H0) that the
*expected price of a house of 2000 square feet is equal to, or less than,
*$120,000. What is the appropriate alternative hypothesis? Use the alpha = 0.01
*level of significance. Obtain the p-value of the test and show its value on a
*sketch. What is your conclusion?
*****************************************************************************

// Null (H0): yhat(2000) = beta1 + 2000*beta2 = 120,000
// Alternative (H1): yhat(2000) > 120,000
// Again, we'll use a t test for this exercise.

// First calculate the point estimate for yhat(2000)
scalar point_yhat_2000 = _b[_cons] + 2000*_b[sqft]

disp point_yhat_2000
```

119136.34

```

// Notice that the point estimate is less than our null of 120000. Already,
// this tells us that we will fail to reject if our alternative is that
// yhat(2000) > 120000 since the rejection region will only cover positive
// t statistics, which will require yhat(2000) > 120000

// Next, we need to calculate the standard error of our estimate.
// Since yhat is a linear combination of b1 and b2, we need information on
// the full variance-covariance matrix from our OLS regression
// We can view the variance-covariance matrix using the following command:
estat vce
```

Covariance matrix of coefficients of **regress** model

e(V)	sqft	_cons
sqft	**5.2944617**	
_cons	**-12335.341**	**32811823**

```
// Note that the diagonal terms (sqft, sqft and _cons, _cons) are equal to
// the square of the Std. Err. reported in the output table

// Next, rather than code the values from variance-covariance matrix terms by
// hand, we can use STATA's stored values to enter these numbers
// Just as we've used scalars before, now we need to define a matrix object,
// and use it to store a saved value from the regression.
matrix define varmat_38a = e(V)

// Next, to extract elements from a matrix, we use the syntax matname[i, j]
// where we want to find the row i and column j value of matrix matname
// From here, we can calculate the standard error of our estimate of yhat
// as follows: se_yhat_2000 = sqrt( var(_cons) + (2000^2)*var(sqft) + 2*2000*cov(_co
> ns,sqft))
scalar se_yhat_2000 = sqrt(varmat_38a[2,2] + (2000^2)*varmat_38a[1,1] + 2*2000*varma
> t_38a[2,1])
```

```
disp se_yhat_2000
  2155.9934

// Calculate the t stat for the null of yhat(2000) = 120000
scalar tstat_38b = (point_yhat_2000 - 120000)/se_yhat_2000

// Compare to the critical value for the 1-sided t-test
// This is the same critical value tc that we had in part (a)
disp tstat_38b
  -.4005866

disp tc_38_1side
  2.3327943


// Given this estimate, we fail to reject the null that yhat(2000) <= 120000

// We still need to calculate the p-value for this test. To be clear about the
// p-value that I'm calculating, I note in the name that we are looking at a test
// where rejecting the null requires values to fall on the right side of the
// distribution
scalar pval_yhat_2000_rt = 1 - t(e(df_r), tstat_38b) // could also do ttail(e(df_r),
>  tstat_38b)

disp pval_yhat_2000_rt
  .65556399


// Comparing this p-value to our alpha, we once again fail to reject the null
// that yhat(2000) <= 120000.

////////////////////////// STATA Command Alternatives \\\\\\\\\\\\\\\\\\\\\\\\\\
>
// For comparison, I also calculate the p-value for the 2-sided test, which
// corresponds to the automatic output of the STATA commands discussed below.
// Note for the 2-sided test, to get the correct value we need to make sure that
// the input for t CDF is a negative value. I use the if, else syntax for STATA
// to assign the value of check_sign to make sure I get the correct result.
if tstat_38b < 0 {
                scalar check_sign = 1
}

        else {
                scalar check_sign = -1
}


scalar pval_yhat_2000_2side = 2*t(e(df_r), check_sign*tstat_38b)


// In addition to calculating the test statistics by hand, we could have STATA
// do a bunch of the work for us using the following commands: lincom or test
// (1) lincom calculates the point estimate and standard error for a linear
//     combination of our beta estimates. It also reports a confidence interval,
//     together with the t-test and p-value for the null that the calculated
//     value is equal to zero.
// syntax: lincome exp
// where exp = c1*var1 + c2*var2 or c1*var2 - c2*var2
```

```
// note that var1 and var2 are the names of RHS variables used in the most recent
// regresion, as reported in the STATA output. lincom stores results in return list
lincom _cons + 2000*sqft

 ( 1)  2000*sqft + _cons = 0

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   119136.3   2155.993    55.26   0.000     114901.8    123370.8
------------------------------------------------------------------------------


// We could also feed in _cons + 2000*sqft - 120000 to lincom. Notice that this
// won't affect the estimate of the standard error, but rather than give us the
// point estimate for yhat(2000) it will give us the the numerator in our
// t-statistic, as well as giving us the t-stat and p-value for the 2-sided t-test
// that yhat(2000) = 120000 (i.e. yhat(2000) - 120000 = 0)
lincom _cons + 2000*sqft - 120000

 ( 1)  2000*sqft + _cons = 120000

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -863.6621   2155.993    -0.40   0.689    -5098.168    3370.844
------------------------------------------------------------------------------

disp tstat_38b
-.4005866

disp pval_yhat_2000_2side
.68887203


// Notice that the t-statistic reported here matches the t-statistic we found
// by hand earlier. The p-value, however, matches that for the 2-sided test
// and not the right-side test.

// (2) test allows you to test (potentially many) linear combinations of the
//     beta estimates against values of your choosing. Although test reports an
//     F-statistic, in the case of testing just a single linear combination, we
//     can recover the t-stat by noting that in the single-restriction case
//     t-stat = sqrt(F-stat). This still poses the issue of what the sign of the
//     t-stat is though, but we know that its sign will match the sign of
//     the point estimate minus its null value.
test _cons + 2000*sqft = 120000

 ( 1)  2000*sqft + _cons = 120000

       F(  1,   580) =    0.16
            Prob > F =    0.6889

disp sqrt(r(F))
.4005866

disp tstat_38b
-.4005866

disp pval_yhat_2000_2side
.68887203
```

```

// We can see that the tstat we calculated earlier matches the absolute value//
// of the square root of the F-stat here. As with lincom above, the automatically
// reported p-value corresponds to the one for the 2-sided test.

******************************************************************************
*Part C: Based on the estimated results from part (a), construct a 95% interval
*estimate of the expected price of a house of 2000 square feet.
******************************************************************************

// To construct the 2-sided t-test, we need 3 elements
// (1) The point estimate for our variable of interest
// (2) The standard error for our variable of interest
// (3) The critical value for the t distribution associated with a 2-sided test
//     at our desired level of significance.
// From these values, we can then calculate the low- and high-points of the
// confidence interval as:
// LOW = point estimate - tc_lvl_2side * se
// HIGH = point estimate + tc_lvl_2side * se

scalar tc_95_2side = -1*invt(e(df_r), 0.05/2) //alpha = 0.05, could also do invttail
> (e(df_r),0.05/2)

scalar yhat_2000_ci95low = point_yhat_2000-tc_95_2side*se_yhat_2000

scalar yhat_2000_ci95high = point_yhat_2000+tc_95_2side*se_yhat_2000

disp yhat_2000_ci95low
114901.83

disp yhat_2000_ci95high
123370.84


// Compare these by-hand estimates to what STATA generated using lincom
lincom _cons + 2000*sqft

 ( 1)  2000*sqft + _cons = 0
```

price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	**119136.3**	**2155.993**	**55.26**	**0.000**	**114901.8**	**123370.8**

```

// To adjust the confidence level in the lincom command, use the level option
// For example, to construct the 99% confidence interval, you would say:
// lincom _cons + 2000*sqft, level(99)

******************************************************************************
*Part D: For the traditional-style houses, estimate the quadratic regression
*model PRICE = alpha1 + alpha2*SQFT^2 + e. Test the null hypothesis that the
*marginal effect of an additional square foot of living area in a home with
*2000 square feet of living space is $75 against the alternative that the effect
*is less than $75. Use the alpha = 0.01 level of significance. Repeat the same
*test for a home of 4000 square feet of living space. Discuss your conclusions.
******************************************************************************

gen sqft_sqr = sqft^2
```

```
reg price sqft_sqr
```

Source	SS	df	MS
Model	2.5786e+12	1	2.5786e+12
Residual	1.2322e+12	580	2.1245e+09
Total	3.8108e+12	581	6.5591e+09

Number of obs	=	582
F(1, 580)	=	1213.74
Prob > F	=	0.0000
R-squared	=	0.6767
Adj R-squared	=	0.6761
Root MSE	=	46093

price	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
sqft_sqr	.0120632	.0003463	34.84	0.000	.0113832	.0127433
_cons	68710.05	2873.195	23.91	0.000	63066.91	74353.18

```
// Marginal effect is 2*x*b2 where x = 2000 or 4000

// Calculate by hand: sqft = 2000
scalar margeff_quad_2000 = 2*_b[sqft_sqr]*2000

scalar se_margeff_quad_2000 = 2*2000*_se[sqft_sqr]

scalar tstat_38d_2000 = (margeff_quad_2000 - 75)/se_margeff_quad_2000

// Compare tstat to the critical value for a left-hand side test at alpha = 0.01
disp tstat_38d_2000
-19.311502

disp -1*tc_38_1side
-2.3327943

// Conclusion: since our t-statistic is less than the critical value for the
// LHS test, we reject the null that marginal effect is $75 at sqft=2000 in favor
// of the alternative that the marginal effect is less than $75

// Calculate by hand: sqft = 4000
scalar margeff_quad_4000 = 2*_b[sqft_sqr]*4000

scalar se_margeff_quad_4000 = 2*4000*_se[sqft_sqr]

scalar tstat_38d_4000 = (margeff_quad_4000 - 75)/se_margeff_quad_4000

// Compare tstat to the critical value for a left-hand side test at alpha = 0.01
disp tstat_38d_4000
7.7636657

disp -1*tc_38_1side
-2.3327943

// Conclusion: fail to reject the null that the marginal effect is $75 at
// sqft = 4000


////////////////////////// STATA Command Alternatives \\\\\\\\\\\\\\\\\\\\\\\\\
>
// Marginal effect is 2*b2*x where x = 2000 or 4000 and null is = 75
// Compare the t-stat to what we calculated above
lincom 2*sqft_sqr*2000 - 75

 ( 1)  4000*sqft_sqr = 75
```

price	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
(1)	-26.74707	1.385033	-19.31	0.000	-29.46736	-24.02678

```
lincom 2*sqft_sqr*4000 - 75

 ( 1)  8000*sqft_sqr = 75

------------------------------------------------------------------------------
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   21.50587   2.770066     7.76   0.000     16.06528    26.94645
------------------------------------------------------------------------------


// Using the # syntax in regress together with the margins command
reg price c.sqft#c.sqft

      Source |       SS           df       MS      Number of obs   =       582
-------------+----------------------------------   F(1, 580)       =   1213.74
       Model |  2.5786e+12         1  2.5786e+12   Prob > F        =    0.0000
    Residual |  1.2322e+12       580  2.1245e+09   R-squared       =    0.6767
-------------+----------------------------------   Adj R-squared   =    0.6761
       Total |  3.8108e+12       581  6.5591e+09   Root MSE        =     46093

-------------------------------------------------------------------------------
        price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+----------------------------------------------------------------
c.sqft#c.sqft |   .0120632   .0003463    34.84   0.000     .0113832    .0127433
              |
        _cons |   68710.05   2873.195    23.91   0.000     63066.91    74353.18
-------------------------------------------------------------------------------

margins, dydx(sqft) at(sqft=(2000 4000))

Conditional marginal effects                    Number of obs     =        582
Model VCE    : OLS

Expression   : Linear prediction, predict()
dy/dx w.r.t. : sqft

1._at        : sqft            =        2000

2._at        : sqft            =        4000

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
sqft         |
         _at |
          1  |   48.25293   1.385033    34.84   0.000     45.53264    50.97322
          2  |   96.50586   2.770066    34.84   0.000     91.06528    101.9464
------------------------------------------------------------------------------


// Compare point estimates and standard errors to what we calculated earlier
matrix define margins_est = r(b) // r(b) is a matrix of point estimates, in order

                                                                   // given in the mar
> gins command
matrix define var_margins_est = r(V) // r(V) is the variance-covariance matrix

                                                                          // for the
> reported point estimates, so that
```

```
                                                                          // the vari
  > ance of an individual estimate
                                                                          // is on th
  > e diagonal of the r(V) matrix

disp (margins_est[1,1]-75)/sqrt(var_margins_est[1,1]) // t-stat at sqft=2000
  -19.311503

disp (margins_est[1,2]-75)/sqrt(var_margins_est[2,2]) // t-stat at sqft=4000
  7.7636653


********************************************************************************
*Part E: For the traditional-style houses, estimate the log-linear regression
*model ln(PRICE) = gamma1 + gamma2*SQFT + e. Test the null hypothesis that the
*marginal effect of an additional square foot of living area in a home with
*2000 square feet of living space if $75 against the alternative that the effect
*is less than $75. Use the alpha = 0.01 level of significance. Repeat the same
*test for a home of 4000 square feet of living space. Discuss your conclusions.
********************************************************************************

gen ln_price = log(price) // generate variable for log price

reg ln_price sqft // log-linear regression
```

Source	SS	df	MS			
				Number of obs	=	**582**
				F(1, 580)	=	**880.41**
Model	**76.4414878**	**1**	**76.4414878**	Prob > F	=	**0.0000**
Residual	**50.3587111**	**580**	**.086825364**	R-squared	=	**0.6028**
				Adj R-squared	=	**0.6022**
Total	**126.800199**	**581**	**.218244749**	Root MSE	=	**.29466**

ln_price	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
sqft	**.0004132**	**.0000139**	**29.67**	**0.000**	**.0003859**	**.0004406**
_cons	**10.79894**	**.0346705**	**311.47**	**0.000**	**10.73084**	**10.86703**

```

// Calculate marginal effect by hand using formula exp(b1 + b2*x)*b2
// where x = 2000 or 4000
scalar margeff_log_2000 = exp(_b[_cons]+_b[sqft]*2000)*_b[sqft]

scalar margeff_log_4000 = exp(_b[_cons]+_b[sqft]*4000)*_b[sqft]

matrix define varmat_loglin = e(V)


// Since we're using a non-linear transformation of the beta values, we need to
// use some different techniques, which are incorporated into the following
// functions: nlcom and testnl
// nlcom is analogous to lincom, while testnl is analogus to test
//
// You will notice below that in the output for nlcom STATA refers to a z value
// which reminds us that (1) this result is relying on completely asymptotic
// results, rather than making a finite sample adjustment, and (2) that we
// should compare that value to a normal distribution. Similarly, testnl reports
// results for a chi2 statistic, rather than an F-distribution.
// Once again, the reported p-values correspond to a 2-sided test.
```

```
//
// The underlying method for working with non-linear functions of the coefficients
// is called the delta method, and it is discussed briefly in section 5.6.3 of
// the textbook. The underlying result is that we can estimate the variance for
// a non-linear function of the data as follows:
// Lambda = f(b1, b2)
// Var(Lambda) = (dLambda/db1)^2 * var(b1) + (dLambda/db2)^2 * var(b2)
//                  + 2 * (dLambda/db1) * (dLambda/db2) * cov(b1,b2)
//
// Below we show how to do this by hand, and then use nlcom and testnl:

// Calculating variance of marginal effect estimate by hand
scalar partial_b1_2000 = exp(_b[_cons]+_b[sqft]*2000)*_b[sqft]

scalar partial_b2_2000 = exp(_b[_cons]+_b[sqft]*2000)*(1+_b[sqft]*2000)

scalar var_margeff_log_2000 = (partial_b1_2000^2)*varmat_loglin[2,2] + ///
>                                                                 (partial_b2_2000^2)*
> varmat_loglin[1,1] + ///
>                                                                 (2*partial_b1_2000*p
> artial_b2_2000)*varmat_loglin[2,1]


scalar partial_b1_4000 = exp(_b[_cons]+_b[sqft]*4000)*_b[sqft]

scalar partial_b2_4000 = exp(_b[_cons]+_b[sqft]*4000)*(1+_b[sqft]*4000)

scalar var_margeff_log_2000 = (partial_b1_4000^2)*varmat_loglin[2,2] + ///
>                                                                 (partial_b2_4000^2)*
> varmat_loglin[1,1] + ///
>                                                                 (2*partial_b1_4000*p
> artial_b2_4000)*varmat_loglin[2,1]


// Pick the critical value for our test. Should be negative since we are doing
// a left-side test.
scalar zc_95_lt = invnormal(0.01)


// Calculate the point estimates using nlcom
nlcom exp(_b[_cons]+_b[sqft]*2000)*_b[sqft]
```

_nl_1: **exp(_b[_cons]+_b[sqft]*2000)*_b[sqft]**

ln_price	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	**46.2433**	**1.459765**	**31.68**	**0.000**	**43.38221**	**49.10439**

```
nlcom exp(_b[_cons]+_b[sqft]*4000)*_b[sqft]
```

_nl_1: **exp(_b[_cons]+_b[sqft]*4000)*_b[sqft]**

ln_price	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	**105.677**	**3.663876**	**28.84**	**0.000**	**98.49592**	**112.858**

```
// Calculate the t value for null = 75 using nlcom
nlcom exp(_b[_cons]+_b[sqft]*2000)*_b[sqft]-75

       _nl_1:  exp(_b[_cons]+_b[sqft]*2000)*_b[sqft]-75
```

ln_price	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	**-28.7567**	**1.459765**	**-19.70**	**0.000**	**-31.61779**	**-25.89561**

```
nlcom exp(_b[_cons]+_b[sqft]*4000)*_b[sqft]-75

       _nl_1:  exp(_b[_cons]+_b[sqft]*4000)*_b[sqft]-75
```

ln_price	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
_nl_1	**30.67698**	**3.663876**	**8.37**	**0.000**	**23.49592**	**37.85805**

```
// Compare the z-scores here to the critical value:
disp zc_95_lt
  -2.3263479

// Do the same exercise using testnl. Compare the sqrt of the chi2 stat to the
// t stat we calculated earlier.
testnl exp(_b[_cons]+_b[sqft]*2000)*_b[sqft]=75 // sqft = 2000

  (1)  exp(_b[_cons]+_b[sqft]*2000)*_b[sqft] = 75

               chi2(1) =      388.07
           Prob > chi2 =        0.0000

disp sqrt(r(chi2))
  19.699539

testnl exp(_b[_cons]+_b[sqft]*4000)*_b[sqft]=75 // sqft = 4000

  (1)  exp(_b[_cons]+_b[sqft]*4000)*_b[sqft] = 75
       warning: derivative with respect to sqft coefficient is near zero,
                derivative treated as zero

               chi2(1) =       70.10
           Prob > chi2 =        0.0000

disp sqrt(r(chi2))
  8.3728216


/* Discussion:
  > The point estimate for the log-linear regression is quite similar to the
  > point estimate in the quadratic case for sqft=2000. Unsurprisingly, in both cases
  > we reject the null that the marginal effect is $75 in favor of the alternative that
  > the effect is less than $75. The difference between the two approaches is larger
  > in the sqft=4000 case, but in the log-linear case we once again fail to reject
  > the null given that the point estimate is above $75. This discussion would be
  > largely the same if we implemented the adjustment to the point estimate for the
  > log-linear regression.
  > */
```

```

//////////////////////////////////////\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
> ////////////////////////////// Question 3.12 \\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
> //////////////////////////////////////\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\\
>
* How does the relationship between experience and wages change over a lifetime?
* How does sample size affect inference in OLS?
clear all

use cps4_small.dta, clear


scalar alpha_312 = 0.05 // set the alpha for this section at 0.05

********************************************************************************
*Part A: Create a new variable called EXPER30 = EXPER - 30. Construct a scatter
*diagram with WAGE on the vertical axis and EXPER30 on the horizontal axis. Are
*any patterns evident?
********************************************************************************
gen exper30 = exper - 30

twoway scatter wage exper30

graph export "Q 3-12 Wage Exper30 Scatter.pdf", replace
(file Q 3-12 Wage Exper30 Scatter.pdf written in PDF format)


/* Discussion:
> For all levels of experience, there is a large mass in the 0-20 range for wages
> but the upper level of wages seems to have a parabolic shape (i.e. rising,
> and then falling).
> */

********************************************************************************
*Part B: Estimate by least squares the quadratic model
*WAGE = gamma1+gamma2*(EXPER30)^2 + e. Are the coefficient estimates
*statistically significant? Test the null that gamma2 >= 0 against the
*alternative that gamma2 < 0 at the alpha = 0.05 level of signifiance. What
*conclusion do you draw?
********************************************************************************
gen exper30_sqr = exper30^2

reg wage exper30_sqr

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =     49.96
       Model |  7845.5768         1  7845.5768   Prob > F        =    0.0000
    Residual |  156719.851       998  157.033919   R-squared       =    0.0477
-------------+----------------------------------   Adj R-squared   =    0.0467
       Total |  164565.428       999  164.730158   Root MSE        =    12.531

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 exper30_sqr |  -.0138283   .0019564    -7.07   0.000    -.0176674   -.0099892
       _cons |   23.06694   .5265962    43.80   0.000     22.03358    24.10031
------------------------------------------------------------------------------


scalar tc_312b_1side = invt(e(df_r),alpha_312) // critical value for left-hand test
```

```
// Since we are working with the null that gamma2 = 0, we can use the t-stat
// stata automatically reported with the regression.
disp tc_312b_1side
```

-1.6463819

```

/* Discussion - the estimate for the beta on exper30_sqr is statistically
> different from zero at the 95% confidence level. We can see this from the the
> fact that the P value is approximately 0 and that 0 is not in the 95% confidence
> interval. Since the 2-sided test is more aggressive on either side than the 1-sided
> test, passing the 2-sided guarantees that we will pass the 1-sided test. To be
> certain, we compare our t-stat to the critical value for the 1-sided test.
> */

*****************************************************************************
*Part C: Using the estimation in part (b), compute the estimated marginal effect
*of experience upon wage for a person with 10 years' experience, 30 years'
*experience and 50 years' experience. Are these slopes significantly different
*from zero at the alpha = 0.05 level of significance?
*****************************************************************************

lincom 2*exper30_sqr*(10-30) //when exper=30, exper30 = -20
```

(1) **- 40*exper30_sqr = 0**

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	**.5531314**	**.0782551**	**7.07**	**0.000**	**.399568**	**.7066948**

```
lincom 2*exper30_sqr*(30-30) //when exper=30, exper30 = 0
```

(1) **= 0**

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	**0**	(omitted)				

```
lincom 2*exper30_sqr*(50-30) //when exper=50, exper30 = 20
```

(1) **40*exper30_sqr = 0**

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	**-.5531314**	**.0782551**	**-7.07**	**0.000**	**-.7066948**	**-.399568**

```

/* Discussion:
> Notice that for exper = 10 and exper = 50, the t-stat has the same magnitude
> as the beta coefficient itself. This follows from the fact that (1) we are only
> evaluating a scalar multiple of the beta coefficient, and (2) that the fixed
> null for the 3 tests is 0. In addition, notice that given the definition of
> exper30 and our choice of regression, we've basically assumed that the marginal
> value at exper=30 will be zero. Given this, we would fail to reject the null
> that the marginal effect is equal to zero for any finite variance.
>
> Comment: why is (absolute value of) the t-stat not affected across the three
> cases b2 = 0, 2*(-20)*b2, 2*20*b2? Recall that the t-stat is given as:
> t-stat = (b2-b2_null)/se(b2). In addition, for any constant c, se(c*b2) =
> abs(c)*se(b2) where abs(c) is the absolute value of c. Then, in the 3 cases, we have
> :
> t-stat = b2/se(b2) = -c*b2/(c*se(b2)) and c*b2/se(b2)
> */
```

```

******************************************************************************
*Part D: Construct 95% confidence interval estimates of each of the slopes in
*part (c) How precisely are we estimating these values?
******************************************************************************

// The lincom command already generated the confidence intervals for us, but
// here we recreate the estimates by hand. For comparison, I also show the
// calcultions for the confidence interval for the beta coefficient itself

scalar tc_312d_2side = invttail(e(df_r),0.05/2) // = (-1)*invt(e(df_r),0.05/2)


scalar beta2 = _b[exper30_sqr] // useful for comparison later


scalar beta2_cilow = _b[exper30_sqr] - tc_312d_2side*_se[exper30_sqr]

scalar margeff_10_cilow = -40*_b[exper30_sqr] - tc_312d_2side*_se[exper30_sqr]*40

scalar margeff_50_cilow = 40*_b[exper30_sqr] - tc_312d_2side*_se[exper30_sqr]*40


scalar beta2_cihigh = _b[exper30_sqr] + tc_312d_2side*_se[exper30_sqr]

scalar margeff_10_cihigh = -40*_b[exper30_sqr] + tc_312d_2side*_se[exper30_sqr]*40

scalar margeff_50_cihigh = 40*_b[exper30_sqr] + tc_312d_2side*_se[exper30_sqr]*40


disp "Confidence Interval for effect at exper = 10: [" margeff_10_cilow ", " margeff
  > _10_cihigh "]"
 Confidence Interval for effect at exper = 10: [.39956799, .70669478]

disp "Confidence Interval for effect at exper = 50: [" margeff_50_cilow ", " margeff
  > _50_cihigh "]"
 Confidence Interval for effect at exper = 50: [-.70669478, -.39956799]


/* Discussion:
  > Overall, we have a fairly tight estimate of the confidence intervals.
  > */

******************************************************************************
*Part E: Using the estimation result from part (b) create the fitted values
*WAGE_hat = gamma1_hat + gamma2_hat*(EXPER30)^2 where _hat denotes the least
*squares estimates. Plot these fitted values and WAGE on the vertical axis of
*the same graph against EXPER30 on the horizontal axis. Are the estimates in
*part (c) consistent with the graph?
******************************************************************************
predict wage_hat, xb

twoway (scatter wage exper30) (line wage_hat exper30, sort)

graph export "Q 3-12E Fitted Values.pdf", replace
 (file Q 3-12E Fitted Values.pdf written in PDF format)


/* Discussion:
  > Not surprisingly, given that everything is based on the same underlying regression,
  > the results in part (c) are consistent with this graph. The curve of fitted
  > values is symmetric around zero, with wages tending to rise before 30 (exper30 = 0),
  > and falling after 30 (exper30 = 0).
  > */
```

```
*******************************************************************************
*Part F: Estimate the linear regression WAGE = beta1+beta2*EXPER30 + e and the
*linear regression WAGE = alpha1 + alpha2*EXPER + e. What differences do you
*observe between these regressions and why do they occur? What is the estimated
*marginal effect of experience on wage from these regressions? Based on our work
*in parts (b)-(d), is the assumption of constant slope in this model a good one?
*Explain.
*******************************************************************************

reg wage exper30

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =      7.98
       Model |  1306.16677         1  1306.16677   Prob > F        =    0.0048
    Residual |  163259.261       998  163.586434   R-squared       =    0.0079
-------------+----------------------------------   Adj R-squared   =    0.0069
       Total |  164565.428       999  164.730158   Root MSE        =     12.79

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     exper30 |   .0889534   .0314802     2.83   0.005     .0271785    .1507283
       _cons |   20.92629    .419131    49.93   0.000     20.10381    21.74876
------------------------------------------------------------------------------

reg wage exper

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =      7.98
       Model |  1306.16677         1  1306.16677   Prob > F        =    0.0048
    Residual |  163259.261       998  163.586434   R-squared       =    0.0079
-------------+----------------------------------   Adj R-squared   =    0.0069
       Total |  164565.428       999  164.730158   Root MSE        =     12.79

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       exper |   .0889534   .0314802     2.83   0.005     .0271785    .1507283
       _cons |   18.25768   .9273279    19.69   0.000     16.43795    20.07742
------------------------------------------------------------------------------

/* Discussion:
> The only difference between the two sets of output are in the estimates tied
> to the constant term. With exper30, the constant is larger but with a smaller
> standard error. If we think about how OLS works, this should make sense since
> adding or subtracting a constant to x in the true model is equivalent to moving
> the constant term around. Mechanically, for OLS, the b2 estimate only cares about
> deviations in x from is sample mean, so that adding/subtracting a constant gets
> stripped out. Similarly, the b1 estimate moves around to ensure that the point
> (xbar, ybar) is on the best-fit line, so adding/subtracting from xbar just moves
> the b1 estimate around so that we continue to satisfy b1 = ybar - b2*xbar
> */

*******************************************************************************
*Part G: Use the larger data cps4.dta (4838 observations) to repeat parts (b),
*(c), and (d). How much has the larger sample improved the precision of the
*interval estimates in part (d)?
*******************************************************************************
```

```
458 reg wage exper30_sqr // re-run regression so easier to directly compare output

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =     49.96
       Model |   7845.5768         1   7845.5768   Prob > F        =    0.0000
    Residual |  156719.851       998  157.033919   R-squared       =    0.0477
-------------+----------------------------------   Adj R-squared   =    0.0467
       Total |  164565.428       999  164.730158   Root MSE        =    12.531

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  exper30_sqr |  -.0138283   .0019564    -7.07   0.000    -.0176674   -.0099892
       _cons |   23.06694   .5265962    43.80   0.000     22.03358    24.10031
------------------------------------------------------------------------------

459
460 use cps4, clear

461
462 gen exper30 = exper-30

463 gen exper30_sqr = exper30^2

464 reg wage exper30_sqr

      Source |       SS           df       MS      Number of obs   =     4,838
-------------+----------------------------------   F(1, 4836)      =    198.74
       Model |  29995.3377         1  29995.3377   Prob > F        =    0.0000
    Residual |  729879.371     4,836  150.926255   R-squared       =    0.0395
-------------+----------------------------------   Adj R-squared   =    0.0393
       Total |  759874.709     4,837  157.096281   Root MSE        =    12.285

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  exper30_sqr |  -.0123931   .0008791   -14.10   0.000    -.0141165   -.0106696
       _cons |   22.35452   .2366069    94.48   0.000     21.89066    22.81838
------------------------------------------------------------------------------

465
466 /* Discussion:
  > There are 3 changes between the two samples:
  > (1) A different value for the point estimate of beta2 (associated with changes
  > in sample variance/covariance). A different value of the point estimate of
  > sigma_hat^2 (associated with changes in beta2 and beta1).
  > (2) Changes in estimate of var(b2). This can come from either changes in estimate of
  > sigma_hat^2 or changes in sum (xi - xbar)^2 since
  > var_hat(b2) = (sigma_hat^2)/(sum (xi -xbar)^2)
  > (3) A higher degrees of freedom for the regression leads to smaller critical
  > values for t tests (so smaller confidence intervals/easier to reject nulls) and
  > smaller p-values (again, easier to reject null for a given alpha).
  >
  > While effect (1) can be important, the expected effect of these changes
  > should be zero and can have positive or negative effects on point estimates and
  > whether we reject certain null hypotheses. Effects (2) and (3) have a clear
  > direction in which they will affect our estimates from a larger vs. a smaller
  > sample. The gain from effect (3) is shrinking with the size of the sample, as
  > the t-distribution approaches a normal distribution at large degrees of freedom.
  > Effect (2) will tend to be lower var_hat(b2) since sum (xi-xbar)^2
  > always increases with more observations, but with diminishing effects for a given
  > number of new observations (i.e. the effect is larger going from 100 to 200 than
  > it is going from 1000 to 1100)
  > */
```

```

// Compare the t critical values for this section to those from earlier
scalar tc_312g_1side = invttail(e(df_r), alpha_312) // = (-1)*invt(e(df_r),0.05)

scalar tc_312g_2side = invttail(e(df_r), alpha_312/2) // = (-1)*invt(e(df_r),0.05/2)
  >  = invt(e(df_r), 1 - 0.05/2)


scalar zc_1side = (-1)*invnormal(alpha_312)

scalar zc_2side = (-1)*invnormal(alpha_312/2)

disp "CPS_Small 1-sided critical value: " tc_312b_1side "   CPS 1-sided critical val
  > ue: " tc_312g_1side    "Normal 1-sided critical value: " zc_1side
  CPS_Small 1-sided critical value: -1.6463819   CPS 1-sided critical value: 1.6451688No
  > rmal 1-sided critical value: 1.6448536

disp "CPS_Small 2-sided critical value: " tc_312d_2side "   CPS 2-sided critical val
  > ue: " tc_312g_2side    "Normal 2-sided critical value: " zc_2side
  CPS_Small 2-sided critical value: 1.9623438   CPS 2-sided critical value: 1.9604546Nor
  > mal 2-sided critical value: 1.959964


/* Discussion:
  > As suggested by the forces discussed above, the biggest change is to the standard
  > error of b2, which halved between the two cases. While there is some benefit
  > in the confidence interval from the smaller t critical values, most of the shrinking
  > comes from the smaller standard error.
  > */

//Convert log file (smcl) to pdf
```