Topic 7: Multiple Regression Inference

1) Hypothesis tests

-

- Everything from simple regression extends here to multiple regression
- Each OLS estimator $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$ will have a standard error: $SE(\hat{\beta}_0), SE(\hat{\beta}_1), SE(\hat{\beta}_2), \ldots, SE(\hat{\beta}_k)$
- For each $\hat{\beta}_j$: $\hat{\beta}_j \sim N\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right)$

  where $E\left(\hat{\beta}_j\right) = \beta_j$, and $var\left(\hat{\beta}_j\right) = \sigma^2_{\hat{\beta}_j}$
- We don't know $\sigma^2_{\hat{\beta}_j}$, but we can use sample to estimate it.
- The estimator $\hat{\sigma}^2_{\hat{\beta}_j}$ (something messy without matrix algebra)
- The standard error is $\sqrt{\hat{\sigma}^2_{\hat{\beta}_j}}$. Stata will calculate it.

1) Hypothesis tests

- Testing each coefficient separately.
- Follow the same exact procedure we did for $\hat{\beta}_1$ as in Chapter 5.
- Hypothesis: $$H_0 : \beta_j = \beta_j^*$$ $$H_1 : \beta_j \neq \beta_j^*$$
- t-stat: $$t = \frac{\hat{\beta}_j - \beta_j^*}{SE\left(\hat{\beta}_j\right)} \sim N(0,1) \text{ under the null.}$$
- We can pick the significance level $\alpha = 0.01,\ 0.5$ or.0.1.
- Reject if t>1.96 or t<-1.96 if alpha=0.05.
- Calculate the p-value and see if it is <alpha. If so, reject H_0.
- Calculate the confidence interval: $$\hat{\beta}_j \pm Z_{\frac{\alpha}{2}} \cdot SE\left(\hat{\beta}_j\right),\ Z_{\frac{\alpha}{2}} = 1.96 \text{ when } \alpha = 0.05$$

1) Hypothesis tests


```
. reg testscr str el_pct, r

Linear regression                               Number of obs =     420
                                                F(  2,   417) =  223.82
                                                Prob > F      =  0.0000
                                                R-squared     =  0.4264
                                                Root MSE      =  14.464

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -1.101296   .4328472    -2.54   0.011     -1.95213   -.2504616
      el_pct |  -.6497768   .0310318   -20.94   0.000     -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000     668.8754     703.189
------------------------------------------------------------------------------
```

1) Hypothesis tests

- So the results:

$$\hat{TestScore} = 686.0 - 1.10 \times STR - 0.650PctEL$$
$$(8.7) \quad (0.43) \quad\quad (0.031)$$

- We use White (heteroskedasticity-robust) standard errors – for exactly the same reason as in the case of a single regressor.
- Test for beta1=0:

$$t = \frac{-1.10 - 0}{0.43} = -2.54$$

- Test for beta2=0:

$$t = \frac{-0.650 - 0}{0.031} = 20.97$$

2) Joint hypothesis tests

- A joint hypothesis specifies a value for two or more coefficients, that is, it imposes a restriction on two or more coefficients:

  H_0: $\beta_1 = 0$ **and** $\beta_2 = 0$

  H_1: **either** $\beta_1 \neq 0$ **or** $\beta_2 \neq 0$ **or both**

- In general, a joint hypothesis will involve q restrictions. In the example above, $q = 2$, and the two restrictions are $\beta_1 = 0$ and $\beta_2 = 0$.
- A "common sense" idea is to reject if either of the individual t-statistics exceeds 1.96 in absolute value, but this "one at a time" test isn't valid.

2) Joint hypothesis tests

- Why can't we just test the coefficients one at a time?
- Let t_1 and t_2 be the t-statistics:

$$t_1 = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} \text{ and } t_2 = \frac{\hat{\beta}_2 - 0}{SE(\hat{\beta}_2)}$$

- The "one at time" ("common-sense") test is:

  reject H_0: $\beta_1 = \beta_2 = 0$ if $|t_1| > 1.96$ and/or $|t_2| > 1.96$

- What is the probability that this "one at a time" test rejects H_0, when H_0 is actually true? (It should be 5%.)

2) Joint hypothesis tests

- The probability of incorrectly rejecting the null hypothesis using the "one at a time" test

$$= \Pr_{H_0}[|t_1| > 1.96 \text{ and/or } |t_2| > 1.96]$$

$$= \Pr_{H_0}[|t_1| > 1.96, |t_2| > 1.96] + \Pr_{H_0}[|t_1| > 1.96, |t_2| \leq 1.96]$$

$$+ \Pr_{H_0}[|t_1| \leq 1.96, |t_2| > 1.96] \qquad \text{(disjoint events)}$$

$$= \Pr_{H_0}[|t_1| > 1.96] \times \Pr_{H_0}[|t_2| > 1.96]$$

$$+ \Pr_{H_0}[|t_1| > 1.96] \times \Pr_{H_0}[|t_2| \leq 1.96]$$

$$+ \Pr_{H_0}[|t_1| \leq 1.96] \times \Pr_{H_0}[|t_2| > 1.96]$$

(t_1, t_2 are independent by assumption)

$$= .05 \times .05 + .05 \times .95 + .95 \times .05$$

$= .0975 = 9.75\%$ – which is ***not*** the desired 5%!!

2) Joint hypothesis tests

- The size of a test is the actual rejection rate under the null hypothesis.
  - The size of the "common sense" test isn't 5%! Intuitively, the individual testing of coefficient using t-stat rejects H_0 too much because individual test gives too many chances to reject H_0.
  - In fact, its size depends on the correlation between t_1 and t_2.
- Two Solutions:
  - Use a different critical value in this procedure – not 1.96 (this is the "Bonferroni method – see SW App. 7.1) (this method is rarely used in practice however)
  - Use a different test statistic that test both β_1 and β_2 at once: the *F*-statistic (this is common practice)

2) Joint hypothesis tests

- Here we are interested in testing more than one coefficient at a time. For this, we use F-test.
- E.g. $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$

  Is $\beta_1 = \beta_3 = 0$?

$$H_0 : \beta_1 = 0,\ \beta_3 = 0$$

$$H_1 : \beta_1 \neq 0, \text{ or } \beta_3 \neq 0, \text{ or both}$$

- In H_0, the number of restriction is denoted q. Here q=2.
- If we believe the errors are homoskedastic, then an easy way to test q restrictions is to run 2 regressions:
- Unrestricted: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$

  restricted: $Y_i = \beta_0 + \beta_2 X_{2i} + \beta_4 X_{4i} + u_i$

2) Joint hypothesis tests

- Unrestricted: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$

  restricted: $Y_i = \beta_0 + \beta_2 X_{2i} + \beta_4 X_{4i} + u_i$

- Calculate the F-stat:

$$F = \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{SSR_{unrestricted}/(n - k_{unrestricted} - 1)}$$

$SSR_{restricted}$: the sum of squared residuals from the restricted regression

$SSR_{unrestricted}$: the sum of squared residuals from the unrestricted regression

$k_{unrestricted}$: the number of regressors in the unrestricted regression

q : the number of restriction.

2) Joint hypothesis tests

- Unrestricted: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + u_i$

  restricted: $Y_i = \beta_0 + \beta_2 X_{2i} + \beta_4 X_{4i} + u_i$

- Calculate the F-stat:

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)}$$

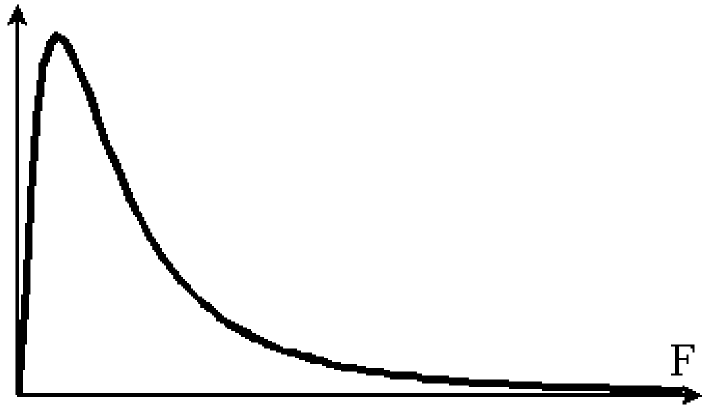
where:

$R^2_{restricted}$ = the R^2 for the restricted regression
$R^2_{unrestricted}$ = the R^2 for the unrestricted regression
q = the number of restrictions under the null
$k_{unrestricted}$ = the number of regressors in the unrestricted regression.

- Compare the fits of the regressions – the R^2's – if the "unrestricted" model fits sufficiently better, reject the null

2) Joint hypothesis tests

- Choose a significance level alpha=0.05
- When the sample size is large, under the null, F-stat follows a $F_{q,\infty}$ distribution. (If a random variable W follows a χ_q^2 distribution, then W/q follows a $F_{q,\infty}$ distribution).



- Find the critical value from the F-distribution. For example, the 5% critical value from $F_{2,\infty}$ distribution is 3.00 (Table 4 in page 795 in SW).
- If F-stat is greater than the critical value, reject the null.

2) Joint hypothesis tests

- Critical Values from $F_{q,\infty}$ distribution.

q	5% critical value
1	3.84
2	3.00
3	2.60
4	2.37
5	2.21

2) Joint hypothesis tests

- Example: are the coefficients on *STR* and *Expn* jointly zero?

  Unrestricted population regression (under H_1):

  $$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

  Restricted population regression (that is, under H_0):

  $$TestScore_i = \beta_0 + \beta_3 PctEL_i + v_i$$

  - The number of restrictions under H_0 is $q = 2$ (*why*?).
  - The fit will be better (R^2 will be higher) in the unrestricted regression (*why*?)

- Expn: Expenditure per Student
  STR: Student Teacher Ratio
  PctEL: Percent Of English Learners;

2) Joint hypothesis tests

- Example:

<u>Restricted regression</u>:

$$\hat{TestScore} = \underset{(1.0)}{644.7} - \underset{(0.032)}{0.671}PctEL,\quad R^2_{restricted} = 0.4149$$

<u>Unrestricted regression</u>:

$$\hat{TestScore} = \underset{(15.5)}{649.6} - \underset{(0.48)}{0.29}STR + \underset{(1.59)}{3.87}Expn - \underset{(0.032)}{0.656}PctEL$$

$$R^2_{unrestricted} = 0.4366,\ k_{unrestricted} = 3,\ q = 2$$

so

$$F = \frac{(R^2_{unrestricted} - R^2_{restricted})/q}{(1 - R^2_{unrestricted})/(n - k_{unrestricted} - 1)} = \frac{(.4366 - .4149)/2}{(1 - .4366)/(420 - 3 - 1)} = \mathbf{8.01}$$

- We reject the null that beta1=0 and beta2=0.

2) Joint hypothesis tests

- Basically what we are doing is seeing if additional explanatory power of the excluded variables is jointly significant.
- In our example, do STR and Expn together explain a large enough portion of variance in TestScore relative to the model with just Pct_EL? The answer is yes. How large is "large enough" ? We look to the F distribution.
- If error terms are heteroskedastic, then the methods won't work and more advanced option is available. Stata will handle either case.

2) Joint hypothesis tests

- Heteroskedasicity and F-test
- Formula for the special case of the joint hypothesis $\beta_1 = \beta_1^*$ and $\beta_2 = \beta_2^*$ in a regression with two regressors:
-

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right)$$

where $\hat{\rho}_{t_1,t_2}$ estimates the correlation between t_1 and t_2.

- Reject when F is large (how large? Look to F distribution)

2) Joint hypothesis tests

- Heteroskedasicity and F-test

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2} \right)$$

- The F-statistic is large when t_1 and/or t_2 is large
- The F-statistic corrects (in just the right way) for the correlation between t_1 and t_2.
- The formula for more than two β's is nasty unless you use matrix algebra.

2) Joint hypothesis tests

- Consider special case that t_1 and t_2 are independent, so $\hat{\rho}_{t_1,t_2} \xrightarrow{p} 0$; in large samples the formula becomes

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1-\hat{\rho}_{t_1,t_2}^2}\right) \cong \frac{1}{2}(t_1^2 + t_2^2)$$

- Under the null, t_1 and t_2 have standard normal distributions that, in this special case, are independent
- The large-sample distribution of the F-statistic is the distribution of the average of two independently distributed squared standard normal random variables.

2) Joint hypothesis tests


```
. reg testscr str expn el_pct

      Source |       SS       df       MS              Number of obs =     420
-------------+------------------------------           F(  3,   416) =  107.45
       Model |  66409.8835     3  22136.6278           Prob > F      =  0.0000
    Residual |  85699.7102   416  206.008919           R-squared     =  0.4366
-------------+------------------------------           Adj R-squared =  0.4325
       Total |  152109.594   419  363.030056           Root MSE      =  14.353

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -.2863993   .4805232    -0.60   0.551    -1.230956    .6581569
        expn |   3.867901   1.412122     2.74   0.006     1.092117    6.643686
      el_pct |  -.6560227   .0391059   -16.78   0.000    -.7328924   -.5791529
       _cons |    649.578   15.20572    42.72   0.000     619.6883    679.4676
------------------------------------------------------------------------------
```

2) Joint hypothesis tests


```
. test (str=0) (expn=0)

 ( 1)  str = 0
 ( 2)  expn = 0

       F(  2,   416) =     8.01
            Prob > F =    0.0004
```

2) Joint hypothesis tests


```
. reg testscr str expn el_pct, r

Linear regression                                      Number of obs =     420
                                                       F(  3,   416) =  147.20
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4366
                                                       Root MSE      =  14.353

------------------------------------------------------------------------------
             |               Robust
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -.2863993   .4820728    -0.59   0.553    -1.234002     .661203
        expn |   3.867901   1.580722     2.45   0.015     .7607024      6.9751
      el_pct |  -.6560227   .0317844   -20.64   0.000    -.7185008   -.5935446
       _cons |    649.578   15.45834    42.02   0.000     619.1917    679.9642
------------------------------------------------------------------------------
```

2) Joint hypothesis tests


```
. test (str=0) (expn=0)

 ( 1)  str = 0
 ( 2)  expn = 0

       F(  2,   416) =    5.43
            Prob > F =    0.0047
```

2) Joint hypothesis tests

- Overall significance of the model
- We also might be interested in testing if the regression model does statistically better than just the mean.
- Unrestricted: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$

  Restricted : $Y_i = \beta_0 + u_i$
- $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$

  H_1 : one or more $\beta_j \neq 0$
- Here q=the number of restrictions= number of explanatory variables of the unrestricted regression (k)
- You can just do a F-test.

2) Joint hypothesis tests

- $$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$
$$v.s.$$
$$TestScore_i = \beta_0 + u_i$$

- What is overall significance test doing for us?
- It is telling us how our model with 4 explanatory variables (including the intercept as an explanatory variable) perform relative to a model with just the intercept?

2) Joint hypothesis tests


```
. test (str=0) (expn=0) (el_pct=0)

 ( 1)   str = 0
 ( 2)   expn = 0
 ( 3)   el_pct = 0

        F(  3,   416) =  147.20
             Prob > F =    0.0000
```

3) Testing Single Restrictions involving multiple coefficient

- We already looked at

$$(i)\ H_0 : \beta_j = 0$$

$$(ii)\ H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$(iii)\ H_0 : \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_k = 0$$

- Now we are interested in linear restriction such as

$$H_0 : \beta_j = \beta_\ell$$

or

$$H_0 : \beta_j = \beta_\ell + 2\beta_r$$

- How? Easy answer is that Stata can do it.

3) Testing Single Restrictions involving multiple coefficient

- $$\widehat{TestScore} = \underset{(15.5)}{649.6} - \underset{(0.48)}{0.29}STR + \underset{(1.59)}{3.87}Expn - \underset{(0.032)}{0.656}PctEL$$
- Suppose we want to know if the magnitude of the impact of STR is PctEL are the same:

$$H_0 : \beta_1 = \beta_3 \text{ v.s. } H_1 : \beta_1 \neq \beta_3$$

- Use F-test with one restriction q=1.
-

```
. test (str=el_pct)

 ( 1)  str - el_pct = 0

       F(  1,   416) =    0.57
            Prob > F =    0.4494
```

3) Testing Single Restrictions involving multiple coefficient

- We can also test this hypothesis “manually” by transforming the regression and doing a t test
- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
- Want to test

$$H_0 : \beta_1 = \beta_3 \text{ v.s. } H_1 : \beta_1 \neq \beta_3$$

-

$$\begin{aligned}
Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \\
&= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + (\beta_3 X_{1i} - \beta_3 X_{1i}) + u_i \\
&= \beta_0 + \beta_1 X_{1i} - \beta_3 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_3 X_{1i} + u_i \\
&= \beta_0 + \underbrace{(\beta_1 - \beta_3)}_{\gamma} X_{1i} + \beta_2 X_{2i} + \beta_3 \underbrace{(X_{3i} + X_{1i})}_{W_i} + u_i \\
&= \beta_0 + \gamma X_{1i} + \beta_2 X_{2i} + \beta_3 W_i + u_i
\end{aligned}$$

3) Testing Single Restrictions involving multiple coefficient

- We can run the following regression:

$$Y_i = \beta_0 + \gamma X_{1i} + \beta_2 X_{2i} + \beta_3 W_i + u_i$$

where $W_i = X_{3i} + X_{1i}$

- and use a basic t test to test $H_0 : \gamma = 0$ v.s. $H_1 : \gamma \neq 0$
- In the example, the original regression:

$$TestScore_i = \beta_0 + \beta_1 STR_i + \beta_2 Expn_i + \beta_3 PctEL_i + u_i$$

- we run the following regression:

$$TestScore_i = \beta_0 + \gamma STR_i + \beta_2 Expn_i + \beta_3 W_i + u_i$$

where $W_i = PctEL_i + STR_i$.

3) Testing Single Restrictions involving multiple coefficient

Econ 3334

```
. gen w=el_pct+str

. reg testscr str expn w, r

Linear regression                                      Number of obs =     420
                                                       F(  3,   416) =  147.20
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4366
                                                       Root MSE      =  14.353
```

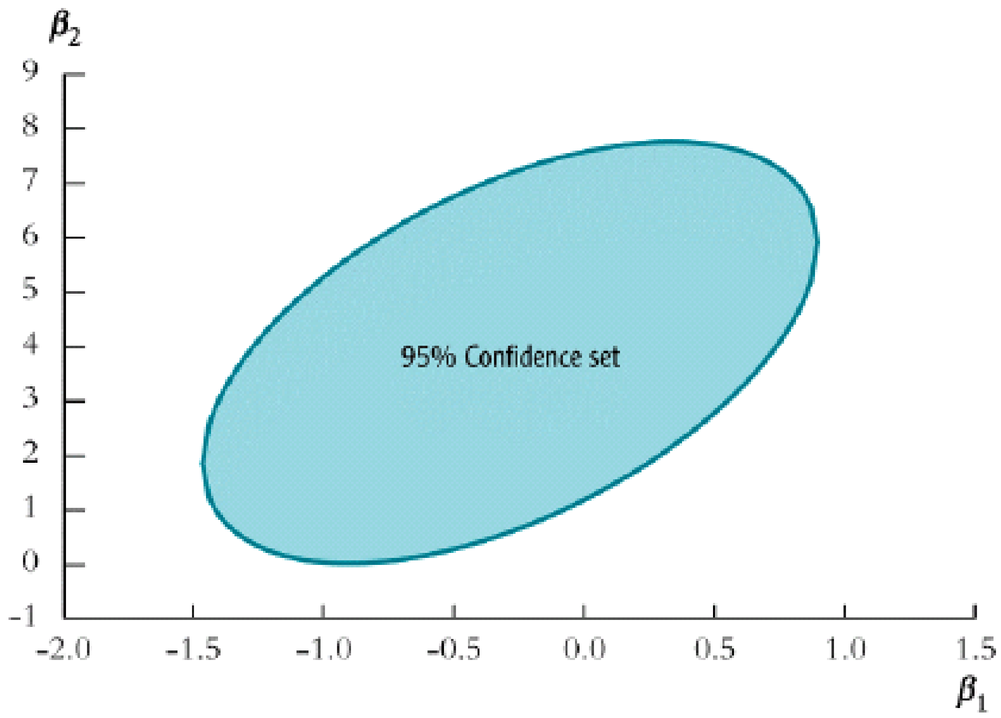
testscr	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
str	.3696233	.4881617	0.76	0.449	-.5899479	1.329195
expn	3.867902	1.580722	2.45	0.015	.7607025	6.975101
w	-.6560227	.0317844	-20.64	0.000	-.7185008	-.5935446
_cons	649.578	15.45834	42.02	0.000	619.1917	679.9642

- $0.76^2 = 0.5776$

4) Confidence Set

➤

FIGURE 5.1 **95% Confidence Set for β_1 and β_2**



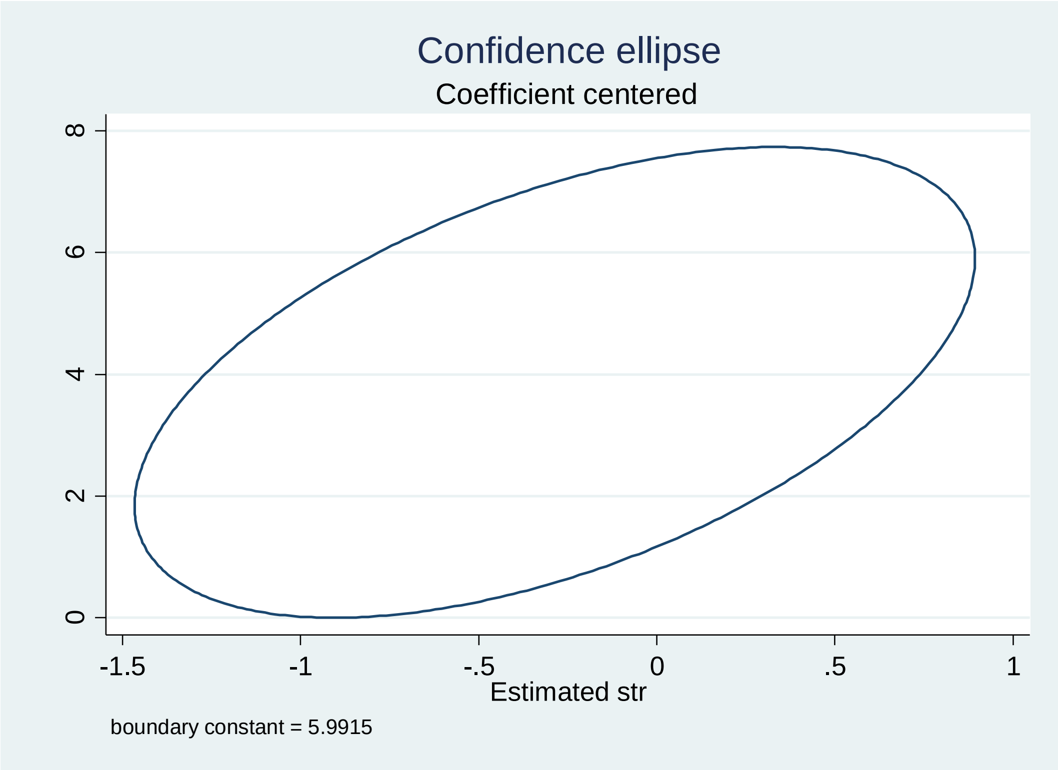
The 95% confidence set for β_1 and β_2 is an ellipse. The ellipse contains the pairs of values of β_1 and β_2 that cannot be rejected using the F-statistic at the 5% significance level.

4) Confidence Set

- In the example, after running the regression:

$$\widehat{TestScore} = \underset{(15.5)}{649.6} - \underset{(0.48)}{0.29}STR + \underset{(1.59)}{3.87}Expn - \underset{(0.032)}{0.656}PctEL$$

- type: ellip expn str, coefs c(chi2)



5) Model Specification

- In multiple regression, if you leave out a variable that is
  1) correlated with at least one of the included regressors
  and
  2) is a determinant of Y,
  then, all the slope estimators will be biased and inconsistent.
- Suppose the TRUE model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$$

- But we do

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_3 X_{3i} + u_i \text{ (left } X_{2i} \text{ out)}$$

- If say X_{2i} is correlated with X_{1i}, then in general both $\hat{\beta}_1$ and $\hat{\beta}_3$ will be biased and inconsistent.

5) Model Specification

- What do you do? Try to find X_2 and include it or at least find some proxy.
  - If you don't have say, mortgage rate, proxy it with Fed Fund's rate
  - If you don't have working experience, proxy it with age.
- How to choose your model:
  - Don't rely on R^2 or adjusted R^2. That is an indication of how you are doing in explaining variation of Y. But it shouldn't be your sole objective.
  - Don't worry too much about significance of individual coefficient.
  - Always let your economic theory dictate your variables.
  - Try different specifications as long as they make sense and see if the results are sensitive.