

I. F TESTS IN MULTIPLE REGRESSION

- So far, we've just examined one individual coefficient at a time. E.g., for the regression model,

$$BWGHT = \beta_0 + \beta_1 CIGS + \beta_2 FAMINC + \beta_3 MOTHEduc + \beta_4 FATHEDUC$$

we've asked questions such as: Holding CIGS, FAMINC, and MOTHEduc constant, does a baby's father's education have any influence on baby's birthweight?

- We may be interested in a number of other kinds of hypotheses, however. For example:

Example 1: Holding CIGS and FAMINC constant, do *both* mother's education and father's education have *any* influence on birthweight? Specifically:

$$H_0: \beta_{motheduc} = 0 \text{ and } \beta_{fatheduc} = 0$$

$$H_1: \beta_{motheduc} \neq 0 \text{ and/or } \beta_{fatheduc} \neq 0$$

Example 2: Or, we might be interested in knowing: Holding CIGS and FAMINC constant, are the effects of mother's education and father's education on birthweight *different*?

$$H_0: \beta_{motheduc} = \beta_{fatheduc}$$

$$H_1: \beta_{motheduc} \neq \beta_{fatheduc}$$

- To conduct these kinds of tests, we **cannot** simply look at the separate *t*-tests for each individual coefficient and then make a conclusion. We need a different kind of test: the *F* test (remember we used the *F*-distribution for ANOVA).
- The key thing to understand is that when you test these hypotheses, you are putting restrictions on the model that you are interested in estimating:

Example 1: Two or more coefficients jointly equal to zero:

$$H_0: \beta_{motheduc} = 0 \text{ and } \beta_{fatheduc} = 0$$

$$H_1: \beta_{motheduc} \neq 0 \text{ and/or } \beta_{fatheduc} \neq 0$$

Unrestricted Model: $BWGHT = \beta_0 + \beta_1CIGS + \beta_2FAMINC + \beta_3MOTHEDUC + \beta_4FATHEDUC$

Restricted Model: $BWGHT = \gamma_0 + \gamma_1CIGS + \gamma_2FAMINC$

The F test that can be used to test this hypothesis is:

$$F = \frac{(SSR_{restricted} - SSR_{unrestricted}) / q}{SSR_{unrestricted} / (n - k - 1)} \quad \text{where}$$

$SSR_{unrestricted}$: Sum of squared residuals from the **Unrestricted** regression (i.e., from the regression model that assumes that the null hypothesis is false)

$SSR_{restricted}$: Sum of squared residuals from the **Restricted** regression (i.e., from the regression model that assumes that the null hypothesis is true)

q : The number of parameters that are restricted (i.e., the number of *additional* parameters that are estimated in the Unrestricted model)

n : The total number of observations in the sample

k : The number of X variables (i.e., regressors) in the **Unrestricted** model

- And, this statistic is distributed as an F with $(q, n-k-1)$ d.f.: $F \sim F_{(q, n-k-1)}$
(see distribution and tabulated values for the F-distribution in back of any textbook)
- THINK about what the F-test is doing. It asks the question: Relatively how much explanatory power do you lose when you drop those variables?

```
. use bwght.dta
. drop if missing(motheduc) | missing(fatheduc)
```

UNRESTRICTED MODEL

```
. regress bwght cigs faminc motheduc fatheduc
```

Source	SS	df	MS			
Model	15827.6593	4	3956.91482	Number of obs =	1191	
Residual	466919.033	1186	393.69227	F(4, 1186) =	10.05	
Total	482746.692	1190	405.669489	Prob > F =	0.0000	
				R-squared =	0.0328	
				Adj R-squared =	0.0295	
				Root MSE =	19.842	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
bwght						
cigs	-.5894954	.1106172	-5.33	0.000	-.8065225	-.3724682
faminc	.0538254	.0366502	1.47	0.142	-.0180811	.1257319
motheduc	-.4379234	.3197377	-1.37	0.171	-1.065238	.1893912
fatheduc	.4936695	.2832896	1.74	0.082	-.0621351	1.049474
_cons	118.0741	3.500291	33.73	0.000	111.2066	124.9415

RESTRICTED MODEL

```
. regress bwght cigs faminc
```

Source	SS	df	MS			
Model	14536.9538	2	7268.47691	Number of obs =	1191	
Residual	468209.738	1188	394.115941	F(2, 1188) =	18.44	
Total	482746.692	1190	405.669489	Prob > F =	0.0000	
				R-squared =	0.0301	
				Adj R-squared =	0.0285	
				Root MSE =	19.852	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
bwght						
cigs	-.5876985	.1090181	-5.39	0.000	-.801588	-.3738091
faminc	.0624684	.0324438	1.93	0.054	-.0011851	.126122
_cons	118.5568	1.234278	96.05	0.000	116.1352	120.9784

$$\text{So, } F = \frac{(SSR_{restricted} - SSR_{unrestricted}) / q}{SSR_{unrestricted} / (n - k - 1)} = \frac{(468,210 - 466,919) / 2}{466,919 / (1,191 - 4 - 1)} = \frac{645.5}{393.6922} = 1.6396$$

Compare this with an F d.f. of (2, 1186): At the 95% confidence level, the F -critical value is 3.00

Make a decision: $1.6396 < 3.00$, so we *fail to reject* the null hypothesis that the effects of both **MOTHEDEC** and **FATHEDUC** are equal to zero.

Note: instead of going through all these calculations, we could have simply asked STATA to calculate the F -test for us:

```

. *** Stata F-Test
. regress bwght cigs faminc motheduc fatheduc

      Source |      SS       df       MS              Number of obs =   1191
-----+-----+-----+-----+-----+-----+-----
      Model | 15827.6593    4 3956.91482          F( 4, 1186) = 10.05
      Residual | 466919.033 1186 393.69227          Prob > F      = 0.0000
-----+-----+-----+-----+-----+-----
                                Adj R-squared = 0.0295
      Total | 482746.692 1190 405.669489          Root MSE     = 19.842

-----+-----
      bwght |      Coef.   Std. Err.      t    P>|t|   [95% Conf. Interval]
-----+-----+-----+-----+-----+-----
      cigs | -5894954   .1106172   -5.33  0.000   -8065225   -3724682
      faminc |  .0538254   .0366502    1.47  0.142   -0180811   .1257319
      motheduc | -4379234   .3197377   -1.37  0.171   -1.065238   .1893912
      fatheduc |  .4936695   .2832896    1.74  0.082   -.0621351   1.049474
      _cons | 118.0741   3.500291   33.73  0.000   111.2066   124.9415

-----+-----
. test (motheduc=0) (fatheduc=0)

(1) motheduc = 0
(2) fatheduc = 0

      F( 2, 1186) = 1.64
      Prob > F = 0.1946

```

Notes:

- In this example, the individual coefficient estimates on MOTHEDEC and FATHEDUC were not statistically significant; AND the F -test showed that they were not jointly statistically significant.

It is possible, however, that the individual t -statistics will not be statistically significant but the joint hypothesis test is statistically significant.

- These kinds of tests can be especially important when dealing with a set of variables that are highly correlated. Why?

Thus, a *conceptual reason* drives the specification of a joint F -test.

- Some studies develop scales that collapse many variables into a single variable. This may be desirable if the $d.f.$ are limited, but if possible, entering all the variables separately and running this kind of F -test is preferable if ample $d.f.$ are available.

- Another kind of test you can run using the F-test is equality of coefficients. For example, is the effect of mother's education and father's education on birthweight the same?:

Example 2: Equality of coefficients

$$H_0: \beta_{motheduc} = \beta_{fatheduc}$$

$$H_1: \beta_{motheduc} \neq \beta_{fatheduc}$$

- Here, too, we think about imposing a *restriction* on the model. If in fact the coefficients on MOTHEDEC and FATHEDUC are equal, then the restricted model would be estimated with *one fewer parameter* than the unrestricted model.

Unrestricted Model: $BWGHT = \beta_0 + \beta_1CIGS + \beta_2FAMINC + \beta_3MOTHEDEC + \beta_4FATHEDUC$

Restricted Model: $BWGHT = \beta_0 + \beta_1CIGS + \beta_2FAMINC + \beta_3(MOTHEDEC + FATHEDUC)$

$$BWGHT = \gamma_0 + \gamma_1CIGS + \gamma_2FAMINC + \gamma_3(TOTALYRSPAREDEC)$$

```
*** F-Test motheduc=fatheduc
. regress bwght cigs faminc motheduc fatheduc

Source |      SS      df    MS          Number of obs = 1191
-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
Model | 15827.6593    4 3956.91482      Prob > F      = 0.0000
Residual | 466919.033  1186 393.69227      R-squared     = 0.0328
-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
Total | 482746.692  1190 405.669489      Adj R-squared = 0.0295
                                          Root MSE    = 19.842

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
bwght |   Coef.  Std. Err.   t  P>|t|   [95% Conf. Interval]
-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
cigs | -0.5894954  .1106172  -5.33  0.000  -0.8065225  -0.3724682
faminc | 0.0538254  .0366502   1.47  0.142  -0.0180811  .1257319
motheduc | -0.4379234  .3197377  -1.37  0.171  -1.065238  .1893912
fatheduc | 0.4936695  .2832896   1.74  0.082  -0.0621351  1.049474
 _cons | 118.0741  3.500291  33.73  0.000  111.2066  124.9415
```

```
. test motheduc=fatheduc

(1) motheduc - fatheduc = 0

F( 1, 1186) = 3.08
Prob > F = 0.0797
```

- Can we proceed as though the effect of mother's education and father's education is the same (i.e., could we just put in "total years of parents' education" into the model?)

- This page shows the kind of coding and thinking behind the “restricted model” that STATA is running in the background of the test above:

```
. gen totpared = motheduc + fatheduc
```

UNRESTRICTED MODEL

```
. regress bwght cigs faminc motheduc fatheduc
```

```
Source |      SS      df      MS      Number of obs = 1191
-----+-----+-----+-----+-----
Model | 15827.6593    4 3956.91482    Prob > F   = 0.0000
Residual | 466919.033  1186 393.69227    R-squared   = 0.0328
-----+-----+-----+-----+-----
Total | 482746.692  1190 405.669489    Adj R-squared = 0.0295
                                     Root MSE   = 19.842
```

```
-----+-----+-----+-----+-----
bwght |   Coef.   Std. Err.   t   P>|t|   [95% Conf. Interval]
-----+-----+-----+-----+-----
cigs | -5894954  .1106172  -5.33  0.000  -8065225  -3724682
faminc | .0538254  .0366502   1.47  0.142  -0180811  .1257319
motheduc | -4379234  .3197377  -1.37  0.171  -1065238  .1893912
fatheduc | .4936695  .2832896   1.74  0.082  -0621351  1.049474
_cons | 118.0741  3.500291  33.73  0.000  111.2066  124.9415
```

RESTRICTED MODEL

```
. regress bwght cigs faminc totpared
```

```
Source |      SS      df      MS      Number of obs = 1191
-----+-----+-----+-----+-----
Model | 14616.9053    3 4872.30176    Prob > F   = 0.0000
Residual | 468129.787  1187 394.380612    R-squared   = 0.0303
-----+-----+-----+-----+-----
Total | 482746.692  1190 405.669489    Adj R-squared = 0.0278
                                     Root MSE   = 19.859
```

```
-----+-----+-----+-----+-----
bwght |   Coef.   Std. Err.   t   P>|t|   [95% Conf. Interval]
-----+-----+-----+-----+-----
cigs | -579494  .1105666  -5.24  0.000  -7964218  -3625663
faminc | .0547747  .0366782   1.49  0.136  -0171867  .1267361
totpared | .064152  .1424803   0.45  0.653  -2153894  .3436934
_cons | 117.1019  3.459139  33.85  0.000  110.3152  123.8886
```

$$F = \frac{(SSR_{restricted} - SSR_{unrestricted}) / q}{SSR_{unrestricted} / (n - k - 1)} = \frac{(468,130 - 466,919) / 1}{466,919 / (1,191 - 4 - 1)} = \frac{1,211}{393.6922} = 3.076$$

- COMPARE THIS F-statistic to the one we calculated through STATA in one step back on page 5. The two ways are equivalent.
- Would it make sense to test the following? Why or why not?

$$H_0: \beta_{cigs} = \beta_{faminc}$$

$$H_1: \beta_{cigs} \neq \beta_{faminc}$$

II. R-SQUARED FORM OF THE F-STATISTIC

- We calculated F -statistics for the previous tests first by using the SSR (the more calculation intensive method); and then by having STATA do the calculations for us.
- For the null hypothesis that the effect of *both* mother's and father's education were equal to zero, we obtained an F -statistic of 1.64, with a p -value of 0.1946.
- Suppose we wanted to test this hypothesis, but we had already shut down STATA, had only the output from the restricted and unrestricted models. There's a shortcut way to calculate the F -statistic (though you sacrifice a little precision):

$$F \equiv \frac{(R_{unrestricted}^2 - R_{restricted}^2) / q}{(1 - R_{unrestricted}^2) / (n - k - 1)} = \frac{0.0328 - 0.0301 / 2}{(1 - 0.0328) / 1,186} = \frac{0.00135}{0.0008155} = 1.655$$

Things to note:

- The order of Unrestricted and Restricted in the numerator is reversed compared to the formula using SSR (above on p. 2). Why does this make sense?
- As with the calculation above, you should ALWAYS get a positive number for the F -statistic: if you don't, something has gone horribly wrong.

III. RELATIONSHIP BETWEEN F AND t STATISTICS

- So what's the relationship between F and t statistics?
- When we look at a t -statistic and its corresponding p -value for a regression coefficient estimate, we're testing the null hypothesis that the true population coefficient for that variable is exactly zero.
- Instead, what if we ran an F -test, imposing a single restriction – that the coefficient on that variable of interest is equal to zero. Would we get a different answer than if we had used a t -test?
- Happily, no. We get completely consistent answers:
- t_{n-k-1}^2 has a $F_{(1, n-k-1)}$ distribution.

```
. regress bwght cigs faminc motheduc fatheduc  
[regression output omitted here]
```

```
. test fatheduc=0
```

```
( 1) fatheduc = 0
```

```
F( 1, 1186) = 3.04  
Prob > F = 0.0817
```

- The p -value here is exactly equal to the p -value on the coefficient estimate for FATHEDUC is the unrestricted model back on p. 4. The F -value is equal to the square of the t -statistic from that output on p. 4: $t^2 = 1.74^2 = 3.03$. This is just slightly off from the 3.04 in the output. Why? rounding.

If started off with the coefficient and s.e., and calculated the t -statistic ourselves, we would get: $\frac{0.49367}{0.28329} = 1.74263$ And, $1.74263^2 = 3.03676$, which is approx = 3.04

IV. F-STATISTIC FOR THE OVERALL REGRESSION

- A special case of the F -test asks the question whether *any* of the coefficients in the model are statistically different from zero. In other words, does the model as a whole help us explain any variation in Y at all?

H0: $\beta_1=0$ and $\beta_2=0$ and $\beta_2=0$ and $\beta_3=0$ and $\beta_4=0$ and... $\beta_k=0$

H1: At least one of the coefficients is not equal to zero.

Unrestricted Model: $BWGHT = \beta_0 + \beta_1 CIGS + \beta_2 FAMINC + \beta_3 MOTHEduc + \beta_4 FATHEDUC$

Restricted Model: $BWGHT = \gamma_0$

UNRESTRICTED

. regress bwght cigs faminc motheduc fatheduc

Source	SS	df	MS			
Model	15827.6593	4	3956.91482	Number of obs =	1191	
Residual	466919.033	1186	393.69227	F(4, 1186) =	10.05	
				Prob > F =	0.0000	
				R-squared =	0.0328	
				Adj R-squared =	0.0295	
Total	482746.692	1190	405.669489	Root MSE =	19.842	

bwght	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
cigs	-.5894954	.1106172	-5.33	0.000	-.8065225	-.3724682
faminc	.0538254	.0366502	1.47	0.142	-.0180811	.1257319
motheduc	-.4379234	.3197377	-1.37	0.171	-1.065238	.1893912
fatheduc	.4936695	.2832896	1.74	0.082	-.0621351	1.049474
_cons	118.0741	3.500291	33.73	0.000	111.2066	124.9415

RESTRICTED

. regress bwght

Source	SS	df	MS			
Model	0	0	.	Number of obs =	1191	
Residual	482746.692	1190	405.669489	F(0, 1190) =	0.00	
				Prob > F =	.	
				R-squared =	0.0000	
				Adj R-squared =	0.0000	
Total	482746.692	1190	405.669489	Root MSE =	20.141	

bwght	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
_cons	119.5298	.5836202	204.81	0.000	118.3848	120.6748

$$F = \frac{(SSR_{restricted} - SSR_{unrestricted}) / q}{SSR_{unrestricted} / (n - k - 1)} = \frac{(482,747 - 466,919) / 4}{466,919 / (1,191 - 4 - 1)} = \frac{3957}{393.6922} = 10.051$$

- This can also be calculated using R-squared from the UNRESTRICTED regression:

$$F = \frac{(R^2) / k}{(1 - R^2) / (n - k - 1)} = \frac{(0.0328) / 4}{(1 - 0.0328) / (1,191 - 4 - 1)} = \frac{0.0082}{0.000816} = 10.049$$