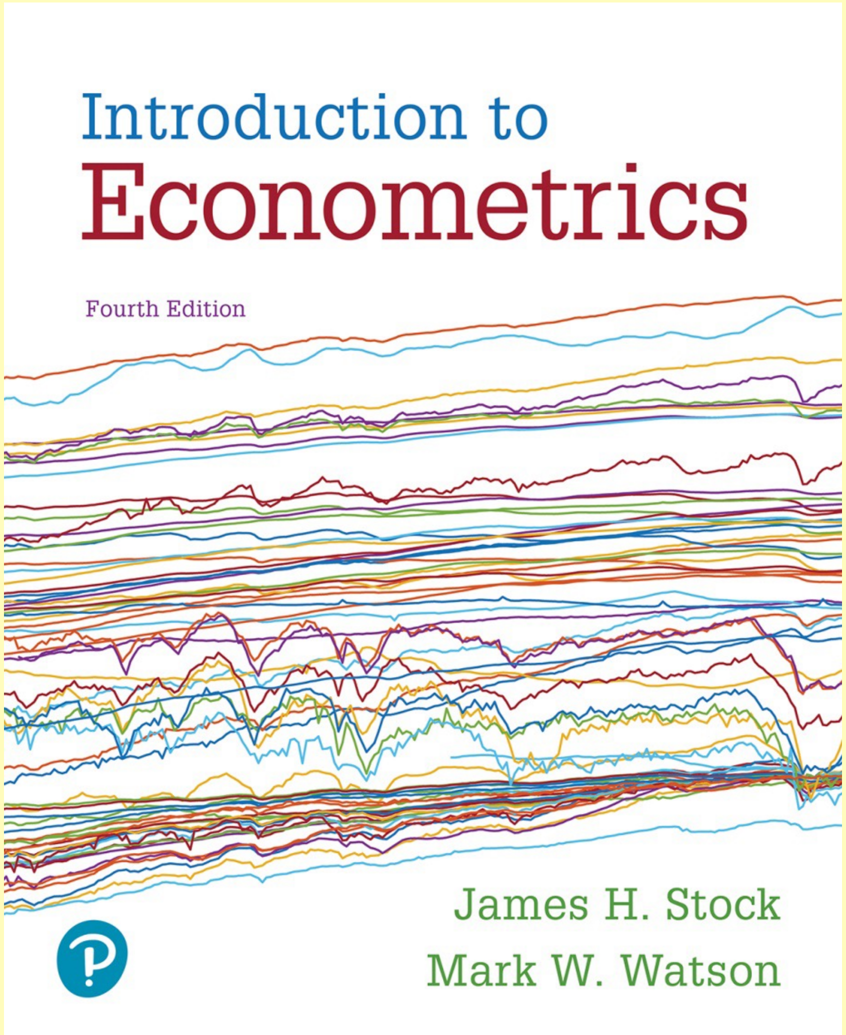
Introduction to Econometrics

Fourth Edition



Chapter 7

Hypothesis Tests and Confidence Intervals in Multiple Regression

Pearson

Outline

1. Hypothesis tests and confidence intervals for one coefficient
2. Joint hypothesis tests on multiple coefficients
3. Other types of hypotheses involving multiple coefficients
4. Confidence sets for multiple coefficients
5. Model specification: how to decide which variables to include in a regression model

χ^2-Distribution

If $x_1, x_2, \dots, x_n$ are independent normal variables with mean zero and variance 1, that is, $x_i \sim IN(0, 1)$, $i = 1, 2, \dots, n$ then

$$Z = \sum_i x_i^2$$

is said to have the χ^2-Distribution with degrees of freedom n, and we will write this as $Z \sim \chi_n^2$.

The subscript n denotes degrees of freedom.

The χ_n^2 distribution is the distribution of the sum of squares of n independent standard normal variables.

If $x_i \sim IN(0, \sigma^2)$, then Z should be defined as $Z = \sum_i \frac{x_i^2}{\sigma^2} = \frac{\sum_i x_i^2}{\sigma^2}$

The χ_n^2 -distribution also has an "additive property," although it is different from the property of the normal distribution and is much more restrictive.

The property is:

If $Z_1 \sim \chi_n^2$; and $Z_2 \sim \chi_m^2$ and Z_1, and Z_2 are independent, then

$$Z_1 + Z_2 \sim \chi_{n+m}^2$$

Note that we need independence and we can consider simple additions only, not any general linear combinations.

Even this limited property is useful in practical applications.

There are many distributions for which even this limited property does not hold.

t-Distribution

If $x \sim N(0, 1)$ and $y \sim \chi_n^2$ and x and y are independent, $Z = \frac{x}{\sqrt{y/n}}$ has a t~distribution with degrees of freedom n.

We write this as $Z \sim t_n$.

The subscript n again denotes the degrees of freedom.

Thus the t-distribution is the distribution of a standard normal variable divided by the square root of an independent averaged χ^2 variable (χ^2 variable divided by its degrees of freedom).

The t-distribution is a symmetric probability distribution like the normal distribution but is flatter than the normal and has longer tails.

As the degrees of freedom n approaches infinity, the t-distribution approaches the normal distribution.

F-Distribution

If $y_1 \sim \chi^2_{n_1}$ and $y_2 \sim \chi^2_{n_2}$ and y_1 and y_2 are independent, $Z = \frac{y_1/n_1}{y_2/n_2}$ has the F-distribution with degrees of freedom (d.f.) n_1 and n_2.

We write this as $Z \sim F_{n_1,n_2}$

The first subscript, n_1 refers to the d.f. of the numerator, and the second subscript, n_2 refers to the d.f. of the denominator.

The F-distribution is thus the distribution of the ratio of two independent averaged χ^2 variables.

If RSS is the residual sum of squares, then

$\hat{\sigma}^2 = \frac{RSS}{n-2}$ **is an unbiased estimator for** σ^2 ($\hat{\sigma}^2$ is variance of residual variable)

Also

$\frac{RSS}{\sigma^2}$ **has a** χ^2**-distribution with degrees of freedom** $(n - 2)$

Further the distribution of RSS is independent of the distributions of $\hat{\alpha}$ and $\hat{\beta}$.

This result can be used to get confidence intervals for σ^2 or to test hypotheses about σ^2.

However, we are still left with the problem of making inferences about α and β.

For this purpose we use the t-distribution.

We know that if we have two variables $x_1 \sim IN(0,1)$ and $x_2 \sim \chi^2$ of freedom k, and x_1 and x_2 are independent, then

$$x = \frac{x_1}{\sqrt{x_2/_k}} = \frac{standard\ normal}{\sqrt{independent\ averaged\chi^2}}$$

has a t-distribution with d.f. k.

In this case $\frac{(\hat{\beta}-\beta)}{\sqrt{\sigma_\varepsilon^2/_{S_{xx}}}} \sim N(0,1)$. (We have subtracted the mean and divided it by the standard deviation. $S_{xx} = \sum(x_i - \bar{x})^2$)

Also, $\frac{RSS}{\sigma^2} \sim \chi^2_{n-2}$ and the two distribution are independent.

Hence we compute the ratio

$$\frac{(\hat{\beta}-\beta)\Big/\sqrt{\sigma^2/S_{xx}}}{\sqrt{\frac{RSS}{(n-2)\sigma^2}}}$$ has a t-distribution with d.f. (n – 2).

Now σ^2 cancels out and writing $\frac{RSS}{n-2}$ as $\hat{\sigma}^2$ we get the result that

$$\frac{(\hat{\beta}-\beta)}{\sqrt{\hat{\sigma}^2/S_{xx}}}$$ has a t-distribution with d.f. (n – 2). $[\sum_{i=1}^{n}(X_i-\bar{X})^2] = S_{xx}$

Note that the variance of $\hat{\beta}$ is σ^2/S_{xx}.

Since σ^2 is not known, we use an unbiased estimator $\frac{RSS}{n-2}$.

Thus $\hat{\sigma}^2/_{S_{xx}}$ is the estimated variance of $\hat{\beta}$ and its square root is called the standard error denoted by SE(B).

We can follow a similar procedure for $\hat{\alpha}$.

We substitute $\hat{\sigma}^2$ for σ^2 in the variance of $\hat{\alpha}$ and take the square root to get the standard error SE($\hat{\alpha}$).

Thus we have the following result:

$\frac{(\hat{\alpha}-\alpha)}{SE(\hat{\alpha})}$ **and** $\frac{(\hat{\beta}-\beta)}{SE(\hat{\beta})}$ **each have a** *t*-**distribution with** d.f. (n – 2). **These distributions can be used to get confidence intervals for α and β and to test hypotheses about** α and β.

$\hat{\sigma}$ is usually called the standard error of the regression.

Hypothesis Tests and Confidence Intervals for a Single Coefficient (SW Section 7.1)

- Hypothesis tests and confidence intervals for a single coefficient in multiple regression follow the same logic and recipe as for the slope coefficient in a single-regressor model.
- $\dfrac{\hat{\beta}_1 - E(\hat{\beta}_1)}{\sqrt{\text{var}(\hat{\beta}_1)}}$ is approximately distributed N(0,1) (CLT).
- Thus hypotheses on β_1 can be tested using the usual t-statistic, and confidence intervals are constructed as $\{\hat{\beta}_1 \pm 1.96 \times \text{SE}\hat{\beta}_1)\}$.
- So too for $\beta_2, \ldots, \beta_k$.

Example: The California class size data

1. $$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR$$

2. $$\widehat{TestScore} = \underset{(8.7)}{686.0} - \underset{(0.43)}{1.10} \times STR - \underset{(0.031)}{0.650} PctEL$$

- The coefficient on *STR* in (2) is the effect on *TestScores* of a unit change in *STR*, holding constant the percentage of English Learners in the district
- The coefficient on *STR* falls by one-half
- The 95% confidence interval for coefficient on *STR* in (2) is {–1.10 ± 1.96 × 0.43} = (–1.95, – 0.26)
- The *t*-statistic testing $\beta_{STR} = 0$ is t = –1.10/0.43 = –2.54, so we reject the hypothesis at the 5% significance level

Standard errors in multiple regression in STATA

```
reg testscr str pctel, robust;
Regression with robust standard errors                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
       pctel |  -.6497768   .0310318   -20.94   0.000     -.710775   -.5887786
       _cons |   686.0322   8.728224    78.60   0.000     668.8754     703.189
------------------------------------------------------------------------------
```

$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.650 PctEL$$

$$(8.7) \quad (0.43) \qquad\qquad (0.031)$$

We use heteroskedasticity-robust standard errors – for exactly the same reason as in the case of a single regressor.

Tests of Joint Hypotheses (SW Section 7.2) (1 of 2)

Let *Expn* = expenditures per pupil and consider the population regression model:

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

The null hypothesis that "school resources don't matter," and the alternative that they do, corresponds to:

$$H_0: \beta_1 = 0 \textbf{\textit{ and }} \beta_2 = 0$$

$$\text{vs. } H_1: \textbf{\textit{either }} \beta_1 \neq 0 \textbf{\textit{ or }} \beta_2 \neq 0 \textbf{\textit{ or both}}$$

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

Tests of Joint Hypotheses (SW Section 7.2) (2 of 2)

- $H_0: \beta_1 = 0$ ***and*** $\beta_2 = 0$
- vs. H_1: ***either*** $\beta_1 \neq 0$ ***or*** $\beta_2 \neq 0$ ***or both***
- A ***joint hypothesis*** specifies a value for two or more coefficients, that is, it imposes a restriction on two or more coefficients.
- 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: the resulting test rejects too often under the null hypothesis (more than 5%)!

Why can't we just test the coefficients one at a time?

Because the rejection rate under the null isn't 5%. We'll calculate the probability of incorrectly rejecting the null using the "common sense" test based on the two individual *t*-statistics. To simplify the calculation, suppose that t_1 and t_2 are independently distributed (this isn't true in general – just in this example). Let t_1 and t_2 be the *t*-statistics:

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

The "one at time" test is:

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

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

Suppose t_1 *and* t_2 *are independent* (for this example)

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

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

$$= 1 - \Pr_{H_0}[|t_1| \leq 1.96 \text{ and } |t_2| \leq 1.96]$$

$$= 1 - \Pr_{H_0}[|t_1| \leq 1.96] \times \Pr_{H_0}[|t_2| \leq 1.96]$$

(because t_1 and t_2 are independent by assumption)

$$= \quad 1 - (.95)^2$$

$$= \quad .0975 = 9.75\% - \text{which is } \boldsymbol{not} \text{ the desired } 5\%!!$$

The *size* of a test is the actual rejection rate under the null hypothesis

- The size of the "common sense" test isn't 5%!
- In fact, its size depends on the correlation between t_1 and t_2 (and thus on the correlation between $\hat{\beta}_1$ and $\hat{\beta}_2$).

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 designed to test *both* β_1 and β_2 at once: the *F*-statistic (this is common practice)

What we want to know is **whether** the unrestricted model provides sufficiently better fit **that** we are willing to reject the null hypothesis.

To implement a test of this nature, we rely on the tools of R^2, i.e., the measures of goodness of fit.

Here, we introduced the concept of total, explained and unexplained sum of squares and used them to conduct a statistic, R^2, **that** provides a summary measures of goodness of fit.

In particular, recall from this discussion the unexplained (residuals) sum of squares is defined to be

$$RSS = \sum_{i=1}^{n} \left(Y_i - \hat{Y}\right)^2$$

and the total sum of squares is defined to be $TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$

$$RSS = \sum_{i=1}^{n}(Y_i - \hat{Y})^2 \; , TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2, \; R^2 = 1 - \frac{RSS}{TSS}$$

If the null hypothesis is true, **then** the RSS, and **hence** the R^2, both in the restricted and unrestricted models would be the same.

Again they may differ somewhat in practice **because** the restricted model is better suited to explaining some of the variation **that** exist in the data.

Therefore, a relevant statistic would compare the RSS or the R^2 in both the restricted and unrestricted models and **whether** the difference is big enough to be statistically significant.

In practice, the statistic we employ to accomplish this is

$$F = \frac{(RSS_{restricted} - RSS_{unrestricted})/_{m}}{RSS_{unrestricted}/_{(n-k-1)}}$$

where $RSS_{unrestricted}$ is the unexplained (or residual) sum of squared residuals in the unrestricted regression; $RSS_{restricted}$ is the unexplained sum of squared residuals in the restricted regression; m is the number of restriction; n-k-1 is the number of observations minus the number of explanatory variables.

We can apply the relationship between RSS and R^2 to convert the format of F statistic to contain R^2 itself.

To do this, notice that $RSS_{unrestricted} = TSS * (1 - R^2_{unrestricted})$

$$RSS_{restricted} = TSS * (1 - R^2_{restricted})$$

When we substitute these expressions into our test statistic and simplify, the test statistic becomes

$$F = \frac{\left(R^2_{unrestricted} - R^2_{restricted}\right)/_{m}}{R^2_{unrestricted}/_{(n-k-1)}}$$

Importantly, the value of F will be close to 0 **when** the null is true, because in this situation $R^2_{restricted} \approx R^2_{unresticted}$.

This suggests **that** values of F **that** are far from 0 would provide evidence favoring the alternative hypothesis.

To define "far" we need to know something of the sampling distribution of F.

Fortunately, we know that F follows the F distribution with m and n-k-1 degrees of freedom.

The *F*-statistic

The *F*-statistic tests all parts of a joint hypothesis at once.

Formula for the special case of the joint hypothesis $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$ 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?)

The F-statistic testing β_1 and β_2

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

- 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.
- This gives the F-statistic a nice large-sample approximate distribution, which is…

Pearson

Large-sample distribution of the *F*-statistic

Consider the *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

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2}t_1t_2}{1-\hat{\rho}_{t_1,t_2}^2}\right) \cong \frac{1}{2}(t_1^2 + t_2^2)$$

- 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.

The chi-squared distribution

The ***chi-squared*** distribution with q degrees of freedom (χ_q^2) is defined to be the distribution of the sum of q independent squared standard normal random variables.

In large samples, F is distributed as χ_q^2 / q.

Selected large-sample critical values of χ_q^2 / q

q	5% critical value	
1	3.84	(*why?*)
2	3.00	(the case q = 2 above)
3	2.60	
4	2.37	
5	2.21	

Computing the p-value using the F-statistic

p-value = tail probability of the χ_q^2/q distribution beyond the F-statistic actually computed.

Implementation in STATA

Use the “test” command after the regression

Example: Test the joint hypothesis that the population coefficients on STR and expenditures per pupil $(expn\_stu)$ are both zero, against the alternative that at least one of the population coefficients is nonzero.

F-test example, California class size data

```
reg testscr str expn_stu pctel, r;
Regression with robust standard errors                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 |  -.2863992   .4820728    -0.59   0.553    -1.234001     .661203
    expn_stu |   .0038679   .0015807     2.45   0.015     .0007607    .0069751
       pctel |  -.6560227   .0317844   -20.64   0.000    -.7185008   -.5935446
       _cons |   649.5779   15.45834    42.02   0.000     619.1917    679.9641
------------------------------------------------------------------------------
                                 NOTES
test str expn_stu;  The test command follows the regression

 ( 1)  str = 0.0     There are q=2 restrictions being tested
 ( 2)  expn_stu = 0.0
       F(  2,   416) =    5.43           The 5% critical value for q=2 is 3.00
       Prob > F =    0.0047   Stata computes the p-value for you
```

More on *F*-statistics

There is a simple formula for the F-statistic that holds <u>only under homoskedasticity</u> (so it isn't very useful) but which nevertheless might help you understand what the F-statistic is doing.

The homoskedasticity-only *F*-statistic

When the errors are homoskedastic, there is a simple formula for computing the "homoskedasticity-only" *F*-statistic:

- Run two regressions, one under the null hypothesis (the "restricted" regression) and one under the alternative hypothesis (the "unrestricted" regression).
- Compare the fits of the regressions – the R^2s – if the "unrestricted" model fits sufficiently better, reject the null

The "restricted" and "unrestricted" regressions

Example: are the coefficients on STR and $Expn$ zero?

<u>Unrestricted population regression (under H_1):</u>

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

<u>Restricted population regression (that is, under H_0):</u>

$$TestScore_i = \beta_0 + \beta_3 PctEL_i + u_i \qquad (why?)$$

- 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?*)

By how much must the R^2 increase for the coefficients on $Expn$ and $PctEL$ to be judged statistically significant?

Simple formula for the homoskedasticity-only F-statistic

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

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.

- The bigger the difference between the restricted and unrestricted R^2s – the greater the improvement in fit by adding the variables in question – the larger is the homoskedasticity-only *F*.

Example

Restricted regression:

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

Unrestricted regression:

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

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

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

Note: Heteroskedasticity-robust F = **5.43**…

The homoskedasticity-only F-statistic – summary

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

- The homoskedasticity-only *F*-statistic rejects when adding the two variables increased the R^2 by "enough" – that is, when adding the two variables improves the fit of the regression by "enough"
- If the errors are homoskedastic, then the homoskedasticity-only F-statistic has a large-sample distribution that is χ^2_q/q.
- But if the errors are heteroskedastic, the large-sample distribution of the homoskedasticity-only F-statistic is <u>not</u> χ^2_q/q

The *F* distribution

Your regression printouts might refer to the "*F*" distribution.

If the four multiple regression LS assumptions hold ***and if***:

5. u_i is homoskedastic, that is, $\text{var}(u|X_1,\ldots,X_k)$ does not depend on X's

6. $u_1,\ldots,u_n$ are normally distributed

then the homoskedasticity-only *F*-statistic has the "$F_{q,n-k-1}$" distribution, where q = the number of restrictions and k = the number of regressors under the alternative (the unrestricted model).

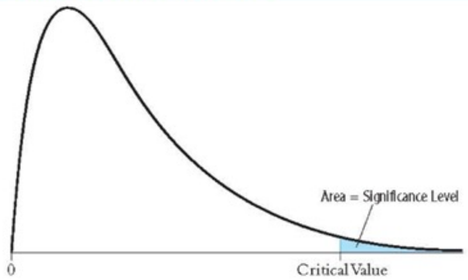
- **The F distribution is to the χ_q^2/q distribution what the t_{n-1} distribution is to the $N(0,1)$ distribution**

The $F_{q,n-k-1}$ distribution (1 of 2)

- The *F* distribution is tabulated many places
- As $n \to \infty$, the $F_{q,n-k-1}$ distribution asymptotes to the χ_q^2/q distribution:
- **The $F_{q,\infty}$ and χ_q^2/q distributions are the same.**
- For q not too big and $n \geq 100$, the $F_{q,n-k-1}$ distribution and the χ_q^2/q distribution are essentially identical.
- Many regression packages (including STATA) compute *p*-values of *F*-statistics using the *F* distribution
- You will encounter the *F* distribution in published empirical work.

The $F_{q,n-k-1}$ distribution (2 of 2)

TABLE 4 Critical Values for the $F_{m,\infty}$ Distribution



Degrees of Freedom	Significance Level 10%	5%	1%
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78
4	1.94	2.37	3.32
5	1.85	2.21	3.02
6	1.77	2.10	2.80
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.18
13	1.52	1.72	2.13
14	1.50	1.69	2.08
15	1.49	1.67	2.04
16	1.47	1.64	2.00
17	1.46	1.62	1.97
18	1.44	1.60	1.93
19	1.43	1.59	1.90
20	1.42	1.57	1.88
21	1.41	1.56	1.85
22	1.40	1.54	1.83
23	1.39	1.53	1.81
24	1.38	1.52	1.79
25	1.38	1.51	1.77
26	1.37	1.50	1.76
27	1.36	1.49	1.74
28	1.35	1.48	1.72
29	1.35	1.47	1.71
30	1.34	1.46	1.70

This table contains the 90th, 95th, and 99th percentiles of the $F_{m,\infty}$ distribution. These serve as critical values for tests with significance levels of 10%, 5%, and 1%.

Another digression: A little history of statistics (1 of 2)

- The theory of the homoskedasticity-only *F*-statistic and the $F_{q,n-k-1}$ distributions rests on implausibly strong assumptions (are earnings normally distributed?)
- These statistics date to the early 20th century… the days when data sets were small and computers were people…
- The *F*-statistic and $F_{q,n-k-1}$ distribution were major breakthroughs: an easily computed formula; a single set of tables that could be published once, then applied in many settings; and a precise, mathematically elegant justification.

Another digression: A little history of statistics (2 of 2)

- The strong assumptions were a minor price for this breakthrough.
- But with modern computers and large samples we can use the heteroskedasticity-robust *F*-statistic and the $F_{q,\infty}$ distribution, which only require the four least squares assumptions (not assumptions #5 and #6)
- This historical legacy persists in modern software, in which homoskedasticity-only standard errors (and *F*-statistics) are the default, and in which *p*-values are computed using the $F_{q,n-k-1}$ distribution.

Summary: the homoskedasticity-only *F*-statistic and the *F* distribution

- These are justified only under very strong conditions – stronger than are realistic in practice.
- *You* should use the heteroskedasticity-robust F-statistic, with χ_q^2/q (that is, $F_{q,\infty}$) critical values.
- For $n \geq 100,$ the F-distribution essentially is the χ_q^2/q distribution.
- For small *n*, sometimes researchers use the *F* distribution because it has larger critical values and in this sense is more conservative.

Summary: testing joint hypotheses

- The "one at a time" approach of rejecting if either of the *t*-statistics exceeds 1.96 rejects more than 5% of the time under the null (the size exceeds the desired significance level)
- The heteroskedasticity-robust *F*-statistic is built into the STATA ("test" command); this tests all *q* restrictions at once.
- For n large, the F-statistic is distributed χ_q^2/q $(= F_{q,\infty})$
- The homoskedasticity-only *F*-statistic is important historically (and thus in practice), and can help intuition, but isn't valid when there is heteroskedasticity

Testing Single Restrictions on Multiple Coefficients (SW Section 7.3) (1 of 2)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i,\ i = 1,\ldots,n$$

Consider the null and alternative hypothesis,

$$H_0: \beta_1 = \beta_2 \text{ vs. } H_1: \beta_1 \neq \beta_2$$

This null imposes a *single* restriction ($q = 1$) on *multiple* coefficients – it is not a joint hypothesis with multiple restrictions (compare with $\beta_1 = 0$ and $\beta_2 = 0$).

Testing Single Restrictions on Multiple Coefficients (SW Section 7.3) (2 of 2)

Here are two methods for testing single restrictions on multiple coefficients:

1. ***Rearrange ("transform") the regression***

Rearrange the regressors so that the restriction becomes a restriction on a single coefficient in an equivalent regression; or,

2. ***Perform the test directly***

Some software, including STATA, lets you test restrictions using multiple coefficients directly

Method 1: Rearrange ("transform") the regression (1 of 2)

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

$$H_0 : \beta_1 = \beta_2 \quad \text{vs.} \quad H_1 : \beta_1 \neq \beta_2$$

Add and subtract $\beta_2 X_{1i}$:

$$Y_i = \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + u_i$$

or

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$$

Where

$$\gamma_1 = \beta_1 - \beta_2$$

$$W_i = X_{1i} + X_{2i}$$

Method 1: Rearrange ("transform") the regression (2 of 2)

(a) Original equation:

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

$$H_0: \beta_1 = \beta_2 \text{ vs. } H_1: \beta_1 \neq \beta_2$$

(b) Rearranged ("transformed") equation:

$$Y_i = \beta_0 + \gamma_1 X_{1i} + \beta_2 W_i + u_i$$

where $\gamma_1 = \beta_1 - \beta_2$ and $W_i = X_{1i} + X_{2i}$

So $H_0: \gamma_1 = 0$ vs. $H_1: \gamma_1 \neq 0$

- These two regressions ((a) and (b)) have the same R^2, the same predicted values, and the same residuals.
- The testing problem is now a simple one: test whether γ_1 = 0 in regression (b).

Method 2: Perform the test directly

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

$$H_0: \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \beta_1 \neq \beta_2$$

Example:

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

In STATA, to test $\beta_1 = \beta_2$ vs. $\beta_1 \neq \beta_2$ (two-sided):

```
regress testscore str expn pctel, r
test str=expn
```

The details of implementing this method are software-specific.

Confidence Sets for Multiple Coefficients (SW Section 7.4) (1 of 2)

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i,\ i = 1,\ldots,n$$

What is a *joint* confidence set for β_1 and β_2?

A 95% ***joint confidence set*** is:

- A set-valued function of the data that contains the true coefficient(s) in 95% of hypothetical repeated samples.
- Equivalently, the set of coefficient values that cannot be rejected at the 5% significance level.

You can find a 95% confidence set as the set of (β_1, β_2) that cannot be rejected at the 5% level using an *F*-test (*why not just combine the two 95% confidence intervals?*).

Confidence Sets for Multiple Coefficients (SW Section 7.4) (2 of 2)

- Let $F(\beta_{1,0}, \beta_{2,0})$ be the (heteroskedasticity-robust) *F*-statistic testing the hypothesis that $\beta_1 = \beta_{1,0}$ and $\beta_2 = \beta_{2,0}$:
- 95% confidence set = $\{\beta_{1,0}, \beta_{2,0}: F(\beta_{1,0}, \beta_{2,0}) < 3.00\}$
- 3.00 is the 5% critical value of the $F_{2,\infty}$ distribution
- This set has coverage rate 95% because the test on which it is based (the test it "inverts") has size of 5%
  - 5% of the time, the test incorrectly rejects the null when the null is true, so 95% of the time it does not; therefore the confidence set constructed as the nonrejected values contains the true value 95% of the time (in 95% of all samples).

Pearson

The confidence set based on the F-statistic is an ellipse

$$\left\{\beta_1, \beta_2: \quad F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1-\hat{\rho}_{t_1,t_2}^2}\right) \le 3.00\right\}$$

Now

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

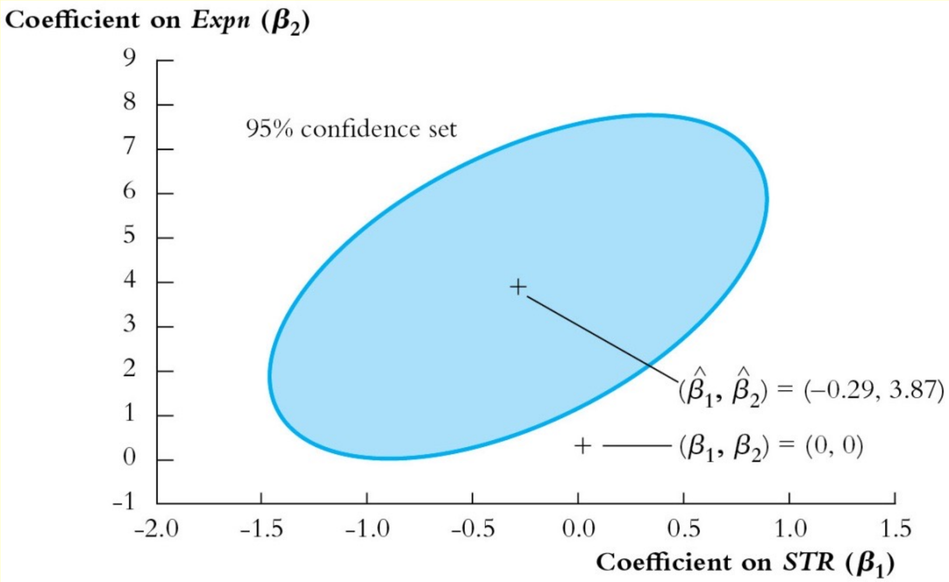
$$= \frac{1}{2(1-\hat{\rho}_{t_1,t_2}^2)} \times$$

$$\left[\left(\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right)^2 + \left(\frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)}\right)^2 + 2\hat{\rho}_{t_1,t_2}\left(\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right)\left(\frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)}\right)\right]$$

This is a quadratic form in $\beta_{1,0}$ and $\beta_{2,0}$ – thus the boundary of the set defined by $F \le 3.00$ is an ellipse.

Confidence set based on inverting the F-statistic

FIGURE 7.1: 95% Confidence Set for Coefficients on *STR* and *Expn* from Equation (7.6)



The 95% confidence set for the coefficients on *STR* (β_1) and *Expn* (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. The point $(\beta_1, \beta_2) = (0, 0)$ is not contained in the confidence set, so the null hypothesis $H_0 : \beta_1 = 0$ and $\beta_2 = 0$ is rejected at the 5% significance level.

Model specification: How to decide what variables to include in a regression (Section 7.5) (1 of 2)

1. Identify the variable of interest
2. Think of the omitted causal effects that could result in omitted variable bias
3. Include those omitted causal effects if you can or, if you can't, include variables correlated with them that serve as control variables.
   - The control variables are effective if the conditional mean independence assumption plausibly holds, that is, if *u* is uncorrelated with *STR* once the control variables are included.
   - This results in a "base" or "benchmark" model.

Model specification: How to decide what variables to include in a regression (Section 7.5) (2 of 2)

4. Also specify a range of plausible alternative models, which include additional candidate variables.
5. Estimate your base model and plausible alternative specifications ("sensitivity checks").
   - Does a candidate variable change the coefficient of interest (β_1)?
   - Is a candidate variable statistically significant?
   - Use judgment, not a mechanical recipe…
   - Don't just try to maximize R^2!

Digression about measures of fit

It is easy to fall into the trap of maximizing the R^2 and $\bar{R}^2$, but this loses sight of our real objective, an unbiased estimator of the class size effect.

- A high R^2 (or $\bar{R}^2$) means that the regressors explain the variation in Y.
- A high R^2 (or $\bar{R}^2$) does *not* mean that you have eliminated omitted variable bias.
- A high R^2 (or $\bar{R}^2$) does *not* mean that you have an unbiased estimator of a causal effect (β_1).
- A high R^2 (or $\bar{R}^2$) does *not* mean that the included variables are statistically significant – this must be determined using hypotheses tests.

Analysis of the Test Score Data Set (SW Section 7.6) (1 of 3)

1. Identify the variable of interest:
   STR

2. Think of the omitted causal effects that could result in omitted variable bias

   Whether the students know English; outside learning opportunities; parental involvement; teacher quality (if teacher salary is correlated with district wealth) – there is a long list!

Analysis of the Test Score Data Set (SW Section 7.6) (2 of 3)

3. Include those omitted causal effects if you can or, if you can't, include variables correlated with them that serve as control variables. The control variables are effective if the conditional mean independence assumption plausibly holds (if *u* is uncorrelated with *STR* once the control variables are included). This results in a "base" or "benchmark" model.

   Many of the omitted causal variables are hard to measure, so we need to find control variables. These include PctEL (both a control variable and an omitted causal factor) and measures of district wealth.

Analysis of the Test Score Data Set (SW Section 7.6) (3 of 3)

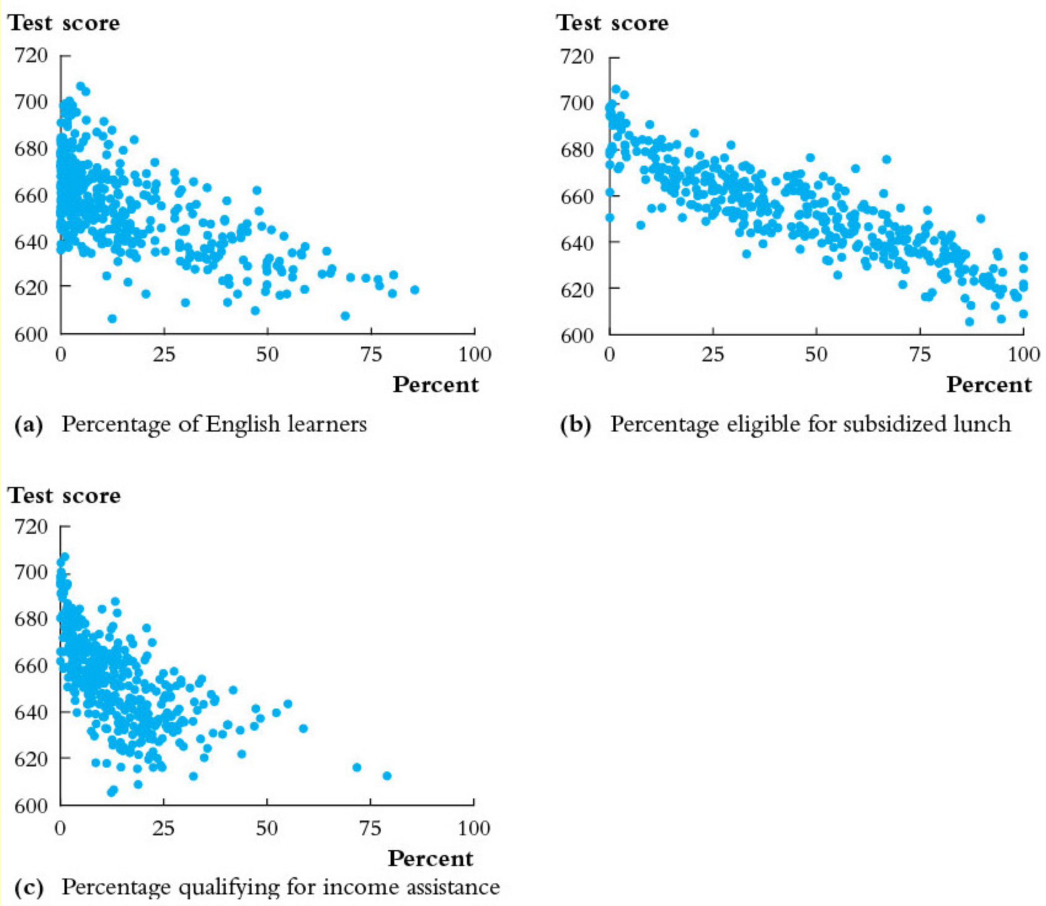
4. Also specify a range of plausible alternative models, which include additional candidate variables.

   It isn't clear which of the income-related variables will best control for the many omitted causal factors such as outside learning opportunities, so the alternative specifications include regressions with different income variables. The alternative specifications considered here are just a starting point, not the final word!

5. Estimate your base model and plausible alternative specifications ("sensitivity checks").

Test scores and California socioeconomic data

FIGURE 7.2: Scatterplots of Test Scores vs. Three Student Characteristics



Presentation of regression results (1 of 2)

- We have a number of regressions and we want to report them. It is awkward and difficult to read regressions written out in equation form, so instead it is conventional to report them in a table.
- A table of regression results should include:
  - estimated regression coefficients
  - standard errors
  - measures of fit
  - number of observations
  - relevant *F*-statistics, if any
  - any other pertinent information, such as confidence intervals for the causal effect of interest
- Find this information in the following table!

Presentation of regression results (2 of 2)

TABLE 7.1 Results of Regressions of Test Scores on the Student-Teacher Ratio and Student Characteristic Control Variables Using California Elementary School Districts

Dependent variable: average test score in the district.

Regressor	(1)	(2)	(3)	(4)	(5)
Student–teacher ratio (X_1)	−2.28 (0.52) [−3.30, −1.26]	−1.10 (0.43) [−1.95, −0.25]	−1.00 (0.27) [−1.53, −0.47]	−1.31 (0.34) [−1.97, −0.64]	−1.01 (0.27) [−1.54, −0.49]
Control variables					
Percentage English learners (X_2)		−0.650 (0.031)	−0.122 (0.033)	−0.488 (0.030)	−0.130 (0.036)
Percentage eligible for subsidized lunch (X_3)			−0.547 (0.024)		−0.529 (0.038)
Percentage qualifying for income assistance (X_4)				−0.790 (0.068)	0.048 (0.059)
Intercept	698.9 (10.4)	686.0 (8.7)	700.2 (5.6)	698.0 (6.9)	700.4 (5.5)
Summary Statistics					
SER	18.58	14.46	9.08	11.65	9.08
$\bar{R}^2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

These regressions were estimated using the data on K–8 school districts in California, described in Appendix 4.1. Heteroskedasticity-robust standard errors are given in parentheses under coefficients. For the variable of interest, the student–teacher ratio, the 95% confidence interval is given in brackets below the standard error.

Summary: Multiple Regression

- Multiple regression allows you to estimate the effect on Y of a change in X_1, holding other included variables constant.
- If you can measure a variable, you can avoid omitted variable bias from that variable by including it.
- If you can't measure the omitted variable, you still might be able to control for its effect by including a control variable.
- There is no simple recipe for deciding which variables belong in a regression – you must exercise judgment.
- One approach is to specify a base model – relying on *a-priori* reasoning – then explore the sensitivity of the key estimate(s) in alternative specifications.