ECON3150/4150 Spring 2015

Lecture 9&10, February 23

Siv-Elisabeth Skjelbred

University of Oslo

Last updated: February 22, 2015

T-test for two groups

In seminar 1 we showed the stata command for ttest:

- The ttest command is used when we want to compare two sample means
- The groups consists of n_1 and n_2 randomly chosen entities and the mean and variance can be computed for each group as normal.
- Two types of tests:
  - Unpaired: We have two separate sets of independent and identically distributed samples. T-test compares the means of the two groups of data to tests whether the two groups are statistically different.
  - Paired: A sample of matched pairs of similar units or one group of units that has been tested twice. The two measurements generally are before and after a treatment intervention. The test is calculated based on the difference between the two sets of paired observations.
- Both assume that the analyzed data is from a normal distribution.
- The unpaired test automatically assume that the variance of the two groups are approximately equal.
- A test can be performed assuming unequal variances. If the variances are equal it is not as powerful as the pooled variance test.

Hypothesis testing in MLRM

Testing a single coefficient

- Instead of testing β_1 we can test any β_j of the regression.

$$H_0 : \beta_j = \beta_{j,0} \text{ vs. } H_1 : \beta_j \neq \beta_{j,0}$$

- A two sided test that the true coefficient β_j on the j^th regressor takes on some specific value $\beta_{j,0}$.
- As in the SLRM you perform hypothesis testing of a single coefficient in three steps:
  - Compute the standard error of β_j
  - Compute the t-statistic
  - Compute the p-value or find the critical t-value.
- Stata automatically reports the t-statistic and the p-value for two-sided test of the null hypothesis $\beta_j = 0$.

Wage and education example

$$\hat{wage} = \underset{(0.77)}{-3.39} + \underset{(0.05)}{0.64}educ + \underset{(0.01)}{0.07}exp$$

$$t^{act} = \frac{0.64 - 0}{0.05} = 12.8$$

With 526 observations n=526. The critical value for the 5% significance level can be found either in the t-table or the Z-table as n is large.

Finding critical values

TABLE 2 Critical Values for Two-Sided and One-Sided Tests Using the Student t Distribution

	Significance Level				
Degrees of Freedom	20% (2-Sided) 10% (1-Sided)	10% (2-Sided) 5% (1-Sided)	5% (2-Sided) 2.5% (1-Sided)	2% (2-Sided) 1% (1-Sided)	1% (2-Sided) 0.5% (1-Sided)
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
26	1.32	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77
28	1.31	1.70	2.05	2.47	2.76
29	1.31	1.70	2.05	2.46	2.76
30	1.31	1.70	2.04	2.46	2.75
60	1.30	1.67	2.00	2.39	2.66
90	1.29	1.66	1.99	2.37	2.63
120	1.29	1.66	1.98	2.36	2.62
∞	1.28	1.64	1.96	2.33	2.58

Values are shown for the critical values for two-sided ($\neq$) and one-sided ($>$) alternative hypotheses. The critical value for the

Finding p-value

Standard Normal Probabilities

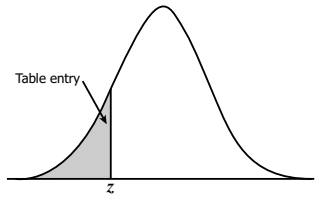


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Finding p-value

From the example last week:

```
1 . reg wage educ, robust

Linear regression                               Number of obs =        935
                                                F(  1,   933) =      95.65
                                                Prob > F      =     0.0000
                                                R-squared     =     0.1070
                                                Root MSE      =     382.32
```

wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	**60.21428**	**6.156956**	**9.78**	**0.000**	**48.1312**	**72.29737**
_cons	**146.9524**	**80.26953**	**1.83**	**0.067**	**-10.57731**	**304.4822**

The constant has a computed t value of 1.83. Since n is large we can use the z-table. The p-value is $2\phi(-1.83)$.

Finding p-value

Standard Normal Probabilities

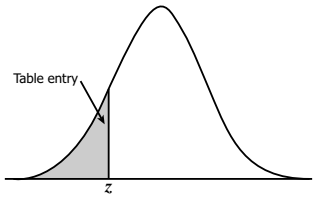


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

$$p = 2 * 0.036$$
$$= 0.0672$$

The Z-table

Remember that:

$$\phi(-Z) = 1 - \phi(Z)$$

Thus the table does not have to show the probability distribution for both positive and negative Z.

Confidence intervals for a single coefficient

Confidence interval for β_j

$$[\hat{\beta}_j - c * SE(\hat{\beta}_j), \hat{\beta}_j + c * SE(\hat{\beta}_j)]$$

where c is the critical value for the given confidence level.

Testing hypotheses about a single linear combination of the parameters

Linear combination of parameters

- Sometimes economic theory suggests relationships between coefficients.
- We can test any linear combination of parameters.
- A linear combination of parameters specifies only one restriction, but the restriction involves multiple parameters.
- The general version of a linear combination of two parameters is:

$$H_0 : \alpha\beta_1 + \gamma\beta_2 = \theta_{1,0}$$

- The linear combination can be tested by the t-statistic

$$t = \frac{\alpha\hat{\beta}_1 + \gamma\hat{\beta}_2 - \theta_{1,0}}{se(\alpha\hat{\beta}_1 + \gamma\hat{\beta}_2)}$$

- Or simply define $\alpha\beta_1 + \gamma\beta_2 = \theta_1$:

$$t = \frac{\hat{\theta}_1 - \theta_{1,0}}{se(\hat{\theta})}$$

Standard error of linear combination

- To find the standard error of $\hat{\theta}_1$ we must first obtain the variance: (for notation simplicity I use $\alpha = 1$ and $\gamma = 1$)

$$Var(\hat{\beta}_1 + \hat{\beta}_2) = Var(\hat{\beta}_1) + Var(\hat{\beta}_2) + 2Cov(\hat{\beta}_1, \hat{\beta}_2)$$

$$SE(\hat{\beta}_1 + \hat{\beta}_2) = \sqrt{Var(\hat{\beta}_1 + \hat{\beta}_2)}$$

The estimator of the standard error is thus given by:

$$SE(\hat{\beta}_1 + \hat{\beta}_2) = \sqrt{[se(\hat{\beta}_1)]^2 + [se(\hat{\beta}_2)]^2 + 2s_{12}}$$

Testing a linear combination

- Neither the standard error of $\hat{\theta}_1$ nor the covariance of the parameters is given in a standard regression.
- To perform hypothesis testing of $\hat{\theta}_1$ (as well as any other linear combination of parameters) you either need:
  - To rewrite the model so that the standard error of the linear combination is given in a standard regression.
  - A statistical software that either computes the sample covariance between the parameters or allow direct testing of linear combinations.

Rewriting model

- An example of a linear combination of parameters is that we may believe that the effect of two variables are the same.
- This can be tested by testing whether the two regression coefficients are equal, for example $\beta_1 = \beta_2$.
- This is equivalent to testing the following linear constraint:

$$H_0 : \beta_1 - \beta_2 = 0$$

- The initial model is then:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u$$

- But it can be rewritten to:

$$Y = \beta_0 + \gamma_1 X_1 + \beta_2 (X_2 - X_1) + u$$

Example rewriting model

- So far we have considered education as one single type of education
- However, there are different types of higher education.
- In Norway we have university college (hyskole) and university, in the US they have a two year college (junior college) and a four year college (university).
- Is the return to one year of education at a junior college the same as one year of education at a university?

Example rewriting model

The model:

$$wage = \beta_0 + \beta_1 jc + \beta_2 univ + \beta_3 exp + u$$

The null hypothesis can be used to specify a new parameter:

$$H_0 : \beta_1 - \beta_2 = \theta_1 \text{ thus } \beta_1 = \theta_1 + \beta_2$$

And inserting this into the original model you can write:

$$\begin{aligned} wage &= \beta_0 + (\theta_1 + \beta_2) jc + \beta_2 univ + \beta_3 exp + u \\ &= \beta_0 + \theta_1 jc + \beta_2 (jc + univ) + \beta_3 exp + u \\ &= \beta_0 + \theta_1 jc + \beta_2 totcoll + \beta_3 exp + u \end{aligned} \tag{1}$$

Thus the parameter of interest is now the coefficient of jc, thus by creating a variable jc+college and running the regression we directly obtain the standard error of θ_1.

Example rewriting model

```
1 . reg wage jc univ exper
```

Source	SS	df	MS
Model	32213.1516	3	10737.7172
Residual	158080.038	6759	23.3880808
Total	190293.19	6762	28.1415543

Number of obs =	6763
F(3, 6759) =	459.11
Prob > F =	0.0000
R-squared =	0.1693
Adj R-squared =	0.1689
Root MSE =	4.8361

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
jc	.6078403	.0767774	7.92	0.000	.4573325	.7583481
univ	.7969221	.0259575	30.70	0.000	.7460373	.8478069
exper	.0415971	.0017705	23.49	0.000	.0381263	.0450678
_cons	3.802714	.236784	16.06	0.000	3.338543	4.266885

```
2 . reg wage jc totcoll exper
```

Source	SS	df	MS
Model	32213.1516	3	10737.7172
Residual	158080.038	6759	23.3880808
Total	190293.19	6762	28.1415543

Number of obs =	6763
F(3, 6759) =	459.11
Prob > F =	0.0000
R-squared =	0.1693
Adj R-squared =	0.1689
Root MSE =	4.8361

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
jc	-.1890818	.0779816	-2.42	0.015	-.3419503	-.0362132
totcoll	.7969221	.0259575	30.70	0.000	.7460373	.8478069
exper	.0415971	.0017705	23.49	0.000	.0381263	.0450678
_cons	3.802714	.236784	16.06	0.000	3.338543	4.266885

Command in stata

```
1 . reg wage jc univ exper
```

Source	SS	df	MS
Model	**32213.1516**	**3**	**10737.7172**
Residual	**158080.038**	**6759**	**23.3880808**
Total	**190293.19**	**6762**	**28.1415543**

Number of obs =	**6763**
F(3, 6759) =	**459.11**
Prob > F =	**0.0000**
R-squared =	**0.1693**
Adj R-squared =	**0.1689**
Root MSE =	**4.8361**

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
jc	**.6078403**	**.0767774**	**7.92**	**0.000**	**.4573325**	**.7583481**
univ	**.7969221**	**.0259575**	**30.70**	**0.000**	**.7460373**	**.8478069**
exper	**.0415971**	**.0017705**	**23.49**	**0.000**	**.0381263**	**.0450678**
_cons	**3.802714**	**.236784**	**16.06**	**0.000**	**3.338543**	**4.266885**

```
2 . lincom jc-univ
```

(1) **jc - univ = 0**

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
(1)	**-.1890818**	**.0779816**	**-2.42**	**0.015**	**-.3419503**	**-.0362132**

Testing multiple linear restrictions

Test of joint hypotheses

- A joint hypothesis is a test that imposes two or more restrictions on the regression coefficients.
- The general joint hypothesis is of the form:

$$H_0 : \beta_j = \beta_{j,0}, \beta_m = \beta_{m,0} ...., \text{for a total of q restrictions}$$

- With the alternative hypothesis:

$$H_1 : \text{One or more of the q restrictions under } H_0 \text{ does not hold.}$$

- If the null hypothesis concerns the value of two coefficients we say that the null hypothesis imposes two restrictions on the multiple regression model.
- If any one of the equalities under the null hypothesis is false the joint null hypothesis itself is false.

Test for exclusion

- A common joint hypothesis test is that a set of q variables all equal to zero.
- For example if we test the hypothesis that the coefficients of x_j and x_k are jointly significant we test:

$$H_0 : \beta_j = \beta_k = 0 \text{ vs.}$$

$$H_1 : \beta_j \neq 0 \text{ and/or } \beta_k \neq 0.$$

- If the null is not rejected we say that x_j and x_k are jointly insignificant which often justifies dropping them from the equation.
- If the null is rejected we say that the variables are jointly statistically significant at the given significance level.
- A joint hypothesis test does not give information about which of the variables has a partial effect on y if the null is rejected.
- Note:
  - The variables can be jointly significant even if all the included variables are individually insignificant.
  - The variables can be jointly insignificant even when one (or more) of the variables are individually significant.

Example joint hypotheses

If the regression model is:

$$price = \beta_0 + \beta_1 assess + \beta_2 lotsize + \beta_3 sqrft + \beta_4 bdrms + u$$

- where assess is the assessed housing value before the house was sold, a null hypothesis may be that none of the other included variables affect price once assess is accounted for.
- Thus means that the hypothesis can be written as: $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$, or in other words, once the assessed housing value is controlled for house statistics such as lot size, house size and number of bedrooms have no partial effect on price.
- The null in this setting has three exclusion restrictions thus q=3.

Test statistic for joint hypothesis

Can we consider the joint test as a repeated t-test where we test each of the variables with the corresponding t-values?

- Assume that we want to test:

$$H_0 : \beta_1 = \beta_{1,0} \text{ and } \beta_2 = \beta_{2,0}$$

- against

$$H_0 : \beta_1 \neq \beta_{1,0} \text{ and/or } \beta_2 \neq \beta_{2,0}$$

- We can regard this as a joint null hypothesis made up of:

$$H_0' : \beta_1 = \beta_{1,0} \text{ and } H_0'' = \beta_2 = \beta_{2,0}$$

- assume as a simplification that the t-tests for the two null hypothesis are stochastically independent with significance level ϵ_1 and ϵ_2.

Significance level for repeated t-test

The overall significance level is then:

$$
\begin{aligned}
& P(\text{reject either } H_0' \text{ or } H_0''|H_0) = \\
1- & P(\text{reject neither } H_0' \text{ nor } H_0''|H_0) = \\
1- & (1-\epsilon_1)(1-\epsilon_2) \\
& \epsilon_1 + \epsilon_2(1-\epsilon_1)
\end{aligned}
$$

If you set the two significance levels equal to each other then:

$$P(\text{reject either } H_0' \text{ or } H_0''|H_0) = \epsilon + \epsilon(1-\epsilon) > \epsilon$$

Thus the significance level of this joint test is larger than the level of each individual test.

Repeated t-test

- The previous illustration indicates that testing the variables individually gives another significance level than the one specified for each hypothesis.
- Then what should constitute rejecting at each significance level?
- The Bonferroni test corrects the individual significance level so that the significance level of the test equals the desired significance level.
- For example if we want an overall significance level of 5% in our example then we should correct ϵ so that it satisfies
  $0.05 = \epsilon + \epsilon(1 - \epsilon)$

The Bonferroni test of a joint hypothesis

- In general the Bonferroni test can be conducted even when the t-statistics are correlated.
- The overall significance level of α is secured by choosing the significance level of each test so that:

$$\epsilon = \frac{\alpha}{m} \text{ (Bonferroni)}$$

- Where m is the number of individual tests.
- However, this is not the preferred method of testing as the t-statistic is calculated without restriction on the other parameters.
- The F-test is the preferred method as it is a better test, but the Bonferroni method may be useful if you only have the regression results and not the data.

The F-test

- The F-test is the preferred method for testing joint hypotheses.
- The F-test requires that you run two regressions.
- Lets call the full model, the model with all the included variables the unrestricted model.

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$$

- The model without the q variables we hypothesize is zero is then called the restricted model. $H_0 : \beta_{k-q+1} = 0, ...\beta_k = 0$
- General:

$$y = \beta_0 + \beta_1 + ... + \beta_{k-q} x_{k-q} + u$$

- Example unrestricted

$$price = \beta_0 + \beta_1 assess + \beta_2 lotsize + \beta_3 sqrft + \beta_4 bdrms + u$$

- Example restricted.

$$price = \beta_0 + \beta_1 assess + u$$

- The restricted model always has fewer parameters than the unrestricted model.

The F-statistic

- To test statistical significance we need to compute a statistic which we know the sampling distribution of under the null hypothesis.
- Under the null hypothesis, and assuming that the OLS assumptions hold, F is distributed as an F random variable with (q,n-k-1) degrees of freedom.
- This is written as $F \sim F_{q,n-k-1}$
- The distribution of $F_{q,n-k-1}$ is tabulated in the appendix for conventional significance levels.
- In large samples the F statistic is distributed $F_{q,\infty}$
- As with the t-statistic we will reject the null hypothesis when F is sufficiently large. $F > c$.

The F-statistic

- The SSR of the restricted model(SSR_R) is greater than the SSR of the unrestricted model(SSR_{UR}) as there are more factors in the error term.
- This difference in SSR can be used to test the joint hypothesis.
- Heuristically, we reject the joint H_0 if SSR_R is significantly larger than SSR_{UR}
- The F-statistic is used for testing whether the increase in SSR from the unrestricted model to the restricted model is large enough to warrant the rejection of the null hypothesis.

$$F \equiv \frac{(SSR_r + SSR_{ur})/q}{SSR_{UR}/(n - k_{ur} - 1)}$$

The F-statistic and R^2

- Since SSR is the main element in R^2 the formula for F can be written in terms of the R^2

$$F = \frac{(R^2_{UR} - R^2_R)/q}{(1 - R^2 UR)/(n - k_{ur} - 1)}$$

Heteroskedasticity and the F-statistic

- Controlling for heteroskedasticity in the regression adjusts the standard errors which makes the statistics that rely on the standard errors reliable.
- However, the robust command does not affect the SSR of the regression.
- The formulas given here is only correct if the errors are homoskedastic.
- However, Stata can compute heteroskedasticity robust F-statistic.

Example homoskedastic F-using formula

```
1 . reg testscr str expn_stu el_pct, robust

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	-.2863992	.4820728	-0.59	0.553	-1.234002	.661203
expn_stu	.0038679	.0015807	2.45	0.015	.0007607	.0069751
el_pct	-.6560227	.0317844	-20.64	0.000	-.7185008	-.5935446
_cons	649.5779	15.45834	42.02	0.000	619.1917	679.9641

```
2 . reg testscr el_pct, robust

Linear regression                               Number of obs =         420
                                                F(  1,   418) =      436.58
                                                Prob > F      =      0.0000
                                                R-squared     =      0.4149
                                                Root MSE      =      14.592
```

testscr	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
el_pct	-.6711562	.0321211	-20.89	0.000	-.7342952	-.6080172
_cons	664.7394	.9740374	682.46	0.000	662.8248	666.6541

```
3 . display (0.4366-0.4149)/2/((1-0.4366)/(420-3-1))
8.0113596
```

Example homoskedastic F-using Stata

```
1 . reg testscr str expn_stu el_pct
```

Source	SS	df	MS
Model	66409.8837	3	22136.6279
Residual	85699.7099	416	206.008918
Total	152109.594	419	363.030056

Number of obs =	420
F(3, 416) =	107.45
Prob > F =	0.0000
R-squared =	0.4366
Adj R-squared =	0.4325
Root MSE =	14.353

testscr	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
str	-.2863992	.4805232	-0.60	0.551	-1.230955	.658157
expn_stu	.0038679	.0014121	2.74	0.006	.0010921	.0066437
el_pct	-.6560227	.0391059	-16.78	0.000	-.7328924	-.5791529
_cons	649.5779	15.20572	42.72	0.000	619.6883	679.4676

```
2 . test str=expn_stu = 0

 ( 1)  str - expn_stu = 0
 ( 2)  str = 0

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

Example heteroskedasticity robust F

```
1 . reg testscr str expn_stu el_pct, robust

  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 |   -.2863992   .4820728    -0.59   0.553    -1.234002     .661203
    expn_stu |    .0038679   .0015807     2.45   0.015     .0007607    .0069751
      el_pct |   -.6560227   .0317844   -20.64   0.000    -.7185008   -.5935446
       _cons |    649.5779   15.45834    42.02   0.000     619.1917    679.9641
------------------------------------------------------------------------------

2 . test str=expn_stu = 0

 ( 1)  str - expn_stu = 0
 ( 2)  str = 0

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

Example joint significance

```
1 . reg lsalary years gamesyr hrunsyr rbisyr bavg
```

Source	SS	df	MS
Model	**308.989247**	**5**	**61.7978493**
Residual	**183.186322**	**347**	**.527914472**
Total	**492.175568**	**352**	**1.39822605**

Number of obs =	**353**
F(5, 347) =	**117.06**
Prob > F =	**0.0000**
R-squared =	**0.6278**
Adj R-squared =	**0.6224**
Root MSE =	**.72658**

lsalary	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
years	**.0688626**	**.0121145**	**5.68**	**0.000**	**.0450354**	**.0926898**
gamesyr	**.0125521**	**.0026468**	**4.74**	**0.000**	**.0073464**	**.0177578**
hrunsyr	**.0144295**	**.016057**	**0.90**	**0.369**	**-.0171517**	**.0460108**
rbisyr	**.0107657**	**.007175**	**1.50**	**0.134**	**-.0033462**	**.0248776**
bavg	**.0009786**	**.0011035**	**0.89**	**0.376**	**-.0011918**	**.003149**
_cons	**11.19242**	**.2888229**	**38.75**	**0.000**	**10.62436**	**11.76048**

```
2 . test  (hrunsyr=0) (rbisyr=0)

 ( 1)  hrunsyr = 0
 ( 2)  rbisyr = 0

       F(  2,   347) =    13.49
            Prob > F =     0.0000
```

- The F statistic is often useful for testing exclusion of a group of variables when the variables in the group are highly correlated.

Testing general linear restrictions

- The examples given here test whether a set of independent variables all equal to zero.
- However, the F-test can be used to test more complicated restrictions.
- The approach is still to compute the unrestricted model and impose the restrictions to obtain the restricted model.
- It is however a bit more complicated to obtain the unrestricted model in these cases as it might involve redefining the dependent variable.
- Note: if the dependent variable is different the R^2 formula cannot be used as $TSS_{UR} \neq TSS_R$

The F-test with q=2 restrictions

If we want to test the joint hypothesis that β_1 and β_2 is both equal to zero we can write:

$$H_0 : \beta_1 = 0 \text{ and } \beta 2 = 0 \text{ vs. } \beta_1 \neq 0 \text{ or } \beta_2 \neq 0.$$

- To test the hypothesis we need to obtain the F-statistic
- In the two restriction case:

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

- Where t_1 is the t-statistic associated with β_1 and t_2 is the t-statistic associated with β_2.
- And $\hat{\rho}_{t_1,t_2} t_1, t_2$ is an estimator of the correlation between the two t-statistics.
- However $\hat{\rho}_{t_1,t_2}$ is not easily obtained, thus the SSR formula is preferred.

Relationship between F and t statistic

- When q=1 the F-statistic tests a single restriction.
- Thus it is analogous to testing a single coefficient.
- The F-statistic is the square of the t-statistic.
- Since t^2_{n-k-2} has an $F_{1,n-k-1}$ distribution the two approaches lead to exactly the same outcome, provided that the alternative is two-sided.
- The t-statistic is more flexible for a single hypothesis as it can also be used for one-sided tests.
- In addition the t statistic is easier obtained than the F statistic

Relationship between F and t-statistic

The following example shows that a variable can be individually significant, but jointly insignificant with another variable.

```
1 . reg wage jc univ ne nc south black hispanic, robust

Linear regression                               Number of obs =      6763
                                                F(  7,  6755) =    118.05
                                                Prob > F      =    0.0000
                                                R-squared     =    0.1114
                                                Root MSE      =    5.0033
```

wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
jc	.5867815	.077754	7.55	0.000	.4343592	.7392038
univ	.7175047	.0286009	25.09	0.000	.6614379	.7735716
ne	.2013285	.2045864	0.98	0.325	-.1997254	.6023824
nc	-.4011147	.1922327	-2.09	0.037	-.7779514	-.0242779
south	-.6561036	.1870899	-3.51	0.000	-1.022859	-.2893485
black	-1.211106	.1966389	-6.16	0.000	-1.59658	-.8256322
hispanic	-.2937563	.2782542	-1.06	0.291	-.8392222	.2517096
_cons	9.474505	.1663662	56.95	0.000	9.148375	9.800635

```
2 . test nc=hispanic=0

 ( 1)  nc - hispanic = 0
 ( 2)  nc = 0

       F(  2,  6755) =      2.37
            Prob > F =      0.0932
```

The p-value of the F-statistic

- If you a large sample you can use the large sample $F_{q,\infty}$ approximation to compute the p-value.
- The p-value is useful since the F distribution depends on the degrees of freedom in the numerator and denominator and thus it is hard to get a feel for how strong or weak the evidence is against the null by simply looking at the value of the F statistic and one or two critical values.

$$p - value = PR[F_{q,\infty} > F^{act}]$$

- The p-value is still the probability of observing a value of F at least as large as we did, given that the null hypothesis is true.

Confidence sets for multiple coefficients

- A 95% confidence set for two or more coefficients is a set that contains the true values of these coefficients in 95% of randomly drawn samples.
- Thus a confidence set is the generalization to two or more coefficients of a confidence interval for a single coefficient.
- The 95% confidence set contains the set of values not rejected at the 5% significance level by the F-statistic.
- A confidence set gives the same conclusion about joint statistical significance as the F-test.

Confidence sets for multiple coefficients

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

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

Now

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

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

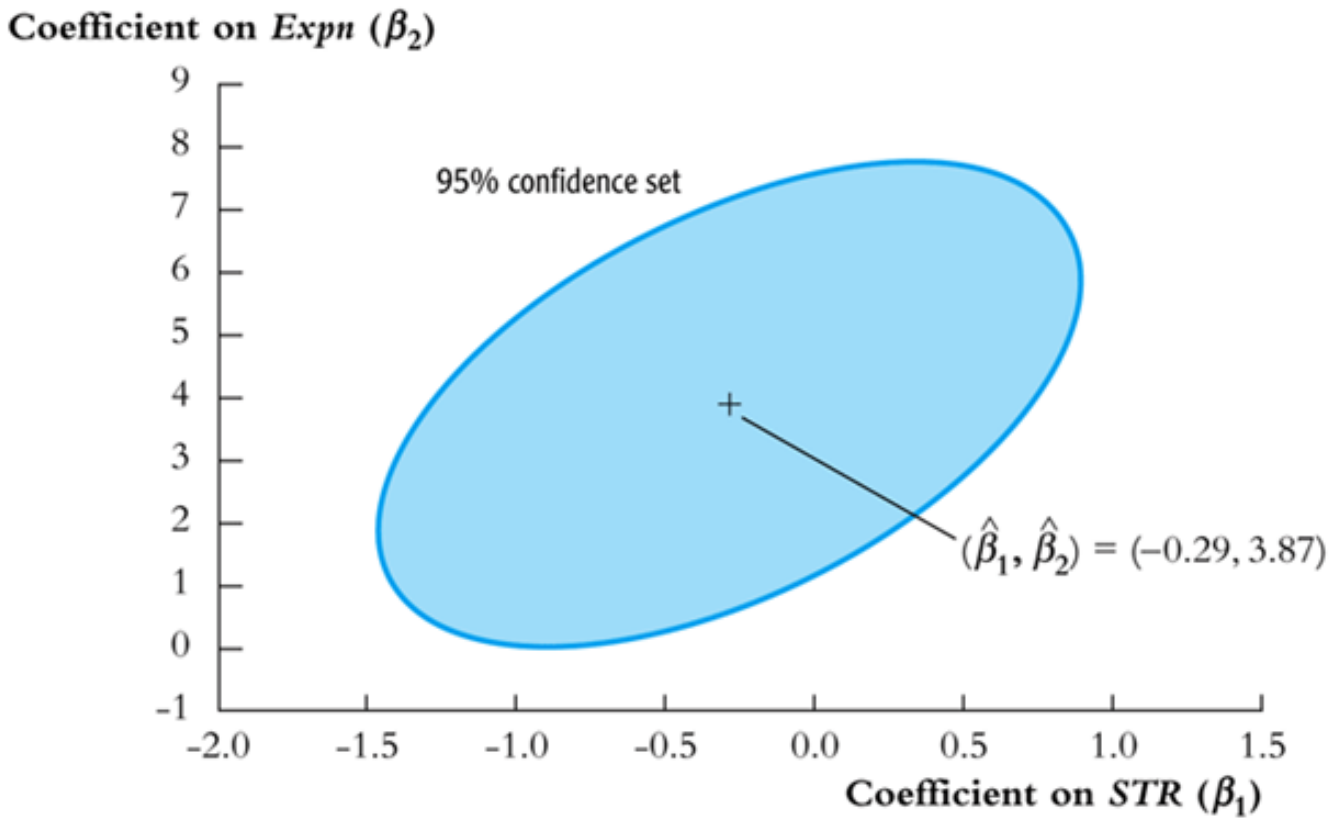
$$\left[\left(\frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)}\right)^2 + \left(\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right)^2 + 2\hat{\rho}_{t_1,t_2}\left(\frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}\right)\left(\frac{\hat{\beta}_2 - \beta_{2,0}}{SE(\hat{\beta}_2)}\right)\right]$$

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

Confidence sets for multiple coefficients

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 F-statistic of the regression

- Stata automatically reports the F statistic of the regression.
- This is the statistic of the test that **all** the slope coefficients are zero.
- Under this null hypothesis none of the regressors explains any of the variation in Y_i.
- The restricted model is then:

$$Y = \beta_0 + u$$

- And the F-statistic is then:

$$\frac{R^2/k}{(1-R^2)/(n-k-1)}$$

- This F-statistic determines the overall significance of the regression.

Non-nested models

- The restricted model is a nested version of the unrestricted as all the parts in the restricted model is included in the unrestricted model.
- Two variables can be individually insignificant in the full model, but jointly significant.
- This suggests that the variables are highly correlated, which means that if you included only one of them in the regression it might be individually significant.
- However, which should you include?
- In the general case there might be many variables you find jointly significant, then you could specify multiple models, one for each of the dependent variables, but how should you choose between the models?

Non-nested models

$$lsalary = \beta_0 + \beta_1 years + \beta_2 bavg + \beta_3 gamesyr + \beta_4 hrunsyr + u$$

$$lsalary = \beta_0 + \beta_1 years + \beta_2 bavg + \beta_3 gamesyr + \beta_4 rbisyr + u$$

- Where salary is yearly salary of a baseball player, bavg is batting average, hrunsyr is homeruns per year and rbisyr is runs batted in per year.
- If both of hrunsyr and rbisyr are included in the same regression they are individually insignificant as they are so strongly correlated, while they are significant if they are included separately.
- The adjusted R-squared can serve as an indicator for which model is to prefer.
- Note: The dependent variable of the two models must be on the same functional form.

Testing linear restrictions

- The t test is used when you have only one restriction such as:
  - $\beta_1 = \beta_{1,0}$
  - $\beta_1 + \beta_2 = \theta_{1,0}$
- The F test is used whenever you have multiple restrictions such as:
  - $\beta_1 = 0$ and $\beta_2 = 0$

Variable selection

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 so serve as control variables.
4. Sensitivity check your model by alternative specifications.

Control variables

Control variable

A control variable W is a variable that is correlated with, and controls for, an omitted causal factor in the regression of Y on X, but which itself does not necessarily have a causal effect on Y.

$$TestScore = \beta_0 + \beta_1 STR + \beta_2 PctEL + \beta_3 LchPct$$

- STR is the student teacher ratio and the variable of interest
- PctEL is the percentage of english learners and probably has a direct causal effect, but it also serves as a control as it is correlated with outside learning opportunities.
- LctPct might have cauasal effect and is correlqated with and controls for income-related outside learning opportunities.

Control variables

- An effective control variable is one which, when included in the regression, makes the error term uncorrelated with the variable of interest.
- Holding constant the control variable(s), the variable of interest is as if randomly assigned.
- Among individuals (entities) with the same value of the control variable(s), the variable of interest is uncorrelated with the omitted determinants of Y.

Control variables

- Control variables are selected because they are correlated with omitted factors, that means that they are likely to be biased.
- This means that the zero conditional mean assumption will not hold.
- However, the control variable is effective if the mean of u does not depend on the variables of interest given the control variable. This is called the conditional mean independence.

Conditional mean independence

Conditional mean independence

$E(u|X,W)=E(u|W)$

- The zero conditional mean assumption ensures unbiased estimates however it a strong assumption and can be replaced by the weaker conditional mean independence assumption.
- If the conditional mean independence assumption holds the variable of interest has a causal interpretation (but the control variables are potentially biased).

Sensitivity check

- Start with specifying your base specification which contains the variables of primary interest and the control variables suggested by economic theory and expert judgement.
- Develop a list of candidate alternative specifications.
- If the estimates of the coefficients are numerically similar across the alternative specifications, then this provides evidence that the estimates from your base specification are reliable.
- If they are not similar this often is evidence that the original specification had omitted variable bias.

Selecting variables to include in your model

- Note that an increase in adjusted R^2 does not necessarily mean that an added variable is statistically significant. Thus perform an hypothesis test using the t-statistic.
- A high R^2 does not mean that the model is correctly specified or that there is no omitted variable bias. Thus you need to do reasoning and testing alternative specifications independent of R^2.
- The question of what constitutes the right set of regressors is difficult as you must weigh issues of omitted variable bias, data availability, data quality and economic theory.

Reporting regression results

Reporting regression results

- The estimated OLS coefficients should always be reported and along with it the standard errors and the number of observations used in estimation.
- Some authors prefer to report the t-statistic, but the standard errors are to prefer.
- For the key variables in an analysis you should interpret the estimated coefficients.
- The economic importance of the estimates of the key variables should be discussed.
- The R-squared from the regression should always bee included.
- If only a couple of models are being estimated the results can be summarized in equation form, but in many cases a table (or multiple tables) is to prefer.
- The dependent variable should be indicated clearly in the table, and the independent variables should be listed in the first column.

Reporting regression results

- You should think about the scale of the variables so that it easy to read and interpret your regression results.
- It is common to indicate significance levels with stars.
- If there is any relevant F-statistic then you should report this.

California test score data set

- The book throughout the first chapters use a data set constituting the tests scores of Californian students.
- The primary interest is in establishing whether the student teacher ratio (STR) has a causal effect on the student tests scores.
- Factor such as outside learning opportunities are correlated with STR and provides potential of OVB.
- These factors are not directly measurable, but we can include control variables that are correlated with these omitted factors.
- If the control variables are adequate in the sense that the conditional mean independence assumption holds, then we can give the coefficient a causal interpretation.

California test score data set

- Potential background variables:
  - Percentage of students who are still learning English.
  - The percentage of students who are eligible for a subsidized or free lunch.
  - The percentage of students whose faimilies qualify for a California income assistance program.

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)	−1.10* (0.43)	−1.00** (0.27)	−1.31** (0.34)	−1.01** (0.27)
Percent English learners (X_2)		−0.650** (0.031)	−0.122** (0.033)	−0.488** (0.030)	−0.130** (0.036)
Percent eligible for subsidized lunch (X_3)			−0.547** (0.024)		−0.529** (0.038)
Percent on public 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
$\overline{R}^2$	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. The individual coefficient is statistically significant at the *5% level or **1% significance level using a two-sided test.

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.