Applied Statistics and Econometrics

Lecture 7

Saul Lach

September 2017

Outline of Lecture 7

1. **Empirical example: Italian labor force survey (LFS):**
   1. Wages and education
   2. Wages and gender
   3. Wages and labor experience
2. Using "dummy variables".
   1. Regional differences in wages.
3. Testing joint hypotheses (SW 7.2)
   1. Testing for regional differences in wages
4. Regression specification (SW 7.5)

Italian labor force survey

- The Italian Labour Force Survey (LFS) provides individual level data on labour market variables (employment status, type of work, age, job search, etc.).

```
## RETRIC ETAM DETIND TISTUD REG SG11 SG16
##  1    1530    50     2     10    10     2     1
##  6    1600    61     2     10    14     2     1
##  7    1500    46     2     10    17     2     1
## 10    2800    43     2     3     1      1     1
## 11    1300    33     2     4     4      1     2
## 12     940    38     1     3     4      2     2
## 16    1700    57     2     5     18     1     1
## 21    2180    32     2     5     1      1     1
## 25    1470    52     2     4     10     1     1
## 26     700    50     2     5     10     2     1
```


LFS

The variables in the LFS:

1. `RETRIC` - Net monthly wage
2. `ETAM` - Age
3. `DETIND` - Temporary/full time worker
4. `TISTUD` - Educational attainment
5. `SG24b` - Educational attainment (>BA)
6. `REG` - Region
7. `SG11` - Gender
8. `SG16` - Italian Citizen

RETRIC

RETRIC **Retribuzione netta del mese scorso escluse altre mensilità (tredicesima, quattordicesima, ecc.) e voci accessorie non percepite regolarmente tutti i mesi (premi di produttività annuali, arretrati, indennità per missioni, straordinari non abituali, ecc.)**

Fino a 250 euro	250
260	260
270	270
----------	------
----------	------
----------	------
2980	2980
2990	2990
3000 euro o più	3000

TISTUD

TISTUD **Titolo di studio a 10 modalità**

Nessun titolo	1
Licenza elementare / Attestato di valutazione finale	2
Licenza media (dall'anno 2007 denominata "Diploma di Istruzione secondaria di I grado") o avviamento professionale (conseguito non oltre all'anno 1965)	3
Diploma di qualifica professionale di scuola secondaria superiore (di II grado) di 2-3 anni che non permette l'iscrizione all'Università /Attestato IFP di qualifica professionale triennale (operatore)/ Diploma professionale IFP di tecnico (quarto anno) (dal 2005)	4
Diploma di maturità / Diploma di istruzione secondaria superiore (di II grado) di 4-5 anni che permette l'iscrizione all'Università/Certificato di specializzazione tecnica superiore IFTS (dal 2000)/ Diploma di tecnico superiore ITS (corsi biennali) (dal 2013)	5
Diploma di Accademia (Belle Arti, Nazionale di arte drammatica, Nazionale di Danza), Istituto superiore Industrie artistiche, Conservatorio di musica statale, Istituto di Musica Pareggiato	6
Diploma universitario di due/tre anni, Scuola diretta a fini speciali, Scuola parauniversitaria	7
Laurea di primo livello (triennale)	8
Laurea specialistica/magistrale biennale	9
Laurea di 4-6 anni: laurea del vecchio ordinamento o laurea specialistica/magistrale a ciclo unico	10

SG24

SG24B. Ha conseguito un titolo di studio post-laurea, post-diploma accademico AFAM o dottorato di ricerca?

- Master universitario di I livello/ Diploma accademico di perfezionamento o Master di I livello/ Diploma accademico di specializzazione di I livello 1 *(passare a SG25)*
- Master universitario di II livello/ Diploma accademico di perfezionamento o Master di II livello/ Diploma accademico di specializzazione di II livello 2 *(passare a SG25)*
- Diploma di specializzazione universitaria 3*(passare a SG25)*
- Dottorato di ricerca/Diploma accademico di formazione alla ricerca AFAM 4*(passare a SG25)*
- Nessuno di questi 5*(passare a SG25)*

REGION

ALLEGATO: REGIONI

Piemonte	*01*	*Marche*	*11*
Valle d'Aosta	*02*	*Lazio*	*12*
Lombardia	*03*	*Abruzzo*	*13*
Trentino Alto Adige	*04*	*Molise*	*14*
Veneto	*05*	*Campania*	*15*
Friuli Venezia Giulia	*06*	*Puglia*	*16*
Liguria	*07*	*Basilicata*	*17*
Emilia Romagna	*08*	*Calabria*	*18*
Toscana	*09*	*Sicilia*	*19*
Umbria	*10*	*Sardegna*	*20*

GENDER

SG11. Sesso del componente

- Maschio 1
- Femmina 2

CITIZEN

SG16. Cittadinanza italiana

- Sì 1
- No 2 *(passare a SG17)*

Wages and education in Italy

- Using the labor force survey we will try to get a sense of the relationship between wages and education in Italy by estimating a linear regression model.
- We won't be able to interpret these estimates as causal effects because we are not able to control for omitted variables that are unobservable (e.g., ability) and, therefore, the estimates are likely to be subject to OVB.
- Nevertheless, the exercise has some merit as we will able to learn something about the correlation between wages and education.

Measuring education in the LFS

- We measure education by the variable TISTUD now (in Lecture 5 we used educ_lev).
- We first recode the categorical variable TISTUD into numbers using:

```
recode tistud (1=0)(2=5)(3=13)
(4=10)(5=13)(6=16)(7=15)(8=16)(9=19) (10=19), gen(educ)

tab educ if    retric~=.
```

RECODE of tistud	Freq.	Percent	Cum.
0	142	0.54	0.54
5	700	2.68	3.22
10	2,289	8.76	11.98
13	18,040	69.05	81.03
15	364	1.39	82.42
16	821	3.14	85.57
19	3,771	14.43	100.00
Total	26,127	100.00	

Wages and years of education

$$\widehat{RETRIC} = 627.96 + 50.51 \times educ$$

```
. reg retric educ, robust

Linear regression                               Number of obs     =     26,127
                                                F(1, 26125)       =   1817.47
                                                Prob > F          =    0.0000
                                                R-squared         =    0.0819
                                                Root MSE          =    500.67
```

retric	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	50.50485	1.184677	42.63	0.000	48.18282	52.82689
_cons	627.9557	15.51411	40.48	0.000	597.5472	658.3642

Gender gap

Gender gap refers to systematic differences in the outcomes that men and women achieve in the labor market.

The simplest gender gap model is $wage = \beta_0 + \beta_1 female + u$

$$\widehat{RETRIC} = 1444.54 - 291.36 \times female$$

```
. reg retric female,robust

Linear regression                               Number of obs     =     26,127
                                                F(1, 26125)       =   2212.72
                                                Prob > F          =    0.0000
                                                R-squared         =    0.0775
                                                Root MSE          =    501.86
```

retric	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-291.3592	6.193919	-47.04	0.000	-303.4997	-279.2188
_cons	1444.535	4.423072	326.59	0.000	1435.866	1453.205

Wages and (potential) work experience

- Potential experience is the maximum years of experience of an individual in the job market.
- Actual experience is usually unavailable. Potential experience is defined as "current age - age at graduation". We do not observe age at graduation in the LFS, so we will use AGE to proxy for experience.

Wages and (potential) work experience

$$\widehat{RETRIC} = 806.98 + 11.37 \times age$$

```
. reg retric etam, robust

Linear regression                               Number of obs     =     26,127
                                                F(1, 26125)       =    1593.01
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0564
                                                Root MSE          =     507.57
```

retric	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
etam	11.3671	.2848005	39.91	0.000	10.80887	11.92532
_cons	806.9773	12.11921	66.59	0.000	783.223	830.7316

Multiple regression model

$$
\begin{aligned}
wage &= \beta_0 + \beta_1 educ + \beta_2 age + \beta_3 female + u \\
\widehat{wage} &= 127.03 + 57.79 \times educ + 12.76 \times age - 334.89 \times female
\end{aligned}
$$

```
. reg retric educ etam female,robust

Linear regression                               Number of obs     =     26,127
                                                F(3, 26123)       =    2323.47
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2462
                                                Root MSE          =     453.67

------------------------------------------------------------------------------
             |               Robust
      retric |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   57.90433   1.083718    53.43   0.000     55.78018    60.02847
        etam |   12.76084   .2579002    49.48   0.000     12.25534    13.26634
      female |  -334.7123   5.662991   -59.11   0.000    -345.8121   -323.6126
       _cons |   125.7257   19.19582     6.55   0.000     88.10087    163.3506
------------------------------------------------------------------------------
```


Where are we?

1. Empirical example: Italian labor force survey (LFS):
   1. Wages and education
   2. Wages and gender
   3. Wages and labor experience
2. **Using "dummy variables".**
   1. Regional differences in wages.
3. Testing joint hypotheses (SW 7.2)
   1. Testing for regional differences
4. Regression specification (SW 7.5)

Regional differences in wages

- Are there regional differences in wages?
- Consider using the variable RIP3

RIP3 **Ripartizione geografica in 3 classi**

- Nord
- Centro
- Mezzogiorno

- RIP3 represents **categorical** data.
- Need to recode RIP3 for use in a regression.

Regional differences: using dummy variables

- To examine regional differences in wages, we suggest the following regression:

$$wage_i = \beta_0 + \beta_1 South_i + \beta_2 Center_i + \beta_3 North_i + u_i$$

where South, Center and North are variables created from the original RIP3 variable:

1. $South_i$ is a variable taking value 1 if individual i resides in the South of Italy, and 0 otherwise.
2. $Center_i$ is a variable taking value 1 if individual i resides in the Center of Italy, and 0 otherwise.
3. $North_i$ is a variable taking value 1 if individual i resides in the North of Italy, and 0 otherwise.

- These are examples of "**dummy variables**" (also called binary or indicator variables).

The dummy variable trap

- The suggested model is

$$wage_i = \beta_0 + \beta_1 South_i + \beta_2 Center_i + \beta_3 North_i + u_i$$

- This regression, however, is . . . **wrong**: it will not "run" because of **perfect multicollinearity**.
- Because the regional categories are exclusive and exhaustive they must add to one. That is:

$$South_i + Center_i + North_i = 1$$

- This means that the intercept (the "variable" having 1 in all n observations) and *South*, *Center* and *North* are **perfectly collinear**.
  - Perfect collinearity: one regressor is a perfect (exact) linear combination of other regressors.
- This is an example of the **dummy variable trap**.

The dummy variable trap

Note that the dummies always add up to one.

```
. l South Center North in 1/10
```

	South	Center	North
1.	0	1	0
2.	0	0	1
3.	0	0	1
4.	0	0	1
5.	1	0	0
6.	1	0	0
7.	1	0	0
8.	0	0	1
9.	0	0	1
10.	0	0	1

```
. g all=South+Center+North

. sum all
```

Variable	Obs	Mean	Std. Dev.	Min	Max
all	101,916	1	0	1	1

The dummy variable trap in Stata

If dummy variables for **all** regions (categories) **and** the intercept are included in the regression, Stata automatically drops one category in order to avoid perfect multicollinearity and be able to run the regression.

```
. reg retric South Center North,robust
note: South omitted because of collinearity

Linear regression                               Number of obs     =     26,127
                                                F(2, 26124)       =     145.75
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0106
                                                Root MSE          =     519.75
```

retric	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
South	0	(omitted)				
Center	68.02011	9.78381	6.95	0.000	48.8433	87.19691
North	133.64	7.995994	16.71	0.000	117.9674	149.3126
_cons	1216.154	6.765793	179.75	0.000	1202.893	1229.416

The dummy variable trap

- In general, if you have a set of multiple binary (dummy) variables, which are mutually exclusive and exhaustive – that is, there are multiple categories (e.g., South, Center and North) and every observation falls in one and only one category – and include all these dummy variables **and** a constant in the regression, you will have perfect multicollinearity (and fall into the dummy variable trap).
- Solution: omit one of the categories to avoid the dummy variable trap (can also omit the constant/intercept but this is not recommended).

Avoiding the dummy variable trap: omit a category

- Suppose we omit **North** from the equation,

$$wage_i = \beta_0 + \beta_1 South_i + \beta_2 Center_i + u_i$$

- There is no perfect collinearity between regressors now.
- What is the interpretation of β_0 and β_1 in this regression?

$$\begin{aligned} E\left(wage|South = 1, Center = 0\right) &= \beta_0 + \beta_1 \\ E\left(wage|South = 0, Center = 0\right) &= \beta_0 \end{aligned}$$

- β_0 is the average wage of workers in the North.
- β_1 is the difference in average wage between workers in the South and workers in the North,

$$\beta_1 = \underbrace{E\left(wage|South = 1, Center = 0\right)}_{\text{avg wage in the South}} - \underbrace{E\left(wage|South = 0, Center = 0\right)}_{\text{avg wage in the North}}$$

- Similarly, β_2 is the difference in average wage between workers in the Center and workers in the North.

Regression with regional dummy variables, North omitted

$$wage_i = \beta_0 + \beta_1 South_i + \beta_2 Center_i + u_i$$

Workers in the South (Center) earn, on average, 133.6 (65.6) euro **less** than in the North (the omitted category).

```
. reg retric South Center,robust

Linear regression                               Number of obs     =     26,127
                                                F(2, 26124)       =     145.75
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0106
                                                Root MSE          =     519.75

------------------------------------------------------------------------------
             |               Robust
      retric |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       South |   -133.64   7.995994   -16.71   0.000    -149.3126   -117.9674
      Center |  -65.6199   8.252694    -7.95   0.000    -81.79563   -49.44417
       _cons |   1349.794   4.261451   316.75   0.000     1341.442    1358.147
------------------------------------------------------------------------------
```

Avoiding the dummy variable trap: omit another category

- Suppose we omit **South** from the equation,

$$wage_i = \beta_0 + \beta_2 Center_i + \beta_3 North_i + u_i$$

- What is the interpretation of β_2 and β_3 in this regression?

$$\begin{aligned} E\left(wage \middle| Center = 1, North = 0\right) &= \beta_0 + \beta_2 \\ E\left(wage \middle| Center = 0, North = 0\right) &= \beta_0 \end{aligned}$$

- β_0 is the average wage of workers in the South.
- β_2 is the difference in average wage between workers in the Center and workers in the South.
- β_3 is the difference in average wage between workers in the North and workers in the South.

Regression with regional dummy variables, South omitted

$$wage_i = \beta_0 + \beta_2 Center_i + \beta_3 North_i + u_i$$

Workers in the Center (North) earn, on average, 68 (133.6) euro **more** than in the South (the omitted category).

```
. reg retric Center North,robust

Linear regression                               Number of obs     =     26,127
                                                F(2, 26124)       =     145.75
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0106
                                                Root MSE          =     519.75

------------------------------------------------------------------------------
             |               Robust
      retric |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Center |   68.02011    9.78381     6.95   0.000      48.8433    87.19691
       North |     133.64   7.995994    16.71   0.000     117.9674    149.3126
       _cons |   1216.154   6.765793   179.75   0.000     1202.893    1229.416
------------------------------------------------------------------------------
```

Interpretation of coefficients of dummy variables

- Note that although the interpretation of the coefficients depends on the choice of the **omitted category** (sometime called the reference group/category), **their economic meaning do not change when the omitted category is changed:**
  - Workers in the South earn 133.6 euro less than in the North ($\hat{\beta}_1$ in first regression, $\hat{\beta}_3$ in second regression).
  - Workers in the Center earn 65.6 euro less than in the North ($\hat{\beta}_2$ in first regression, $\hat{\beta}_2 - \hat{\beta}_3$ in second regression).
  - Workers in the South earn 68 euro less than in the Center ($\hat{\beta}_1 - \hat{\beta}_2$ in first regression, $-\hat{\beta}_2$ in second regression).

Categorical variables and dummy variables

- If a categorical variable (e.g., RIP3) represents p (e.g., 3) categories we define a set of $p-1$ dummy variables and use them in the regression, along with the intercept.
- Doing this avoids the dummy variable trap.
- The coefficients of the $p-1$ dummy variables represent the difference between the average y in a category and the average y in the omitted category (or reference category), holding other things (regressors) constant.
- You can choose the omitted (reference) category according to your needs or the software does it for you automatically.

Avoiding the dummy variable trap: omit the intercept?

- We could have also omitted the intercept altogether and run

$$wage_i = \beta_1 South_i + \beta_2 Center_i + \beta_3 North_i + u_i$$

- There is no perfect collinearity between regressors now.
- What is the interpretation of the $\beta's$ in this regression?

$$\begin{aligned} E\left(wage|South = 1, Center = 0, North = 0\right) &= \beta_1 \\ E\left(wage|South = 0, Center = 1, North = 0\right) &= \beta_2 \\ E\left(wage|South = 1, Center = 0, North = 1\right) &= \beta_3 \end{aligned}$$

- Omitting the intercept, however, is **not recommended and never done in practice!**

Wages and education with controls

```
. reg retric educ etam female Center South,robust

Linear regression                               Number of obs     =     26,127
                                                F(5, 26121)       =    1535.35
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2647
                                                Root MSE          =      448.1

------------------------------------------------------------------------------
             |               Robust
     retric |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       educ |   58.36051   1.081296    53.97   0.000     56.24111    60.4799
       etam |   13.10358   .2559897    51.19   0.000     12.60183    13.60534
     female |   -342.248   5.613383   -60.97   0.000    -353.2506   -331.2455
     Center |   -95.4606   7.079332   -13.48   0.000    -109.3365   -81.58472
      South |  -174.0507   6.833489   -25.47   0.000    -187.4447   -160.6566
      _cons |   166.3771   19.17763     8.68   0.000     128.7879    203.9663
------------------------------------------------------------------------------
```

Wages and education with controls

$$\begin{aligned}\widehat{wage} &= 166.4 + 58.4 \times educ + 13.1 \times ETAM - 342.3 \times female \\ &\quad -95.5 \times Center - 174.1 \times South\end{aligned}$$

- Individuals living in the South make, on average, **174.1 euro less** than comparable individuals living in the North.
  - "Comparable" here refers to the observed regressors (education, age and gender).
  - We say that we are "**holding other factors constant**" or **"controlling for other factors"** in this comparison between South and North.
  - For 174.1 to be interpreted as a causal effect we need to assume that other unobserved factors in u are also the same on average between the regions. That is, we need to assume Assumption #1.
- Similarly, individuals living in the Center make, on average, **95.5 euro less** than comparable individuals living in the North, holding other factors constant.

Where are we?

1. Empirical example: Italian labor force survey (LFS):
   1. Wages and education
   2. Wages and gender
   3. Wages and labor experience
2. Using "dummy variables".
   1. Regional differences in wages.
3. **Testing joint hypotheses (SW 7.2)**
   1. Testing for regional differences
4. Regression specification (SW 7.5)

Are wages different across regions? Joint hypothesis

- To answer this question we want to compare the wages of "similar" individuals residing in different regions in Italy.
- The multiple regression can help us answer this question,

$$wage = \beta_0 + \beta_1 educ + \beta_2 ETAM + \beta_3 female + \beta_4 Center + \beta_5 South + u$$

- The hypothesis that there are no wage differences between the three regions in Italy is the **joint hypothesis**

$$\begin{aligned} H_0 &: \beta_4 = 0 \textbf{ and } \beta_5 = 0 \\ vs\ H_1 &: either\ \beta_4 \neq 0\ or\ \beta_5 \neq 0\ or\ both \end{aligned}$$

- If either of the equalities under H_0 is false, then the joint null hypotheses itself is false. Thus, H_1 is that at least one of the equalities in H_0 does not hold.
- A joint hypothesis specifies a value for two or more coefficients. That is, it imposes a **restriction** on two or more coefficients.
  - Previously, we used the t-statistic to test single parameters only.

Joint hypothesis

- In general, a joint hypothesis involves q restrictions:

$$\begin{aligned} H_0 &: \beta_j = \beta_{j,0},\ \beta_m = \beta_{m,0}, \ldots \text{ for a total of } q \text{ restrictions} \\ vs\ H_1 &: \text{one or more of the q restrictions under } H_0 \text{ does not hold} \end{aligned}$$

- In the regional example, $q = 2$ and the restrictions are $\beta_4 = 0$ and $\beta_5 = 0$.

Common sense does not work

- A "common sense" approach is to reject H_0 if either of the individual t-statistics exceeds 1.96 in absolute value.
- But this "common sense" approach doesn't work! It rejects the null when the null is indeed correct in more than 5% of the cases.
- Suppose that in a multiple regression model we want to test

$$\begin{aligned} H_0 &: \beta_1 = 0 \textbf{ and } \beta_2 = 0 \\ \textit{vs } H_1 &: \textit{either } \beta_1 \neq 0 \textit{ or } \beta_2 \neq 0 \textit{ or both} \end{aligned}$$

  and use the t-statistics

$$t_1 = \frac{\hat{\beta}_1 - 0}{SE\left(\hat{\beta}_1\right)}, \quad t_2 = \frac{\hat{\beta}_2 - 0}{SE\left(\hat{\beta}_2\right)}$$

- The "one at a time" test is: reject $H_0 : \beta_1 = \beta_2 = 0$ if $|t_1| > 1.96$ and/or $|t_2| > 1.96$.

Common sense does not work

- What is the probability that using this "one at a time" test rejects H_0, when H_0 is actually true? (The answer **should** be 5%).
- To simplify the computations we assume that t_1 and t_2 are independent (this is not true in general).
- Note that H_0 is **not** rejected only if both $|t_1| \leq 1.96$ and $|t_2| \leq 1.96$.
- Because of independence, the probability of not rejecting H_0 is:

$$\begin{aligned} \Pr\left[|t_1| \leq 1.96 \textit{ and } |t_2| \leq 1.96\right] &= \Pr\left[|t_1| \leq 1.96\right] \times \Pr\left[|t_2| \leq 1.96\right] \\ &= 0.95 \times 0.95 = 0.9025 \end{aligned}$$

- Thus the probability of rejecting H_0 when it is is true – called the "size" of the test – is

$$1 - .95^2 = 0.0975$$

  which is not the desired 5%!

Common sense test rejects too often

- The "one at a time" test rejects H_0 too often because it gives too many chances: if you fail to reject using t_1, you try again using t_2.
- When t_1 and t_2 are correlated, the computation is more complicated but the message is similar: the "one at a time" test has the wrong size (not equal to desired significance level).

Solutions

- How do we construct a test with a 5% size ?

1. 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.
2. Use a different test statistic designed to test β_1 and β_2 at once: the F test. This test is common practice and the focus of this Section.

The F statistic

- The F statistic tests all parts of a joint hypothesis **at once**.
- We start with a joint hypothesis involving $q = 2$ **restrictions.**
- The formula of the F statistic for testing the joint hypothesis when $q = 2$ is

$$\begin{aligned} H_0 &: \beta_1 = \beta_{1,0} \text{ and } \beta_2 = \beta_{2,0} \\ \text{vs } H_1 &: \text{either } \beta_1 \neq \beta_{1,0} \text{ or } \beta_2 \neq \beta_{2,0} \text{ or both} \end{aligned}$$

is

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

where r_{t_1,t_2} is the sample correlation coefficient between t_1 and t_2, and

$$t_1 = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE\left(\hat{\beta}_1\right)}, \quad t_2 = \frac{\hat{\beta}_2 - \beta_{2,0}}{SE\left(\hat{\beta}_2\right)}$$

The F statistic

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

- The F statistic corrects (in just the right way) for the correlation between t_1 and t_2.
- The F statistic is large when t_1 and/or t_2 is large.
- We therefore reject H_0 when F is large, i.e., above its critical value.
- In large samples, the distribution of F is approximated by the $F_{2,\infty}$ distribution which is tabulated (SW, Appendix Table 4).
- This is the distribution of a chi-squared variable divided by its degrees of freedom (See SW chapter 2.4).
- Equivalently, we say that the large-sample distribution of F is $\chi^2_2/2$.

The F statistic in general

- In general, the F statistic can be used to test q restrictions

$$\begin{aligned} H_0 &: \beta_j = \beta_{j,0},\ \beta_m = \beta_{m,0}, \ldots \text{ for a total of } q \text{ restrictions} \\ \text{vs } H_1 &: \text{one or more of the q restrictions under } H_0 \text{ does not hold} \end{aligned}$$

- These restrictions need not be simple as here (e.g., $\beta_j = \beta_{j,0}$) but can be more complicated (such as linear functions of the $\beta's$). For example,
-

$$\begin{aligned} H_0 &: \beta_1 = \beta_{1,0}, \\ &\quad \beta_2 + \beta_3 - .5\beta_7 = 1 \\ \text{vs } H_1 &: \text{one or more of the restrictions under } H_0 \text{ does not hold} \end{aligned}$$

- The general formula for the F-statsitic is complicated and uses matrix algebra (See SW 18.3) but it is usually computed by the statistical software.

The F statistic in general

- The large sample distribution of such F statistic is $F_{q,\infty}$ – equivalently, $\chi^2_q/q-$ and, again, we reject H_0 when F is above the critical value
- Selected large-sample critical values from the χ^2_q/q (Table 3) or $F_{q,\infty}$ (Table 4) distribution:

q	5% critical value	10% critical value
1	3.84	2.71
2	3.00	2.30
3	2.60	2.08
4	2.37	1.94
5	2.21	1.85

The F test in Stata

- 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.
- Stata command: `test str=expn_stu=0` or `test str expn_stu` follows the `regress` command (`test` applies to last regression run)

The F test in Stata

Compare value of F statistic to the critical value (3); Stata prints the p-value of the test.

```
. 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
------------------------------------------------------------------------------

. test str=expn_stu=0

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

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

Additional comments on the F statistic 1

- When $q = 1$, the null hypothesis has a single restriction, and the F statistic is identical to the square of the t statistic (the critical value is the square of the critical value used for the t statistic).

```
. test str

 ( 1)  str = 0

       F(  1,   416) =    0.35
            Prob > F =    0.5528
```


Additional comments on the F statistic 2

- A common null hypothesis is that **all** the slopes in the regression model - k restrictions – are zero:

$$\begin{aligned} H_0 &: \beta_1 = 0,\ \beta_2 = 0, \ldots, \beta_k = 0 \\ vs \quad H_1 &: \beta_j \neq 0 \text{ for at least on } j = 1, \ldots, k \end{aligned}$$

- In this case, the F statistic is sometimes called the "**overall**" regression F statistic and is usually printed the in the regression output (along with its p-value).

```
Linear regression                               Number of obs     =     26,127
                                                F(5, 26121)       =   1535.35
                                                Prob > F          =    0.0000
                                                R-squared         =    0.2647
                                                Root MSE          =     448.1
```

Additional comments on the F statistic 3

- The F statistic presented is valid wether there is heteroskedasticity or homoskedasticity (implicitly reflected in the choice of formula used for the variance of the estimators, note the robust option in the regress command).
  - We say that we use a **heteroskedasticity-robust** F statistic.
- There is a **homoskedasticity-only** version of the F statistic.

The homoskedasticity-only version of the F statistic

- The main (only?) reason to present this homoskedasticity-only F statistic is that it can be calculated in a very intuitive way.
- It links the F statistic to the improvement in the fit of the regression when H_0 is relaxed: if the increase in R^2 when the q restrictions under H_0 are **not** imposed is large enough, we reject H_0.
  - Of course, we could just compute it with the previous formula using homoskedasticity-only standard errors, but this will not generate any additional intuition.
- It turns out that the homoskedasticity-only F statistic under can be computed from the $R^2 s$ of two regressions:
  1. One regression is specified under H_0 (the "restricted" regression)
  2. And the other one is specified under H_1 (the "unrestricted" regression).
- And then compare these two $R^2 s$ as shown below

The homoskedasticity-only version of the F statistic

- We test the hypothesis that the coefficients of STR and exp_stu are zero.
- The **unrestricted** regression (that is, under H_1) is:

$$testscore = \beta_0 + \beta_1 STR + \beta_2 \exp n\_stu + \beta_3 el\_pct + u$$

- The **restricted** regression (that is, under H_0):

$$testscore = \beta_0 + \beta_3 e \ln\_pct + u$$

- The number of restrictions under H_0 is $q = 2$.
- 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_stu` and `el_pct` to be judged statistically significant?
- The formula for the F statistic provides the answer.

The homoskedasticity-only version of the F statistic

- Given the $R^2 s$ of the restricted and unrestricted regression the F statistic assuming homoskedastic errors u has a very simple and intuitive formula:

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

- The larger the increase in R^2, the greater the improvement in fit obtained by adding the variables restricted under H_0.
- A large enough increase in R^2 is evidence against H_0 and leads us to reject it.
- 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, then its large-sample distribution is **not** χ^2_q / q.

The homoskedasticity-only version of the F statistic

```
. reg testscr el_pct  //restricted

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(1, 418)       =    296.40
       Model |  63109.5688         1  63109.5688   Prob > F        =    0.0000
    Residual |  89000.0248       418   212.91872   R-squared       =    0.4149
-------------+----------------------------------   Adj R-squared   =    0.4135
       Total |  152109.594       419  363.030056   Root MSE        =    14.592

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      el_pct |  -.6711562   .0389837   -17.22   0.000    -.7477847   -.5945277
       _cons |   664.7394   .9406415   706.69   0.000     662.8905    666.5884
------------------------------------------------------------------------------

. reg testscr str expn_stu el_pct // unrestricted

      Source |       SS           df       MS      Number of obs   =       420
-------------+----------------------------------   F(3, 416)       =    107.45
       Model |  66409.8837         3  22136.6279   Prob > F        =    0.0000
    Residual |  85699.7099       416  206.008918   R-squared       =    0.4366
-------------+----------------------------------   Adj R-squared   =    0.4325
       Total |  152109.594       419  363.030056   Root MSE        =    14.353

------------------------------------------------------------------------------
     testscr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         str |  -.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
------------------------------------------------------------------------------

. di ((0.4366-0.4149)/2)/((1-0.4366)/416)
8.0113596

. test str expn_stu

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

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


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

- These are justified only under very strong conditions – stronger than are realistic in practice.
- Yet, they are widely used.
- You should use the heteroskedasticity-robust F statistic, with χ^2_q/q (that is, $F_{q,\infty}$) critical values.
  - For $n \geq 120$, the F distribution essentially is the χ^2_q/q distribution.
- For small n, the F distribution isn't necessarily a "better" approximation to the sampling distribution of the F statistic – only if the strong homoskedasticity conditions are true.

Summary: testing joint hypotheses

- The "common-sense" 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 in into STATA ("test" command after a "reg" command with the `robust` option), R and other software; this tests all q restrictions at once.
- For n large, F is distributed as χ^2_q/q (that is, $F_{q,\infty}$).
- The homoskedasticity-only F statistic is important historically (and thus in practice), and is intuitively appealing, but invalid when there is heteroskedasticity (the usual case).

Where are we?

1. Empirical example: Italian labor force survey (LFS):
   1. Wages and education
   2. Wages and gender
   3. Wages and labor experience
2. Using "dummy variables".
   1. Regional differences in wages.
3. Testing joint hypotheses (SW 7.2)
   1. Testing for regional differences
4. **Regression specification (SW 7.5)**

Control variables

- We want to estimate the causal effect of class size on test scores.
- If we could run an **experiment**, we would randomly assign students (and teachers) to different sized classes. Then STR would be independent of all the things that go into u, so $E(u|STR) = 0$ and the OLS slope estimator in the regression of Testscore on STR will be an unbiased estimator of the desired causal effect.
- But with **observational** data, u depends on additional factors (visit to museums, parental involvement, knowledge of English etc.).
- If those factors are observed (e.g. el_pct), then include them in the regression.
- But usually we can't observe all these omitted factors (e.g., parental involvement in homework)...leading to an Omitted Variable Bias.
- In this case, we include **"control variables"** which are variables correlated with these omitted factors, but which themselves are not necessarily causal.

Variables of interest and control variables

- **A control variable W is a variable that is correlated with, and controls for, an omitted factor in the regression of Y on X, but which itself does not necessarily have a causal effect on Y.**
- Note that we implicitly differentiate between the **variable(s) of interest** (STR) and the other **control variables** (el_pct, etc.): not all $X_j's$ are the same!
- How do we know that including control variable(s) avoids the problem of Omitted Variable Bias (OVB)?
- We don't....but by adding enough controls it is **hoped** that LSA #1 holds so that OLS is unbiased.
- This is the **regression specification** problem....more on this below.

An example from the California testscore data

$$\widehat{testscr} = \underset{(5.6)}{700.2} + \underset{(0.27)}{0.998}str - \underset{(0.033)}{0.122}el\_pct - \underset{(0.024)}{0.547}meal\_pct$$

el_pct = percent English Learners in the school district
meal_pct = percent of students receiving a free/subsidized lunch (only students from low-income families are eligible)

- Which variable is the variable of interest?
- Which variables are control variables? Do they have a causal interpretation? What do they control for?

An example from the California testscore data

- STR is the variable of interest.
- el_pct probably has a direct causal effect (school is tougher if you are learning English!). But it is also a control variable: immigrant communities tend to be less affluent and often have fewer outside learning opportunities, and el_pct is correlated with those omitted causal variables. **el_pct is both a possible causal variable and a control variable.**
- meal_pct might have a causal effect (eating lunch helps learning); it also is correlated with and controls for income-related outside learning opportunities. **meal_pct** is both a possible causal variable and a control variable.

Control variables need not be causal

- The coefficients of the control variables generally do not have a causal interpretation.
- For example, does the coefficient on `meal_pct` have a causal interpretation?
  - If so, then we should be able to boost test scores (by a lot!) by simply eliminating the school lunch program.
  - This does not make much sense.
  - The estimated coefficient of `meal_pct` is not the causal effect of the lunch program (which we expect it to be non-negative) because `meal_pct` is correlated with other income-related variables not controlled for in the regression, i.e., variables that are in u.
- That is, strictly speaking, LSA #1 $E(u|X_1, ..., X_k) = 0$ does not hold when a control variable (e.g., `meal_pct`) is correlated with u.
  - This is what leads to a biased estimate of the causal effect of a lunch program.

Conditional mean independence

- So we do not believe that LSA #1 holds in the testscore data.
- Can we reformulate LSA # 1 in such a way that it ensures that OLS is an unbiased and consistent estimator of the causal effect of the variables of interest, but not necessarily of the causal effect of the control variables?
- Consider the assumption of **conditional mean independence**: given the control variable(s) (W), the mean of u doesn't depend on the variable of interest (X),

$$E(u|X, W) = E(u|W) \quad \text{(conditional mean independence)}$$

- Controlling for W, X can be treated **as if** it were randomly assigned, in the sense that $E(u|X, W)$ does not vary with X.
- Although $E(u|X, W)$ can vary with W which is not allowed under LSA #1.
- Conditional mean independence is the reformulation of Least Squares Assumption # 1 (in fact, it relaxes it).

Conditional mean independence

- Consider the regression model

$$Y = \beta_0 + \beta_1 X + \beta_2 W + u$$

where X is the variable of interest and W is the control variable.

- **Conditional mean independence** (in addition to LSA #2, #3 and #4) ensures that:

1. β_1 has a causal interpretation.
2. $\hat{\beta}_1$ is unbiased and consistent. (see SW Appendix 7.2)
3. The estimated coefficient of the control variable, $\hat{\beta}_2$, is in general biased (and not consistent) because it can suffer from OVB. (see SW Appendix 7.2)

Beta1 has a causal interpretation

- The change in expected Y resulting from a change in X, holding W constant, is:

$$E(Y|X = x + \Delta x, W = w) - E(Y|X = x, W = w)$$

- Is this difference equal to β_1?
- The answer is YES! under the conditional mean independence assumption.
- Thus, β_1 measures the effect on Y of a change in X and of nothing else since W is hold fixed and conditional mean independence is assumed.
- Thus β_1 has a causal interpretation.

Proof that beta1 has a causal interpretation

$$E(Y|X = x + \Delta x, W = w) = \beta_0 + \beta_1 (X + \Delta x) + \beta_2 W + E(u|X = x + \Delta x, W = w)$$
$$E(Y|X = x, W = w) = \beta_0 + \beta_1 X + \beta_2 W + E(u|X = x, W = w)$$

so that the difference equals

$$\beta_1 \Delta x + E(u|X = x + \Delta x, W = w) - E(u|X = x, W = w)$$

and under conditional mean independence:

$$E(u|X = x + \Delta x, W = w) = E(u|X = x, W = w) = E(u|W = w)$$

which proves the result that the change in expected Y resulting from a change in X, holding W constant, is $\beta_1 \Delta x$.

Three interchangeable statements about what makes an effective control variable

1. Effective control variables are those which, when included in the regression, makes the error term uncorrelated with the variable of interest.
2. Holding constant the control variable(s), the variable of interest is **"as if"** randomly assigned.
3. Among individuals (entities) with the same value of the control variable(s), the variable of interest is uncorrelated with the omitted determinants of Y.

- In short, an effective W is one that ensures that **conditional mean independence** holds.
- Thus, X and W are not treated the same way: we can choose the controls so that conditional mean independence holds but X is given to us by the nature of the problem under study.

Implications for regression (model) specification

1. Identify the variable of interest (e.g., STR).
2. Think of the omitted factors that could result in omitted variable bias.
3. Include those omitted factors if you can (e.g., el_pct). If you can't, include variables correlated with them that serve as control variables (e.g., meal_pct).
4. The control variable(s) are effective if the conditional mean independence plausibly holds.
5. This results in a "**base**" or "**benchmark**" model.
6. Also specify a range of plausible alternative models, which include additional candidate control variables.
7. Estimate base and alternative models and conduct "sensitivity checks":
    1. Does a candidate variable change the coefficient of interest (β_1)?
    2. Is a candidate variable statistically significant?
    3. Use judgment, not a mechanical recipe...and certainly don't just try to maximize R^2!

Regression specification

	(1)	(2)	(3)	(4)	(5)
	reg1	reg2	reg3	reg4	reg5
VARIABLES	testscr	testscr	testscr	testscr	testscr
str	-2.28***	-1.10**	-1.00***	-1.31***	-1.01***
	(0.52)	(0.43)	(0.27)	(0.34)	(0.27)
el_pct		-0.65***	-0.12***	-0.49***	-0.13***
		(0.031)	(0.033)	(0.030)	(0.036)
meal_pct			-0.55***		-0.53***
			(0.024)		(0.038)
calw_pct				-0.79***	-0.048
				(0.068)	(0.059)
Constant	699***	686***	700***	698***	700***
	(10.4)	(8.73)	(5.57)	(6.92)	(5.54)
Observations	420	420	420	420	420
R-squared	0.051	0.426	0.775	0.629	0.775

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1