Econometrics Lecture 3

EC2METRIE

Dr. Anna Salomons

Utrecht School of Economics (U.S.E.)

28 November 2016

This class

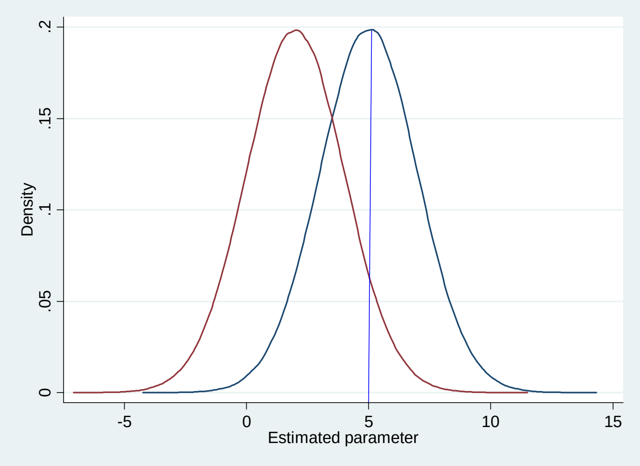
- **Hypothesis testing**
  - For 1 population parameter: t-test
  - For >1 population parameter: F-test
- **Omitted variable bias** (violation of OLS assumption $Corr(\varepsilon_i, X_i) = 0$)

Assumptions 1-4

To get that $E(\widehat{\beta}_k) = \beta_k$:

1. **Population model is linear in parameters** (and the error term is additive)
2. **Error term has a zero population mean**: $E(\varepsilon_i) = 0$
3. **All independent variables are uncorrelated with the error term**: $Corr(\varepsilon_i, X_i) = 0$
4. **No perfect (multi)collinearity** between independent variables (and no variable is a constant)

Sampling distribution: unbiasedness



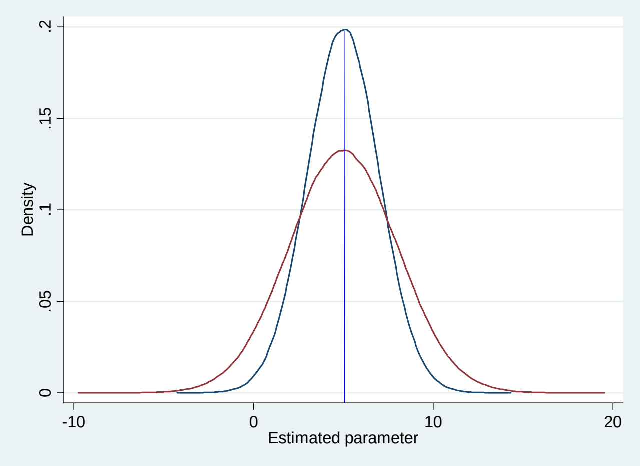
Assumptions 5-6

To get that $E\left(\widehat{Var}(\widehat{\beta})\right) = Var(\widehat{\beta})$, we need assumptions 1-4 plus:

5. **No serial correlation**: errors are not correlated with each other across different observations, $Corr(\varepsilon_i, \varepsilon_j) = 0$.
6. **No heteroskedasticity**: error term has constant variance, $Var(\varepsilon_i) = \sigma^2$ (where σ^2 is a constant).

Together, assumptions 1-6 make OLS the Best (i.e. minimum variance) Linear Unbiased Estimator (BLUE).

Sampling distribution: minimum variance



Sampling distribution: normality

- To be able to perform hypothesis tests, we need **one more property of the sampling distribution,** other than unbiasedness and minimum variance: **normality**.
- Note that last week we have already been drawing the sampling distributions as normal distributions
  - Bell-shaped, symmetric
  - The location and shape of any particular normal distribution depend on its mean and its variance

Normal distributions

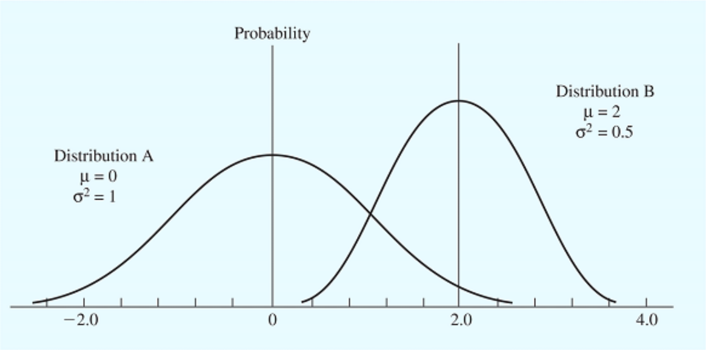


Figure 4.3 Normal Distributions

Although all normal distributions are symmetrical and bell-shaped, they do not necessarily have the same mean and variance. Distribution A has a mean of 0 and a variance of 1, whereas distribution B has a mean of 2 and a variance of 0.5. As can be seen, the whole distribution shifts when the mean changes, and the distribution gets fatter as the variance increases.

Assumption 7: normality of the error term

7. $\varepsilon \backsim Normal(0, \sigma^2)$: This means we assume that the **error term is normally distributed**, with a mean of zero (cf. assumption 2) and a constant variance (cf. assumption 6) of σ^2.

Under assumptions 1-7, it holds that

$$\widehat{\beta}_k \backsim Normal\left(\beta_k, Var(\widehat{\beta}_k)\right)$$

I.e., the **sampling distribution of $\widehat{\beta_k}$ is normally distributed** with a mean of β_k (the true population parameter- cf. assumptions 1-4) and a variance of $Var(\widehat{\beta}_k)$ (cf. assumptions 5-6).

Reason normality of error gives normality of sampling distribution

- Can be shown that $\widehat{\beta}_k$ is a linear combination of error terms ε_i
- We assume that the errors are normally distributed (assumption 7), $\varepsilon \backsim \mathit{Normal}$
- We had already assumed the errors are independent (assumption 5, and given random sampling)
- $\widehat{\beta}_k$ is thus a linear combination of independent normally distributed terms
- A statistical theorem tells us any such linear combination will itself also be normally distributed: hence, $\widehat{\beta}_k \backsim \mathit{Normal}$

How important is assumption 7?

- **In large samples**: **not very important**, as the **Central Limit Theorem** (CLT) tells us that sampling distributions will be normally distributed for large sample sizes.
- **In small(er) samples**: **important**, since the CLT might not apply and assumption 7 gives us normality of the sampling distribution.

Why do we need assumption 7 for hypothesis testing?

- Assumption 7 means we know the entire shape of the sampling distribution: we use this for hypothesis testing
- Let's see how by explaining the **procedure of hypothesis testing**:
  - About 1 population parameter (t-test)
  - About >1 population parameter (F-test)

Hypothesis testing: an example

Ashley Madison profiles by gender

Hypothesis testing: example

Research question: does people's gender impact the number of extra-marital affairs they have?

Population model:

$$naffairs_i = \beta_0 + \beta_1 male_i + \varepsilon_i$$

where: *naffairs* = nr of affairs; *male* = dummy variable for men

Hypotheses:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}$$

Hypothesis testing

$$naffairs_i = \beta_0 + \beta_1 male_i + \varepsilon_i$$

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}$$

We estimate the population model in a sample:

$$naffairs_i = \widehat{\beta}_0 + \widehat{\beta}_1 male_i + e_i$$

Summary statistics for our example

```
. descr naffairs male

              storage  display     value
variable name   type   format      label      variable label
---------------------------------------------------------------------------
naffairs        byte   %9.0g                  number of affairs within last year
male            byte   %9.0g                  =1 if male

. sum naffairs male

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    naffairs |       601    1.455907    3.298758          0         12
        male |       601    .4758735    .4998336          0          1

.
. sum naffairs if male==0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    naffairs |       315    1.419048    3.309264          0         12

. sum naffairs if male==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    naffairs |       286    1.496503    3.292467          0         12
```

Bivariate regression for our example

```
. reg naffairs male

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  1,   599) =    0.08
       Model |   .899313     1    .899313              Prob > F      =  0.7740
    Residual |  6528.18222   599  10.8984678           R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.0015
       Total |  6529.08153   600  10.8818026           Root MSE      =  3.3013

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0774559   .2696384     0.29   0.774    -.4520956    .6070073
       _cons |   1.419048   .1860062     7.63   0.000     1.053744    1.784351
------------------------------------------------------------------------------
```

$$\widehat{\beta}_1 = 0.077$$

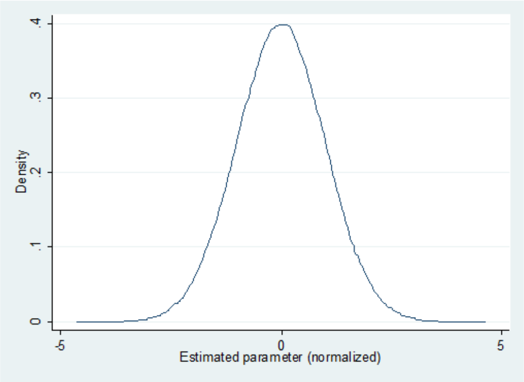
Hypothesis testing

- **How high is the chance** (=probability) **that our estimated parameter**, $\widehat{\beta}_1 = 0.077$, **comes from a population with** $\boldsymbol{\beta_1 = 0}$?
- **If very low probability, reject $\mathbf{H_0}$**- but if high probability, do not reject H_0.
- To find the probability, we standardize our sampling distribution:

$$
\begin{aligned}
\widehat{\beta}_1 &\backsim Normal\left(\beta_1, Var(\widehat{\beta}_1)\right) \\
\frac{\widehat{\beta}_1 - \beta_1}{sd(\widehat{\beta}_1)} &\backsim Normal\,(0, 1)
\end{aligned}
$$

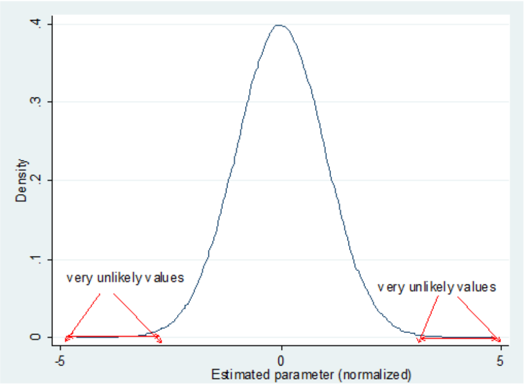
Hypothesis testing

What does the sampling distribution now look like if H_0 is true? A normal distribution with $E(\widehat{\beta}_1) = 0$ and variance of 1



Hypothesis testing

Which values of $\widehat{\beta}_1$ from this sampling distribution are very unlikely if H_0 is true (i.e. $\beta_1 = 0$)? (The ones indicated by arrows)



Hypothesis testing

This is the test statistic:

$$\frac{\widehat{\beta}_1 - \beta_1}{sd(\widehat{\beta}_1)} \backsim Normal\,(0, 1)$$

- Note that this **requires us to know sd**$(\widehat{\boldsymbol{\beta}}_{\mathbf{1}}) = \sqrt{Var(\widehat{\beta}_1)}$ – but remember from last week that we do not as the error variance σ^2 is **unknown**!
- Instead, we have to estimate $sd(\widehat{\beta}_1)$, **replacing it by se**$(\widehat{\boldsymbol{\beta}}_{\mathbf{1}}) = \sqrt{\widehat{Var}(\widehat{\beta}_1)}$. (Formula was discussed last week)

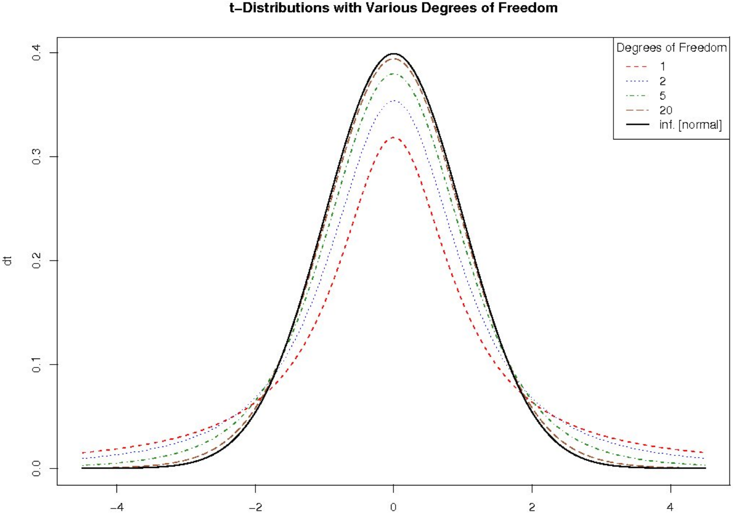
Hypothesis testing

We then obtain the following **test statistic** (replacing the standard deviation of $\widehat{\beta}_1$ by its estimate, the standard error):

$$\frac{\widehat{\beta}_1 - \beta_1}{se(\widehat{\beta}_1)} \backsim t_{n-k-1}$$

- Instead of a normal distribution, this follows a **t-distribution with n-k-1 degrees of freedom.**
- It is therefore known as the **t-statistic**.

Difference between normal and t-distributions



Difference between normal and t-distributions

- The **t-distribution** looks like a normal distribution, but with **fatter tails**
  - This reflects the additonal sampling uncertainty from estimating $sd(\widehat{\beta}_1)$
- As the number of **degrees of freedom increases**, the t-distribution becomes **increasingly similar to the normal distribution**.
- See Studenmund Appendix Table B1.

Using the t-statistic

Step 1 Formulate hypotheses H_0 and H_A

Step 2 Choose a significance level, α

Step 3 Calculate t-statistic, $t = \frac{\widehat{\beta}_k - \beta_k}{se(\widehat{\beta}_k)}$

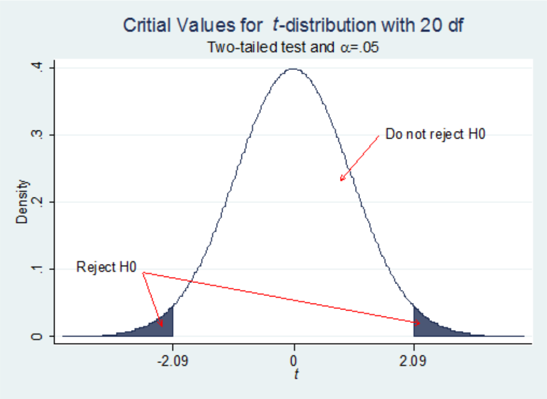
Step 4 Find critical value from t-distribution with $n-k-1$ degrees of freedom and significance level α (for one-sided H_A) or $\alpha/2$ (for two-sided H_A).

Step 5 Compare t-statistic to critical value[1].

Decision rule:

- For two-sided H_A : Reject H_0 if $|t| > t_c$
- For one-sided H_A : Reject H_0 if $|t| > t_c$ and t has the sign hypothesized in H_A

Using the t-statistic



The shaded areas together have a probability of 5% (i.e. the chosen significance level).

The t-statistic in Stata

```
. reg naffairs male

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  1,   599) =    0.08
       Model |   .899313     1    .899313              Prob > F      =  0.7740
    Residual |  6528.18222   599  10.8984678           R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.0015
       Total |  6529.08153   600  10.8818026           Root MSE      =  3.3013

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0774559   .2696384     0.29   0.774    -.4520956    .6070073
       _cons |   1.419048   .1860062     7.63   0.000     1.053744    1.784351
------------------------------------------------------------------------------
```

$$t = \frac{\widehat{\beta}_1 - \beta_1}{se(\widehat{\beta}_1)} = \frac{0.0775 - 0}{0.2696} = 0.29$$

Example of using the t-stat: 2-sided Ha

$$naffairs_i = \beta_0 + \beta_1 male_i + \varepsilon_i$$

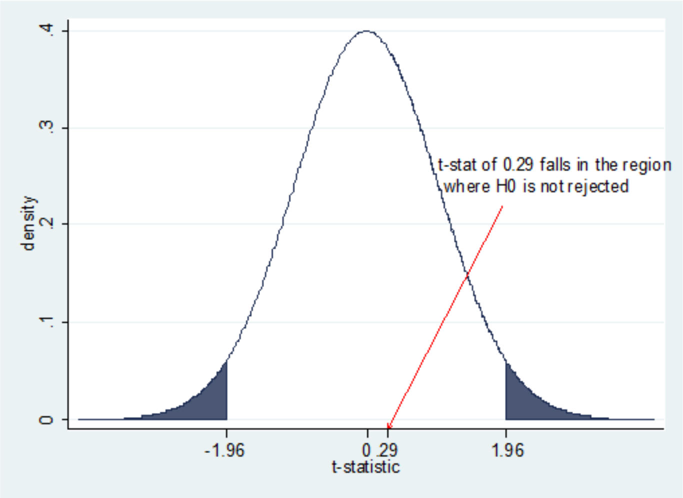
1. Hypotheses:

$$H_0 : \beta_1 = 0 \qquad H_A : \beta_1 \neq 0$$

2. $\alpha = 0.05$
3. $t = 0.29$ (see Stata output).
4. $t_c = t_{n-k-1,\alpha/2} = t_{599,0.025} = 1.96$ (Studenm Table B1).
5. $0.29 < 1.96$, hence do not reject H_0. **Conclusion**: no significant difference between the number of affairs of men and women.

Important: we **cannot interpret the size of the coefficient**, since it is not significantly different from zero!

Illustration of the 2-sided t-test in our example



Example of using the t-stat: 1-sided Ha

$$naffairs_i = \beta_0 + \beta_1 male_i + \varepsilon_i$$

1. Hypotheses:

$$\begin{aligned} H_0 &: \beta_1 \leq 0 \\ H_A &: \beta_1 > 0 \end{aligned}$$

2. $\alpha = 0.05$
3. $t = 0.29$ (see Stata output).
4. $t_c = t_{n-k-1,\alpha} = t_{599,0.05} = 1.645$ (Table B1).
5. $0.29 < 1.645$, hence do not reject H_0. **Conclusion**: no significant difference between the number of affairs of men and women.

Alternatives to using the t-statistic

- Can also perform the hypothesis test by using the **p-value** or the **confidence interval**
- This will of course lead to the **exact same conclusion** about H_0 (for a given chosen confidence level α)
- **Stata also reports these alternatives**

P-value and confidence interval in Stata

```
. reg naffairs male

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  1,   599) =    0.08
       Model |   .899313     1   .899313               Prob > F      =  0.7740
    Residual |  6528.18222   599  10.8984678           R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.0015
       Total |  6529.08153   600  10.8818026           Root MSE      =  3.3013

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0774559   .2696384     0.29   0.774    -.4520956    .6070073
       _cons |   1.419048   .1860062     7.63   0.000     1.053744    1.784351
------------------------------------------------------------------------------
```

Using the p-value

Step 1 Formulate hypotheses H_0 and H_A

Step 2 Choose a significance level, α

Step 3 Calculate t-statistic, $t = \frac{\widehat{\beta}_k - \beta_k}{se(\widehat{\beta}_k)}$

Step 4 Obtain p-value[2] for t-statistic,
$Pr\left(T_{n-k-1} \geq \left|\frac{\widehat{\beta}_k - \beta_k}{se(\widehat{\beta}_k)}\right|\right)$ for one-sided H_A and
$Pr\left(|T_{n-k-1}| \geq \left|\frac{\widehat{\beta}_k - \beta_k}{se(\widehat{\beta}_k)}\right|\right)$ for two-sided H_A,

Step 5 Compare p-value to the chosen significance level: reject H_0 if $p\text{-}value < \alpha$

[2] A p-value is the probability of obtaining a test statistic at least as extreme as the one actually observed, if the null hypothesis is true. It is therefore the smallest significance level at which we would reject the null hypothesis.

Example of using the p-value: 2-sided Ha

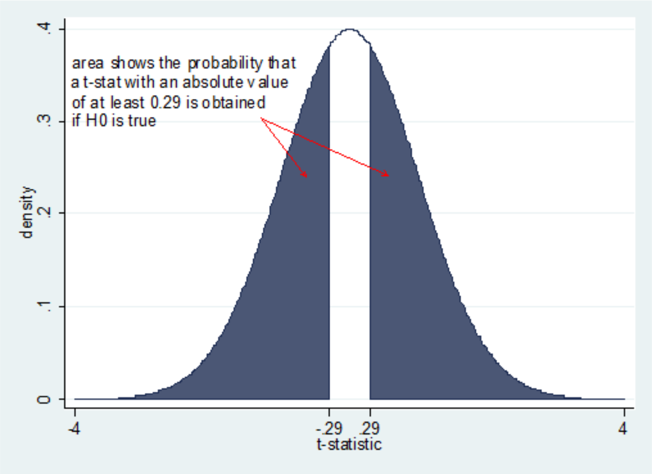
$$naffairs_i = \beta_0 + \beta_1 male_i + \varepsilon_i$$

1. Hypotheses:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}$$

2. $\alpha = 0.05$
3. $t = 0.29$ (see Stata output).
4. $p\text{-}value = 0.774$ (see Stata output).
5. $0.774 > 0.05$, hence do not reject H_0. **Conclusion**: no significant difference between the number of affairs of men and women.

An illustration of the 2-sided p-value in our example



$$p - value = 0.774$$

Example of using the p-value: 1-sided Ha

$$naffairs_i = \beta_0 + \beta_1 male_i + \varepsilon_i$$

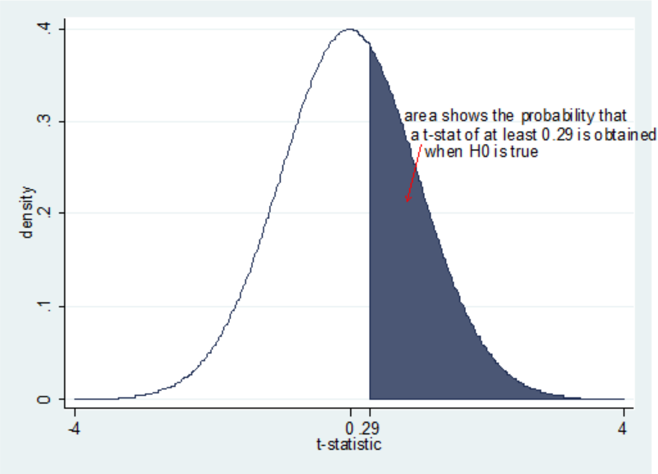
1. Hypotheses:

$$\begin{aligned} H_0 &: \beta_1 \leq 0 \\ H_A &: \beta_1 > 0 \end{aligned}$$

2. $\alpha = 0.05$
3. $t = 0.29$
4. $p - value = 0.77/2 = 0.385$.
5. Since $0.385 > 0.05$, do not reject H_0. **Conclusion**: men do not have significantly more affairs than women.

Note that **Stata by default reports two-sided p-values**- to obtain the one-sided p-value, divide by 2.

An illustration of the one-sided p-value in our example



$$p - value = 0.385$$

The confidence interval

- Can construct a **confidence interval** (CI) for β_k
- A CI of $(1-\alpha) \times 100\%$ is given by:

$$\widehat{\beta}_k \pm t_c \times se(\widehat{\beta}_k)$$

- That is,

$$\Pr\left(\widehat{\beta}_k - t_c \times se(\widehat{\beta}_k) \leq \beta_k \leq \widehat{\beta}_k + t_c \times se(\widehat{\beta}_k)\right) = 1-\alpha$$

- The lower and upper bounds of the CI depend on the critical value of the t-statistic, and on the standard error of the estimated coefficient.

Using the confidence interval

Step 1 Formulate hypotheses H_0 and H_A

Step 2 Choose a significance level, α

Step 3 Find critical t-statistic, t_c (depends on the number of degrees of freedom and the significance level).

Step 4 Construct the $(1-\alpha)\%$ CI for β_k

Step 5 Reject H_0 if the value of β_k hypothesized in H_0 does not lie within the constructed CI.

Note: cannot test a one-sided alternative hypothesis using a CI (since CI is always two-sided).

Example of using the confidence interval

$$naffairs_i = \beta_0 + \beta_1 male_i + \varepsilon_i$$

1. Hypotheses:
$$H_0 : \beta_1 = 0 \qquad H_A : \beta_1 \neq 0$$
2. $\alpha = 0.05$
3. $t_c = t_{n-k-1,\alpha/2} = t_{599,0.025} = 1.96$
4. $\widehat{\beta}_1 \pm t_c \times se(\widehat{\beta}_1) = 0.0775 \pm 1.96 \times 0.2696$
$\Leftrightarrow -0.45 \leq \beta_1 \leq 0.61$ (see Stata output).
5. Since $\beta_1 = 0$ lies in this CI, do not reject H_0. **Conclusion**: no significant difference between the number of affairs of men and women.

Extra example I

Population model:

$$naffairs_i = \beta_0 + \beta_1 male_i + \beta_2 religion_i + \varepsilon_i$$

Research question: does religion have an impact on the number of affairs people have?

Hypotheses:

$$\begin{aligned} H_0 &: \beta_2 = 0 \\ H_A &: \beta_2 \neq 0 \end{aligned}$$

Summary statistics

```
. descr relig

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------
relig           byte   %9.0g                  5 = very relig., 4 =
                                                somewhat, 3 = slightly, 2 =
                                                not at all, 1 = anti

. sum relig

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       relig |       601    3.116473    1.167509          1          5
```

OLS estimates

```
. reg naffairs male relig

      Source |       SS           df       MS      Number of obs   =       601
-------------+----------------------------------   F(2, 598)       =      6.43
       Model |  137.408874         2  68.7044371   Prob > F        =    0.0017
    Residual |  6391.67266       598  10.6884158   R-squared       =    0.0210
-------------+----------------------------------   Adj R-squared   =    0.0178
       Total |  6529.08153       600  10.8818026   Root MSE        =    3.2693

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0847845   .2670351     0.32   0.751    -.4396562    .6092252
       relig |  -.4085623    .114323    -3.57   0.000    -.6330856    -.184039
       _cons |   2.688833   .4002195     6.72   0.000     1.902827     3.47484
------------------------------------------------------------------------------
```

Hypothesis test: t-statistic

1. Hypotheses:
$$H_0 : \beta_2 = 0 \qquad H_A : \beta_2 \neq 0$$
2. $\alpha = 0.05$
3. Find t-stat: $t = -3.57$ (in Stata output)
4. Find critical t-stat: $t_c = t_{n-k-1,\alpha/2} = t_{598,0.025} = 1.96$
5. Compare actual t-stat to critical value: $|-3.57| > 1.96$, hence reject H_0. **Conclusion**: religious people have a different number of affairs, ceteris paribus on gender.

Since we **find a statistically significant effect, we can interpret the estimated coefficient**: for each point that people are more religious (on the 5-point scale), they have 0.41 fewer affairs in a year, holding gender constant.

Hypothesis test: p-value & CI

```
. reg naffairs male relig

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  2,   598) =    6.43
       Model |  137.408874     2  68.7044371           Prob > F      =  0.0017
    Residual |  6391.67266   598  10.6884158           R-squared     =  0.0210
-------------+------------------------------           Adj R-squared =  0.0178
       Total |  6529.08153   600  10.8818026           Root MSE      =  3.2693

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0847845   .2670351     0.32   0.751    -.4396562    .6092252
       relig |  -.4085623    .114323    -3.57   0.000    -.6330856    -.184039
       _cons |   2.688833   .4002195     6.72   0.000     1.902827     3.47484
------------------------------------------------------------------------------
```

- P-value is 0.000, and $0.000 < 0.05$, hence reject H_0
- 95% CI is $-0.633 \leq \beta_2 \leq -0.184$; $\beta_2 = 0$ does not lie within this CI, hence reject H_0.

Of course, same conclusion as with using the t-statistic.

Extra example II

Population model:

$$naffairs_i = \beta_0 + \beta_1 male_i + \beta_2 religion_i + \varepsilon_i$$

Research question: do people who are more religious have fewer affairs?

Hypotheses:

$$\begin{aligned} H_0 &: \beta_2 \geq 0 \\ H_A &: \beta_2 < 0 \end{aligned}$$

Hypothesis test: t-statistic

1. Hypotheses:
$$H_0 : \beta_2 \geq 0 \qquad H_A : \beta_2 < 0$$
2. $\alpha = 0.05$
3. Find t-stat: $t = -3.57$ (in Stata output)
4. Find critical t-stat: $t_c = t_{n-k-1,\alpha/2} = t_{598,0.05} = 1.345$
5. Compare actual t-stat to critical value: $|-3.57| > 1.345$, and $-3.57 < 0$, hence reject H_0.

Conclusion: more religious people have significantly fewer affairs than less religious people, cet. par. (Interpretation of coefficient is the same as before.)

Extra example III

Population model:

$$naffairs_i = \beta_0 + \beta_1 male_i + \beta_2 religion_i + \varepsilon_i$$

Research question: do people who are 1 point higher on the religiousness scale have 1 less affair per year?

Hypotheses:

$$\begin{aligned} H_0 &: \beta_2 = -1 \\ H_A &: \beta_2 \neq -1 \end{aligned}$$

OLS estimates

```
. reg naffairs male relig

      Source |       SS       df       MS              Number of obs =     60
-------------+------------------------------           F(  2,   598) =    6.4
       Model |  137.408874     2  68.7044371           Prob > F      =  0.001
    Residual |  6391.67266   598  10.6884158           R-squared     =  0.021
-------------+------------------------------           Adj R-squared =  0.017
       Total |  6529.08153   600  10.8818026           Root MSE      =  3.269

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval
-------------+----------------------------------------------------------------
        male |   .0847845   .2670351     0.32   0.751    -.4396562    .609225
       relig |  -.4085623    .114323    -3.57   0.000    -.6330856    -.18403
       _cons |   2.688833   .4002195     6.72   0.000     1.902827    3.4748
------------------------------------------------------------------------------
```

Hypothesis test: t-statistic

1. Hypotheses:

$$\begin{aligned} H_0 &: \beta_2 = -1 \\ H_A &: \beta_2 \neq -1 \end{aligned}$$

2. $\alpha = 0.05$
3. Find t-stat: $t = \frac{\widehat{\beta}_2 - \beta_2}{se(\widehat{\beta}_2)} = \frac{-0.4086-(-1)}{0.1143} = 5.17$
4. Find critical t-stat: $t_c = t_{n-k-1,\alpha/2} = t_{598,0.025} = 1.96$
5. Compare actual t-stat to critical value: $|5.17| > 1.96$, hence reject H_0. **Conclusion:** people who are 1-point more religious do not have exactly 1 less affair a year.

Hypothesis test: p-value

- The p-value reported in Stata's regression output is always for the null hypothesis that the true coefficient is zero.
- Use Stata's *test* command for any other hypothesis tests, such as $H_0 : \beta_2 = -1, \;\; H_A : \beta_2 \neq -1$

```
. reg naffairs male relig

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  2,   598) =    6.43
       Model |  137.408874     2  68.7044371           Prob > F      =  0.0017
    Residual |  6391.67266   598  10.6884158           R-squared     =  0.0210
-------------+------------------------------           Adj R-squared =  0.0178
       Total |  6529.08153   600  10.8818026           Root MSE      =  3.2693

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0847845   .2670351     0.32   0.751    -.4396562    .6092252
       relig |  -.4085623    .114323    -3.57   0.000    -.6330856    -.184039
       _cons |   2.688833   .4002195     6.72   0.000     1.902827     3.47484
------------------------------------------------------------------------------

. test relig==-1

 ( 1)  relig = -1

       F(  1,   598) =   26.76
            Prob > F =    0.0000
```

Which method to choose? It doesn't matter, but:

- Examining the **p-value is easiest**
  - Vis-a-vis t-stat: does not require looking up a critical t-statistic.
  - Vis-a-vis CI: can also be used for significance levels other than 5% (the CI given by Stata is a 95% CI by default).

- However, when we have a one-sided H_A, we have to be more careful:
  - Use the t-statistic (and compare it to the critical t-statistic for a one-sided H_A)
  - Or use the p-value, but p-values reported in Stata are by default for two-sided alternatives: you have to divide these by 2 to obtain the one-sided p-value!

Summary

- We have covered **hypothesis testing about one population parameter**: can use t-stat, p-value or CI.
- However, sometimes we want to test hypothesis that are about more than one population parameter at the same time, i.e. **joint hypotheses**.
- For this, we will need a **different test statistic**: the **F-statistic**.

Why test joint hypotheses?

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

- **Exclusion restrictions**: we want to test whether we can leave out a group of independent variables from the model, or not. For instance:

$$\begin{aligned} H_0 &: \beta_2 = \beta_3 = 0 \\ H_A &: H_0 \text{ not true} \end{aligned}$$

- Economic theory tells us a **certain relationship has to exist** between coefficients. For instance:

$$\begin{aligned} H_0 &: \beta_2 + \beta_3 = 1 \\ H_A &: H_0 \text{ not true} \end{aligned}$$

Example: estimating a Mincer model

$$hrlywage_i = \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 age_i^2 + \varepsilon_i$$

Research question: does age have an impact on hourly wages? Use $\alpha = 0.05$.

$$H_0 : \beta_2 = \beta_3 = 0 \qquad H_A : H_0 \text{ not true}$$

Age enters this equation twice: need to **assess both coefficients simultaneously**.

OLS regression results on next slide show that neither *age* nor *age*^{2} has an individually significant impact on wages: their p-values are above 0.05. But it turns out we **should not use t-statistics to answer this research question**.

Example: estimating a Mincer model

```
. descr hrwage educ age agesq

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
hrwage          float  %9.0g                  hourly wage
educ            byte   %9.0g                  years of schooling
age             byte   %9.0g                  age in years
agesq           int    %9.0g                  age^2

. reg hrwage educ age agesq

      Source |       SS       df       MS              Number of obs =     151
-------------+------------------------------           F(  3,   147) =    8.96
       Model |  452.987817     3  150.995939           Prob > F      =  0.0000
    Residual |  2476.05932   147   16.843941           R-squared     =  0.1547
-------------+------------------------------           Adj R-squared =  0.1374
       Total |  2929.04714   150  19.5269809           Root MSE      =  4.1041

------------------------------------------------------------------------------
      hrwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5842392   .1198999     4.87   0.000      .347289    .8211894
         age |   .4195173   .2447848     1.71   0.089    -.0642346    .9032692
       agesq |  -.0039981    .002883    -1.39   0.168    -.0096955    .0016994
       _cons |  -11.84532   5.269089    -2.25   0.026    -22.25827   -1.432374
------------------------------------------------------------------------------
```

Why not use t-statistics for testing joint hypotheses?

$$hrlywage_i = \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 age_i^2 + \varepsilon_i$$
$$H_0 : \beta_2 = \beta_3 = 0 \qquad H_A : H_0 \text{ not true}$$

Using the two **t-statistics** for $\widehat{\beta}_2$ and $\widehat{\beta}_3$ would lead to a **smaller probability of a Type I error α than we decided on**. (I.e. we are being "too strict".)

- Recall that $Pr(Type\ I\ error) = Pr(H_0 = rejected | H_0 = true)$.
- $\alpha_1 = 0.05$: type I error probability for $H_0 : \beta_2 = 0$;
  $\alpha_2 = 0.05$: type I error probability for $H_0 : \beta_3 = 0$
- Type 1 error probability for $H_0 : \beta_2 = \beta_3 = 0$ is $\alpha_1 \times \alpha_2 = 0.05 \times 0.05 = 0.025$
- Hence H_0 will not be rejected often enough when using sequential t-tests: **need F-test**!

Interlude: a reminder on Type I and Type II errors

		POPULATION	
		H_0 true	H_A true
SAMPLE	H_0 not rejected	Accurate $1-\alpha$ ☺	Type II error β ☹
	H_0 rejected	Type I error α ☹	Accurate $1-\beta$ ☺

Why not use t-statistics for testing joint hypotheses?

$$\begin{aligned} hrlywage_i &= \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 age_i^2 + \varepsilon_i \\ H_0 &: \beta_2 = \beta_3 = 0 \qquad H_A : H_0 \text{ not true} \end{aligned}$$

- **Using** the two **t-statistics** for $\widehat{\beta}_2$ and $\widehat{\beta}_3$ would lead to a **smaller probability of a Type I error α than we decided on**. (I.e. we are being "too strict".)

- Sequential t-tests **do not take the correlation between explanatory variables into account:** *age* & age^2 strongly correlated.

Testing joint hypotheses

- **Hypothesis testing with the F-statistic is very similar to with the t-statistic**- only the test statistic itself (and its distribution) is different.
- We again need **assumption 7**, normality of the error term.
- Unlike with the t-test, it is **impossible to test one-sided alternative hypotheses with the F-test**

The F-test

Step 1 Define hypotheses, H_0 and H_A

Step 2 Choose a significance level α

Step 3 Estimate the restricted and unrestricted models, where the restricted model is the one obtained if H_0 is true.

Step 4 Compare the residual sum of squares (RSS) across the two models by calculating the F-statistic:

$$\frac{(RSS_M - RSS)/M}{RSS/(n-k-1)} \backsim F_{M,n-k-1}$$

Step 5 Find the critical value for the F-statistic, $F_c = F_{M,n-k-1,\alpha}$.

Step 6 Compare the observed statistic to the critical value- reject H_0 if $F > F_c$

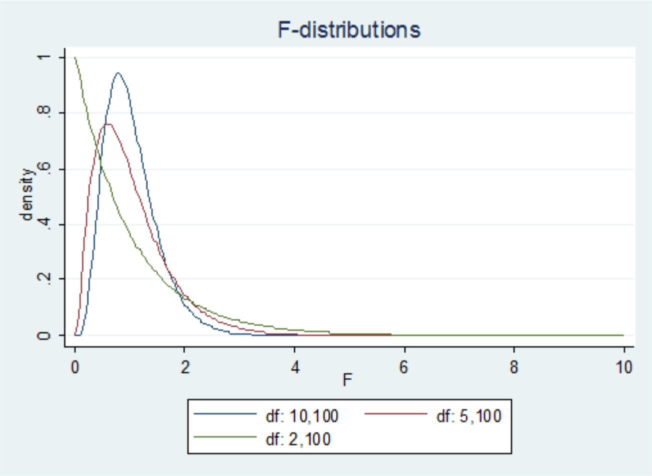
The F-statistic

$$\frac{(RSS_M - RSS)\,/\,M}{RSS\,/\,(n-k-1)} \backsim F_{M,n-k-1}$$

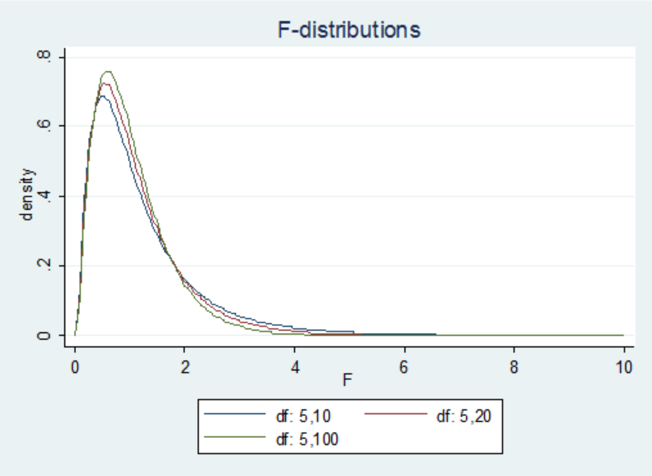
- RSS_M : residual sum of squares from the restricted model
- RSS : residual sum of squares from the unrestricted model
- M : number of restrictions in the null hypothesis
- n : number of observations in the sample
- k : number of parameters in the unrestricted model

This statistic **follows an F-distribution** with M numerator degrees of freedom and $n-k-1$ denominator degrees of freedom if H_0 is true.

Interlude: some examples of F-distributions



Interlude: some examples of F-distributions



F-test steps 1 & 2: our example

$$\begin{aligned} hrlywage_i &= \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 age_i^2 + \varepsilon_i \\ H_0 &: \beta_2 = \beta_3 = 0 \\ H_A &: H_0 \text{ not true} \\ \alpha &= 0.05 \end{aligned}$$

Restricted model (with $M = 2$, i.e. 2 restrictions):

$$hrlywage_i = \beta_0 + \beta_1 educ_i + \varepsilon_i$$

Unrestricted model:

$$hrlywage_i = \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 age_i^2 + \varepsilon_i$$

Step 3: estimating the restricted and unrestricted models

```
. reg hrwage educ

      Source |       SS       df       MS              Number of obs =     151
-------------+------------------------------           F(  1,   149) =   17.34
       Model |  305.278788     1  305.278788           Prob > F      =  0.0001
    Residual |  2623.76835   149  17.6091836           R-squared     =  0.1042
-------------+------------------------------           Adj R-squared =  0.0982
       Total |  2929.04714   150  19.5269809           Root MSE      =  4.1963

------------------------------------------------------------------------------
      hrwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .4877584   .1171456     4.16   0.000     .2562771    .7192396
       _cons |  -.8489289   1.527422    -0.56   0.579    -3.867134    2.169276
------------------------------------------------------------------------------

. reg hrwage educ age agesq

      Source |       SS       df       MS              Number of obs =     151
-------------+------------------------------           F(  3,   147) =    8.96
       Model |  452.987817     3  150.995939           Prob > F      =  0.0000
    Residual |  2476.05932   147   16.843941           R-squared     =  0.1547
-------------+------------------------------           Adj R-squared =  0.1374
       Total |  2929.04714   150  19.5269809           Root MSE      =  4.1041

------------------------------------------------------------------------------
      hrwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5842392   .1198999     4.87   0.000      .347289    .8211894
         age |   .4195173   .2447848     1.71   0.089    -.0642346    .9032692
       agesq |  -.0039981    .002883    -1.39   0.168    -.0096955    .0016994
       _cons |  -11.84532   5.269089    -2.25   0.026    -22.25827   -1.432374
------------------------------------------------------------------------------
```

$RSS_M = 2623.77 \qquad RSS = 2476.06$

F-test step 4: our example

- We want to quantify the added value of *age* and age^2 to the equation: can do that by **comparing the residual sum of squares from the restricted model, $\mathbf{RSS_M}$, to that of the unrestricted model, RSS**.
- But note that **by construction of OLS, $\mathbf{RSS_M} > \mathbf{RSS}$**!
- Therefore, to reject H_0 (i.e. reject the restricted model in favor of the unrestricted model), the difference in RSS should be "big enough". This is exactly what the **F-statistic** tells us.

F-test step 4: our example

$$H_0 : \beta_2 = \beta_3 = 0 \qquad H_A : H_0 \text{ not true}$$

- The **F-statistic** is

$$F = \frac{(RSS_M - RSS)\,/\,M}{RSS\,/\,(n-k-1)} \backsim F_{M,n-k-1}$$

- Here we have 2 restrictions (M=2):
  - Restriction 1: $\beta_2 = 0$
  - Restriction 2: $\beta_3 = 0$
- This statistic **follows an F-distribution** with M numerator degrees of freedom and $n-k-1$ denominator degrees of freedom if H_0 is true.

F-test step 4: our example

- The **F-statistic formula** is

$$F = \frac{\left(RSS_M - RSS\right)/M}{RSS/\left(n-k-1\right)} \backsim F_{M,n-k-1}$$

- Here the value is:

$$F = \frac{\left(2623.77 - 2476.06\right)/2}{2476.06/\left(151-3-1\right)} = 4.38$$

F-test steps 5 & 6: our example

- This F-statistic of 4.38 needs to be **compared to a critical value** obtained from the relevant F-distribution: here, with 2 numerator df and 147 denominator df and $\alpha = 0.05$.
- $F_c = F_{2,147,0.05} = 3.00$ (see Studenmund Appendix Table B2)
- 4.38>3.00: the **F-statistic is larger than the critical value, so we reject the null hypothesis** that the restricted model is true.
- **Conclusion**: *age* and *age*^{2} have a jointly significant impact on wages, holding constant education. (Note that this is despite the effects being *individually* insignificant!)

The F-statistic in Stata

```
. reg hrwage educ age agesq

      Source |       SS       df       MS              Number of obs =     151
-------------+------------------------------           F(  3,   147) =    8.96
       Model |  452.987817     3  150.995939           Prob > F      =  0.0000
    Residual |  2476.05932   147   16.843941           R-squared     =  0.1547
-------------+------------------------------           Adj R-squared =  0.1374
       Total |  2929.04714   150  19.5269809           Root MSE      =  4.1041

------------------------------------------------------------------------------
      hrwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5842392   .1198999     4.87   0.000      .347289    .8211894
         age |   .4195173   .2447848     1.71   0.089    -.0642346    .9032692
       agesq |  -.0039981    .002883    -1.39   0.168    -.0096955    .0016994
       _cons |  -11.84532   5.269089    -2.25   0.026    -22.25827   -1.432374
------------------------------------------------------------------------------

. test age agesq

 ( 1)  age = 0
 ( 2)  agesq = 0

       F(  2,   147) =    4.38
            Prob > F =    0.0141
```

Stata gives us the probability value, which we can compare directly to our chosen significance level: 0.01<0.05 hence reject H_0.

More applications of the F-test

Take the following unrestricted model:

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

- The example used the **F-test for exclusion restrictions**, e.g. $H_0 : \beta_2 = \beta_3 = 0$

But the F-test has more applications which are natural extensions of the one above:

- **Model F-test:** $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$
- Other **multiple linear restrictions**, e.g.: $H_0 : \beta_1 = \beta_2 = 1$; $H_0 : \beta_3 = 2 * \beta_2$

Model F-test

The **model F-test** is an F-test on the joint impact of **all independent variables** included in the model. (Note that the intercept is still allowed to be non-zero.)

$$hrlywage_i = \beta_0 + \beta_1 educ_i + \beta_2 age_i + \beta_3 age_i^2 + \varepsilon_i$$

$$\begin{aligned} H_0 &: \beta_1 = \beta_2 = \beta_3 = 0 \\ H_A &: H_0 \text{ not true} \end{aligned}$$

- The model F-test establishes **whether a regression exists**
- It is **reported in Stata regression output**, together with its p-value.

The model F-test

```
. reg hrwage educ age agesq

      Source |       SS       df       MS              Number of obs =     151
-------------+------------------------------           F(  3,   147) =    8.96
       Model |  452.987817     3  150.995939           Prob > F      =  0.0000
    Residual |  2476.05932   147   16.843941           R-squared     =  0.1547
-------------+------------------------------           Adj R-squared =  0.1374
       Total |  2929.04714   150  19.5269809           Root MSE      =  4.1041

------------------------------------------------------------------------------
      hrwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5842392   .1198999     4.87   0.000      .347289    .8211894
         age |   .4195173   .2447848     1.71   0.089    -.0642346    .9032692
       agesq |  -.0039981    .002883    -1.39   0.168    -.0096955    .0016994
       _cons |  -11.84532   5.269089    -2.25   0.026    -22.25827   -1.432374
------------------------------------------------------------------------------
```

All independent variables (educ, age, agesq) are jointly statistically significant since 0.00<0.05.

F-test for other multiple linear restrictions: example

Cobb-Douglas production function:

$$Y = \gamma Labor^{\alpha} Capital^{\beta} \quad \text{with } \alpha + \beta = 1$$

Log-linearize:

$$\ln Y = \ln \gamma + \alpha \ln Labor + \beta \ln Capital$$

We can write this as the following **population model**:

$$\ln Y_i = \beta_0 + \beta_1 \ln Labor_i + \beta_2 \ln Capital_i + \varepsilon_i$$

To **test for constant returns to scale**:

$$\begin{aligned} H_0 &: \beta_1 + \beta_2 = 1 \\ H_A &: H_0 \text{ not true} \end{aligned}$$

F-test for constant returns to scale

```
. reg lnY lnL lnK

      Source |       SS       df       MS              Number of obs =      30
-------------+------------------------------           F(  2,    27) =   94.71
       Model |  4.52531097     2  2.26265549           Prob > F      =  0.0000
    Residual |  .645039833    27  .023890364           R-squared     =  0.8752
-------------+------------------------------           Adj R-squared =  0.8660
       Total |  5.17035081    29  .178287959           Root MSE      =  .15457

------------------------------------------------------------------------------
         lnY |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lnL |   .4144814   .0404376    10.25   0.000     .3315102    .4974526
         lnK |    .347066   .0418764     8.29   0.000     .2611427    .4329894
       _cons |   .4500011   .2789965     1.61   0.118    -.1224526    1.022455
------------------------------------------------------------------------------

. test lnL + lnK = 1

 ( 1)  lnL + lnK = 1

       F(  1,    27) =   18.32
            Prob > F =    0.0002
```

0.00<0.05: H_0 rejected, hence constant returns to scale rejected.

The F-test versus the t-test

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i$$

- The **t-test** can only be used for null hypotheses involving a **single population parameter**.
- The **F-test** must be used for null hypotheses involving **multiple population parameters.**
- But the F-test can also be used to test against a two-sided H_A whenever the t-test is used: in that case $t^2 = F$ with identical associated p-values.

F-test versus t-test

```
. reg naffairs male

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  1,   599) =    0.08
       Model |   .899313       1     .899313           Prob > F      =  0.7740
    Residual |  6528.18222   599  10.8984678           R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.0015
       Total |  6529.08153   600  10.8818026           Root MSE      =  3.3013

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        male |   .0774559   .2696384     0.29   0.774    -.4520956    .6070073
       _cons |   1.419048   .1860062     7.63   0.000     1.053744    1.784351
------------------------------------------------------------------------------

. test male

 ( 1)  male = 0

       F(  1,   599) =    0.08
            Prob > F =    0.7740
```

$$t^2 = F$$
$$0.29^2 = 0.08$$

This class

- We have discussed **hypothesis testing** (Studenmund Chapter 5)
  - For 1 population parameter: t-test
  - For >1 population parameter: F-test
- Now we turn to **omitted variable bias** (a violation of OLS assumption $Corr(\varepsilon_i, X_i) = 0$) (Studenmund Chapter 6)

Assumptions 1-4

To get that $E(\widehat{\beta}_k) = \beta_k$:

1. Population model is linear in parameters (and the error term is additive)
2. Error term has a zero population mean: $E(\varepsilon_i) = 0$
3. **All independent variables are uncorrelated with the error term**: $Corr(\varepsilon_i, X_i) = 0$. Note that this has to hold for all k independent variables.
4. No perfect (multi)collinearity between independent variables (and no variable is a constant)

Violation of assumption 3: omitted variable bias

$$Corr(\varepsilon_i, X_i) = 0$$

- Assumption 3 is violated when **we exclude a relevant variable**, that is, a variable that has a partial effect on Y and is therefore in the error term ε, **and this variable is also correlated with the included independent variable(s)**, such that $Corr(\varepsilon_i, X_i) \neq 0$.
- We will see that this causes **biased estimates**, $E(\widehat{\beta}_k) \neq \beta_k$
- This bias is called **omitted variable bias** (OVB)

Conditions for omitted variable bias

The omitted variable X_m must satisfy two conditions to get OVB:

1. X_m is in the error term, i.e. X_m is a determinant of Y; **and**
2. X_m is correlated with the included regressor(s) X_k, i.e. $Corr(X_{ki}, X_{mi}) \neq 0$

Taken together, these two conditions give us $Corr(\varepsilon_i, X_i) \neq 0$

Omitted variable bias: mechanics

- Assume that the **true population model**, which satisfies all OLS assumptions, is

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

  Estimating this model with OLS yields unbiased estimates.

- But **now assume we omit $\mathbf{X_2}$**, such that

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \varepsilon_i^* \\ \varepsilon_i^* &= \beta_2 X_{2i} + \varepsilon_i \end{aligned} \tag{2}$$

- Note that the error term ε^* now contains the effect of the omitted variable X_2.

Omitted variable bias: mechanics

- **We omitted $\mathbf{X_2}$**, such that

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \varepsilon_i^* \\ \varepsilon_i^* &= \beta_2 X_{2i} + \varepsilon_i \end{aligned} \tag{2}$$

- The error term ε^* contains the effect of the omitted variable X_2.
- We know this model violates assumption 3 that $Corr(\varepsilon_i^*, X_{1i}) = 0$, if $Corr(X_{1i}, X_{2i}) \neq 0$.
- As a result, OLS estimates of equation 2 will be **biased**. But we can **find the direction and size of the bias**!

The omitted variable bias formula

The **misspecified model**, where X_2 is the omitted variable:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \varepsilon_i^* \\ \varepsilon_i^* &= \beta_2 X_{2i} + \varepsilon_i \end{aligned} \tag{2}$$

The **relationship between the omitted and the included variable** is given by

$$X_{2i} = \alpha_0 + \alpha_1 X_{1i} + u_i$$

Estimates of equation 2:

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + e_i$$

We can show that $E(\widehat{\beta}_1) \neq \beta_1$ but instead:

$$\mathbf{E}(\widehat{\boldsymbol{\beta}}_{\mathbf{1}}) = \boldsymbol{\beta}_1 + \boldsymbol{\alpha}_1 \boldsymbol{\beta}_2 \tag{OVB formula}$$

See appendix for proof.

OVB: direction & size of the bias

The **OVB formula** is:

$$\underbrace{E(\widehat{\beta}_1)}_{E(\text{estimated coefficient})} = \underbrace{\beta_1}_{\text{true coefficient}} + \underbrace{\alpha_1 \beta_2}_{\text{bias}}$$

This shows that that **the bias depends on**:

- α_1 : the bivariate relationship between the included and the omitted variable: sign and size
- β_2 : the partial effect of the omitted variable on the dependent variable Y: sign and size

OVB: direction of the bias

$$\underbrace{E(\widehat{\beta}_1)}_{E(\text{estimated coefficient})} = \underbrace{\beta_1}_{\text{true coefficient}} + \underbrace{\alpha_1 \beta_2}_{\text{bias}}$$

- $\alpha_1 \beta_2 > 0$: **Positive bias** = overestimate the effect of X_1 on Y, such that $\widehat{\beta}_1 > \beta_1$
  - Occurs when α_1 and β_2 have the same sign
- $\alpha_1 \beta_2 < 0$: **Negative bias** = underestimate the effect of X_1 on Y, such that $\widehat{\beta}_1 < \beta_1$
  - Occurs when α_1 and β_2 have opposing signs

OVB: direction of the bias

$$\underbrace{E(\widehat{\beta}_1)}_{E(\text{estimated coefficient})} = \underbrace{\beta_1}_{\text{true coefficient}} + \underbrace{\alpha_1 \beta_2}_{\text{bias}}$$

	$\alpha_1 > 0$	$\alpha_1 < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

OVB: size of the bias

$$\underbrace{E(\widehat{\beta}_1)}_{E(\text{estimated coefficient})} = \underbrace{\beta_1}_{\text{true coefficient}} + \underbrace{\alpha_1 \beta_2}_{\text{bias}}$$

- In absolute terms, the bias is larger when α_1 and β_2 are larger.
- The **bias is zero if**
  - $\alpha_1 = 0$: correlation between X_1 and X_2 is zero, or
  - $\beta_2 = 0$: X_2 has no effect on Y
  - These are the 2 conditions for OVB!

Consequence of OVB

- When we have omitted variable bias, **OLS produces biased estimates**
- This means we can **no longer interpret our coefficients as causal** since not all relevant other variables are held constant!
- Let's illustrate this with some **examples**.

Affairs

Let's say we want to find the **effect of having kids on the number of affairs**, and we estimate the following sample model

$$nraffairs_i = \widehat{\beta}_0 + \widehat{\beta}_1 kids_i + e_i$$

where *kids* is a dummy variable, $=1$ if the person has kids.

We expect that people with kids have more affairs than people without kids, and choose the following hypotheses about the unobserved population model:

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 \neq 0 \end{aligned}$$

Affairs

```
. sum kids

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        kids |       601    .7154742    .4515641          0          1

. reg naffairs kids

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  1,   599) =    6.55
       Model |  70.632204     1   70.632204           Prob > F      =  0.0107
    Residual | 6458.44933   599  10.7820523           R-squared     =  0.0108
-------------+------------------------------           Adj R-squared =  0.0092
       Total | 6529.08153   600  10.8818026           Root MSE      =  3.2836

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |   .7598123   .2968627     2.56   0.011      .176794    1.342831
       _cons |   .9122807   .2511034     3.63   0.000     .4191306    1.405431
------------------------------------------------------------------------------
```

$$\widehat{\beta}_1 = 0.76$$

The effect is significant at the 5% level: having kids increases the number of affairs in the last year by 0.76 affair.

Affairs

But what if the true population model is not

$$nraffairs_i = \beta_0 + \beta_1 kids_i + \varepsilon_i^*$$

but rather

$$nraffairs_i = \beta_0 + \beta_1 kids_i + \beta_2 yrsmarried_i + \varepsilon_i$$

This means we have omitted the variable years of marriage. We should have estimated:

$$nraffairs_i = \widehat{\beta}_0 + \widehat{\beta}_1 kids_i + \widehat{\beta}_2 yrsmarried_i + e_i$$

Affairs

```
. sum yrsmarr

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     yrsmarr |       601    8.177696    5.571303       .125         15

. reg naffairs kids yrsmarr

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  2,   598) =   10.82
       Model |  228.017878     2  114.008939           Prob > F      =  0.0000
    Residual |  6301.06365   598  10.5368957           R-squared     =  0.0349
-------------+------------------------------           Adj R-squared =  0.0317
       Total |  6529.08153   600  10.8818026           Root MSE      =  3.2461

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -.0328768   .3580389    -0.09   0.927    -.7360433    .6702896
     yrsmarr |   .1121551   .0290197     3.86   0.000     .0551622     .169148
       _cons |    .562259   .2642378     2.13   0.034     .0433122    1.081206
------------------------------------------------------------------------------
```

The effect is **no longer statistically significant**: having children does not have an impact on the number of affairs, once the duration of the marriage is controlled for!

Affairs: finding the bias with the OVB formula

```
. reg naffairs kids yrsmarr

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  2,   598) =   10.82
       Model |  228.017878     2  114.008939           Prob > F      =  0.0000
    Residual |  6301.06365   598  10.5368957           R-squared     =  0.0349
-------------+------------------------------           Adj R-squared =  0.0317
       Total |  6529.08153   600  10.8818026           Root MSE      =  3.2461

------------------------------------------------------------------------------
    naffairs |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |  -.0328768   .3580389    -0.09   0.927    -.7360433    .6702896
     yrsmarr |   .1121551   .0290197     3.86   0.000     .0551622     .169148
       _cons |    .562259   .2642378     2.13   0.034     .0433122    1.081206
------------------------------------------------------------------------------

. reg yrsmarr kids

      Source |       SS       df       MS              Number of obs =     601
-------------+------------------------------           F(  1,   599) =  292.59
       Model |  6111.64094     1  6111.64094           Prob > F      =  0.0000
    Residual |  12512.0103   599  20.8881642           R-squared     =  0.3282
-------------+------------------------------           Adj R-squared =  0.3270
       Total |  18623.6513   600  31.0394188           Root MSE      =  4.5704

------------------------------------------------------------------------------
     yrsmarr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        kids |   7.067794    .413195    17.11   0.000     6.256307    7.879281
       _cons |   3.120871   .3495039     8.93   0.000     2.434469    3.807273
------------------------------------------------------------------------------
```

Affairs: finding the bias with the OVB formula

$$
\begin{aligned}
nraffairs_i &= \widehat{\beta}_0 + \widehat{\beta}_1 kids_i + \widehat{\beta}_2 yrsmarried_i + e_i \\
\widehat{\beta}_2 &= 0.1121551 \\
yrsmarried_i &= \widehat{\alpha}_0 + \widehat{\alpha}_1 kids_i + u_i \\
\widehat{\alpha}_1 &= 7.067794
\end{aligned}
$$

$$
\begin{aligned}
bias &= \widehat{\alpha}_1 \widehat{\beta}_2 \\
&= 7.067794 \times 0.1121551 \\
&= 0.79269
\end{aligned}
$$

The **bias is positive**, which means we **overestimated the effect of having children on the number of affairs** because we omitted the duration of the marriage from the equation. The size of the overestimation was 0.79 ($= 0.7598123 - -0.0328768$).

Affairs example: intuition

- The first estimate gave the impression that having children has a positive effect on the number of affairs in the past year.
- However, once we controlled for the duration of the marriage, this effect disappeared.
- This is because longer marriages are more likely to have children, and people in longer marriages had more affairs in the past year.
- So what appeared to be the effect of having children on affairs, was actually the effect of being married longer on affairs!

Ability bias in the Mincer equation

Wage regression (so-called Mincer equation):

$$wage_i = \beta_0 + \beta_1 educ_i + \varepsilon_i^*$$

- β_1 : the impact of one additional year of education on wage, i.e. the **rate of return for one year of education**.
- All **other factors that have an influence on the wage are in the error term**- if such factors are **also correlated with education**, we get **omitted variable bias**.
  - **Ability** is one of the factors that also impact the wage: we can think of ability as people's IQ.
  - Ability is also **correlated with education**: people with higher IQ are more likely to get more years of education.

Ability bias in the Mincer equation

Wage regression (so-called Mincer equation):

$$
\begin{aligned}
wage_i &= \beta_0 + \beta_1 educ_i + \varepsilon_i^* \\
wage_i &= \beta_0 + \beta_1 educ_i + \beta_2 IQ_i + \varepsilon_i \\
IQ_i &= \alpha_0 + \alpha_1 educ_i + u_i
\end{aligned}
$$

We expect:

- $\beta_2 > 0$: IQ increases wages (cet. par. on education)
- $\alpha_1 > 0$: higher educated people have higher IQs
- Hence, bias $\alpha_1 \beta_2 > 0$, so we **overestimate the return to education if we do not control for IQ**. This is a very famous example of OVB in economics, called "ability bias".

Ability bias in the Mincer equation

```
. descr wage educ iq

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
wage            float  %9.0g                  Wage in cents per hour
educ            float  %9.0g                  Nr of years of education
iqscore         float  %9.0g                  a normed IQ score

. reg wage educ

      Source |       SS       df       MS              Number of obs =    2061
-------------+------------------------------           F(  1,  2059) =  115.13
       Model |  7596517.54     1  7596517.54           Prob > F      =  0.0000
    Residual |   135851554  2059   65979.385           R-squared     =  0.0530
-------------+------------------------------           Adj R-squared =  0.0525
       Total |   143448071  2060  69634.9861           Root MSE      =  256.86

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   26.70297   2.488607    10.73   0.000     21.82252    31.58342
       _cons |   242.8137   35.10625     6.92   0.000     173.9663    311.6612
------------------------------------------------------------------------------

. reg wage educ iq

      Source |       SS       df       MS              Number of obs =    2061
-------------+------------------------------           F(  2,  2058) =   66.24
       Model |  8675400.98     2  4337700.49           Prob > F      =  0.0000
    Residual |   134772670  2058  65487.2062           R-squared     =  0.0605
-------------+------------------------------           Adj R-squared =  0.0596
       Total |   143448071  2060  69634.9861           Root MSE      =   255.9

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   20.73191   2.882899     7.19   0.000      15.0782    26.38561
     iqscore |   1.725291   .4250631     4.06   0.000     .8916926     2.55889
       _cons |   149.1893   41.89648     3.56   0.000     67.02539    231.3532
------------------------------------------------------------------------------
```

Explaining the obesity epidemic

We want to investigate **whether a lack of exercise causes higher BMI**

$$BMI_i = \beta_0 + \beta_1 exercise_i + \varepsilon_i^* \tag{3}$$

But we **omit an important explanatory variable which is correlated with exercise, food intake** (daily calories consumed):

$$\begin{aligned} BMI_i &= \beta_0 + \beta_1 exercise_i + \beta_2 calories_i + \varepsilon_i \\ calories_i &= \alpha_0 + \alpha_1 exercise_i + u_i \end{aligned}$$

If we estimated equation 3, what direction would we expect for the bias?

$$bias = \alpha_1 \beta_2$$

Explaining the obesity epidemic

$$
\begin{aligned}
BMI_i &= \beta_0 + \beta_1 exercise_i + \varepsilon_i^* \qquad (4) \\
BMI_i &= \beta_0 + \beta_1 exercise_i + \beta_2 calories_i + \varepsilon_i \\
calories_i &= \alpha_0 + \alpha_1 exercise_i + u_i
\end{aligned}
$$

We expect $\beta_2 < 0$, and $\alpha_1 < 0-$ that is, taking in more calories increases BMI (cet. par.) and there is a negative correlation between exercise & calories (i.e. people who exercise more tend to take in fewer calories, e.g. because they're the health-conscious type).

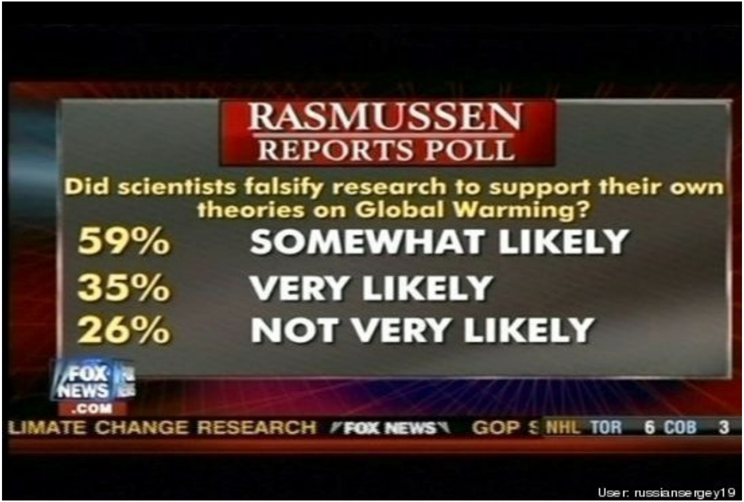
$$bias = \alpha_1 \beta_2 < 0$$

We would expect a **negative bias** in equation 3, such that we find a more negative effect of exercise on BMI than the true effect of exercise on BMI if we omit calorie intake from the equation.

Fox news


FOX40
NEWS
BREAKING NEWS
FOX40.com
REPORTS: OBAMA BIN LADEN DEAD

Fox news, another gem



Does Fox news make you stupid?

NEW ISSUE ARCHIVES
LOG IN TO RS ALL ACCESS

Rolling Stone POLITICS

Study: Watching Fox News Actually Makes You Stupid

POSTED: MAY 24, 1:20 PM ET | By JILLIAN RAYFIELD

Like 1.7k Tweet 235 +1 10 Submit Comment 63

It's not exactly a revelation that Fox News viewers are spectacularly ill informed about current events compared to people who watch other networks. But according to a recent report, the Fox audience knows less even than folks *who don't watch any news at all.*

Researchers from New Jersey's Fairleigh Dickinson University asked about a thousand people five questions on domestic issues (e.g., "Which party has the most seats in the House of Representatives right now?") and five on international ones (e.g., "There have been increasing talks about economic sanctions against Iran. What are these sanctions supposed to do?").

Fox viewers scored the lowest in both categories, getting an average of 1.04 questions right on domestic issues and 1.08 on international, behind people who watch MSNBC (which, on international affairs, also lagged the non-news watchers) *The Daily Show*, and the

Does Fox news make you stupid?

Research question: does watching Fox news *reduce* people's knowledge about what is going on in the world?

$$knowledge_i = \beta_0 + \beta_1 Fox_i + \varepsilon_i$$

where *knowledge* measures respondents score on a test about knowledge of the world; and *Fox* is a dummy variable, =1 if the person watches Fox news, =0 if they do not.

$$\begin{aligned} H_0 &: \beta_1 \geq 0 \\ H_A &: \beta_1 < 0 \end{aligned}$$

The researchers claim that they reject H_0, and **conclude that watching Fox news makes you less informed**.

Is this a causal effect? Suspected omitted variable bias..

$$
\begin{aligned}
knowledge_i &= \beta_0 + \beta_1 Fox_i + \varepsilon_i^* \\
knowledge_i &= \beta_0 + \beta_1 Fox_i + \beta_2 IQ_i + \varepsilon_i \\
IQ_i &= \alpha_0 + \alpha_1 Fox_i + u_i
\end{aligned}
$$

I suspect **IQ is an omitted variable** here, causing a biased $\widehat{\beta}_1$.
What would the bias be?

- $\beta_2 > 0$: smarter people have more knowledge of the world
- $\alpha_1 < 0$: people who watch Fox news are less likely to be smart.
- $\alpha_1\beta_2 < 0$: **negative bias**, hence the estimated effect is more negative than the true effect!

Suspected omitted variable bias

$$E(\widehat{\beta}_1) = \beta_1 + \underbrace{\alpha_1 \beta_2}_{\text{negative}}$$

The researchers found a statistically significant negative effect, e.g. $\widehat{\beta}_1 = -5$.
But the bias term is also negative, so the **true effect could be zero** (of course, it could also still be negative, or even be positive, depending on how large the bias is!). For instance:

$$\underbrace{-5}_{\text{found effect}} = \underbrace{0}_{\text{true effect}} + \underbrace{(-5)}_{\text{bias}}$$

OVB detection & an easy solution

- **Use economic theory** and reasoning to detect OVB:
  - Think of any variables you may have omitted from your model: which variables are relevant for explaining the dependent variable Y? Are these also correlated with the included X variables? If so, you have an OVB problem.
- If possible, **easy solution = include the omitted variable(s)** in your model.

OVB: when the easy solution is impossible

- **If the dataset does not have information the omitted variable(s)**, there is no easy solution.
- Instead: **use the OVB formula** to reason whether the omitted variable(s) would bias your estimate of interest, and if so, in what direction. (This was done in examples 3 & 4: even without being able to estimate the bias, we can still reason to say something about how it would affect our estimates.)

(In a MSc course, you will see there is another estimator you can use in the case of suspected OVB, rather than OLS.)

Summary

- **Omitted variable bias** is a **serious problem** for empirical economic research: it violates one of the OLS assumptions for unbiasedness.
- You cannot rely on the data to tell you everything: you must use economic reasoning to detect and fix omitted variable bias!
- Also see Studenmund's "four important specification criteria" (p. 178 in 6th edition)

Project paper

- Interpretation of **statistical significance** (using t-values / p-values / confidence intervals).
- **Economic significance**: given statistical significance, discuss the magnitude of the effects you find (e.g. is it what you expected from economic reasoning?).
- Interpretation of **F-statistics**.
- Are there important **omitted variables** you can think of? (Note the 2 conditions for OVB!)
- If you can think of such omitted variables: what are the **consequences for the bias of the parameter estimates**? Do you expect you are over- or underestimating the true effect?

Derivation of OVB formula (bivariate case)

True model:

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

Model with variable X_2 omitted:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i^* \tag{2}$$

In equation (2), the OLS estimate of β_1 is given by (see lecture slides week 2):

$$\widehat{\beta}_1 = \frac{Cov(X_{1i}, Y_i)}{Var(X_{1i})}$$

Derivation of OVB formula (bivariate case), cont'd

$$\widehat{\beta}_1 = \frac{Cov(X_{1i}, Y_i)}{Var(X_{1i})}$$

We can work out the covariance term to find the OVB formula:

$$\begin{aligned} Cov(X_{1i}, Y_i) &= Cov(X_{1i}, \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i) \\ &= Cov(X_{1i}, \beta_0) + Cov(X_{1i}, \beta_1 X_{1i}) + Cov(X_{1i}, \varepsilon_i^*) \\ &= 0 + \beta_1 Var(X_{1i}) + Cov(X_{1i}, \varepsilon_i^*) \end{aligned}$$

Derivation of OVB formula (bivariate case), cont'd

Using that $\varepsilon_i^* = \beta_2 X_{2i} + \varepsilon_i$:

$$\begin{aligned} Cov(X_{1i}, Y_i) &= \beta_1 Var(X_{1i}) + Cov(X_{1i}, \beta_2 X_{2i} + \varepsilon_i) \\ &= \beta_1 Var(X_{1i}) + Cov(X_{1i}, \beta_2 X_{2i}) + Cov(X_{1i}, \varepsilon_i) \\ \text{since } Cov(X_{1i}, \varepsilon_i) &= 0: \\ Cov(X_{1i}, Y_i) &= \beta_1 Var(X_{1i}) + Cov(X_{1i}, \beta_2 X_{2i}) \end{aligned}$$

Derivation of OVB formula (bivariate case), cont'd

$$\widehat{\beta}_1 = \frac{Cov(X_{1i}, Y_i)}{Var(X_{1i})}$$

Using the previously found expression for $Cov(X_{1i}, Y_i)$,

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\beta_1 Var(X_{1i}) + Cov(X_{1i}, \beta_2 X_{2i})}{Var(X_{1i})} \\
&= \frac{\beta_1 Var(X_{1i})}{Var(X_{1i})} + \frac{Cov(X_{1i}, \beta_2 X_{2i})}{Var(X_{1i})} \\
&= \beta_1 + \beta_2 \frac{Cov(X_{1i}, X_{2i})}{Var(X_{1i})}
\end{aligned}$$

Derivation of OVB formula (bivariate case), cont'd

$$\widehat{\beta}_1 = \beta_1 + \beta_2 \frac{Cov(X_{1i}, X_{2i})}{Var(X_{1i})}$$

Note that $\frac{Cov(X_{1i}, X_{2i})}{Var(X_{1i})}$ is α_1, the slope coefficient from the bivariate model of X_2 on X_1 :

$$X_{2i} = \alpha_0 + \alpha_1 X_{1i} + u_i$$

Hence, the OVB formula is given by:

$$\widehat{\beta}_1 = \beta_1 + \beta_2 \alpha_1$$

Back to main slides