The simple and multiple regression model

Chapters 2 & 3: Introductory Econometrics 771

Prof Dieter von Fintel

Department of Economics

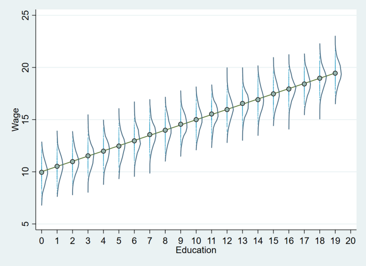
Stellenbosch University

14 February 2024

Overview

1. Conditional expectation $\longrightarrow$ linear regression
2. When does regression have causal or ceteris paribus interpretation?
   - Population vs Sample Regression Functions
3. The Ordinary Least Squares Estimator
   - Derivation
4. Mechanics and interpretation of OLS with multiple regressors
   - Properties of OLS estimators
   - Goodness of fit
   - Partialling out interpretation
5. Expected values and variances of OLS
   - Assumptions to ensure that OLS is unbiased/causal
   - Including too many variables
   - Sample variation in OLS estimates
   - Imperfect multicollinearity
   - Variances in misspecified models
   - Gauss-Markov Theorem

Conditional expectation function



- Different distributions of

  $Y = wage$ at $x = educ = 1 \cdots 20$
  $\rightarrow$ distributions around CEF

  - Deterministic vs statistical
- How to estimate the **slope** of the CEF?

$$\widehat{\beta}_1 = \frac{\partial E(Y|X)}{\partial x}$$

...it quantifies the **relationship** between $Y = wage$ and $X = educ$

Definition of the regression model

"Explain $y = wage$ in terms of $x = educ$"

- Functional form: "Linear Regression"

$$y = \beta_0 + \beta_1 x_{main} + u$$

  - β_0: $y - intercept$ - *"mean wage of individuals with 0 education"*
    - **CONDITIONAL** mean
  - $\beta_1 = \frac{\Delta wage}{\Delta educ}$: slope of a straight line - $\Delta wage$ for one year $\Delta educ$
  - u are unobservables - *social networks, soft skills, ability, motivation, etc*
    - Ceteris paribus???
- Linearity?
  - In parameters, not in variables (more later)
  - A marginal change in x (say, education) has the same impact on y (say, wage), regardless of the level of x
    - Realistic? We will see how to deal with this later

When is β_1 a *ceteris paribus* effect?

- Hold other **observables** (x_{other}) & **unobservables** (ε) constant as x_{main} changes
  - Think of $u = \beta_{other}x_{other} + \varepsilon \Rightarrow y = \beta_0 + \beta_1 x_{main} + \beta_{other}x_{other} + \varepsilon$
    - Split u into "information" and "randomness" that is uncorrelated with x_{main}
  - If x_{other} is part of u (OR: x_{other} **also** determines y) **AND** correlates with x_{main}, cannot "hold it constant" unless somehow "taken out of u"
- **POPULATION REGRESSION FUNCTION**: β_1 is "true" (not necessarily known) relationship if... all relevant x's included (x_{main}); or only "randomness" (ε) is left in u, so that $Cov(u; x_{main}) = Cov(\varepsilon; x_{main}) = 0$

$$y = \beta_0 + \beta_1 x_{main} + u$$

$$\Delta y = \beta_1 \Delta x_{main} + \Delta u$$

$$\frac{\Delta y}{\Delta x_{main}} = \frac{\Delta \beta_0}{\Delta x_{main}} + \beta_1 \frac{\Delta x_{main}}{\Delta x_{main}} + \frac{\Delta u}{\Delta x_{main}}$$

$$\frac{\Delta y}{\Delta x_{main}} = 0 + \beta_1 + \frac{\Delta u}{\Delta x_{main}} \Rightarrow \beta_1 = \frac{\Delta y}{\Delta x_{main}} - \frac{\Delta u}{\Delta x_{main}}$$

$$\beta_1 = \frac{\Delta y}{\Delta x} \textit{ only if } \frac{\Delta u}{\Delta x} = 0 \textit{ or } \frac{\Delta x_{other}}{\Delta x_{main}} = \frac{\Delta \varepsilon}{\Delta x_{main}} = 0$$

State this more formally

- If there is an intercept, it can be shown that $E(u) = 0$... **always**
- Now what must we assume to obtain "ceteris paribus" estimates?
  - No correlation between x and u
  - Generalise this to **non-linear relationships** with conditional expectations: $E(u|x) = E(u)$
    - Mean (non-linear and linear) **INDEPENDENCE**
    - Average of unobservables is the same, regardless of values of x
    - **Concretely:** for regression to have *ceteris paribus* or **causal** interpretation, average motivation/ability/access to education (absorbed in u because it is **not measured/unobserved**) must be the same for people with low and high levels of education (x_{main}) $\rightarrow$ **likely not a good assumption** $\rightarrow$ **estimate of β_1 does not necessarily have causal interpretation**
    - **How could unobservables influence our estimate relative to the true ("unbiased"/causal/population) value?**
  - Often simplified as: $E(u|x) = 0$ because $E(u) = 0$
    - **Zero conditional mean assumption**

POPULATION regression function

y, x and u are random variables

- They have a population distribution
  - A "real" set of values that is partially reflected in our sample
  - $E(y|x)$: how the average value of y changes with x in the population
  - In the population, the β are not random
    - They have no distribution, because one true (unbiased/causal/*ceteris paribus)* population value for them
  - "DATA GENERATING PROCESS": the conditional expectation function is the systematic/deterministic part of PRF, separated from the random component

$$y = \beta_0 + \beta_1 x + u$$
$$E(y|x) = E(\beta_0 + \beta_1 x + u|x)$$
$$= E(\beta_0|x) + E(\beta_1 x|x) + E(u|x)$$
$$= \beta_0 + \beta_1 x + 0$$

because if the PRF is fully specified, there is no remaining relationship between u and x

Model with 2 independent variables

Suppose the Population Regression Function includes `experience` according to theory

$$wage = \beta_0 + \beta_1 education + \beta_2 experience + u$$

- Taking experience out of the error term, and assume this theory is "enough" to characterise the DGP (ie u is now random and unrelated to all the x's)
  - β_1 is *ceteris paribus* effect of `education` on `wage` holding `experience` and `u` fixed
  - β_2 is *ceteris paribus* effect of `experience` on `wage` holding `education` and `u` fixed
- But now we have a better estimate of it; it is a **causal** estimate **IF** we have fully specified the PRF, meaning that $E(u|educ; exper) = 0$
- Had we left `experience` out

$$wage = \tilde{\beta}_0 + \tilde{\beta}_1 education + \tilde{u} \text{ where } \tilde{u} \text{ contains experience}$$

  - If `education` and `experience` are correlated, $E(educ|\tilde{u}) \neq 0$ so that $\tilde{\beta}_1 \neq \beta_1$

If PRF must contain more variables (k of them)

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \cdots + \beta_k x_k + u$$

The **zero conditional mean assumption** extends to:

$$E(u|x_1, x_2, \cdots, x_k) = E(u|\mathbf{x}) = 0$$

- Average of unobservables is zero regardless of **each value** of **each** x_j, for example
  - Average motivation (contained in u) must be zero at $educ = 0$ and $educ = 1$ and... $educ = 20$
  - **AND** average motivation must be zero at $exp = 0$ and $exp = 1$ and... $exp = 40$
  - **AND** similar for **all** other variables in the PRF
- Or simply: independence of **all** the variables and the unobservable population error

Sample Regression Function (SRF)

- Hardly ever have data on the whole population
- Two **main** data reasons for biased estimation (among others)
  1. Not all variables collected (as before): a "**column** problem"
  2. Do not sample whole population: a "**row** problem"
     - Draw *representative* SAMPLE from population
     - Draw inferences about population based on sample
     - Different sub-samples of data from the **same** population, estimate of the PRF (= SRF) is different in each case
     - *Estimate* because know true PRF without full information
     - $\widehat{\beta}$ is therefore also stochastic - a *random variable* $\Rightarrow$ $\widehat{\beta}$ has a distribution
     - (remember the distributions around the slope of the CEF?)
       (**NOTE:** the "hat" emphasises that this is an *estimate* from a sample)

Full information

- Imagine for a moment that *educ* and *age* tell us everything about why people get paid what they do...
- Code simulates a **fake** “population” level dataset that reflects the following PRF:

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u$$

where $\beta_0 = 10$, $\beta_1 = 0.5$, $\beta_2 = 0.1$

STATA CODE

```
 clear
set seed 1234
set obs 60000000
gen educ = int(rnormal()*1.4 + 12)
gen age = int(rnormal()*4+40)
gen exper = max(age - educ - 6 -int(rnormal()*0.1),0)
gen u = 0.1*rnormal()
gen wage = 10 + 0.5*educ +0.1*exper +u
drop age
```

Full information

- Population of $N = 60 million$
- "True" population regression function is

$$wage = \beta_0 + \beta_1 education + \beta_2 exper + u$$

- With full information could estimate β_1 from the PRF without a problem using Ordinary Least Squares (OLS) - more later

Observation	wage	educ	exper	random u
1	18.63818	11	31	0.0381753
2	17.58195	9	30	0.0819504
3	17.20783	11	17	0.0078265
4	18.22533	11	28	-0.0746732
5	18.55296	12	26	-0.0470415
6	17.37125	11	19	-0.0287531
7	17.56123	13	11	-0.038775
8	17.26208	11	19	-0.1379214
9	17.77695	11	23	-0.02305
10	17.60788	8	35	0.1078842
:	:	:	:	:
100	18.40646	13	19	0.0064596
:	:	:	:	:
:	:	:	:	:
1000	18.24569	12	23	-0.054311
:	:	:	:	:
100000	18.15609	11	28	-0.1439148
:	:	:	:	:
10000000	17.86163	13	15	-0.1383734
:	:	:	:	:
60000000	19.64043	15	22	-0.0595725

```
. reg wage educ exper u

      Source |       SS           df       MS      Number of obs   =  60000000
-------------+----------------------------------   F(3, 59999996)  >  99999.00
       Model |  29870427.6         3  9956809.19   Prob > F        =    0.0000
    Residual |  .000018149  59999996  3.0248e-13   R-squared       =    1.0000
-------------+----------------------------------   Adj R-squared   =    1.0000
       Total |  29870427.6  59999999  .497840468   Root MSE        =   5.5e-07

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |         .5   5.27e-11  9.5e+09   0.000           .5          .5
       exper |         .1   1.77e-11  5.6e+09   0.000           .1          .1
           u |          1   7.10e-10  1.4e+09   0.000            1           1
       _cons |         10   8.26e-10  1.2e+10   0.000           10          10
------------------------------------------------------------------------------
```

Don't observe randomness

- Population of $N = 60million$
- Estimate Sample Regression Function
  $\widehat{wage} = \widehat{\beta}_0 + \widehat{\beta}_1 education + \widehat{\beta}_2 exper$
- Only omitting random information (u) gives $\widehat{\beta}_1$ close to population β_1

Observation	wage	educ	exper	random u
1	18.63818	11	31	~~0.0381753~~
2	17.58195	9	30	~~0.0819504~~
3	17.20783	11	17	~~0.0078265~~
4	18.22533	11	28	~~-0.0746732~~
5	18.55296	12	26	~~-0.0470415~~
6	17.37125	11	19	~~-0.0287531~~
7	17.56123	13	11	~~-0.038775~~
8	17.26208	11	19	~~-0.1379214~~
9	17.77695	11	23	~~-0.02305~~
10	17.60788	8	35	~~0.1078842~~
:	:	:	:	~~:~~
100	18.40646	13	19	~~0.0064596~~
:	:	:	:	~~:~~
:	:	:	:	~~:~~
1000	18.24569	12	23	~~-0.054311~~
:	:	:	:	~~:~~
100000	18.15609	11	28	~~-0.1439148~~
:	:	:	:	~~:~~
10000000	17.86163	13	15	~~-0.1383734~~
:	:	:	:	~~:~~
60000000	19.64043	15	22	~~-0.0595725~~

```
. reg wage educ exper

      Source |       SS           df       MS      Number of obs   =  60000000
-------------+----------------------------------   F(2, 59999997)  >  99999.00
       Model |    29270445         2  14635222.5   Prob > F        =    0.0000
    Residual |   599982.53  59999997  .009999709   R-squared       =    0.9799
-------------+----------------------------------   Adj R-squared   =    0.9799
       Total |  29870427.6  59999999  .497840468   Root MSE        =        .1

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5000077   9.59e-06  5.2e+04   0.000     .4999889    .5000265
       exper |    .100001   3.22e-06  3.1e+04   0.000     .0999947    .1000073
       _cons |   9.999876   .0001503  6.7e+04   0.000     9.999582    10.00017
------------------------------------------------------------------------------
```

+Don't observe *exper* (part of PRF)

- Population of $N = 60million$
- Estimate Sample Regression Function

$$\widehat{wage} = \widehat{\beta}_0 + \widehat{\beta}_1 education$$

- Omitting non-random information (*exper*) gives $\widehat{\beta}_1$ not close to true β_1

Observation	wage	educ	~~exper~~	random u
1	18.63818	11	~~31~~	~~0.0381753~~
2	17.58195	9	~~30~~	~~0.0819504~~
3	17.20783	11	~~17~~	~~0.0078265~~
4	18.22533	11	~~28~~	~~-0.0746732~~
5	18.55296	12	~~26~~	~~-0.0470415~~
6	17.37125	11	~~19~~	~~-0.0287531~~
7	17.56123	13	~~11~~	~~-0.038775~~
8	17.26208	11	~~19~~	~~-0.1379214~~
9	17.77695	11	~~23~~	~~-0.02305~~
10	17.60788	8	~~35~~	~~0.1078842~~
:	:	:	:	:
100	18.40646	13	~~19~~	~~0.0064596~~
:	:	:	:	:
:	:	:	:	:
1000	18.24569	12	~~23~~	~~-0.054311~~
:	:	:	:	:
100000	18.15609	11	~~28~~	~~-0.1439148~~
:	:	:	:	:
10000000	17.86163	13	~~15~~	~~-0.1383734~~
:	:	:	:	:
60000000	19.64043	15	~~22~~	~~-0.0595725~~

```
. reg wage educ

      Source |       SS           df       MS      Number of obs   =  60000000
-------------+----------------------------------   F(1, 59999998)  >  99999.00
       Model |  19617090.2         1  19617090.2   Prob > F        =    0.0000
    Residual |  10253337.4  59999998  .170888962   R-squared       =    0.6567
-------------+----------------------------------   Adj R-squared   =    0.6567
       Total |  29870427.6  59999999  .497840468   Root MSE        =    .41339

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .4000193   .0000373   1.1e+04   0.000     .3999461    .4000925
       _cons |   13.34987   .0004327   3.1e+04   0.000     13.34902    13.35072
------------------------------------------------------------------------------

. correl
(obs=60,000,000)

             |     educ    exper        u     wage
-------------+------------------------------------
        educ |   1.0000
       exper |  -0.3357   1.0000
           u |   0.0001   0.0000   1.0000
        wage |   0.8104   0.2635   0.1418   1.0000
```

+take one sample of $n = 1000$

- Sample of **first** $n = 1000$ from population of $N = 60million$
- Estimate **Sample** Regression Function

$$\widehat{wage} = \widehat{\beta}_0 + \widehat{\beta}_1 education$$

- Omitting the $> 59million$ observations gives different $\widehat{\beta}_1$ to before

Observation	wage	educ	exper	random u
1	18.63818	11	~~31~~	~~0.0381753~~
2	17.58195	9	~~30~~	~~0.0819504~~
3	17.20783	11	~~17~~	~~0.0078265~~
4	18.22533	11	~~28~~	~~-0.0746732~~
5	18.55296	12	~~26~~	~~-0.0470415~~
6	17.37125	11	~~19~~	~~-0.0287531~~
7	17.56123	13	~~11~~	~~-0.038775~~
8	17.26208	11	~~19~~	~~-0.1379214~~
9	17.77695	11	~~23~~	~~-0.02305~~
10	17.60788	8	~~35~~	~~0.1078842~~
:	:	:	~~:~~	:
100	18.40646	13	~~19~~	~~0.0064596~~
:	:	:	~~:~~	:
:	:	:	~~:~~	:
1000	18.24569	12	~~23~~	~~-0.054311~~
:	:	:	:	:
~~100000~~	~~18.15609~~	~~11~~	28	~~-0.1439148~~
:	:	:	:	:
~~10000000~~	~~17.86163~~	~~13~~	15	~~-0.1383734~~
:	:	:	:	:
~~60000000~~	~~19.64043~~	~~15~~	22	~~-0.0595725~~

```
. reg wage educ if _n<=1000

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =   1828.98
       Model |   331.79645         1   331.79645   Prob > F        =    0.0000
    Residual |  181.048127       998  .181410948   R-squared       =    0.6470
-------------+----------------------------------   Adj R-squared   =    0.6466
       Total |  512.844576       999  .513357934   Root MSE        =    .42592

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .3990291   .0093304    42.77   0.000     .3807197    .4173386
       _cons |   13.36459   .1080214   123.72   0.000     13.15261    13.57656
------------------------------------------------------------------------------
```

+take 2^{nd} sample of $n = 1000$

- Sample of **last** $n = 1000$ from population of $N = 60million$
- Estimate Sample Regression Function

$$\widehat{wage} = \widehat{\beta}_0 + \widehat{\beta}_1 education$$

- Omitting the $> 59million$ observations gives a different $\widehat{\beta}_1$ to before (but with good sample design, it may not be that far away)

Observation	wage	educ	~~exper~~	random u
~~1~~	~~18.63818~~	~~11~~	~~31~~	~~0.0381753~~
~~2~~	~~17.58195~~	~~9~~	~~30~~	~~0.0819504~~
~~3~~	~~17.20783~~	~~11~~	~~17~~	~~0.0078265~~
~~4~~	~~18.22533~~	~~11~~	~~28~~	~~-0.0746732~~
~~5~~	~~18.55296~~	~~12~~	~~26~~	~~-0.0470415~~
~~6~~	~~17.37125~~	~~11~~	~~19~~	~~-0.0287531~~
~~7~~	~~17.56123~~	~~13~~	~~11~~	~~-0.038775~~
~~8~~	~~17.26208~~	~~11~~	~~19~~	~~-0.1379214~~
~~9~~	~~17.77695~~	~~11~~	~~23~~	~~-0.02305~~
~~10~~	~~17.60788~~	~~8~~	~~35~~	~~0.1078842~~
~~⋮~~	~~⋮~~	~~⋮~~	~~⋮~~	~~⋮~~
~~100~~	~~18.40646~~	~~13~~	~~19~~	~~0.0064596~~
~~⋮~~	~~⋮~~	~~⋮~~	~~⋮~~	~~⋮~~
~~⋮~~	~~⋮~~	~~⋮~~	~~⋮~~	~~⋮~~
~~1000~~	~~18.24569~~	~~12~~	~~23~~	~~-0.054311~~
~~⋮~~	~~⋮~~	~~⋮~~	~~⋮~~	~~⋮~~
~~100000~~	~~18.15609~~	~~11~~	~~28~~	~~-0.1439148~~
~~⋮~~	~~⋮~~	~~⋮~~	~~⋮~~	~~⋮~~
~~10000000~~	~~17.86163~~	~~13~~	~~15~~	~~-0.1383734~~
⋮	⋮	⋮	~~⋮~~	~~⋮~~
60000000	19.64043	15	~~22~~	~~-0.0595725~~

```
. reg wage educ if _n> _N-1000

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =   1725.41
       Model |  318.019237         1  318.019237   Prob > F        =    0.0000
    Residual |  183.947044       998  .184315675   R-squared       =    0.6335
-------------+----------------------------------   Adj R-squared   =    0.6332
       Total |  501.966281       999  .502468749   Root MSE        =    .42932

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .3932891   .0094682    41.54   0.000     .3747092    .4118689
       _cons |    13.4136   .1091353   122.91   0.000     13.19944    13.62776
------------------------------------------------------------------------------
```

- **Random** sample of $n = 1000$ from population of $N = 60million$
- Estimate Sample Regression Function

$$\widehat{wage} = \widehat{\beta}_0 + \widehat{\beta}_1 education$$

- Omitting the $> 59million$ observations gives a different $\widehat{\beta}_1$ to before (but with good sample design, it may not be that far away)

```
. sample 1000, count
(59,999,000 observations deleted)

. reg wage educ

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =   1789.27
       Model |  325.887479         1  325.887479   Prob > F        =    0.0000
    Residual |  181.769655       998  .182133923   R-squared       =    0.6419
-------------+----------------------------------   Adj R-squared   =    0.6416
       Total |  507.657135       999    .5081653   Root MSE        =    .42677

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .4107828   .0097112    42.30   0.000      .391726    .4298395
       _cons |   13.22154   .1127518   117.26   0.000     13.00028     13.4428
------------------------------------------------------------------------------
```

In summary

- We usually have **column** problems (omitted variables) that give us $\widehat{\beta} \neq \beta$
- We usually observe **one set of rows** that deviates from the population
  - Omitting rows can add to the column problem if the sample is **non-randomly** collected
  - Omitting rows is less problematic with random sampling
  - If we were to observe a **different** set of rows in our sample, we would get a different $\widehat{\beta}$ (even ignoring the column problems)
    - Our sample regression function therefore has **stochastic** estimates of $\widehat{\beta}$ with a distribution

SRF – an illustration using census

- "Population" - note: we are ignoring **column** problems for now

```
. reg l_inc educ

      Source |       SS           df       MS      Number of obs   = 1,540,893
-------------+----------------------------------   F(1, 1540891)   >  99999.00
       Model |  1034722.29         1  1034722.29   Prob > F        =    0.0000
    Residual |  2097529.44 1,540,891  1.36124453   R-squared       =    0.3303
-------------+----------------------------------   Adj R-squared   =    0.3303
       Total |  3132251.73 1,540,892  2.03275228   Root MSE        =    1.1667

------------------------------------------------------------------------------
       l_inc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   .2096349   .0002404   871.85   0.000     .2091636    .2101062
       _cons |   7.782751   .0020603  3777.40   0.000     7.778713    7.786789
------------------------------------------------------------------------------
```

2x Random samples: 0.05% observations

Source	SS	df	MS			
Model	447.772298	1	447.772298	Number of obs	=	770
Residual	1058.74018	768	1.37856794	F(1, 768)	=	324.81
				Prob > F	=	0.0000
Total	1506.51247	769	1.95905393	R-squared	=	0.2972
				Adj R-squared	=	0.2963
				Root MSE	=	1.1741

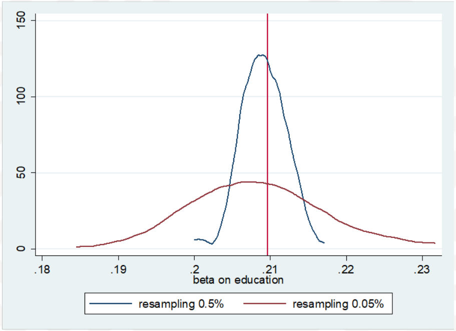
l_inc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
education	.195089	.0108248	18.02	0.000	.1738394	.2163387
_cons	7.923942	.0921992	85.94	0.000	7.742949	8.104934

Source	SS	df	MS			
Model	584.256764	1	584.256764	Number of obs	=	770
Residual	948.791361	768	1.23540542	F(1, 768)	=	472.93
				Prob > F	=	0.0000
Total	1533.04812	769	1.99356063	R-squared	=	0.3811
				Adj R-squared	=	0.3803
				Root MSE	=	1.1115

l_inc	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
education	.2185987	.010052	21.75	0.000	.1988661	.2383313
_cons	7.710336	.0878119	87.81	0.000	7.537956	7.882716

“Distribution” of $\widehat{\beta}_1$ from 100 different SRFs

- Sample 0.5% from population (larger sample size *n*)
- Sample 0.05% from population (smaller sample size *n*)
  - distribution is wider in smaller samples
  - In Chapter 4: use distribution to assess the validity of our estimates



Variable	Obs	Mean	Std. Dev.	Min	Max
b_50	100	.2091172	.002918	.200028	.2170477
b_5	100	.2082802	.0088448	.1844666	.2316327

Deriving OLS Estimates

- We do not know population parameters **or** the distribution
- Need to find an mathematical estimators to approximate these from a sample
- Ordinary Least Squares Estimator
  - Carl Friedrich Gauss, University of Göttingen
    - An official partner to our Economics Department

  - Approach is to find the best fitting line that minimises the sum of squared residuals ($\sum_{i=1}^{N} \widehat{u}_i^2$)

Obtaining OLS estimates

- Take the following SRF

$$
\begin{aligned}
y_i &= \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} \cdots + \widehat{\beta}_k x_{ki} + \widehat{u}_i = \widehat{y}_i + \widehat{u}_i \\
\textbf{SAMP. resid.} &= \widehat{u}_i = y_i - \widehat{y}_i \\
&= y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} \cdots + \widehat{\beta}_k x_{ki}\right) \\
\textbf{POP. unobs.} &= u_i = y_i - \left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \cdots + \beta_k x_{ki} + \cdots + \beta_{(k+j)} x_{(k+j)i}\right)
\end{aligned}
$$

- **SAMPLE** residual not the same as **POPULATION** unobservable, unless can control for all x_j: $\widehat{u} \neq u$
- Minimise sum of squared **residuals** using optimisation techniques
- Get the fitted model to be as close to the data as possible

$$
min \sum_{i=1}^{n} \widehat{u}_i^2 = min \sum_{i=1}^{n} \left[\widehat{y}_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} \cdots + \widehat{\beta}_k x_{ki}\right)\right]^2
$$

- Minimisation with multivariate algebra in Appendix E and SunLearn

Express the OLS model in matrix and vector notation:

$$\boldsymbol{y} = X\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{u}} = \widehat{\beta}_0 + \widehat{\beta}_1 \boldsymbol{x}_1 + \cdots + \widehat{\beta}_k \boldsymbol{x}_k + \widehat{\boldsymbol{u}}$$

where $\underbrace{\boldsymbol{y}}_{n\times 1} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ is the dependent variable vector

$$\underbrace{X}_{n\times(k+1)} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

is the matrix of explanatory variables, the first col is to estimate the intercept,

$\underbrace{\widehat{\boldsymbol{\beta}}}_{(k+1)\times 1} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix}$ is the coefficient vector and $\underbrace{\widehat{\boldsymbol{u}}}_{n\times 1} = \begin{bmatrix} \widehat{u}_1 \\ \widehat{u}_2 \\ \vdots \\ \widehat{u}_n \end{bmatrix}$ is the residual vector

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT forward together sonke siya phambili saam vorentoe

$$\Rightarrow \widehat{\mathbf{u}} = \mathbf{y} - X\widehat{\boldsymbol{\beta}}$$

$$\widehat{\mathbf{u}}'\widehat{\mathbf{u}} = \left(\mathbf{y} - X\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - X\widehat{\boldsymbol{\beta}}\right) = \widehat{u}_1 \times \widehat{u}_1 + \widehat{u}_2 \times \widehat{u}_2 + \cdots \widehat{u}_n \times \widehat{u}_n = \sum_{i=1}^{n} \widehat{u}_i^2$$

$$= \underbrace{\mathbf{y}'\mathbf{y}}_{(1\times n)(n\times 1)} - \underbrace{\widehat{\boldsymbol{\beta}}'X'\mathbf{y}}_{(1\times k+1)(k+1\times n)(n\times 1)} - \underbrace{\mathbf{y}'X\widehat{\boldsymbol{\beta}}}_{(1\times n)(n\times k+1)(k+1\times 1)} + \underbrace{\widehat{\boldsymbol{\beta}}'X'X\widehat{\boldsymbol{\beta}}}_{(1\times k+1)(k+1\times n)(n\times k+1)(k+1\times 1)}$$

$$= \mathbf{y}'\mathbf{y} - 2\widehat{\boldsymbol{\beta}}'X'\mathbf{y} + \widehat{\boldsymbol{\beta}}'X'X\widehat{\boldsymbol{\beta}}$$

$$\frac{\partial \widehat{\mathbf{u}}'\widehat{\mathbf{u}}}{\partial \widehat{\boldsymbol{\beta}}'} = -2X'\mathbf{y} + 2X'X\widehat{\boldsymbol{\beta}} = 0$$

$$X'X\widehat{\boldsymbol{\beta}} = X'\mathbf{y}$$

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$$

IF $(X'X)$ is invertible: X has full column rank (no perfect linear relationships)

SIMPLE REGRESSION: $\widehat{\beta}_1 = \frac{Cov(y;x_1)}{Var(x_1)} \Rightarrow Var(x_1) \neq 0$

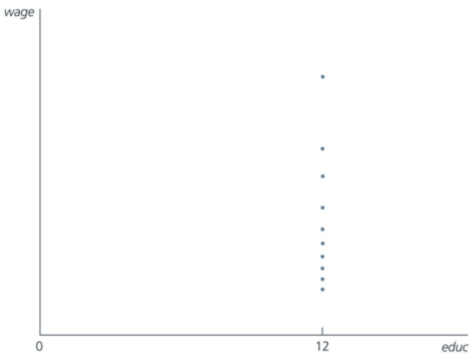
MULTIPLE REGRESSION: typical element of $\widehat{\boldsymbol{\beta}}$ is $\widehat{\beta}_j = \frac{Cov(y;\tilde{x}_j)}{Var(\tilde{x}_j)} \Rightarrow Var(x_j) \neq 0$, where $\tilde{x}_j$ is "partialled out" (later)

If $Var(x_j) = 0$

- No estimate if all values of x_j are the same (denominator of $\widehat{\beta}_1$)

FIGURE 2.3

A scatterplot of wage against education when $educ_i = 12$ for all i.



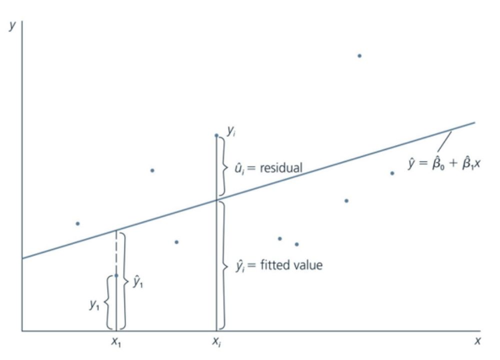
Fitted Values and Residuals

$$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i + \widehat{u}_i = \widehat{y}_i + \widehat{u}_i$$
$$\widehat{u}_i = y_i - \widehat{y}_i$$

NOTE: with the hat they are *predictions* and *residuals* (not the population error term)

FIGURE 2.4

Fitted values and residuals.



Properties of OLS on Any Sample of Data

- OLS is an estimat***or*** (a mathematical rule) that uses a sample to find estimat***es*** for $E(y|x)$ - not reaching the population estimate exactly
- OLS estimates differ for each sample used: How well does it perform on the **specific** sample available to researcher?
- $\widehat{salary}$ = regression line
- $\hat{u}$ = residuals
  - Negative: function overpredicts
  - Positive: function underpredicts

TABLE 2.2

Fitted Values and Residuals for the First 15 CEOs

obsno	*roe*	*salary*	*salaryhat*	*uhat*
1	14.1	1095	1224.058	−129.0581
2	10.9	1001	1164.854	−163.8542
3	23.5	1122	1397.969	−275.9692
4	5.9	578	1072.348	−494.3484
5	13.8	1368	1218.508	149.4923
6	20.0	1145	1333.215	−188.2151
7	16.4	1078	1266.611	−188.6108
8	16.3	1094	1264.761	−170.7606

Properties of OLS

- Algebraic
  - Residuals sum to zero or average to zero
    - By implication, the average of actual y values equals the average of fitted values

$$\sum_{i=1}^{n} \widehat{u}_i^2 = \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i^2 = 0$$

  - **Sample** covariance between **residuals** and variables is zero
    - Does not imply $Cov(u;x) = 0$ in **population**
    - $Cov(\widehat{u};x) = 0$ in **sample** does not that imply satisfying $E(u|x) = 0$ in the population
    - OLS estimation imposes this assumption on the sample; we get it “wrong” (ie we get bias) if it does not also hold in the population

$$Cov(\widehat{u};x_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}\widehat{u}_i = 0 \text{ for } j = 1\cdots k$$

  - $(\bar{\boldsymbol{x}};\bar{y}) = (\bar{x}_1, \bar{x}_2, \cdots, \bar{x}_k, \bar{y})$ is always on the regression line

Properties of OLS

- Total sum of squares (SST)
  - The total variation in y
- Explained sum of squares (SSE)
  - The variation in y explained by the model
- Residual sum of squares (SSR)
  - The variation in y that is not explained, and contained in residuals

$$SST = SSE + SSR$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\sigma_y^2 = \frac{SST}{n-1} = \text{variance of } y$$

Goodness of fit

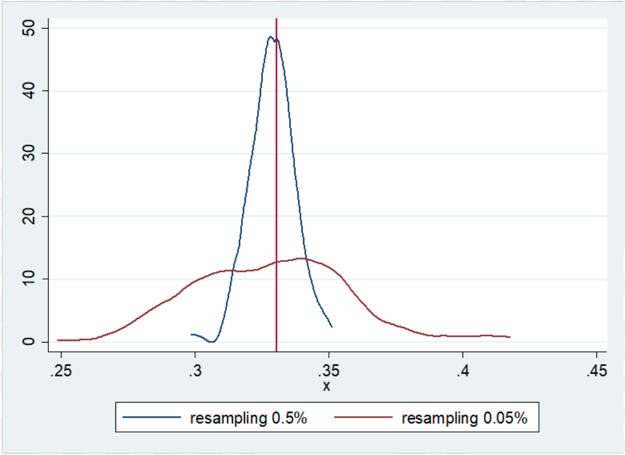
- Small residuals: model fits the specific **sample** data well
  - Small SSR means a “better” **sample** fit
  - Could get a different R^2 in a different sample
- R^2 is a measure of **sample** fit
  - Not how well the data fits the population
  - Not how well the model fits the population
  - Ratio of explained variance to total variance in **sample**

$$SST = SSE + SSR$$

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \quad where \quad 0 \leq R^2 \leq 1$$

  - Adding more variables: $SSR \downarrow \Rightarrow R^2 \uparrow$ **as soon as you add more (even irrelevant) variables to the model**
  - Also, the squared correlation coefficient between y and $\widehat{y}$
    - Intuitively, how related is the prediction from the model to the observed data

R^2 from our SRF experiment



- Small probability of drawing sample with low or high R^2

Goodness of fit

- We tend to obtain low R^2 in cross section analyses
- Does this mean we have a bad equation?
  - No, we just have a lot that is unexplained by the factor we have included in the model
  - We may still have the correct relationship between x and y if zero-conditional mean assumption holds.
- Be cautious to think a high R^2 means you have a good model
  - More later

"Partialling Out" interpretation of OLS

- Consider 2 variable case

$$y_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i} + \widehat{u}_i$$

- Suppose we have a second regression which removes the overlap between x_1 and x_2

$$x_{1i} = \widehat{\alpha}_0 + \widehat{\alpha}_1 x_{2i} + \widehat{r}_i$$

  - $Cov(\widehat{r}; x_2) = 0$ by properties of OLS - x_2 is "partialled out"
  - $\widehat{r}$ is a "new version" of x_1 that removes x_2
- In next slide we show that $\widehat{\beta}_1 = \frac{Cov(\widehat{r},y)}{Var(\widehat{r})}$ or the regression of r on y
- In other words: $\widehat{\beta}_1$ measures the effect of x_1 on y after removing their shared correlation with x_2
  - Holding x_2 constant, *ceteris paribus*

Partialling out

Vector notation, no $\widehat{\beta}_0$ for simplicity $\boldsymbol{y} = \widehat{\beta}_1 \boldsymbol{x}_1 + \widehat{\beta}_2 \boldsymbol{x}_2 + \widehat{\boldsymbol{u}}$ Stacking the explanatory vectors in columns gives $X = [\boldsymbol{x_1} \ \ \boldsymbol{x_2}]$

By matrix multiplication $X'X = \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1'\boldsymbol{x}_1 & \boldsymbol{x}_1'\boldsymbol{x}_2 \\ \boldsymbol{x}_2'\boldsymbol{x}_1 & \boldsymbol{x}_2'\boldsymbol{x}_2 \end{bmatrix}$

Recall that $X'X\widehat{\boldsymbol{\beta}} = X'\boldsymbol{y} \Rightarrow$ "stacked" version of the OLS equations:

$$\begin{bmatrix} \boldsymbol{x}_1'\boldsymbol{x}_1 & \boldsymbol{x}_1'\boldsymbol{x}_2 \\ \boldsymbol{x}_2'\boldsymbol{x}_1 & \boldsymbol{x}_2'\boldsymbol{x}_2 \end{bmatrix} \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1'\boldsymbol{y} \\ \boldsymbol{x}_2'\boldsymbol{y} \end{bmatrix}$$

$$\begin{aligned} \text{Write out first row as: } \boldsymbol{x}_1'\boldsymbol{x}_1\widehat{\beta}_1 + \boldsymbol{x}_1'\boldsymbol{x}_2\widehat{\beta}_2 &= \boldsymbol{x}_1'\boldsymbol{y} \\ \boldsymbol{x}_1'\boldsymbol{x}_1\widehat{\beta}_1 &= \boldsymbol{x}_1'\boldsymbol{y} - \boldsymbol{x}_1'\boldsymbol{x}_2\widehat{\beta}_2 \\ \widehat{\beta}_1 &= (\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1'\boldsymbol{y} - (\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1'\boldsymbol{x}_2\widehat{\beta}_2 \\ &= (\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1'\left(\boldsymbol{y} - \boldsymbol{x}_2\widehat{\beta}_2\right) \\ &= (\boldsymbol{x}_1'\boldsymbol{x}_1)^{-1}\boldsymbol{x}_1'\left(\widehat{\boldsymbol{r}}_1\right) \end{aligned}$$

To get $\widehat{\beta}_1$, run a **simple** OLS on "partialled out" x_1; similar for $\widehat{\beta}_2$

Illustration

- How can we see the "partial effect interpretation" of the coefficient on education?

```
. reg lwage educ exp

      Source |       SS           df       MS      Number of obs   =    23436
-------------+----------------------------------   F(  2, 23433)   =  4912.28
       Model |  8688.99642         2  4344.49821   Prob > F        =   0.0000
    Residual |  20724.4959     23433  .884414965   R-squared       =   0.2954
-------------+----------------------------------   Adj R-squared   =   0.2953
       Total |  29413.4923     23435  1.25510955   Root MSE        =   .94043

------------------------------------------------------------------------------
      lwage1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .184121   .0018693    98.50   0.000     .1804569     .187785
         exp |   .0262852    .000575    45.71   0.000     .0251581    .0274123
       _cons |   -.324087   .0274771   -11.79   0.000    -.3779439   -.2702301
------------------------------------------------------------------------------
```

Remove shared effect of `educ` and `exper`

- "Purify" the overlap from `educ` to get "educ only"

```
. reg educ exp if e(sample)==1

      Source |       SS       df       MS              Number of obs =   23436
-------------+------------------------------           F(  1, 23434) =10575.67
       Model |  114220.396     1  114220.396           Prob > F      =  0.0000
    Residual |   253094.23 23434  10.8003              R-squared     =  0.3110
-------------+------------------------------           Adj R-squared =  0.3109
       Total |  367314.626 23435  15.6737626           Root MSE      =  3.2864

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         exp |   -.171537    .001668  -102.84   0.000    -.1748064   -.1682676
       _cons |   13.03032    .044434   293.25   0.000     12.94322    13.11741
------------------------------------------------------------------------------

. predict r, res
(290 missing values generated)
```

- An aside: `if e(sample)==1` limits sample to same observations used in previous estimates
- An aside: after we have run estimates, we can store certain aspects of the model as variables with `predict`, in this case we create the variable `r` which is the `res`(iduals) from the regression

Use "purified" educ (residuals from previous equation)

```
. reg lwage r

      Source |       SS           df       MS      Number of obs   =    23436
-------------+----------------------------------   F(  1, 23434)   =  9651.03
       Model |  8580.02727         1  8580.02727   Prob > F        =   0.0000
    Residual |   20833.465    23434  .889027269   R-squared       =   0.2917
-------------+----------------------------------   Adj R-squared   =   0.2917
       Total |  29413.4923    23435  1.25510955   Root MSE        =   .94288

------------------------------------------------------------------------------
      lwage1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           r |    .184121   .0018742    98.24   0.000     .1804474    .1877945
       _cons |   1.951492   .0061591   316.85   0.000      1.93942    1.963565
------------------------------------------------------------------------------
```

The **simple** regression with the "purified" educ, gives us almost identical estimates to the **multiple** regression that included both `educ` and `exper`

Comparison: Simple & Multiple Regression

$$\text{Simple Regression: } \tilde{y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{1i}$$

$$\text{Multiple Regression: } \widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{1i} + \widehat{\beta}_2 x_{2i}$$

$$\text{Can be compared by: } \tilde{\beta}_1 = \widehat{\beta}_1 + \widehat{\beta}_2 \widehat{\delta}$$

$$\text{where } \widehat{\delta} \text{ is the coefficient of regressing } x_2 \text{ on } x_1$$

- Multiple regression simplifies to simple regression only if
  - $Cov(x_2; y) = 0$ or $\widehat{\beta}_2 = 0$
  - $Cov(x_2; x_1) = 0$ or $\widehat{\delta} = 0$
- We will use this formula to argue about bias in estimating models that have omitted variables

Illustration: simple vs multiple

```
reg lwage educ

      Source |       SS       df       MS              Number of obs =   23436
-------------+------------------------------           F(  1, 23434) = 7102.12
       Model |  6841.02182     1  6841.02182           Prob > F      =  0.0000
    Residual |  22572.4705 23434  .963235916           R-squared     =  0.2326
-------------+------------------------------           Adj R-squared =  0.2325
       Total |  29413.4923 23435  1.25510955           Root MSE      =  .98145

------------------------------------------------------------------------------
      lwage1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1364713   .0016194    84.27   0.000     .1332972    .1396454
       _cons |   .7192268   .0159658    45.05   0.000     .6879327    .7505208
------------------------------------------------------------------------------

reg lwage educ exp

      Source |       SS       df       MS              Number of obs =   23436
-------------+------------------------------           F(  2, 23433) = 4912.28
       Model |  8688.99642     2  4344.49821           Prob > F      =  0.0000
    Residual |  20724.4959 23433  .884414965           R-squared     =  0.2954
-------------+------------------------------           Adj R-squared =  0.2953
       Total |  29413.4923 23435  1.25510955           Root MSE      =  .94043

------------------------------------------------------------------------------
      lwage1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |    .184121   .0018693    98.50   0.000     .1804569     .187785
         exp |   .0262852    .000575    45.71   0.000     .0251581    .0274123
       _cons |   -.324087   .0274771   -11.79   0.000    -.3779439   -.2702301
------------------------------------------------------------------------------
```

Find $\widehat{\delta}$ and put it all together...

```
. reg exp educ if e(sample)==1

      Source |       SS       df       MS              Number of obs =   23436
-------------+------------------------------           F(  1, 23434) =10575.67
       Model |  1207072.88     1  1207072.88           Prob > F      =  0.0000
    Residual |  2674681.49 23434  114.136788           R-squared     =  0.3110
-------------+------------------------------           Adj R-squared =  0.3109
       Total |  3881754.37 23435  165.639188           Root MSE      =  10.683

------------------------------------------------------------------------------
         exp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |  -1.812791   .0176276  -102.84   0.000    -1.847342   -1.778239
       _cons |     39.692    .173795   228.38   0.000     39.35135    40.03265
------------------------------------------------------------------------------
```

$$\tilde{\beta}_1 = \widehat{\beta}_1 + \widehat{\beta}_2 \widehat{\delta}$$
$$0.1364713 = 0.184121 + 0.0262852 \times -1.812791$$

- Note differences due to rounding
- What does this tell us?

Expected Values & Variances of $\widehat{\beta}_{OLS}$

Up to now: used a "formula" to find out a relationship between x and y

- But the result depends on the **one** sample drawn from many possible samples that make up the population
- Different estimates of $\widehat{\beta}$, depending on our sample
  - OLS estimates are therefore also random variables with a distribution
    - Which have both expected values and a variances
  - Objective:
    - Show under which circumstances OLS is unbiased and efficient at estimating (unknown) population model
    - For this we need assumptions
    - SLR 1-4 (Simple Linear Regression)
    - MLR 1-4 (Multiple Linear Regression)

- **SLR1/MLR1** – Linearity in all $k+1$ *parameters*
  - Linear relationship between (perhaps non-linearly transformed) variables ($\widehat{\beta}$ to the power 1)
  - Must assume a population model: $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u$
  - If the PRF were non-linear in *parameters*, OLS is not the right estimator
- **SLR2/MLR2** – Random sampling (*"the row problem"*)
  - A **random** sample from the population for these random variables $\{(x_{ij}; y_i) : i = 1, 2, \cdots n \text{ and } j = 1, \cdots, k\}$
  - Sample size = n; number of variables = k
  - PRF "holds" for **each** unit in the sample $\Rightarrow$ add a sub-script: $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + u_i$ for $i = 1, \cdots, n$

OLS assumptions

- **SLR3** – Sample variation of explanatory variable
  - Any explanatory variable (x_j) may not be the same value for all observations (i)
  - Otherwise impossible to compute OLS estimate $\widehat{\beta} = (X'X)^{-1}X'\boldsymbol{y}$
  - $Var(x) \neq 0 \Leftrightarrow (X'X)$ cannot be inverted
  - NO INFORMATION in variable to distinguish between units of analysis
- **MLR3** - No **perfect** multicollinearity - cannot estimate if this fails
  - No **exact** linear relationship among independent variables $\Leftrightarrow (X'X)$ cannot be inverted
    - Eg including expenditure in Rands and expenditure in Dollars in same model
    - Eg including expenditure A, expenditure B and total expenditure (A+B)
    - NO NEW INFORMATION by adding a variable
  - Column vector (**1** in X) to estimate $\widehat{\beta}_0$ is constant, so that $Var(x_j) \neq 0$
  - Need more observations than regressors
    - Can you draw a unique straight line through one datapoint?

An example of perfect multicollinearity

- One variable can be expressed as an exact linear combination of other variables in the model
- $Potential\ Experience = Age - Education - 6$
  by Mincer's (1974) definition and a possible PRF:

$$\begin{aligned} log(wage) &= \beta_0 + \beta_1 Exper + \beta_2 Educ + \beta_3 Age + u \\ &= \beta_0 + \beta_1 (Age - Educ - 6) + \beta_2 Educ + \beta_3 Age + u \\ &= (\beta_0 - 6\beta_1) + (\beta_1 + \beta_3) Age + (\beta_2 - \beta_1) Educ + u \\ &= \alpha_1 + \alpha_2 Age + \alpha_3 Educ + u \end{aligned}$$

- Possible to estimate α_j, but impossible to find unique solutions for β_j

Perfect Multicollinearity

```
. reg lwage educ exp age
note: educ omitted because of collinearity

      Source |       SS           df       MS      Number of obs   =    23,436
-------------+----------------------------------   F(2, 23433)     =   4912.23
       Model |  8688.92651         2  4344.46325   Prob > F        =    0.0000
    Residual |  20724.5658    23,433  .884417948   R-squared       =    0.2954
-------------+----------------------------------   Adj R-squared   =    0.2953
       Total |  29413.4923    23,435  1.25510955   Root MSE        =    .94043

------------------------------------------------------------------------------
      lwage1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |          0  (omitted)
         exp |  -.1578357   .0016206   -97.40   0.000    -.1610121   -.1546593
         age |   .1841194   .0018693    98.49   0.000     .1804553    .1877834
       _cons |  -1.428745   .0377775   -37.82   0.000    -1.502791   -1.354699
------------------------------------------------------------------------------
```

- In technical terms, this is the same as saying that $(X'X)$ cannot be inverted.
    - X not of **full column rank**
    - Cannot calculate $\widehat{\beta} = (X'X)^{-1}X'\mathbf{y}$ unless we drop a variable
- `STATA` simply drops a variable of its choice to "make it work": no need to test for perfect multicollinearity

Multicollinearity

- Are *exper* and *exper*^{2} perfectly multicollinear?
  - No!
  - Multico**linear**ity implies perfect linear relationships
  - These variables are perfectly non-linearly correlated
- This has nothing to do with the error term

OLS assumptions

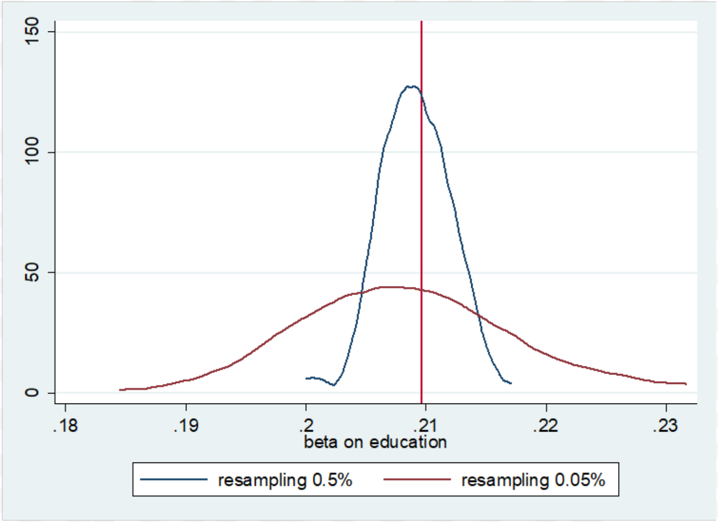
- **SLR4/MLR4** – Zero Conditional Mean (*"the column problem"*)
  - $E(u|\boldsymbol{x}) = 0$
  - Implies independence of u and x, as before

Does OLS have causal interpretation?

Is OLS unbiased/causal? ($E(\widehat{\beta}_j | x_1, x_2, \cdots, x_k) = \beta_j$ for all $j = 1, \cdots, k$)

- Yes! **IF ALL THE ASSUMPTIONS HOLD!**
  - SLR1: if your PRF is linear, OLS is a good way of estimating it $\rightarrow$ if PRF is non-linear one obviously cannot fit straight lines through data
    - Could introduce non-linear variables
    - Or would have to move to non-linear estimators, which do not fit straight lines
  - SLR2: random sampling solves the "row problem"
  - SLR3: you **cannot** estimate OLS without variation
  - SLR4: zero conditional mean solves the "column problem"
  - The *estimator* is unbiased
    - *Specific estimates* may not exactly reflect the population, if we use a sample that produces $\widehat{\beta}_1$ that is in the tail of the population distribution of *all possible estimates*
    - But the average of ALL possible estimates using a representative sample will be the true population value under the assumptions

Distribution: $\widehat{\beta}$ 100 SRFs same population



Variable	Obs	Mean	Std. Dev.	Min	Max
b_50	100	.2091172	.002918	.200028	.2170477
b_5	100	.2082802	.0088448	.1844666	.2316327

Theorem E.1

Show that MLR4 gives unbiased/causal estimates of the population β

1. Substitute PRF into OLS estimator
2. Take conditional expectations

$$
\begin{aligned}
\widehat{\beta} &= (X'X)^{-1}X'\mathbf{y} \\
&= (X'X)^{-1}X'(X\beta + \mathbf{u}) \\
&= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\mathbf{u} \\
&= \beta + (X'X)^{-1}X'\mathbf{u} \\
\Rightarrow E(\widehat{\beta}|X) &= \beta + (X'X)^{-1}X'E(\mathbf{u}|X) \\
&= \beta \text{ if and only if } E(\mathbf{u}|X) = \mathbf{0}
\end{aligned}
$$

Omitted Variable Bias: a simple case

Stellenbosch University / Iyunivesithi / Universiteit — forward together / sonke siya phambili / saam vorentoe

$$PRF : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$SRF : y = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{u}$$

- Population model includes x_2 ($\beta_2 \neq 0$), but when omitted (perhaps because there is no data), SRF restricted to $\widehat{\beta}_2 = 0$ in sample
- Violation of MLR4 - biased estimate of β_1 - **how large is the bias?**
  - Use what we know about relationship between simple and multiple regression

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

where δ is the regression coefficient of x_2 on x_1

$$Estimate ='' Truth'' + bias$$

Direction of Bias

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

- **UPWARD BIAS**: $\beta_2\delta > 0$

$$\left.\begin{array}{l} \beta_2 > 0; \delta > 0 \\ \beta_2 < 0; \delta < 0 \end{array}\right\} \quad \beta_1 > 0 : \widehat{\beta}_1 \text{“too positive”} \quad \beta_1 < 0 : \widehat{\beta}_1 \text{“not as negative”}$$

- **DOWNWARD BIAS**: $\beta_2\delta < 0$

$$\left.\begin{array}{l} \beta_2 > 0; \delta < 0 \\ \beta_2 < 0; \delta > 0 \end{array}\right\} \quad \beta_1 > 0 : \widehat{\beta}_1 \text{“not as positive”} \quad \beta_1 < 0 : \widehat{\beta}_1 \text{“too negative”}$$

TABLE 3.2

Summary of Bias in $\tilde{\beta}_1$ when x_2 Is Omitted in Estimating Equation (3.40)

	$\text{Corr}(x_1,x_2) > 0$	$\text{Corr}(x_1,x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Omitted Variable Bias : examples

$$PRF : log(wage) = \beta_0 + \beta_1 education + \beta_2 ability + u$$
$$SRF : log(wage) = \widehat{\beta}_0 + \widehat{\beta}_1 education + \widehat{u}$$

A classical example from the literature

- "Ability bias" in estimating $\beta_1 > 0$
- What direction is the bias likely to take?
  - How are education and "ability" **likely** to be correlated? ($\delta > 0$)
  - How are wages and "ability" **likely** to be correlated? ($\beta_2 > 0$)
    - NOTE: this is a theoretical argument, because we do not observe "ability" and we argue about unobserved population relationships

$\beta_1 > 0$ and $\beta_2\delta > 0 \Rightarrow$ effect of educ "too positive" if "ability" omitted

Omitted Variable Bias : examples

$$PRF : crime = \beta_0 + \beta_1 expenditure + \beta_2 past\ crime + u$$
$$SRF : crime = \widehat{\beta}_0 + \widehat{\beta}_1 expenditure + \widehat{u}$$

Does expenditure on policing reduce crime?

- What direction is the bias likely to take?
  - How are expenditure and past crime **likely** to be correlated? ($\delta > 0$)
  - How are current crime and past crime **likely** to be correlated? ($\beta_2 > 0$)

$\beta_2 \delta > 0$ and $\beta_1 < 0 \Rightarrow$ effect of expenditure is "not as negative" if we omit "past crime"

Including irrelevant variables in a model

$$PRF : y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$
$$SRF : y = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \widehat{u}$$

- **Omitting** variables results in bias ("missing column"); does adding **too many variables** have similar effect?
  - Short answer: **no effect** on **bias**; but risk of increasing standard errors
- x_3 is not part of PRF (ie $\beta_3 = 0$ in population)
  - $\widehat{\beta}_3$ will average to zero across all random samples
    - But it is possible that we draw a sample where it is large and significant
  - Overspecification is not serious for bias of $\widehat{\beta}_1$ and $\widehat{\beta}_2$
  - Variance: will discuss later

Variance of OLS estimators

- Want to know "how far" estimates are from population value on average
  - Variance of the estimator
  - Standard error of the estimator
    - Remember the estimator is also a random variable
  - **BUT** we don't observe the variation
    - In real life: only observe one estimate from one sample
  - Can be calculated under assumptions MLR1-MLR4, but need to add another assumption to simplify the calculation

Variance of OLS estimators

Add assumption **SLR5/MLR5**: Homoskedasticity

- u has same variance given **any** values of **all** explanatory variables
  - But also constant variance of y across different values of $\boldsymbol{x}$

$$Var(u|\boldsymbol{x}) = Var(y|\boldsymbol{x}) = \sigma^2$$

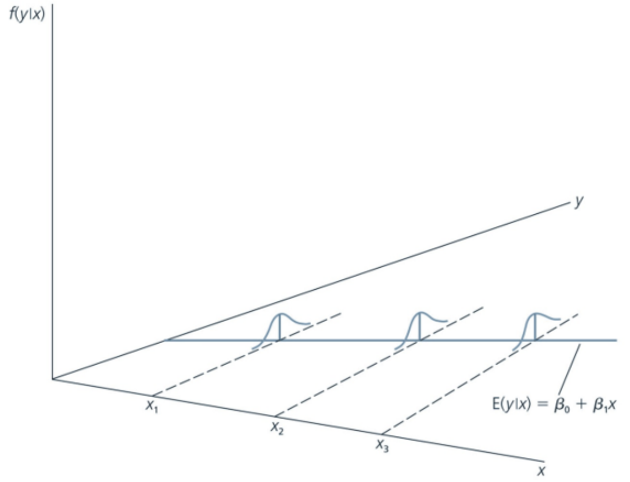
- Allows us to calculate standard errors for $\widehat{\beta}$ simply **and efficiently**, **even if we do not observe the distribution** of $\widehat{\beta}$
- The assumption is NOT the same as $E(u|X) = 0$
- MLR5 can easily be violated
  - Eg at high education you have wider interests and greater variation in wages
    - Low levels generally constant (low) wages

Homoskedastic errors

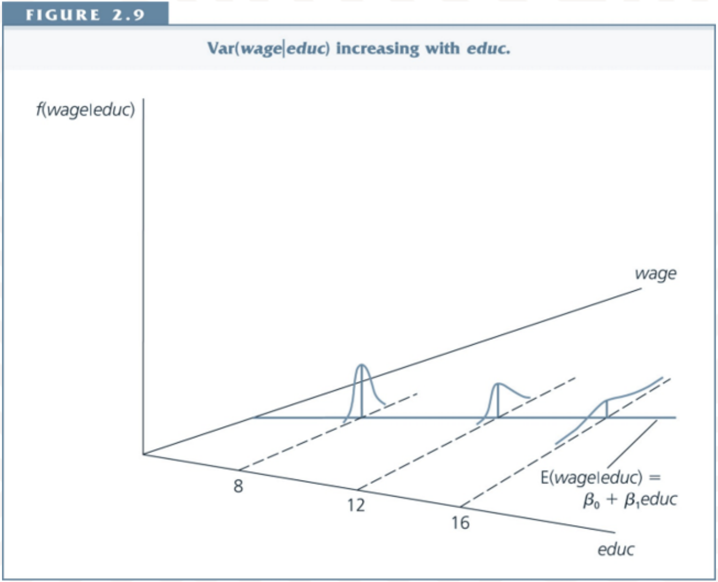
Stellenbosch University Iyunivesithi Universiteit — forward together, sonke siya phambili, saam vorentoe

FIGURE 2.8

The simple regression model under homoskedasticity.



Heteroskedastic errors



Theorem E.2

Homoskedasticity in matrix form.

- Diagonals: same variance for each observation; off-diags: no autocorrelation

$$Var(\mathbf{u}|X) = \begin{bmatrix} \sigma^2 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & \sigma^2 \end{bmatrix} = I_n\sigma^2$$

$$\text{Then: } \widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (X'X)^{-1}X'\mathbf{u}$$

$$\begin{aligned} Var\left(\widehat{\boldsymbol{\beta}}|X\right) &= Var\left(\boldsymbol{\beta} + (X'X)^{-1}X'X|\mathbf{u}\right) \\ &= Var\left((X'X)^{-1}X'\mathbf{u}|X\right) \text{ because } \boldsymbol{\beta} \text{ is not random} \\ &= (X'X)^{-1}X'Var(\mathbf{u}|X)X(X'X)^{-1} \\ &= (X'X)^{-1}X'I_n\sigma^2X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

Estimating the error variance

- We do not know σ^2 because it is the variance of **population** errors u, which we do not observe
- However, an unbiased estimator for σ^2 comes from *sample* residuals $SSR = \sum_{i=1}^{n} \widehat{u}_i^2$

$$\widehat{\sigma}^2 = s^2 = \frac{SSR}{n-(k+1)}$$

  - Standard error of regression (square root of estimated variance)
  - Also estimates the standard error of y once effect of x is removed

Standard Deviation vs Standard Error

- Standard deviation: if we knew σ^2 estimated from u
- Standard error is an estimate of the standard deviation ($\widehat{\sigma}^2$ estimated from residuals $\widehat{u}$)
  - Because we do not have population errors
  - It is therefore in itself a random variable, because it differs by sample

Gauss-Markov Assumptions

- MLR1-MLR5 are the **Gauss-Markov assumptions** for cross section data with random sampling
  - Change slightly for time series data
- All G-M assumptions are required to get OLS standard errors
  - **MLR1-MLR4**: to establish whether $\widehat{\beta}_j$ is biased or not
  - **MLR1-4 plus MLR5** is required for variance calculations

Under **MLR5**:

$Var\left(\widehat{\beta}|X\right) = \sigma^2 \left(X'X\right)^{-1}$ with diagonal elements $Var(\widehat{\beta}_j) = \frac{\sigma^2}{SST_{x_j}\left(1-R_j^2\right)}$

where $SST_{x_j} = \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2$ is the variation in x_j
and R_j^2 is the fit of the regression of x_j on all other covariates
Summarised in $(X'X)^{-1}$

The components of OLS Variances

Under **MLR5**:

$$Var\left(\widehat{\beta}|X\right) = \sigma^2\left(X'X\right)^{-1} \text{ with diagonal elements } Var(\widehat{\beta}_j) = \frac{\sigma^2}{SST_{x_j}\left(1-R_j^2\right)}$$

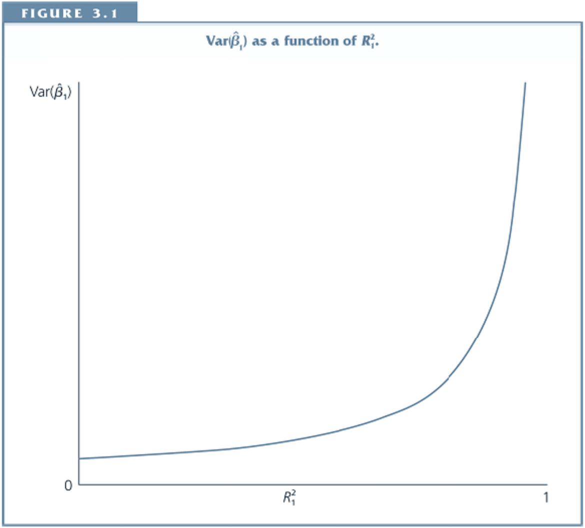
3 changes determine whether OLS estimates are more/less efficient when adding/dropping a variable

- $\uparrow \sigma^2 = \frac{SSR}{n-k-1} \Rightarrow \uparrow Var\left(\widehat{\beta}|X\right)$
  - Cannot reduce SSR by $\uparrow n$, but can do so by $\uparrow k$ (number of variables)
- $\uparrow SST_{x_j} \Rightarrow \downarrow Var\left(\widehat{\beta}|X\right)$
  - Non-experimental analysis: cannot "introduce" variation in x_j, unless $\uparrow n$
- **IMPERFECT** multicollinearity $\uparrow R_j^2 \Rightarrow \downarrow (1 - R_j^2) \Rightarrow \uparrow Var\left(\widehat{\beta}|X\right)$
  - $(X'X)^{-1}$ captures both the variation within each x_j (which is SST_j) and the variation between the explanatory variables (R_j^2)

OLS Variance: Imperfect Multicollinearity

- The strength of the linear relationship among the independent variables (R_j^2)
  - R_j^2 is the R^2 of $x_j = \widehat{\alpha}_0 + \widehat{\alpha}_1 x_1 + \cdots + \widehat{\alpha}_{j-1} x_{j-1} + \widehat{\alpha}_{j+1} x_{j+1} + \cdots \widehat{\alpha}_k x_k + \widehat{u}$
  - If $R_j^2 \rightarrow 1$ (=**perfect** multicollinearity), $Var(\widehat{\beta}_j) \rightarrow \infty$
    - Same as not being able to estimate the coefficient at all (**MLR3** fails)
  - When R_j^2 moves close to 1 (but $R_j^2 \neq 1$), large $Var(\widehat{\beta}_j)$, but does not violate the perfect multicollinearity assumption
  - Strong interrelationships between x's make it difficult to distinguish which of the variables is "doing the work" in explaining y
    - The uncertainty is reflected in higher standard errors

Multicollinearity and variances of estimates



Solving multicollinearity?

- Drop variables?
  - But omitted variable bias is the trade-off!
- Collect more data?
  - Higher n increases variation in x, and can reduce correlation between x's
- Detection:
  - $VIF = \frac{1}{1-R_j^2} > 10$ is "too high" - rule of thumb, but an "arbitrary threshold"
- If one variable is not highly correlated with other controls
  - It's variance remains unaffected (low R_j^2)

Variances in Misspecified Models

Trade-off between bias and variance

- If population model contains many collinear variables:
  - Include all variables to avoid omitted variable bias
    - Cannot solve this by increasing n
  - But at the cost of high variance
    - **Can** solve this by increasing n ($\uparrow SST_{x_j}$; $\downarrow \sigma^2$)
- Ideally: have a large sample size to mitigate against collinearity and specify all variables in the PRF in the sample model

Example

Use the famous `auto.dta` dataset on car prices in `STATA`
Suppose for some reason the following PRF is important for a research question:

$$ln(price) = \beta_0 + \beta_1 length + \beta_2 weight + \beta_3 foreign + u$$

```
. correl ln_price length_m weight_k foreign
(obs=74)

             | ln_price length~s weight~g  foreign
-------------+------------------------------------
    ln_price |   1.0000
length_met~s |   0.4589   1.0000
   weight_kg |   0.5405   0.9460   1.0000
     foreign |   0.0870  -0.5702  -0.5928   1.0000
```

- Length and weight are strongly correlated with each other, and also with price
- Foreign is weakly correlated with price, but strongly negatively related to length and weight

Simple regressions

```
. reg ln_price length_m

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     19.21
       Model |  2.3640604         1   2.3640604   Prob > F        =    0.0000
    Residual |  8.85947268        72  .123048232   R-squared       =    0.2106
-------------+----------------------------------   Adj R-squared   =    0.1997
       Total |  11.2235331        73  .153747029   Root MSE        =    .35078

------------------------------------------------------------------------------
    ln_price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
length_metres |   2.020501   .4609645     4.38   0.000     1.101585    2.939417
       _cons |   7.121762   .3489118    20.41   0.000     6.426219    7.817305
------------------------------------------------------------------------------

. reg ln_price weight_kg

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =     29.71
       Model |  3.27831499         1  3.27831499   Prob > F        =    0.0000
    Residual |  7.94521809        72  .110350251   R-squared       =    0.2921
-------------+----------------------------------   Adj R-squared   =    0.2823
       Total |  11.2235331        73  .153747029   Root MSE        =    .33219

------------------------------------------------------------------------------
    ln_price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   weight_kg |   .0005453   .0001001     5.45   0.000     .0003459    .0007448
       _cons |   7.817322   .1559096    50.14   0.000     7.506521    8.128122
------------------------------------------------------------------------------

. reg ln_price foreign

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =      0.55
       Model |  .085003065         1  .085003065   Prob > F        =    0.4609
    Residual |    11.13853        72  .154701806   R-squared       =    0.0076
-------------+----------------------------------   Adj R-squared   =   -0.0062
       Total |  11.2235331        73  .153747029   Root MSE        =    .39332

------------------------------------------------------------------------------
    ln_price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   .0741515   .1000347     0.74   0.461    -.1252639     .273567
       _cons |   8.618587   .0545439   158.01   0.000     8.509856    8.727319
------------------------------------------------------------------------------
```

Example

	(1)	(2)	(3)	(4)	(5)	(6)
	ln_price	ln_price	ln_price	ln_price	ln_price	ln_price
length_met~s	2.02050***				3.31760***	-1.94830
	(0.46096)				(0.49639)	(1.06693)
weight_kg		0.00055***		0.00092***		0.00134***
		(0.00010)		(0.00010)		(0.00025)
foreign			0.07415	0.53527***	0.44027***	0.52982***
			(0.10003)	(0.08441)	(0.09607)	(0.08311)
_cons	7.12176***	7.81732***	8.61859***	7.09086***	6.01581***	7.92509***
	(0.34891)	(0.15591)	(0.05454)	(0.16989)	(0.39181)	(0.48647)
r2	0.21063	0.29209	0.00757	0.54804	0.39082	0.56859
N	74	74	74	74	74	74
ssr	8.85947	7.94522	11.13853	5.07258	6.83712	4.84193

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

- (1), (2) and (3) confirm correlations, but notably $se(\widehat{\beta}_{foreign}) > \widehat{\beta}_{foreign}$ (noise > signal)
- (1) and (5): $SSR \downarrow$, SST_{length} and $SST_{foreign}$ unchanged
  - but $se(\widehat{\beta}_{length}) \uparrow$ because of strong collinearity with `foreign`
  - and $se(\widehat{\beta}_{foreign}) \downarrow$ so that effect of SSR dominates collinearity with `length`

Example

	(1) ln_price	(2) ln_price	(3) ln_price	(4) ln_price	(5) ln_price	(6) ln_price
length_met~s	2.02050***				3.31760***	-1.94830
	(0.46096)				(0.49639)	(1.06693)
weight_kg		0.00055***		0.00092***		0.00134***
		(0.00010)		(0.00010)		(0.00025)
foreign			0.07415	0.53527***	0.44027***	0.52982***
			(0.10003)	(0.08441)	(0.09607)	(0.08311)
_cons	7.12176***	7.81732***	8.61859***	7.09086***	6.01581***	7.92509***
	(0.34891)	(0.15591)	(0.05454)	(0.16989)	(0.39181)	(0.48647)
r2	0.21063	0.29209	0.00757	0.54804	0.39082	0.56859
N	74	74	74	74	74	74
ssr	8.85947	7.94522	11.13853	5.07258	6.83712	4.84193

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

- (2) and (4): $SSR \downarrow$, SST_{weight} and $SST_{foreign}$ unchanged
  - similar to before
- (2) and (4): $\beta_{foreign} > 0$, $\delta_{foreign;weight} < 0$, so that simpler regression was downward biased
  - Controlling for foreign $\uparrow \widehat{\beta}_{weight}$

Example

	(1) ln_price	(2) ln_price	(3) ln_price	(4) ln_price	(5) ln_price	(6) ln_price
length_met~s	2.02050***				3.31760***	-1.94830
	(0.46096)				(0.49639)	(1.06693)
weight_kg		0.00055***		0.00092***		0.00134***
		(0.00010)		(0.00010)		(0.00025)
foreign			0.07415	0.53527***	0.44027***	0.52982***
			(0.10003)	(0.08441)	(0.09607)	(0.08311)
_cons	7.12176***	7.81732***	8.61859***	7.09086***	6.01581***	7.92509***
	(0.34891)	(0.15591)	(0.05454)	(0.16989)	(0.39181)	(0.48647)
r2	0.21063	0.29209	0.00757	0.54804	0.39082	0.56859
N	74	74	74	74	74	74
ssr	8.85947	7.94522	11.13853	5.07258	6.83712	4.84193

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

- (5) and (6): $SSR \downarrow$; SST_{length}, SST_{weight} and $SST_{foreign}$ unchanged
  - But the very high collinearity between **`weight`** and **`length`** make the latter standard error grow very large
- (5) and (6): $\beta_{length} > 0$, $\delta_{foreign;length} > 0$, so that simpler regression was perhaps upward biased
  - Controlling for **`foreign`** $\downarrow \widehat{\beta}_{weight}$: it a large negative value
  - But does it make sense? (It is not statistically significant - next chapter)

VIFs

```
. estat vif
```

Variable	VIF	1/VIF
weight_kg	9.92	0.100839
length_met~s	9.53	0.104932
foreign	1.54	0.647716
Mean VIF	7.00	

- In the final regression we detect high levels of multicollinearity
- What if weight **and** length matter in the PRF, but we cannot distinguish their effects in a small sample of $n = 74$ with high collinearity?

QUESTION

What would happen if we added a variable that was not correlated to any other x's?

1. To coefficients?
2. To standard errors?

Gauss-Markov Theorem

Stellenbosch UNIVERSITY IYUNIVESITHI UNIVERSITEIT
forward together sonke siya phambili saam vorentoe

- Why use OLS? - it is unbiased under **MLR1-4**
  - But there are other unbiased linear estimators for β
- OLS is BLUE - *Best Linear Unbiased Estimator*
  - "Best" - it has the smallest variance (most efficient) if we assume **MLR5**
- Gauss-Markov Theorem
  - Among all linear unbiased estimators, the OLS estimator has smallest variance - given that MLR1-MLR5 hold
- Homoskedasticity got us "best"
  - Heteroskedasticity doesn't affect bias of coefficients, but biases the standard errors that we calculated because we do not observe all samples
  - We no longer have the "best" estimator if MLR 5 fails