Econometrics Lecture 2

EC2METRIE

Dr. Anna Salomons

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

21 November 2016

This class

- **Bivariate regression** (simple regression) analysis: regression with 1 explanatory variable
- **Multivariate regression** (multiple regression) analysis: regression with >1 explanatory variable
- **Inference** about the population

Overview

Population: purely theoretical

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

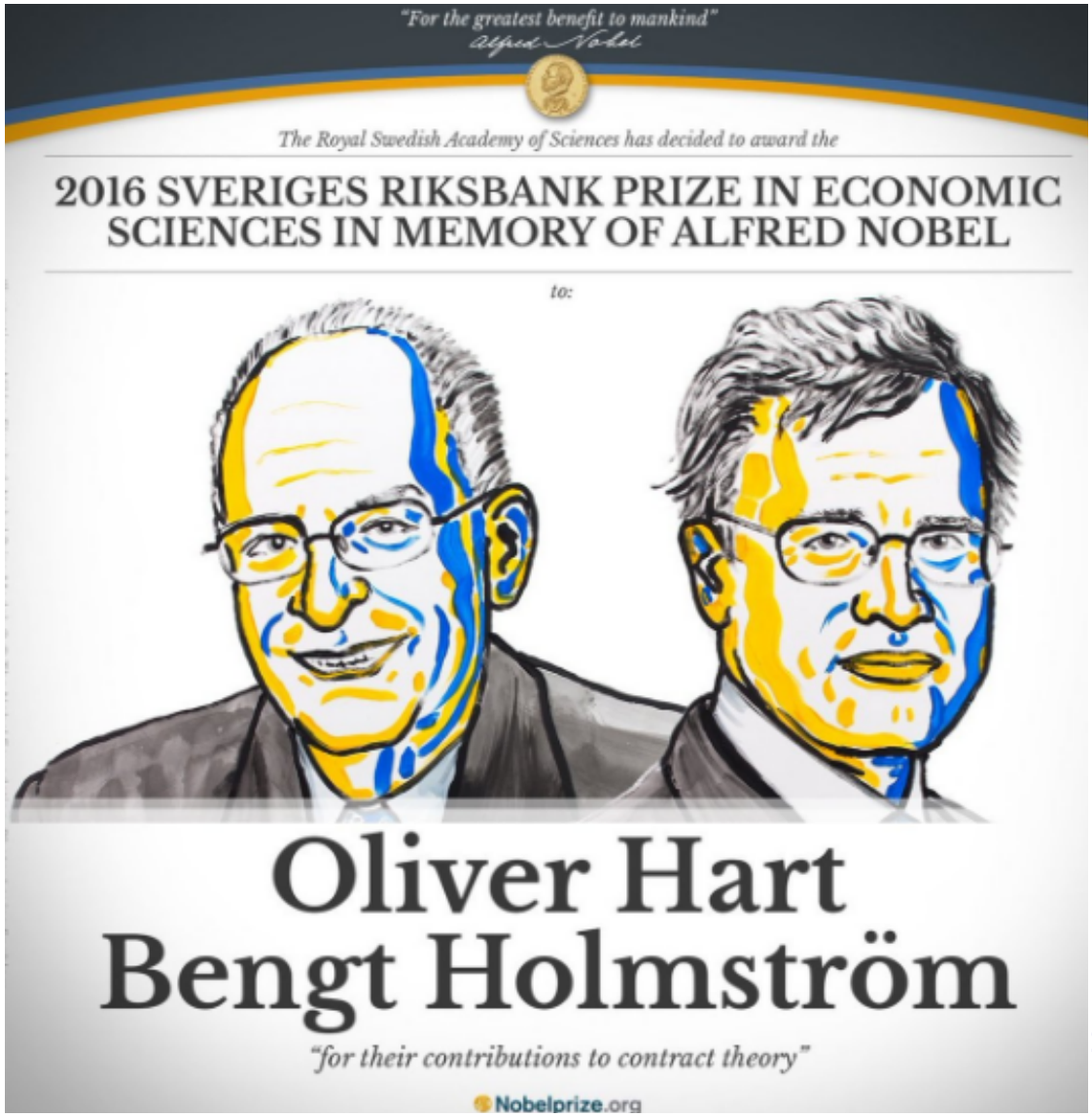
Sample: empirical counterpart

$$Y_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + e_i \tag{2}$$

Estimation = how to estimate equation 2Bivariate regression analysis in a sample? (Short answer: using the OLS estimator)

Inference = how to use equation 2Bivariate regression analysis to say something about equation 1Bivariate regression analysis? (Short answer: using the concept of a sampling distribution)

Contract theory



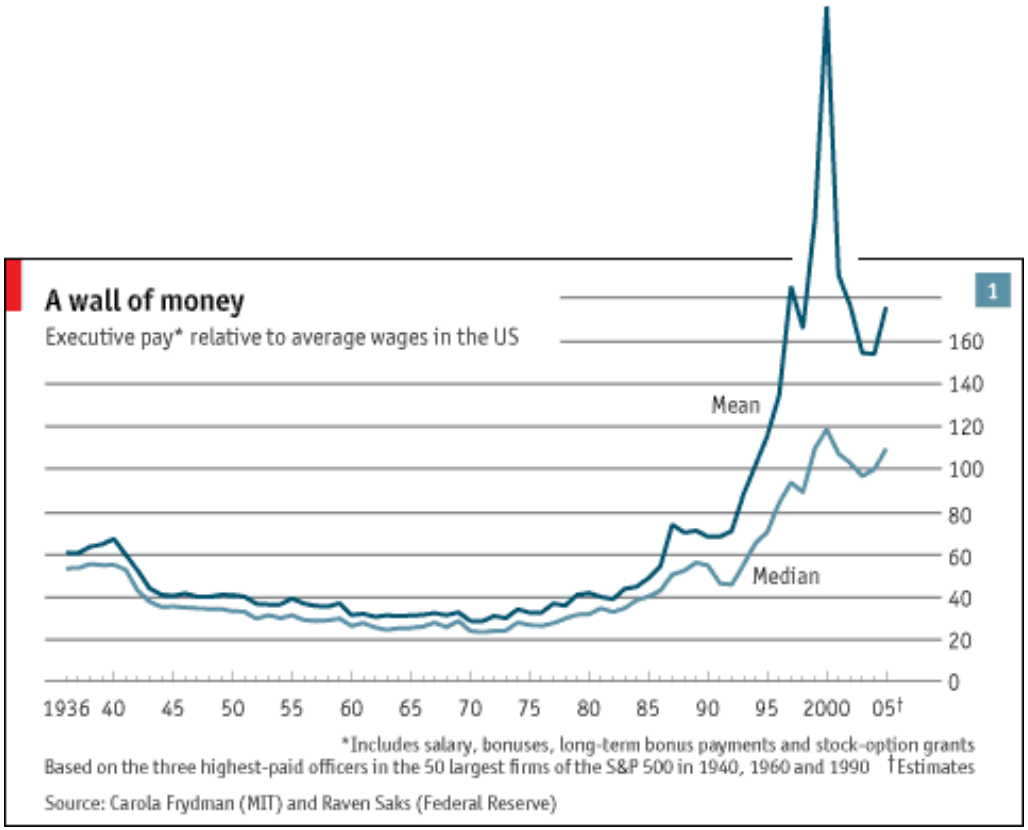
Top income inequality: a hot research topic

Income inequality

- Continental **European countries have less income inequality than the United States**: e.g. the median wage and the wage at the 90th percentile are much closer together here than in the US.
- Since the 1980s, there has been a tendency for inequality to increase across many advanced countries.
- Part of this increase that has garnered a lot of attention is the **rapid rise of incomes at the very top of the distribution**.
- Some of this can be accounted for by **skyrocketing pay for top managers**[1]

[1] Here's an interesting talk on this topic by Harvard professor Richard Freeman:
http://www.lse.ac.uk/newsAndMedia/videoAndAudio/channels/publicLecturesAndEvents/player.aspx?id=1457

Example: CEO pay



The dark blue line shows average CEO pay divided by average US wages: this shows the average CEO earned about 30 times as much as the average worker in 1965, and over 150 times as much in 2005 (and even 320 times at its peak).

Business as usual after the Great Recession

Table 1. Real Income Growth by Groups

	Average Income Real Growth (1)	Top 1% Incomes Real Growth (2)	Bottom 99% Incomes Real Growth (3)	Fraction of total growth (or loss) captured by top 1% (4)
Full period 1993-2012	**17.9%**	**86.1%**	**6.6%**	**68%**
Clinton Expansion 1993-2000	31.5%	98.7%	20.3%	45%
2001 Recession 2000-2002	-11.7%	-30.8%	-6.5%	57%
Bush Expansion 2002-2007	16.1%	61.8%	6.8%	65%
Great Recession 2007-2009	-17.4%	-36.3%	-11.6%	49%
Recovery 2009-2012	6.0%	31.4%	0.4%	95%

Computations based on family market income including realized capital gains (before individual taxes).
Incomes exclude government transfers (such as unemployment insurance and social security) and non-taxable fringe benefits.
Incomes are deflated using the Consumer Price Index.
Column (4) reports the fraction of total real family income growth (or loss) captured by the top 1%.
For example, from 2002 to 2007, average real family incomes grew by 16.1% but 65% of that growth accrued to the top 1% while only 35% of that growth accrued to the bottom 99% of US families.
Source: Piketty and Saez (2003), series updated to 2012 in August 2013 using IRS preliminary tax statistics for 2012.

Example: CEO pay

Research question: **is CEO pay in 1999 linked to 1998 profits**?

$$pay_i = \beta_0 + \beta_1 profit_i + \varepsilon_i$$

where pay_i is the pay (salary+bonuses) of the CEO of the i^{th} company and $profit_i$ the profit of the i^{th} company.

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

Economic theory supporting H_A:

- rewards for creating shareholder value;
- performance pay (give CEO incentive to increase company's profits)

Sample

We use a sample of Fortune500 companies, containing the following information:

- CEO pay (salary + bonuses) in 1999 measured in thousands of \$;
- profits of 1998 in millions of \$.

The sample regression model is:

$$pay_i = \widehat{\beta}_0 + \widehat{\beta}_1 profit_i + e_i$$

Summary statistics

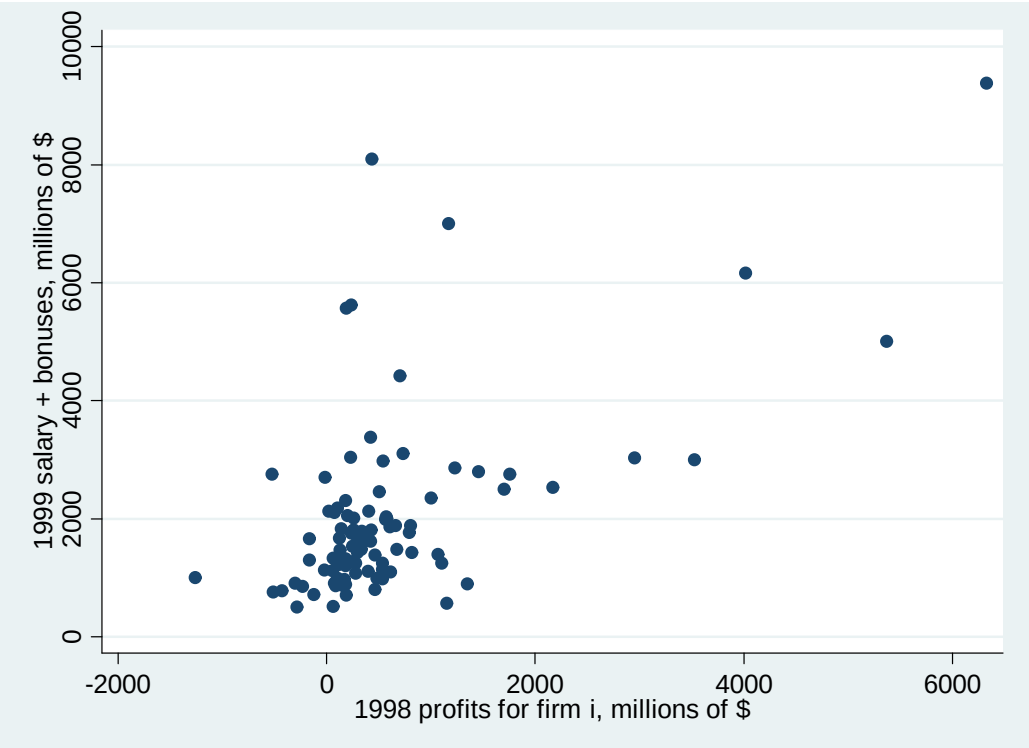
```
. describe pay profits

              storage  display     value
variable name   type   format      label      variable label
------------------------------------------------------------------------------
pay             int    %8.0g                  1999 CEO salary + bonuses, thousands of $
profits         float  %8.0g                  1998 profits for firm i, millions of $

. sum pay profits

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         pay |        100     1992.28    1527.201         500       9375
     profits |        100     570.283    1051.385    -1260.5       6328
```

$$\overline{pay} = \$1,992,280 \qquad s_{pay} = \$1,527,000$$

$$\overline{profits} = \$570,000,000 \qquad s_{profit} = \$1,050,000,000$$

Note: not just profits, there are also some losses!

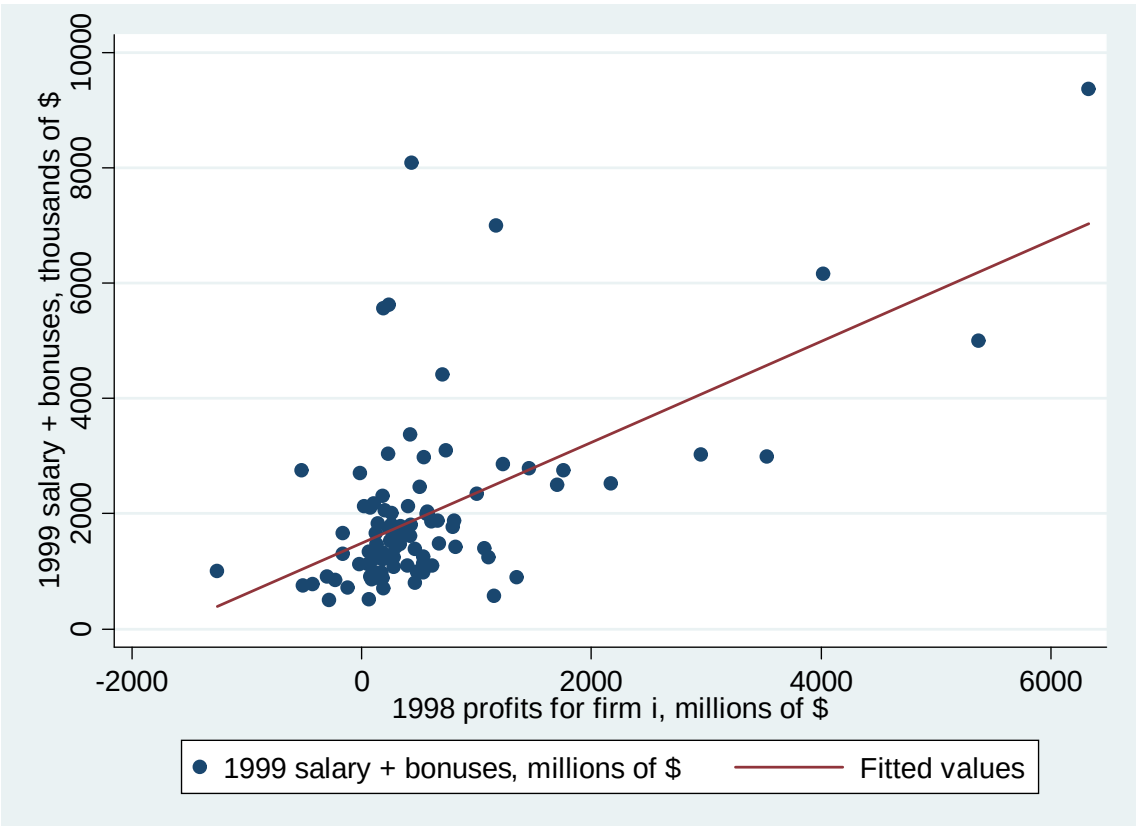
Sample data in a scatter diagram



We have pay_i and $profit_i$—how can we find unknowns $\widehat{\beta}_0, \widehat{\beta}_1$?

$$pay_i = \widehat{\beta}_0 + \widehat{\beta}_1 profit_i + e_i$$

Data + fitted OLS regression line



Ordinary Least Squares (OLS)

- OLS is an estimator: it **gives us values for the unknown coefficients** $\widehat{\beta}_0, \widehat{\beta}_1$
- OLS has certain **desirable statistical properties**: under certain assumptions (to be discussed) it is **BLUE** (best linear unbiased estimator).
- Work-horse of empirical research

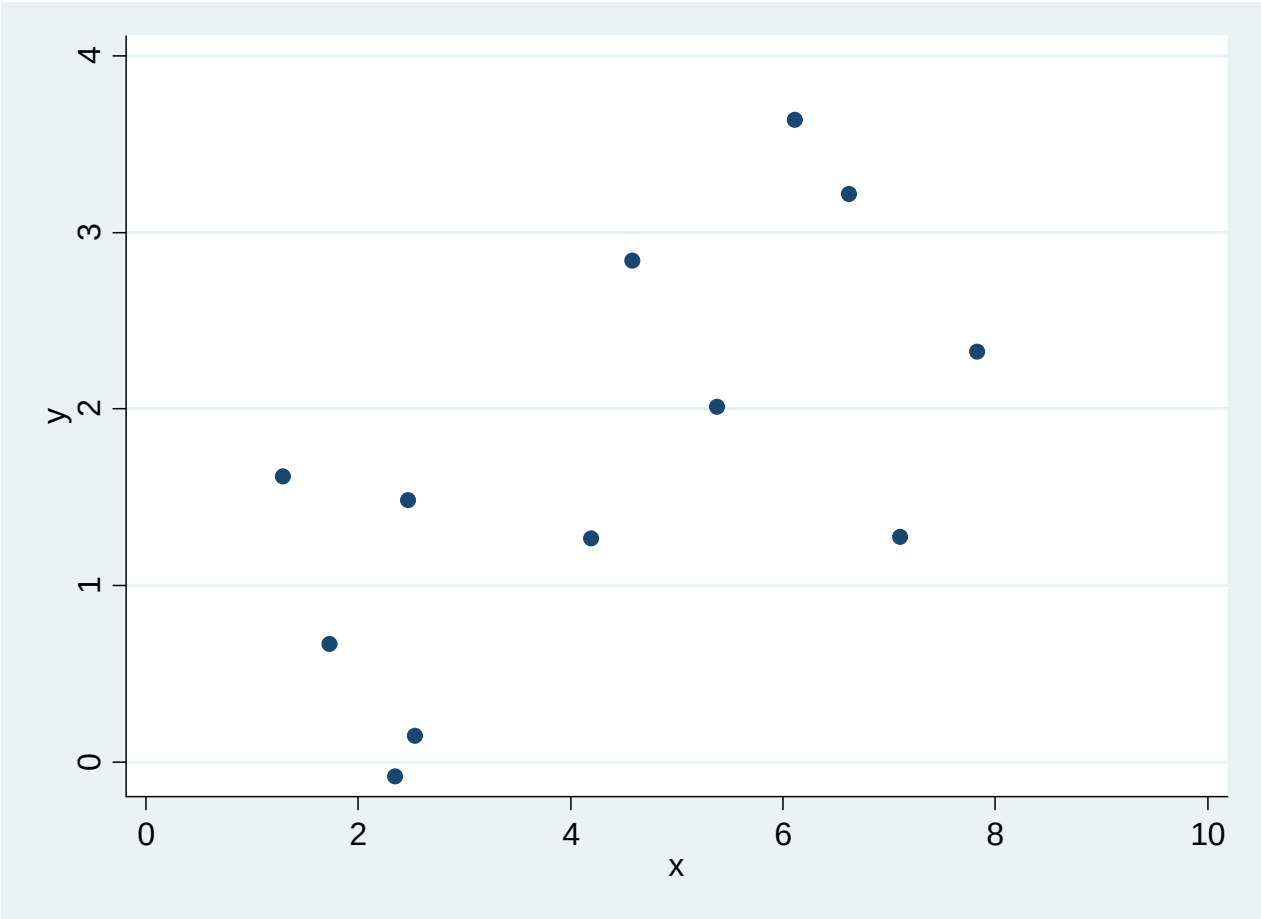
Mechanics of OLS

- OLS estimator chooses $\widehat{\beta}_0, \widehat{\beta}_1$ such that the **sum of squared deviations** (i.e. vertical distances) **from the regression line is minimized** (hence "least squares"):

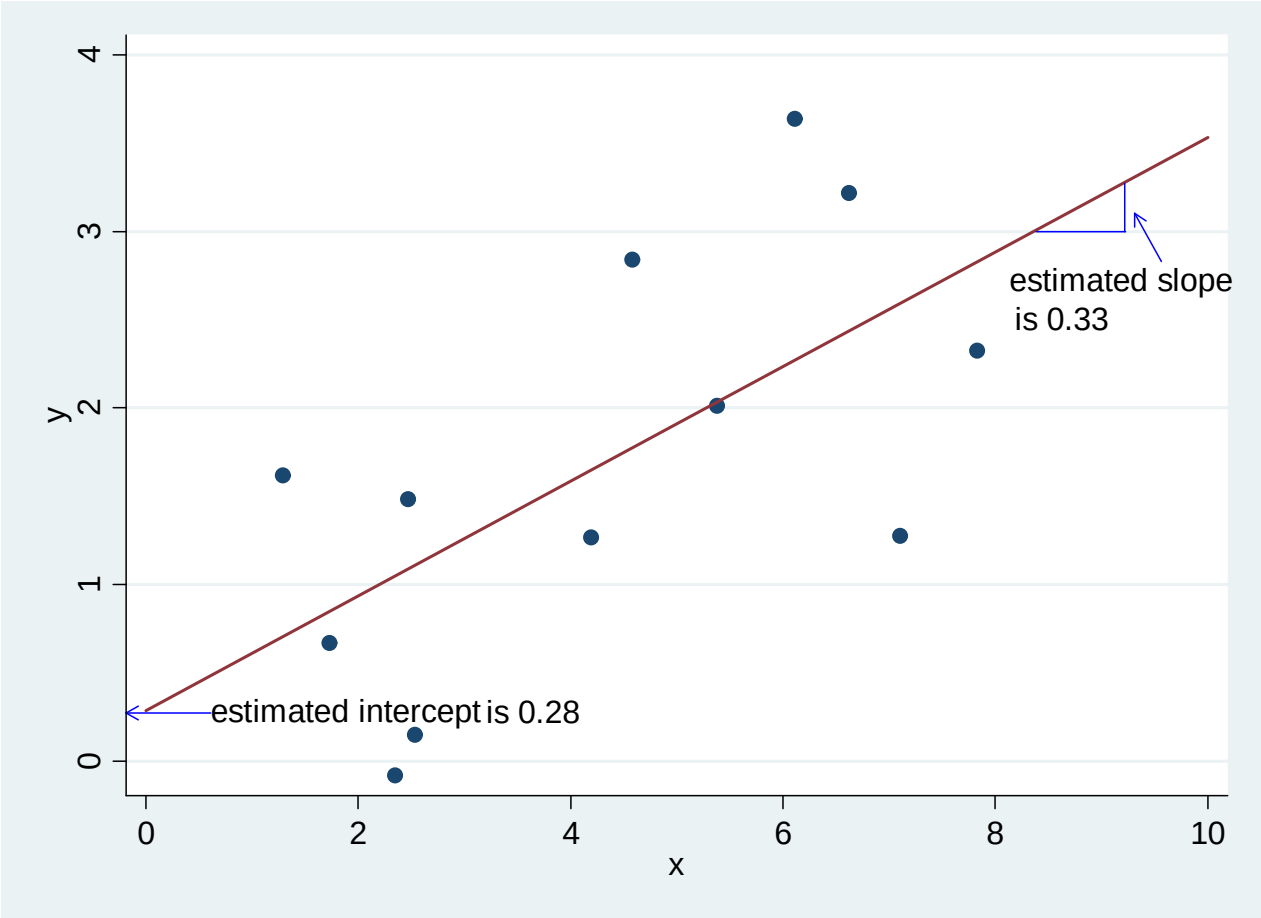
$$\begin{aligned} Y_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + e_i \\ \widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} \\ e_i &= Y_i - \widehat{Y}_i \end{aligned}$$

- Residual $e_i > 0$ if data point lies above the regression line; $e_i < 0$ if it lies below the regression line.

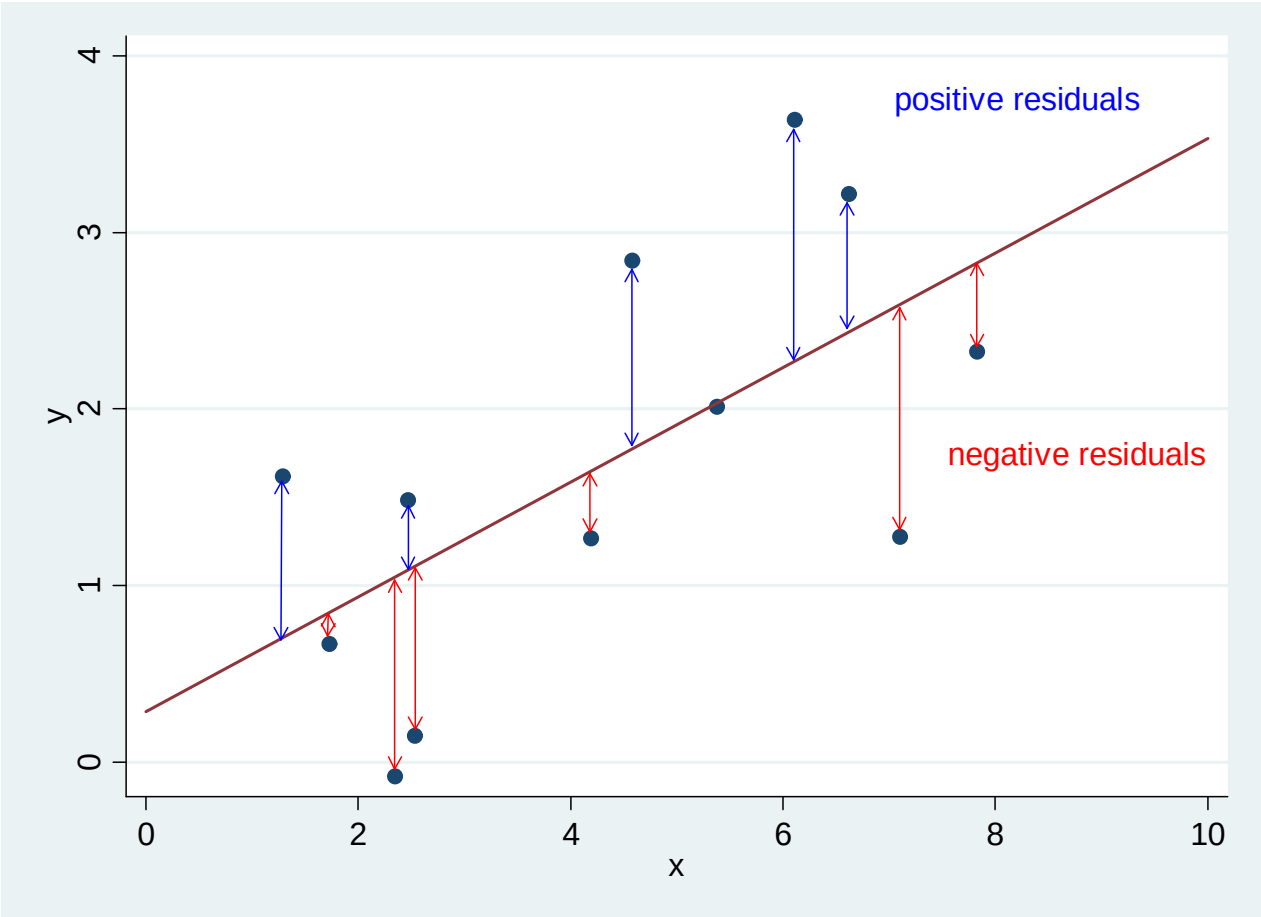
An illustration of OLS



An illustration of OLS



An illustration of OLS



Mechanics of OLS

- OLS estimator chooses $\widehat{\beta}_0, \widehat{\beta}_1$ such that the **sum of squared deviations from the regression line is minimized**:

$$\begin{aligned} Y_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i + e_i \\ \widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 X_i \\ &\Leftrightarrow \quad e_i = Y_i - \widehat{Y}_i \end{aligned}$$

- Sum of squared residuals, $\sum_{i=1}^{n} e_i^2$, in terms of $\widehat{\beta}_0, \widehat{\beta}_1$:

$$\begin{aligned} e_i &= Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \\ \sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \right)^2 \end{aligned}$$

Mechanics of OLS

Sum of squared residuals:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \right)^2$$

- Now find $\widehat{\beta}_0, \widehat{\beta}_1$ such that $\sum_{i=1}^{n} e_i^2$ is minimized.
- This can be done by taking the first order partial derivatives wrt $\widehat{\beta}_0$ and $\widehat{\beta}_1$, and setting the derivatives to zero to find the minimum.

OLS: constant and slope

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \right)^2$$

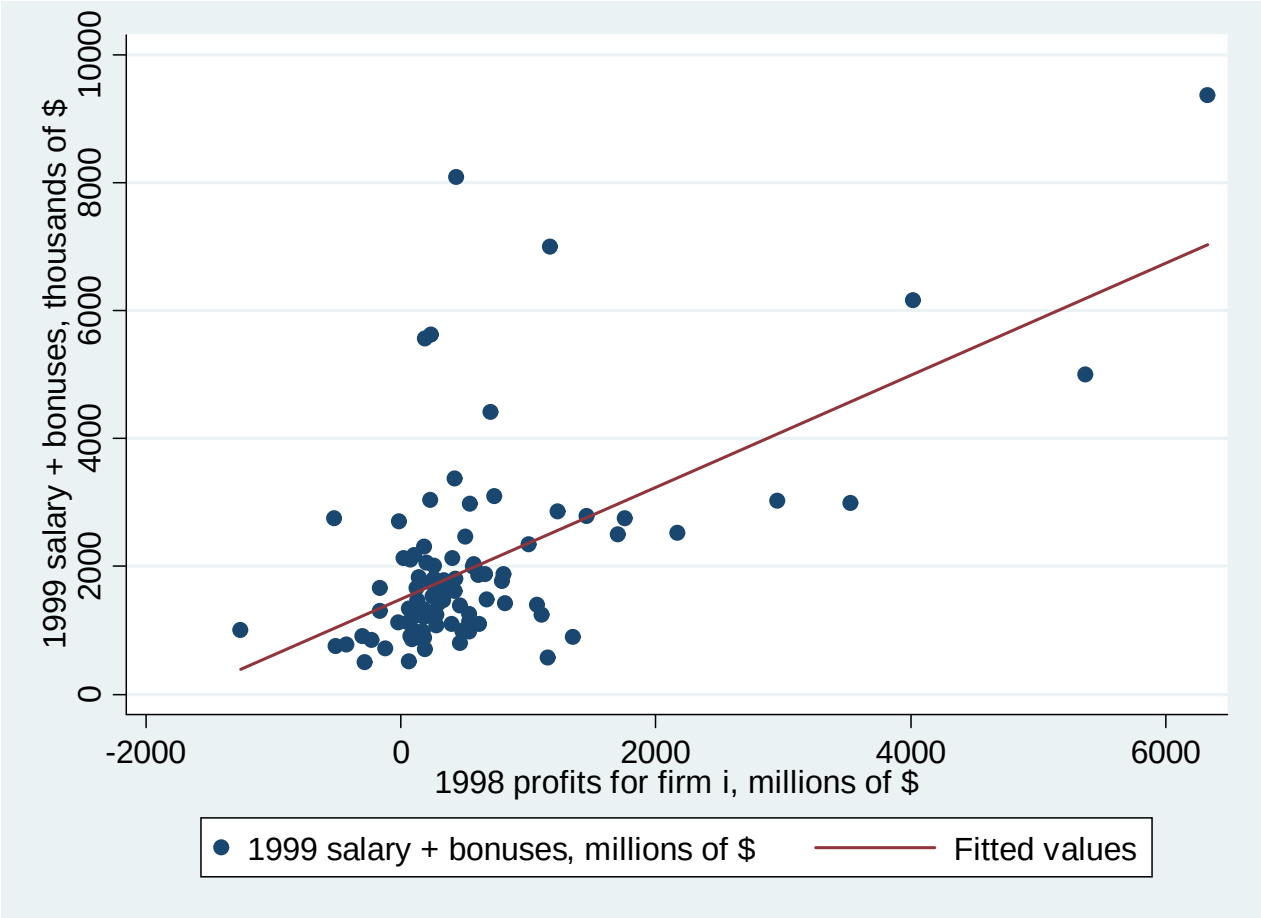
$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \widehat{\beta}_0} = 0, \qquad \frac{\partial \sum_{i=1}^{n} e_i^2}{\partial \widehat{\beta}_1} = 0$$

$$\Leftrightarrow \quad \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} \qquad \text{(intercept)}$$

$$\Leftrightarrow \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X} \right) \left(Y_i - \overline{Y} \right)}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2} = \frac{Cov(X_i, Y_i)}{Var(X_i)} \quad \text{(slope)}$$

Click for proof

Back to our example



Bivariate OLS estimates

```
. reg pay profits
```

Source	SS	df	MS
Model	83853950.7	1	83853950.7
Residual	147047911	98	1500488.89
Total	230901862	99	2332342.04

Number of obs	= 100
F(1, 98)	= 55.88
Prob > F	= 0.0000
R-squared	= 0.3632
Adj R-squared	= 0.3567
Root MSE	= 1224.9

pay	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
profits	.8753509	.1170946	7.48	0.000	.6429805	1.107721
_cons	1493.082	139.5137	10.70	0.000	1216.222	1769.943

$$\widehat{\beta}_0 = 1493.052 \qquad \widehat{\beta}_1 = 0.875$$

$$\begin{aligned} pay_i &= \widehat{\beta}_0 + \widehat{\beta}_1 profit_i + e_i \\ pay_i &= 1493.052 + 0.875 profit_i + e_i \end{aligned}$$

Calculation OLS estimates

```
. sum pay, det

        1999 CEO salary + bonuses, thousands of $
-------------------------------------------------------------
      Percentiles      Smallest
 1%         505.5            500
 5%         730.5            511
10%         865.5            566       Obs                 100
25%          1118            700       Sum of Wgt.         100

50%        1590.5                      Mean           1992.28
                        Largest         Std. Dev.      1527.201
75%          2154           6163
90%          3067           7000       Variance       2332342
95%        5592.5           8088       Skewness       2.641081
99%        8731.5           9375       Kurtosis       10.87568
```

```
. sum profits, det

        1998 profits for firm i, millions of $
-------------------------------------------------------------
      Percentiles      Smallest
 1%       -891.75         -1260.5
 5%       -292.65            -523
10%         -70.9         -510.6       Obs                 100
25%         128.7            -427       Sum of Wgt.         100

50%         289.5                       Mean           570.283
                        Largest         Std. Dev.      1051.385
75%         592.2           3527
90%        1291.7           4023       Variance       1105411
95%          2564           5372       Skewness       3.282631
99%          5850           6328       Kurtosis       15.96671
```

```
. corr pay profits, cova
(obs=100)

             |      pay  profits
-------------+------------------
         pay |  2.3e+06
     profits |   967623   1.1e+06
```

$$\widehat{\beta}_1 = \frac{Cov(X_i, Y_i)}{Var(X_i)} = \frac{967623}{1105411} = 0.875$$

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X} = 1992.28 - 0.875(570.283) = 1493$$

Interpretation of OLS estimates

$$\widehat{pay}_i = 1493.052 + 0.875 profit_i$$

- When profits increase with \$1 million, CEO pay increases with 875 dollars.
  - Alternatively: when profits increase with \$1 billion, CEO pay increases with \$875,000.
  - (Note that sd of profits is \$1051 million (i.e. \$1.05 billion); and sd of pay is \$1.5 million)
- If profits are zero, CEO pay is predicted to be \$1.49 million.
- Note that it makes economic sense to estimate the elasticity of of CEO pay to profits: we learn how to do this later in this course.

R-squared: goodness of fit

- OLS also gives us a measure of **goodness of fit of the regression**: R^2
- R^2 is the fraction of total variation in the dependent variable explained by the independent variable(s):

$$\begin{aligned} R^2 &= \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} \\ &= \frac{\sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2} \end{aligned}$$

where TSS=total sum of squares; ESS=explained sum of squares; RSS=residual sum of squares.

R-squared in our example

```
. reg pay profits
```

Source	SS	df	MS
Model	83853950.7	1	83853950.7
Residual	147047911	98	1500488.89
Total	230901862	99	2332342.04

Number of obs	100
F(1, 98)	55.88
Prob > F	0.0000
R-squared	0.3632
Adj R-squared	0.3567
Root MSE	1224.9

pay	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
profits	.8753509	.1170946	7.48	0.000	.6429805	1.107721
_cons	1493.082	139.5137	10.70	0.000	1216.222	1769.943

$$R^2 = 1 - \frac{147047911}{230901862} = 0.36$$

So variation in profits explains 36% of the variation in CEO pay across these 100 companies. This implies profits are indeed important for pay, but there are other factors that matter as well. (Although the R^2 is a useful statistic, see Studenmund for a caution about focusing too much on the R^2)

Root mean squared error (not in Studenmund)

- OLS also gives us a measure of the **average size of a residual**, in units of Y: **root mean squared error**

$$\sqrt{MSE} = \widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-k-1}}$$
$$= \sqrt{\frac{e_1^2 + e_2^2 + .. + e_n^2}{n-k-1}}$$

- where **n-k-1** is the **number of degrees of freedom**: n=nr of observations in the sample, k=number of independent variables.
- Root MSE is sometimes also called the standard error of the regression.

Root MSE

```
. reg pay profits
```

Source	SS	df	MS
Model	83853950.7	1	83853950.7
Residual	147047911	98	1500488.89
Total	230901862	99	2332342.04

Number of obs	100
F(1, 98)	55.88
Prob > F	0.0000
R-squared	0.3632
Adj R-squared	0.3567
Root MSE	1224.9

pay	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
profits	.8753509	.1170946	7.48	0.000	.6429805	1.107721
_cons	1493.082	139.5137	10.70	0.000	1216.222	1769.943

$$\text{Root MSE} = \sqrt{\frac{e_1^2 + e_2^2 + .. + e_n^2}{n-k-1}}$$

$$= \sqrt{\frac{147047911}{100-1-1}} = 1224.9$$

Research question: multiple regression

- Bivariate regression: Are CEO salaries (including bonuses) in 1999 linked to 1998 profits? $pay_i = \beta_0 + \beta_1 profit_i + \varepsilon_i$
- **Human capital theory** predicts pay does not only depend on profits but also on **the experience of the CEO with the firm**: to reflect this, we include **tenure** (=how many years the CEO has been at the firm) in the population model.

$$pay_i = \beta_0 + \beta_1 profit_i + \beta_2 tenure_i + \varepsilon_i$$

- For this **multivariate regression**, we also use OLS to calculate the unknown $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2$ (calculations omitted here)

$$pay_i = \widehat{\beta}_0 + \widehat{\beta}_1 profit_i + \widehat{\beta}_2 tenure_i + e_i$$

Why do we need multiple regression?

- To **better explain & understand economic processes**
  - E.g. demand for a good depends not only on its own price but also on the price of substitutes, as well as on income.

- We are **often interested in partial effects** (ceteris paribus effects): i.e. the impact of one independent variable on the dependent variable, holding some other independent variable(s) constant.
  - This is the power of multiple regression analysis: despite having non-experimental data, it allows us to do what natural scientists are able to do in a controlled laboratory setting: **keeping certain other factors fixed**!

Partial effects: example

E.g.: estimate the impact of one more year of education on wages. But we want to hold work experience constant. Use multivariate regression of wages on education and experience.

$$\widehat{\text{wage}}_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 \text{education}_i \qquad \text{(simple)}$$

$$\widehat{\text{wage}}_i = \widehat{\beta}_0 + \widehat{\beta}_1 \text{education}_i + \widehat{\beta}_2 \text{experience}_i \qquad \text{(multiple)}$$

- Important: in the simple regression, experience is in the error term; whereas for multiple regression, it is not.
- Generally $\widehat{\alpha}_1 \neq \widehat{\beta}_1$ unless $\widehat{\beta}_2 = 0$ or $Cov(\text{educ,exper}) = 0$.

Multiple regression: ceteris paribus interpretation

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + .. + \widehat{\beta}_k X_{ki}$$

$$\widehat{\Delta Y}_i = \widehat{\beta}_1 \Delta X_{1i} + \widehat{\beta}_2 \Delta X_{2i} + .. + \widehat{\beta}_k \Delta X_{ki}$$

Interpretation of $\widehat{\beta}_1$:
the impact of $\Delta X_{1i} = 1$ on $\widehat{\Delta Y}_i$, holding fixed $X_{2i}, ...X_{ki}$ such that $\Delta X_{2i} = .. = \Delta X_{ki} = 0$.

CEO pay: multiple regression

```
. reg pay profits tenure
```

Source	SS	df	MS
Model	104506598	2	52253298.8
Residual	126395265	97	1303043.96
Total	230901862	99	2332342.04

Number of obs	100
F(2, 97)	40.10
Prob > F	0.0000
R-squared	0.4526
Adj R-squared	0.4413
Root MSE	1141.5

pay	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
profits	.8979159	.109266	8.22	0.000	.6810532	1.114779
tenure	62.59064	15.72176	3.98	0.000	31.3873	93.79397
_cons	1004.525	178.7805	5.62	0.000	649.6951	1359.355

$$pay_i = \widehat{\beta}_0 + \widehat{\beta}_1 profit_i + \widehat{\beta}_2 tenure_i + e_i$$

$$pay_i = 1004.53 + 0.898 profit_i + 62.59 tenure_i + e_i$$

CEO pay: multiple regression interpretation

$$\widehat{pay}_i = 1004.53 + 0.898 profit_i + 62.59 tenure_i$$

- When profits are \$1 million higher, the CEO earns \$898 more, holding CEO tenure constant.
- When the CEO has been with the firm 1 year longer, his/her pay is \$62,590 higher, holding profits constant.
- A new CEO (i.e. who has been at the company for less than 1 year) for a company making zero profits earns \$1,004,000 a year.

Multiple regression: fitted values

Fitted values, or predicted values, are the $\widehat{Y}_i$. To find the predicted value of Y for any particular observation i, we just fill in the particular values that the X variables take on for that observation i.

- **Example**: **US box office revenues** (*rev*, in \$millions) of movie i depend on the movie's Rotten Tomatoes score (RT_i), budget in \$millions (B_i), the number of US theaters the movie could be viewed in (T_i), whether or not it is a sequel (S_i, 0=no 1=yes) and whether or not it's rated PG13 (0=no, 1=yes). This is the estimated equation:

$$\widehat{rev_i} = -80 + 0.6RT_i + 0.5B_i + 0.025T_i + 50S_i + 20PG13_i$$

reported by Cinemath (http://collider.com/box-office-reviews-statistics/173667/)

The Big Short

Multiple regression: fitted values

$$\widehat{rev_i} = -80 + 0.6RT_i + 0.5B_i + 0.025T_i + 50S_i + 20PG13_i$$

- Let's predict the **revenues for this particular movie, i.e. i =The Big Short**.
  - Rotten Tomatoes score = 88%, RT=88
  - Budget = \$28 million, B=28
  - Nr of US theaters it played in, T= 2229
  - Not a sequel, hence S=0
  - Rated PG13, hence PG13=1

$$\widehat{rev}_{bigshort} =$$

$$= -80 + 0.6 \times 88 + 0.5 \times 28 + 0.025 \times 2229 + 50 \times 0 + 20 \times 1$$

$$\approx \$63 \text{ million}$$

Multiple regression: fitted values

In reality, The Big Short made \$70 million in the US rather than the \$63 million predicted from the model. The difference between actual and predicted The Big Short box office takings is the residual $e_{bigshort}$

- **Positive residual** since
  $e_{bigshort} = rev_{bigshort} - \widehat{rev}_{bigshort} = 70 - 63 = 7$
- Hence the model somewhat underpredicts U.S. box office takings for this particular movie.

Algebraic properties of OLS

Algebraic properties of OLS are those properties which are **automatic outcomes of all regressions estimated with OLS** (i.e. of minimizing the sum of squared residuals)- they **tell us nothing about the population**!

1. **Residuals have a mean of zero**

$$\frac{1}{n}\sum_{i=1}^{n} e_i = \overline{e} = 0$$

2. **Residuals are uncorrelated with all the independent variables X_k**

$$Corr(X_{ki}, e_i) = Cov(X_{ki}, e_i) = 0$$

Note that these two properties can be written in one equation as:

$$E(e_i|X_{1i}, X_{2i}..X_{ki}) = 0$$

Algebraic properties of OLS shown in our example

```
. reg pay profits tenure
```

Source	SS	df	MS
Model	104506598	2	52253298.8
Residual	126395265	97	1303043.96
Total	230901862	99	2332342.04

Number of obs	100
F(2, 97)	40.10
Prob > F	0.0000
R-squared	0.4526
Adj R-squared	0.4413
Root MSE	1141.5

pay	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
profits	.8979159	.109266	8.22	0.000	.6810532	1.114779
tenure	62.59064	15.72176	3.98	0.000	31.3873	93.79397
_cons	1004.525	178.7805	5.62	0.000	649.6951	1359.355

```
. predict uhat, resid

. sum uhat
```

Variable	Obs	Mean	Std. Dev.	Min	Max
uhat	100	-5.40e-06	1129.92	-2974.253	5504.917

```
. corr uhat profits tenure
(obs=100)
```

	uhat	profits	tenure
uhat	1.0000		
profits	0.0000	1.0000	
tenure	-0.0000	-0.0519	1.0000

Overview

Population: purely theoretical

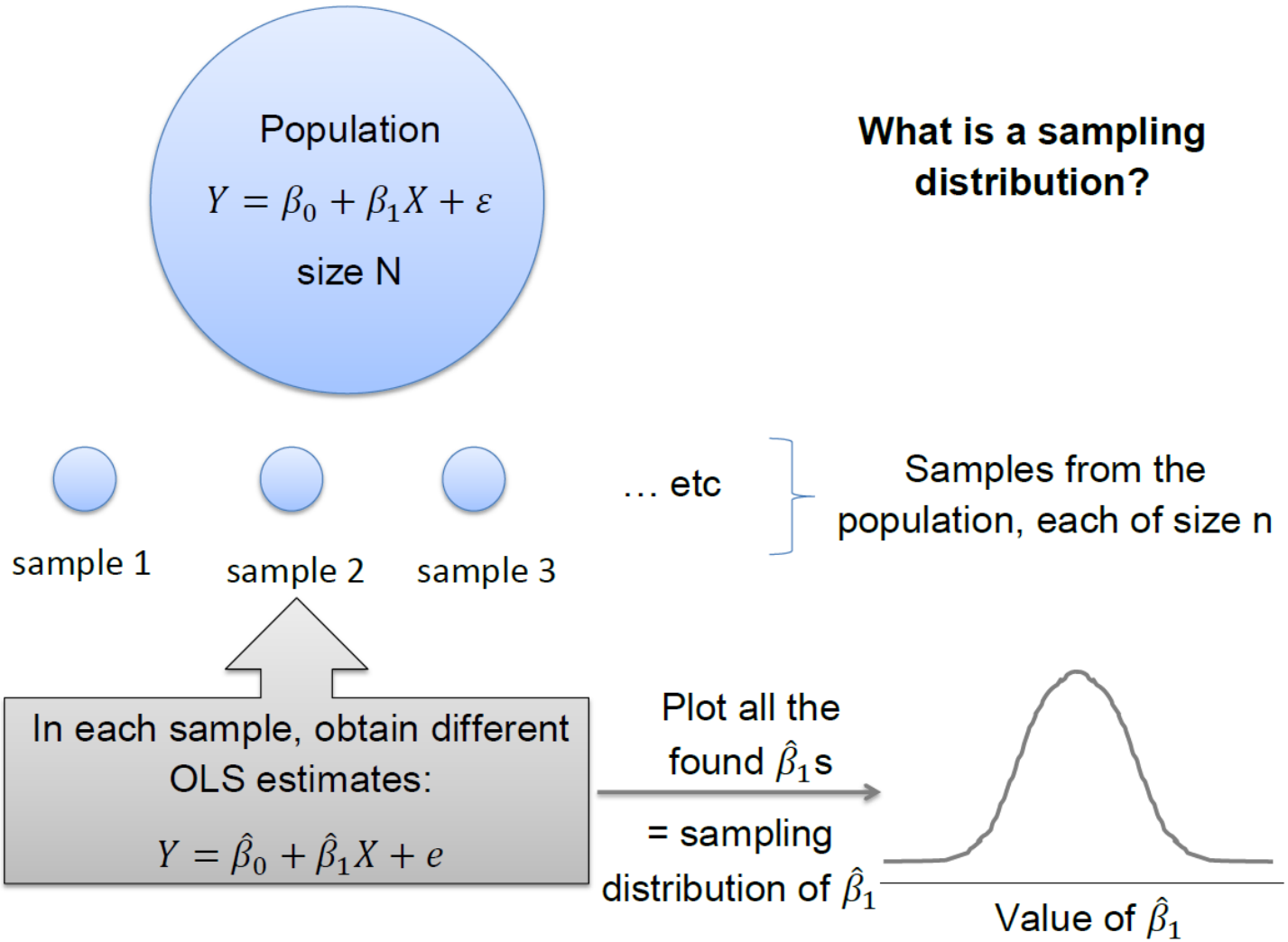
$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

Sample: empirical counterpart

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

- **Estimation**: we have seen that we can estimate the population equation in a sample by using the OLS estimator.
- Now we turn to **inference:** how to use our OLS estimates to say something about the population? (Short answer: by describing the sampling distribution of $\widehat{\beta}_k$)

The sampling distribution



Inference

Inference: using our estimates of the impact of profits on CEO pay *in our sample* ($\widehat{\beta}_1$) to say something about the impact of profits on pay *in the unobserved population* (β_1).

- We use the theoretical concept of the **sampling distribution** of $\widehat{\beta}_1$= the distribution of $\widehat{\beta}_1$ in many different samples from the population
- **Under certain assumptions**, the sampling distribution is centered on the true population β_1 ($E(\widehat{\beta}_1) = \beta_1$) and **we can infer about the population** from the sample.
- But there is always some **uncertainty** about this inference: although the estimator is unbiased on average, *in any particular sample* $\widehat{\beta}_1$ need not equal the true population value β_1. We can quantify this uncertainty by calculating the variance of $\widehat{\beta}_1$.

Statistical properties of OLS

- Under 4 assumptions, **OLS** estimates $\widehat{\beta}_k$ are **unbiased** estimates of the population parameters β_k
- Under 2 additional assumptions, we can obtain an **unbiased estimate of** $Var(\widehat{\beta}_k)$ (measure of sampling uncertainty).
- Combining all these assumptions, OLS is "**BLUE**" - this means OLS is the Best (i.e. most efficient = has the smallest $Var(\widehat{\beta}_k)$) estimator among the set of Linear Unbiased Estimators

Unbiasedness of OLS: assumptions

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)

Unbiasedness: a reminder

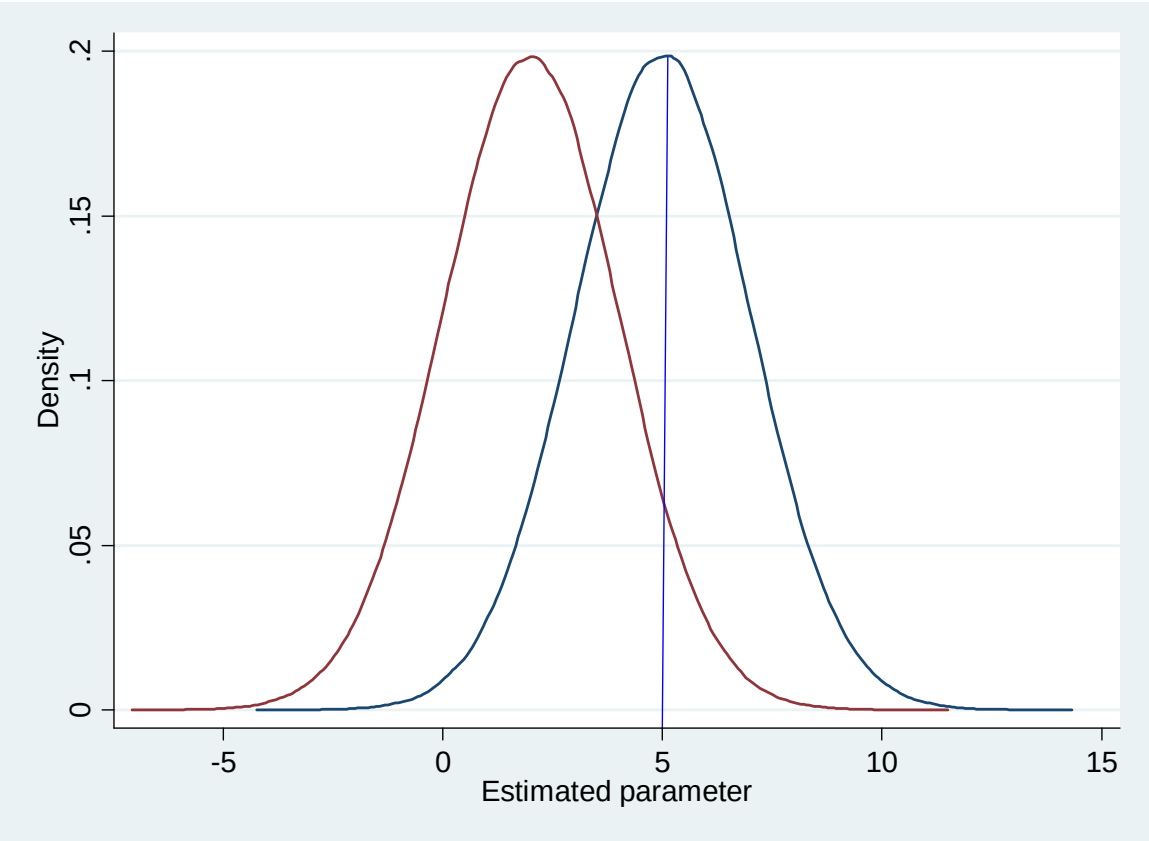
OLS estimator is **unbiased** if the expected value of the estimates produced by the estimator equals the population parameter.

- That is, for some population parameter θ

$$E(\widehat{\theta}) = \theta \qquad \text{(unbiasedness)}$$

- So if we had **many different samples from the population**, and we would use the OLS estimator to calculate the **estimate** $\widehat{\theta}$ **in each of these samples**, the **average value** of these estimates **would equal the population value** θ.
- proof that under the 4 assumptions OLS is an unbiased estimator of coefficients β_k.

Sampling distribution of estimated parameter



Example for when the true population parameter is 5: graph shows sampling distribution for a biased (red) and unbiased (blue) estimator.

A note on unbiasedness vs. consistency

- An estimator is **unbiased** if the average value of the estimator in an infinite number of samples equals the population parameter.
- An estimator is **consistent** if the estimator converges to the population parameter as the size of the sample tends toward infinity. (We can show that the OLS estimator is also consistent under the same 4 assumptions for unbiasedness)
- **Important**: in this course, we will not explicitly distinguish between unbiasedness and consistency of estimators- **you may use the two terms interchangeably**. At the MSc level, we will return to this in more detail.

A closer look at assumption 1

Population model is linear in parameters $\beta_0, \beta_1, .., \beta_k$

- Linear in parameters (can still be nonlinear in $X_1, X_2..X_k$!) - examples:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{1i} + \varepsilon_i \\ Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + \varepsilon_i \\ \ln Y_i &= \beta_0 + \beta_1 \ln X_{1i} + \varepsilon_i \\ Y_i &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i \end{aligned}$$

- NOT linear in parameters (cannot be estimated with OLS) - examples:

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

A closer look at assumption 2

Error term has a zero population mean: $E(\varepsilon_i) = 0$

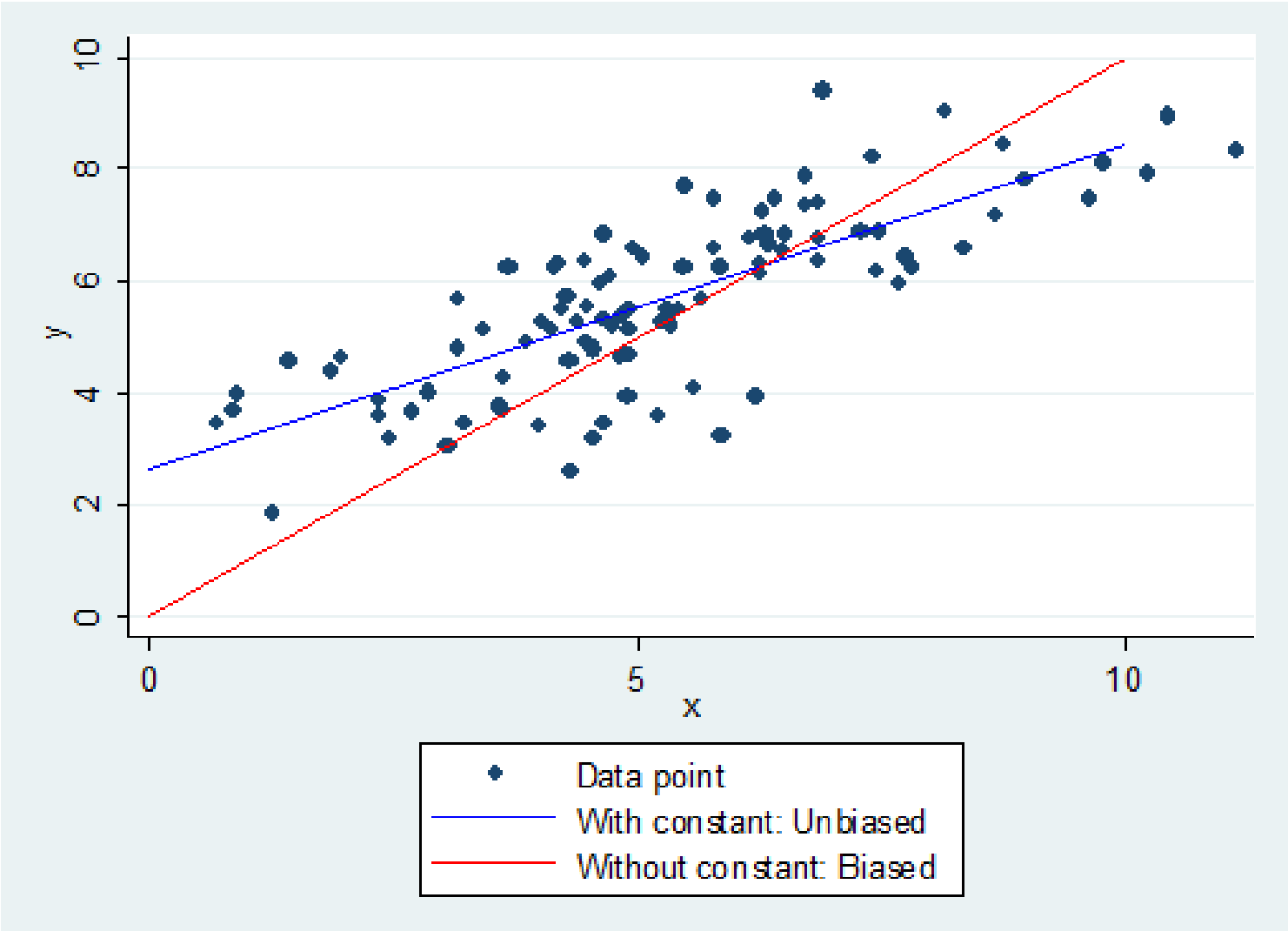
- This **assumption is met as long as a constant** (β_0) **is included** in the model
- This is because the constant will always absorb any non-zero mean of the error term. To see this, write:

$$Y_i = \beta_0 + \beta_1 X_{i1} + u_i$$

where we suppose that $E(u_i) \neq 0$, and denote $\alpha_0 = E(u_i)$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + u_i - \alpha_0 + \alpha_0 \\ &= (\beta_0 + \alpha_0) + \beta_1 X_{i1} + (u_i - \alpha_0) \\ &= \gamma_0 + \beta_1 X_{i1} + \varepsilon_i \qquad \text{where } E(\varepsilon_i) = 0 \end{aligned}$$

Not including a constant leads to biased estimates



A closer look at assumption 3

All independent variables are uncorrelated with the error term: $Corr(\varepsilon_i, X_i) = 0$ holds for all independent variables

- This assumption states that the X variables have to be **exogenous**
- It is the most important assumption: **without it, our estimates do NOT have a causal interpretation**
- **Omitted variable bias** is the most important reason why this assumption can fail (next week, we will discuss this at length!).

A closer look at assumption 4

No perfect (multi)collinearity

- Bivariate case: the OLS solution for $\widehat{\beta}_1$ shows that X cannot be constant, i.e. cannot have $Var(X_i) = 0$:

$$\widehat{\beta}_1 = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

- Multivariate case: we additionally have that **independent variables cannot be perfect linear functions of each other** (no perfect (multi)collinearity)
- This assumption can be observed in the sample: Stata will deal with it automatically by dropping the perfectly collinear variables

A closer look at assumption 4

- Assumption 4 is only about **perfect** collinearity, (extremely) **high correlations between variables do not violate this assumption**
- Examples of **perfectly (multi)collinear variables**:
  - Income measured in Euros and income measured in Dollars
  - Income, consumption and savings, where $income \equiv consumption + savings$
- Examples of **not perfectly collinear variables**:
  - Individuals' height and weight (very high correlation, but not a correlation of 1)
  - Individuals' age and age squared (very high correlation, but not 1: see last week's tutorial)

Assumptions 1-4: summary

All 4 assumptions are needed for OLS to be an unbiased estimator- as soon as 1 fails, $E(\widehat{\beta}) \neq \beta$

A1: Linearity: not extremely restrictive as linear models can still be non-linear in the independent variables

A2: $E(\varepsilon_i) = 0$: guaranteed to hold as long as we include a constant in our model (and therefore, we always will)

A3: $Corr(X_i, \varepsilon_i) = 0$: very strong assumption, and by far the most important - we will get back to it next week

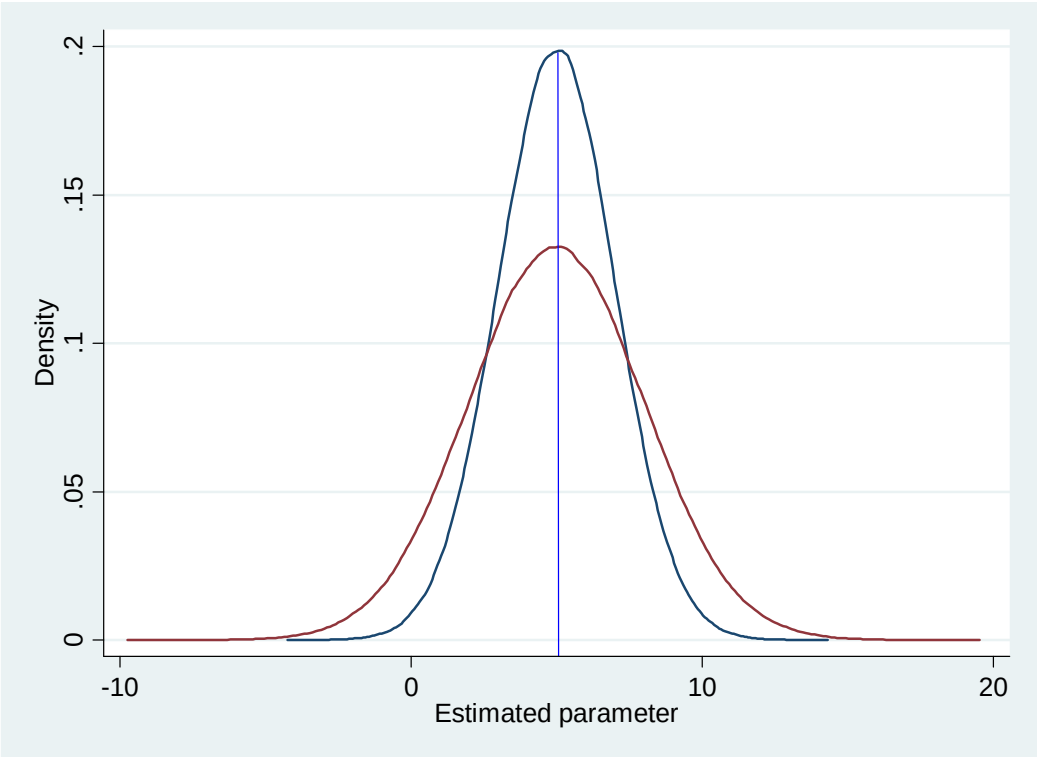
A4: No perfect (multi)collinearity: can be observed

Variance of OLS estimator

- Now we know that the **sampling distribution of our estimate $\widehat{\beta}$ is centered around the true parameter** β, i.e. OLS is unbiased.

- But it **also matters how wide this sampling distribution is**, since we obtain only one sample from the population: the variance of the sampling distribution $Var(\widehat{\beta})$ tells us how far that sample's $\widehat{\beta}$ is likely to be from the true population β.

Variance of the sampling distribution



Both estimators are unbiased, but one has a larger variance (red) than the other (blue). We of course prefer the latter, since in any one sample we are more likely to obtain a $\widehat{\beta}$ close to the true β!

Variance of the OLS estimator: assumptions

We can obtain an unbiased estimate of $Var(\widehat{\beta})$ (i.e. $E\left(\widehat{Var}(\widehat{\beta})\right) = Var(\widehat{\beta})$) using OLS, if **assumptions 1-4 hold, as well as**:

5. **No serial correlation**: errors are not correlated with each other across different observations, $Corr(\varepsilon_i, \varepsilon_j) = 0$– this is mostly important for timeseries and we will discuss it in week 6 of the course.

6. **No heteroskedasticity**: error term has constant variance, $Var(\varepsilon_i) = \sigma^2$ (where σ^2 is a constant).

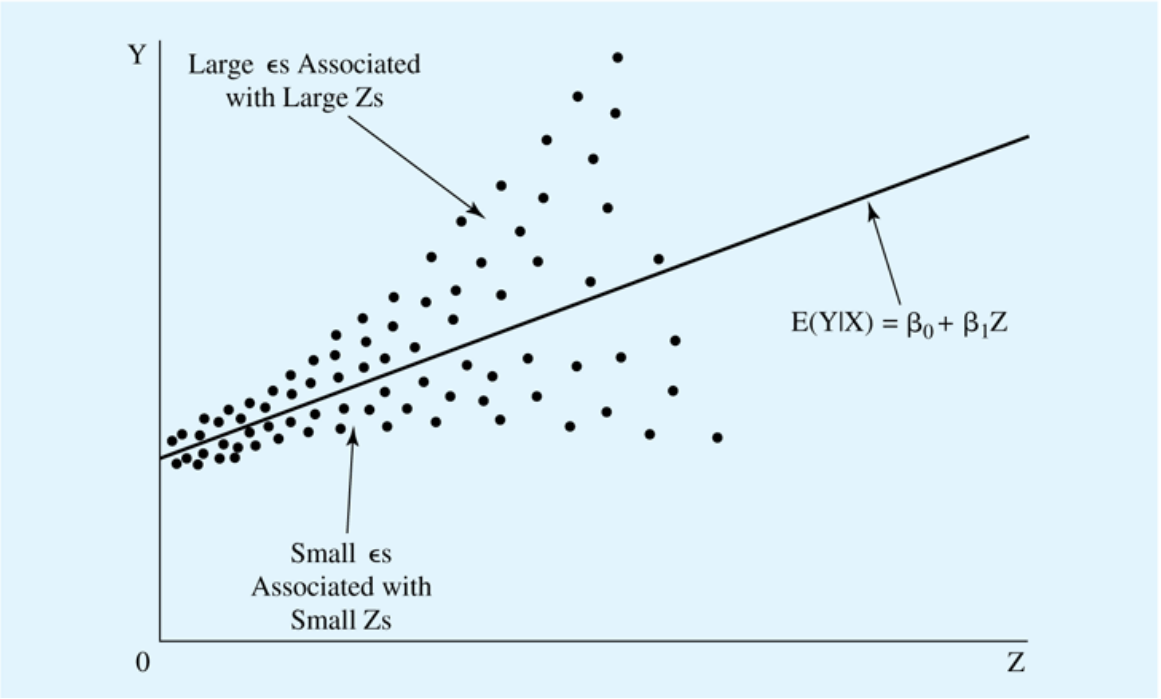
Under all 6 assumptions, OLS is BLUE: the estimator with the smallest variance (i.e. most efficient) among linear unbiased estimators.

A closer look at assumption 6

No heteroskedasticity:

- Errors are assumed to be **homoskedastic**, i.e. have a constant variance
- Although we don't need this assumption to have $E(\widehat{\beta}) = \beta$, we do need it to have **an unbiased estimate of the error variance,** $Var(\varepsilon_i)$
- This unbiased estimate of the error variance, we in turn need for **an unbiased estimate of the variance of** $\widehat{\beta}$, $Var(\widehat{\beta})$

Heteroskedastic errors



We deal with heteroskedasticity in more detail in week 5 of the course.

Variance of the OLS estimator (not in Studenmund!)

Under A1-A6, the **variance of the OLS estimate** of β_k is:

$$\begin{aligned} Var(\widehat{\beta}_k) &= \frac{Var(\varepsilon_i)}{(1-R_k^2)\sum_{i=1}^{n}\left(X_{ki}-\overline{X}_k\right)^2} \\ &= \frac{\sigma^2}{(1-R_k^2)\,TSS_{X_k}} \end{aligned}$$

where:

- σ^2 is the variance of the error term;
- TSS_{X_k} is the total sum of squares for independent variable X_k
- R_k^2 is the R^2 from an auxiliary regression of X_k on all other independent variables, e.g.
  - $X_1 = \alpha_0 + \alpha_1 X_2 + \alpha_2 X_3 + ..\alpha_k X_k + v$ gives R_1^2;
  - $X_2 = \delta_0 + \delta_1 X_1 + \delta_2 X_3 + ..\delta_k X_k + \nu$ gives R_2^2

Intuition behind the variance of the OLS estimator

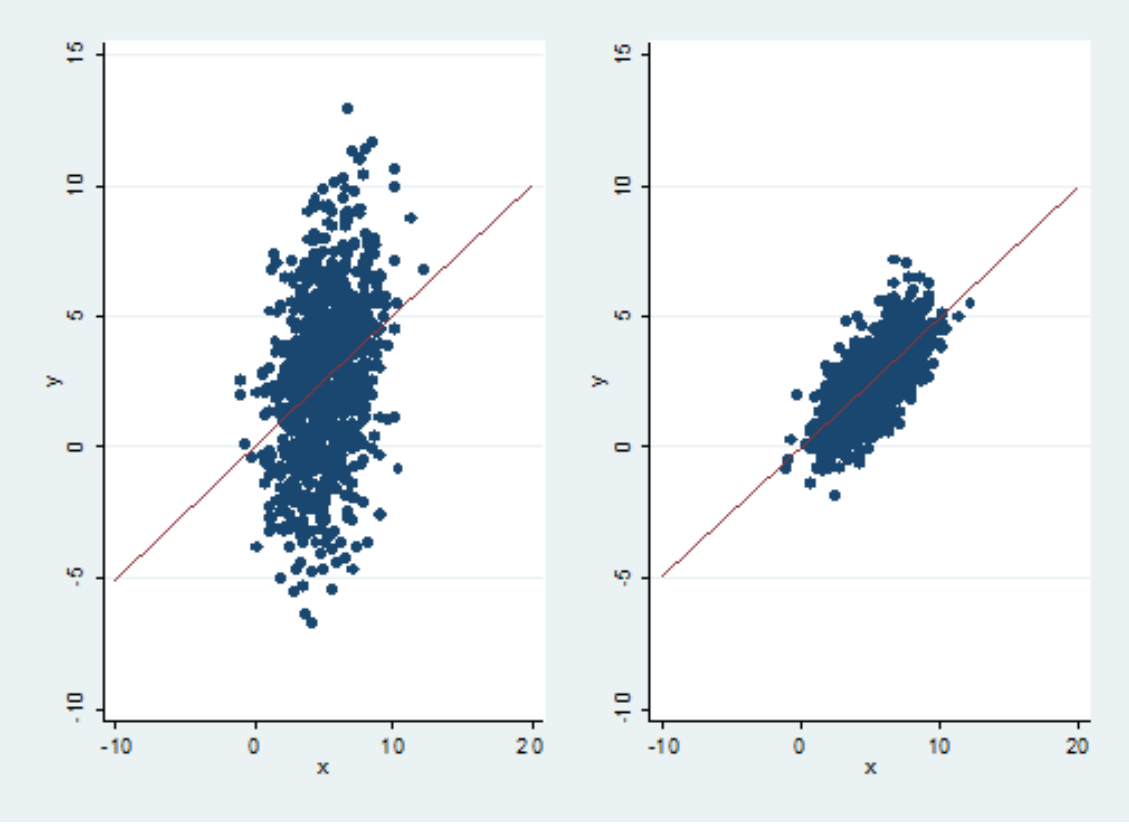
$$Var(\widehat{\beta}_k) = \frac{\sigma^2}{(1 - R_k^2)\, TSS_{X_k}}$$

The **larger Var**$(\widehat{\beta}_k)$, the **larger the "sampling uncertainty"**- that is, the higher the chance our found $\widehat{\beta}_k$ in any particular sample is far from the true population β_k.
The **sampling uncertainty is smaller (i.e. more precision on our estimate** $\widehat{\beta}_k$**)**:

- **The smaller** σ^2: that is, the less error
- **The larger** TSS_{X_k} : that is, the more variation in X_k
- **The smaller** R_k^2 : that is, more variation in X_k that is not shared with other regressors

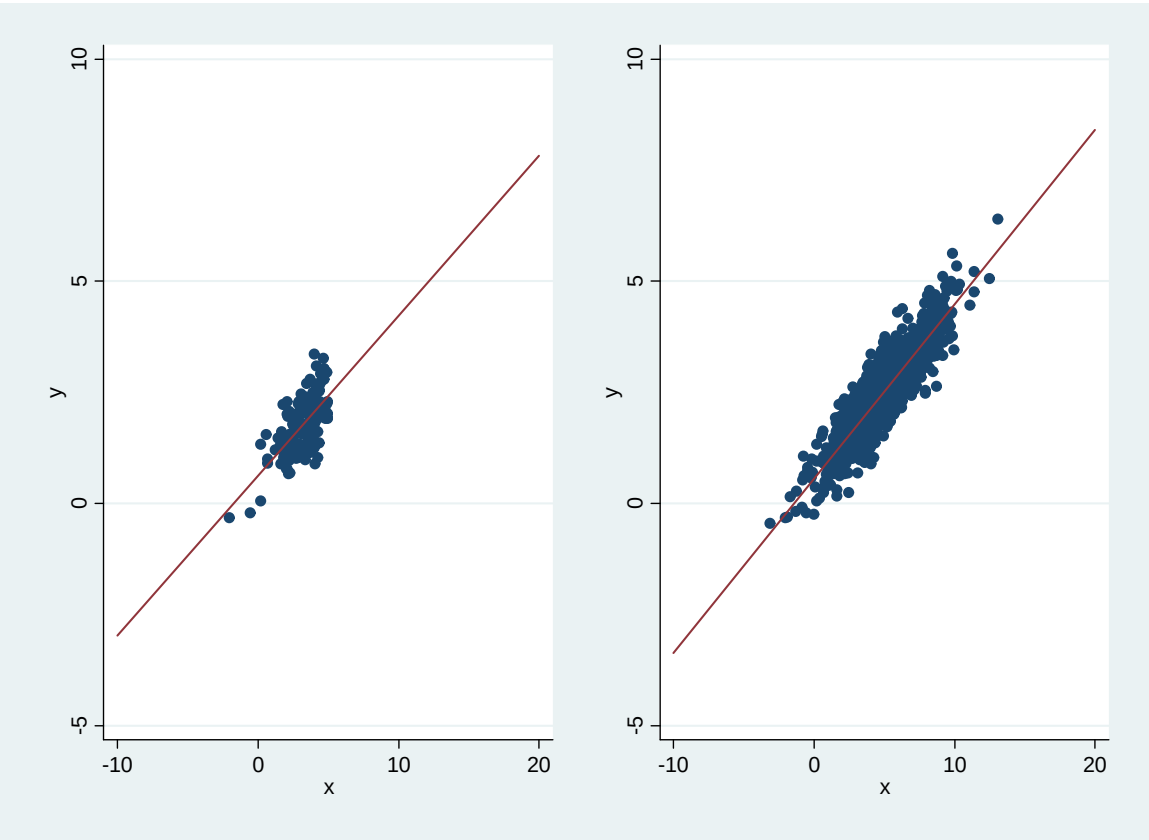
Intuition: large versus small error variance

A smaller error variance σ^2 gives higher precision:



Intuition: less versus more variation in X

More variation in X (which can be obtained by having a larger sample) gives higher precision:



Intuition: less versus more "unique" variation in X

The smaller R_k^2, the smaller $Var(\widehat{\beta}_k)$ (i.e. more precision):

- A **high $\mathbf{R_k^2}$** (close to 1) indicates that **much of the variation in the independent variable X_k is shared with other independent variables** included in the regression.
- This means that, when X_k changes, the other independent variable(s) often also change(s)- it's therefore difficult to find the effect of X_k on Y while holding the other X variables constant (=partial effect)
- This phenomenon is called multicollinearity: unlike for perfect multicollinearity, multicollinearity does not cause bias in our OLS estimates, but it does increase their variances. (We address this in more detail in week 5)

Variance of the OLS estimator

The **variance of the OLS estimate** $\widehat{\beta}_k$:

$$Var(\widehat{\beta}_k) = \frac{\sigma^2}{(1 - R_k^2)\ TSS_{X_k}}$$

- But there is a problem! We **never observe** σ^2, the variance of the unobserved error term.
- We do, however, observe **residuals**- we can **use these to construct an estimate of** σ^2 (with $E(\widehat{\sigma}^2) = \sigma^2$):

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n - k - 1} = \frac{e_1^2 + e_2^2 + .. + e_n^2}{n - k - 1}$$

Variance of the OLS estimator

Hence we obtain (note the added hat on $\widehat{Var}$ which indicates we *estimated* the variance of $\widehat{\beta}_k$!):

$$\begin{aligned}\widehat{Var}(\widehat{\beta}_k) &= \frac{\widehat{\sigma}^2}{(1-R_k^2)\,TSS_{X_k}} \\ &= \frac{\sum_{i=1}^{n} e_i^2}{(n-k-1)\,(1-R_k^2)\,TSS_{X_k}}\end{aligned}$$

The **standard error of the estimated parameter** is:

$$se(\widehat{\beta}_k) = \widehat{\sigma}_{\widehat{\beta}_k} = \frac{\widehat{\sigma}}{\sqrt{(1-R_k^2)\,TSS_{X_k}}}$$

This is reported in Stata output.

Standard error of the OLS estimates in our example

```
. reg pay profits tenure
```

Source	SS	df	MS
Model	104506598	2	52253298.8
Residual	126395265	97	1303043.96
Total	230901862	99	2332342.04

Number of obs	100
F(2, 97)	40.10
Prob > F	0.0000
R-squared	0.4526
Adj R-squared	0.4413
Root MSE	1141.5

pay	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
profits	.8979159	.109266	8.22	0.000	.6810532	1.114779
tenure	62.59064	15.72176	3.98	0.000	31.3873	93.79397
_cons	1004.525	178.7805	5.62	0.000	649.6951	1359.355

$$se(\widehat{\beta}_0) = 178.78; \quad \widehat{Var}(\widehat{\beta}_0) = \left[se(\widehat{\beta}_0)\right]^2 = 31962$$

$$se(\widehat{\beta}_1) = 0.10927; \quad \widehat{Var}(\widehat{\beta}_1) = \left[se(\widehat{\beta}_1)\right]^2 = 0.0199$$

$$se(\widehat{\beta}_2) = 15.722; \quad \widehat{Var}(\widehat{\beta}_2) = \left[se(\widehat{\beta}_2)\right]^2 = 247.18$$

Summary

Today, we've seen:

- How to use OLS to estimate a population model in a sample
- How sampling distributions can be used to infer about the population from a sample, and under which assumptions such inference in valid.
- In this week's tutorial, we further deepen our understanding by showing exactly where the discussed Stata output comes from.

What we covered this week, highlighted in Stata output

```
. reg pay profits tenure
```

Source	SS	df	MS
Model	104506598	2	52253298.8
Residual	126395265	97	1303043.96
Total	230901862	99	2332342.04

Number of obs	100
F(2, 97)	40.10
Prob > F	0.0000
R-squared	0.4526
Adj R-squared	0.4413
Root MSE	1141.5

pay	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
profits	.8979159	.109266	8.22	0.000	.6810532	1.114779
tenure	62.59064	15.72176	3.98	0.000	31.3873	93.79397
_cons	1004.525	178.7805	5.62	0.000	649.6951	1359.355

Note that MS = mean square, which is given by SS (=sum of squares) divided by df (=the degrees of freedom)

Next weeks of this course

- **Hypothesis testing** in bivariate/multivariate regression
- Different **model specifications** (quadratic, logs, dummies as independent variables, interaction terms, ..)
- **Diseases, Diagnostics & Cures**
  - **Disease** = violation of one of the OLS assumptions: what are the effects on the estimator (unbiasedness of $\widehat{\beta}$ and of ?
  - **Diagnostics** = how can we tell whether the particular assumption has been violated?
  - **Cure** = how can we fix the problem?

Project paper

- Why is it necessary to use a multivariate equation for your research question (i.e. which factors do you want to keep fixed- for now, focus on the ones on which your dataset has information)?
- Write down the multivariate population model & estimate it in your sample.
- Give a careful interpretation of the estimated regression parameters.
- Interpret the R-squared.
- Calculate correlations among regressors.

Project paper

- Start considering whether the 4 assumptions for an unbiased estimator of the regression parameters β are met (more on this in coming weeks).
- Start considering whether the 2 additional assumptions for unbiased estimator of σ^2 are met (more on this in coming weeks).

Bivariate regression: deriving the OLS intercept

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \right)^2 \qquad \text{(proof)}$$

$$\frac{\partial \sum_{i=1}^{n} e_i}{\partial \widehat{\beta}_0} = \sum_{i=1}^{n} -2 \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \right) = 0$$

$$= -2 \sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \right) = 0$$

$$= -2 \left(\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \widehat{\beta}_0 - \sum_{i=1}^{n} \widehat{\beta}_1 X_i \right) = 0$$

Bivariate regression: deriving the OLS intercept (continued)

$$
\begin{aligned}
\frac{\partial \sum_{i=1}^{n} e_i}{\partial \widehat{\beta}_0} &= \sum_{i=1}^{n} Y_i - n\widehat{\beta}_0 - \sum_{i=1}^{n} \widehat{\beta}_1 X_i = 0 \\
&\quad \text{Divide both sides by n} \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i - \widehat{\beta}_0 - \frac{1}{n}\sum_{i=1}^{n} \widehat{\beta}_1 X_i = 0 \\
&= \overline{Y} - \widehat{\beta}_0 + \widehat{\beta}_1 \overline{X} = 0 \\
\Leftrightarrow \quad & \widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}
\end{aligned}
$$

Bivariate regression: deriving the OLS slope

$$
\begin{aligned}
\sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i \right)^2 \\
\text{using that } \widehat{\beta}_0 &= \overline{Y} - \widehat{\beta}_1 \overline{X} \\
\sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} \left(Y_i - \left[\overline{Y} - \widehat{\beta}_1 \overline{X} \right] - \widehat{\beta}_1 X_i \right)^2 \\
&= \sum_{i=1}^{n} \left(Y_i - \overline{Y} + \widehat{\beta}_1 \overline{X} - \widehat{\beta}_1 X_i \right)^2 \\
&= \sum_{i=1}^{n} \left(Y_i - \overline{Y} - \widehat{\beta}_1 \left(X_i - \overline{X} \right) \right)^2
\end{aligned}
$$

Bivariate regression: deriving the OLS slope (continued)

$$
\begin{aligned}
\sum_{i=1}^{n} e_i^2 &= \sum_{i=1}^{n} \left(Y_i - \overline{Y} - \widehat{\beta}_1 \left(X_i - \overline{X} \right) \right)^2 \\
\frac{\partial \sum_{i=1}^{n} e_i}{\partial \widehat{\beta}_1} &= -2 \sum_{i=1}^{n} \left(X_i - X \right) \left(Y_i - \overline{Y} - \widehat{\beta}_1 \left(X_i - X \right) \right) = 0 \\
&= \sum_{i=1}^{n} \left(\left(X_i - X \right) \left(Y_i - \overline{Y} \right) - \widehat{\beta}_1 \left(X_i - \overline{X} \right)^2 \right) = 0 \\
&= \sum_{i=1}^{n} \left(X_i - \overline{X}_1 \right) \left(Y_i - \overline{Y} \right) - \sum_{i=1}^{n} \widehat{\beta}_1 \left(X_i - \overline{X} \right)^2 = 0
\end{aligned}
$$

Bivariate regression: deriving the OLS slope (continued)

$$\Leftrightarrow \quad \sum_{i=1}^{n} \widehat{\beta}_1 \left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

$$\Leftrightarrow \quad \widehat{\beta}_1 \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)$$

$$\Leftrightarrow \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

Back to main slides

Proof of OLS unbiasedness under assumptions 1-4

$$E\left(\widehat{\beta}_1\right) \stackrel{?}{=} \beta_1 \qquad \text{(proof)}$$

Step 1: Replacing $\widehat{\beta}$ with its OLS definition:

$$E\left(\widehat{\beta}_1\right) = E\left(\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)$$

This step is only valid if **assumption 4** is met:
$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \neq 0$, which for the multivariate case implies no perfect (multi)collinearity

Proof of OLS unbiasedness under assumptions 1-4

$$E\left(\widehat{\beta}_1\right) = E\left(\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i-1}^{n}\left(X_i - \overline{X}\right)^2}\right)$$

Step 2: Using that $\left(Y_i - \overline{Y}\right)$ is $\left[\beta_1\left(X_i - \overline{X}\right) + \left(\varepsilon_i - \overline{\varepsilon}\right)\right]$:

$$E\left(\widehat{\beta}_1\right) = E\left(\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left[\beta_1\left(X_i - \overline{X}\right) + \left(\varepsilon_i - \overline{\varepsilon}\right)\right]}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)$$

This step is only valid if **assumption 1** is met: the population model is linear in parameters (and the error term is additive).

Proof of OLS unbiasedness under assumptions 1-4

We can rewrite the expression (without making any further assumptions):

$$
\begin{aligned}
E\left(\widehat{\beta}_1\right) &= E\left(\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left[\beta_1\left(X_i-\overline{X}\right)+\left(\varepsilon_i-\overline{\varepsilon}\right)\right]}{\sum_{i-1}^{n}\left(X_i-\overline{X}\right)^2}\right) \\
&= E\left(\frac{\sum_{i=1}^{n}\beta_1\left(X_i-\overline{X}\right)^2+\left(X_i-\overline{X}\right)\left(\varepsilon_i-\overline{\varepsilon}\right)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\right) \\
&= \beta_1+E\left(\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)\left(\varepsilon_i-\overline{\varepsilon}\right)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\right)
\end{aligned}
$$

Proof of OLS unbiasedness under assumptions 1-4

$$E\left(\widehat{\beta}_1\right) = \beta_1 + E\left(\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(\varepsilon_i - \overline{\varepsilon}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right)$$

Step 3: Simplifying this to:

$$E\left(\widehat{\beta}_1\right) = \beta_1$$

This step is only valid if:

$$E\left(\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(\varepsilon_i - \overline{\varepsilon}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}\right) = 0$$

and for this, **assumptions 2 and 3** have to be met: $E(\varepsilon_i|X) = 0$.

Back to main slides