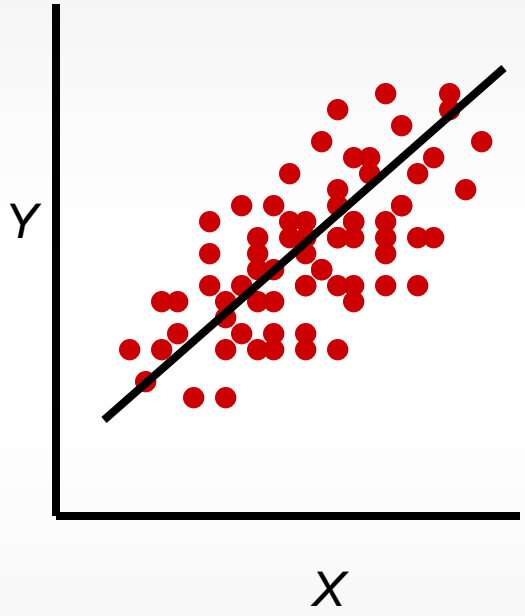
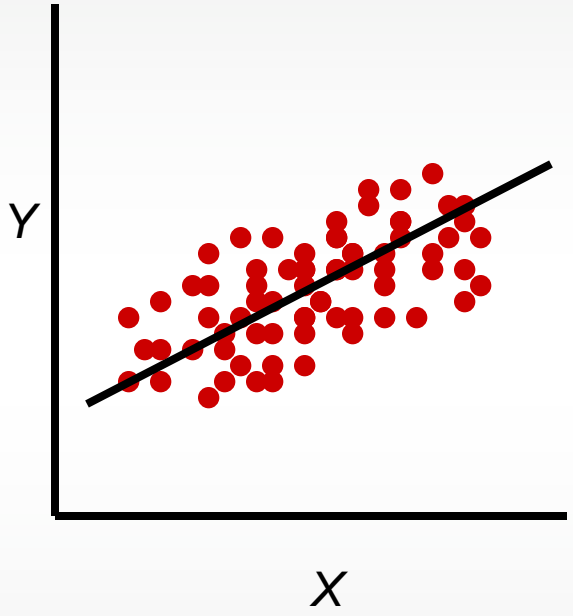
1. Strength of Relationships

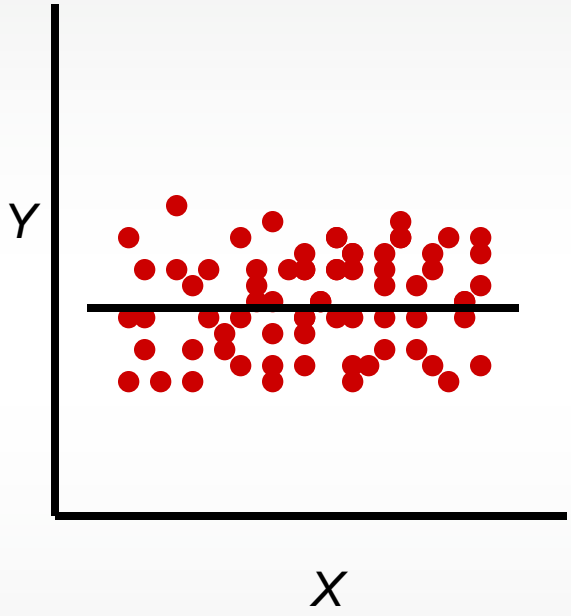
- Which graph represents the "strongest" relationship between X and Y?
- Describe the relationship in Graph C.



Graph A



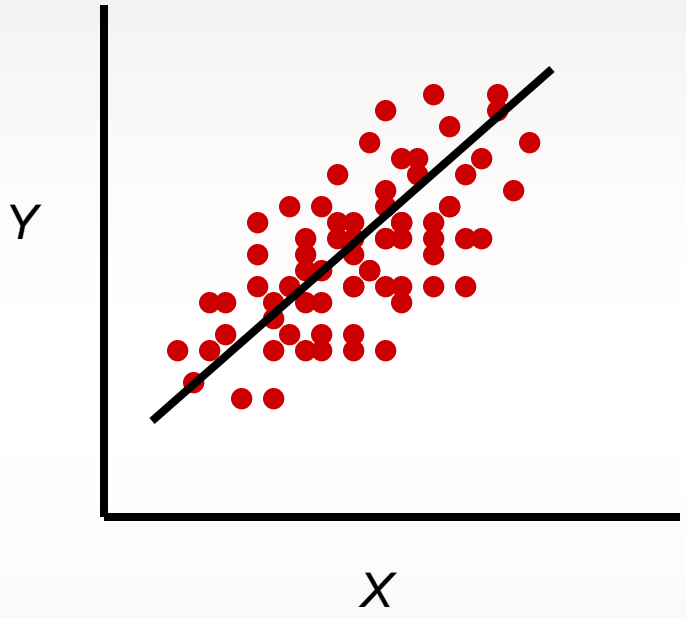
Graph B



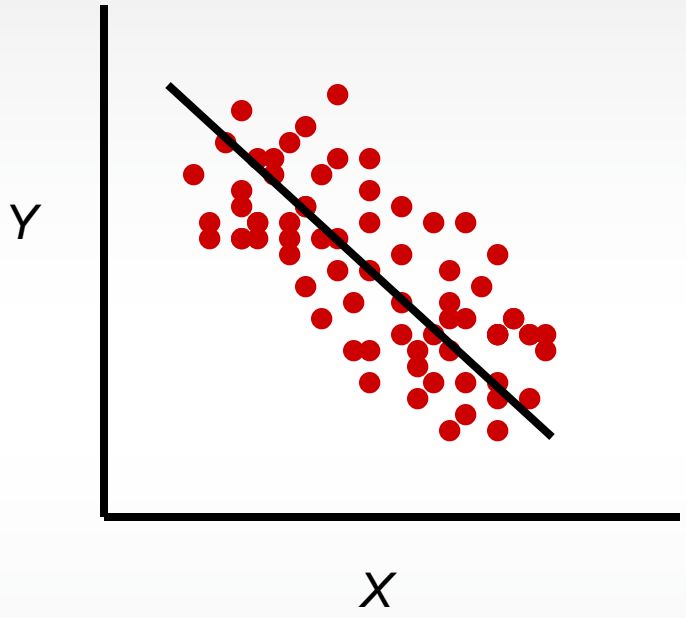
Graph C

Aside: Direction of Relationship

Positive or negative relationship?



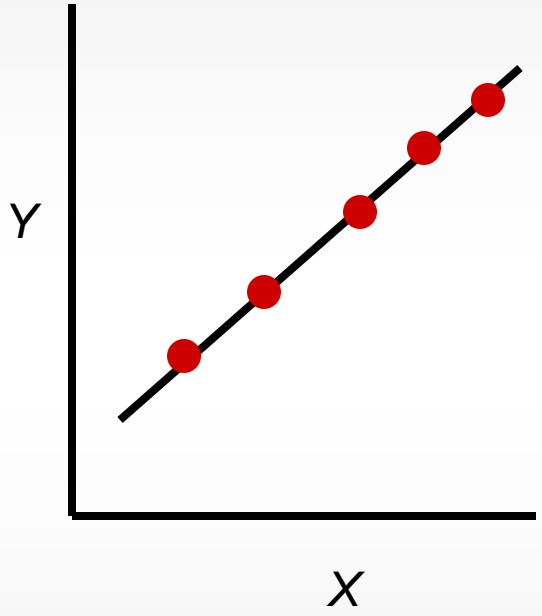
Graph A



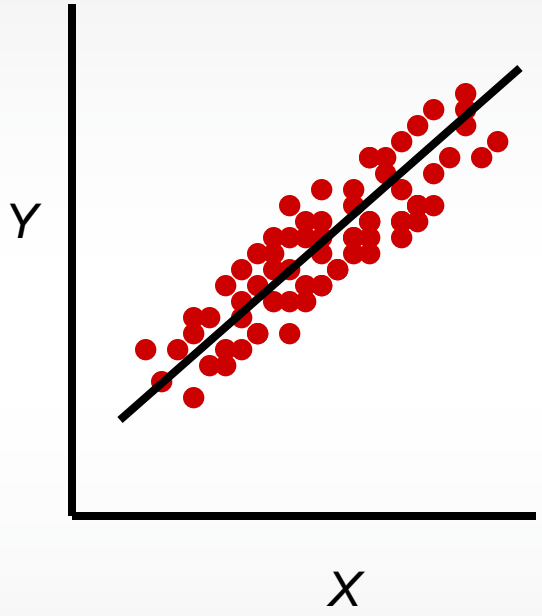
Graph B

2. Goodness of Fit

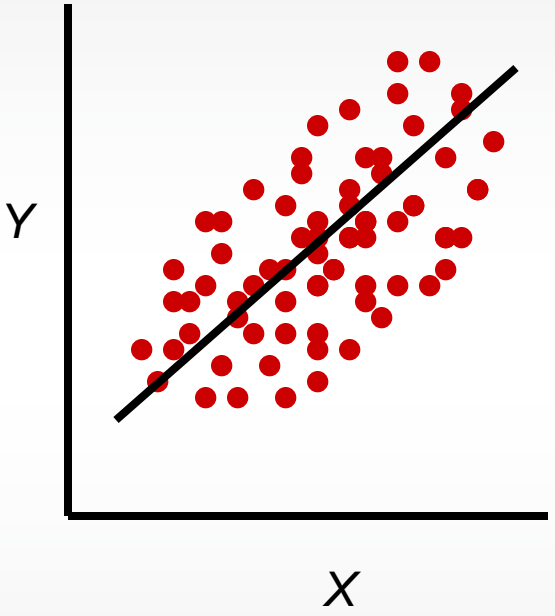
- Which regression provides the best "fit" of the data?
- Describe Graph A.



Graph A



Graph B



Graph C

Population Regression and Causality in Linear Regression

- **When is there a causal interpretation?**
- **Conditional mean independence assumption**

$$E(u|x) = 0$$

The explanatory variable must not contain information about the mean of the unobserved factors; related to **endogeneity/exogeneity**

- **Example: wage equation**

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

e.g. intelligence …

The conditional mean independence assumption is unlikely to hold because individuals with more education will also be more intelligent on average.

Population Regression

- **Population regression function (PFR)**
  - The conditional mean independence assumption implies that

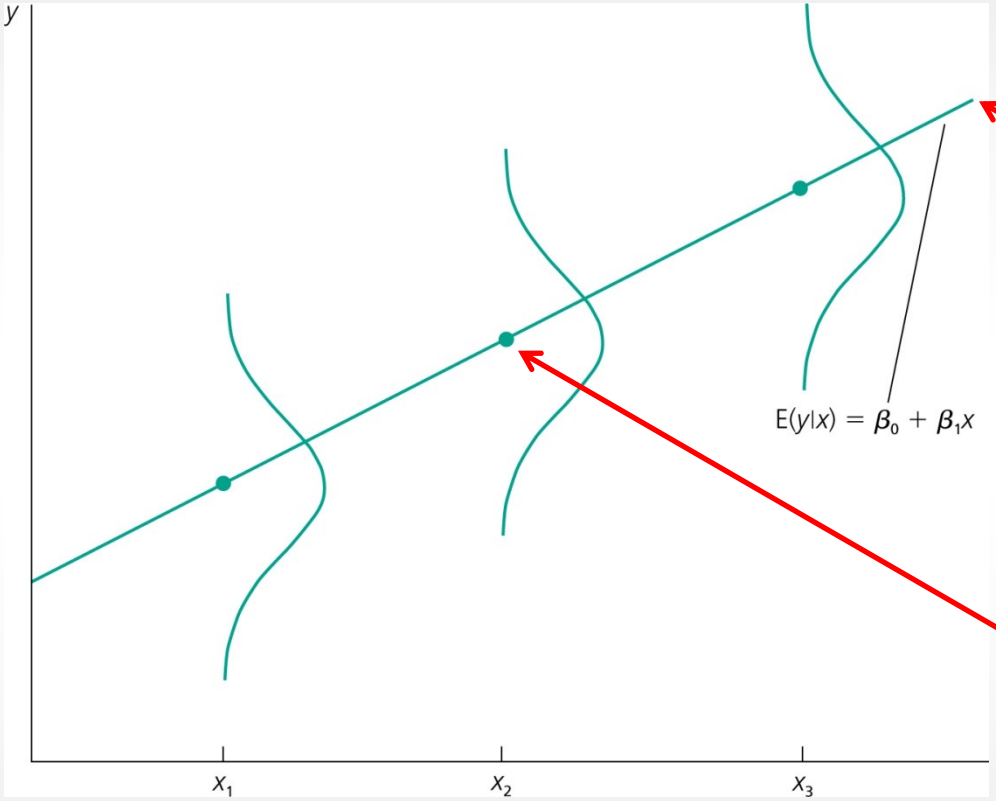
$$E(y|x) = E(\beta_0 + \beta_1 x + u|x)$$

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

$$= \beta_0 + \beta_1 x$$

  - This means that the average value of the dependent variable can be expressed as a linear function of the explanatory variable

Population Regression



Population regression function

For individuals with $x = x_2$, the average value of y is $\beta_0 + \beta_1 x$

Simple Linear Regression

Simple Regression Equation (scalar approach):

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

y_i = dependent variable; i subscripts units of analysis from 1, 2, 3, …n (where n = sample size)

x_i = independent variable for ith unit of analysis

β_0 = intercept (or y-intercept)

β_1 = slope

u_i = *residual* (estimate of error) for the ith unit of analysis

Simple Linear Regression

- **Properties of OLS on any sample of data**
- **Fitted values and residuals**

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Fitted or predicted values

$$\hat{u}_i = y_i - \hat{y}_i$$

Deviations from regression line (= residuals)

- **Algebraic properties of OLS regression**

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

Deviations from regression line sum up to zero

$$\sum_{i=1}^{n} x_i \hat{u}_i = 0$$

Correlation between deviations and regressors is zero

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

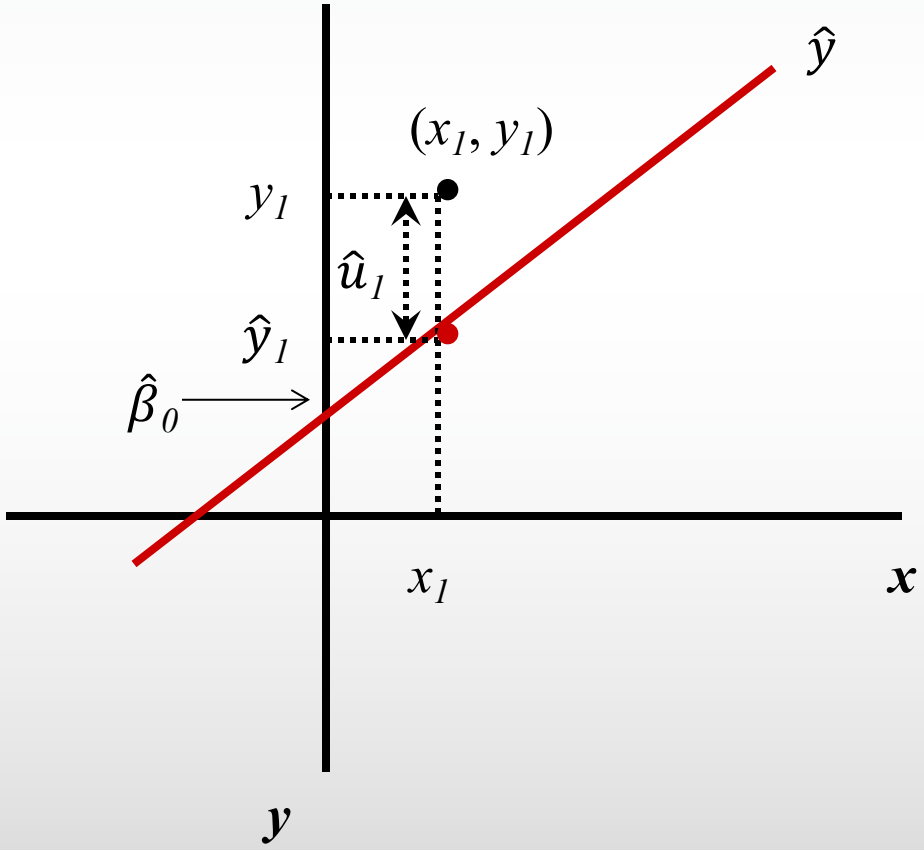
Sample averages of y and x lie on regression line

Simple Linear Regression

(1) $y_i = \beta_0 + \beta_1 x_{1i} + u_i$

(2) $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$

(3) $u_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i})$

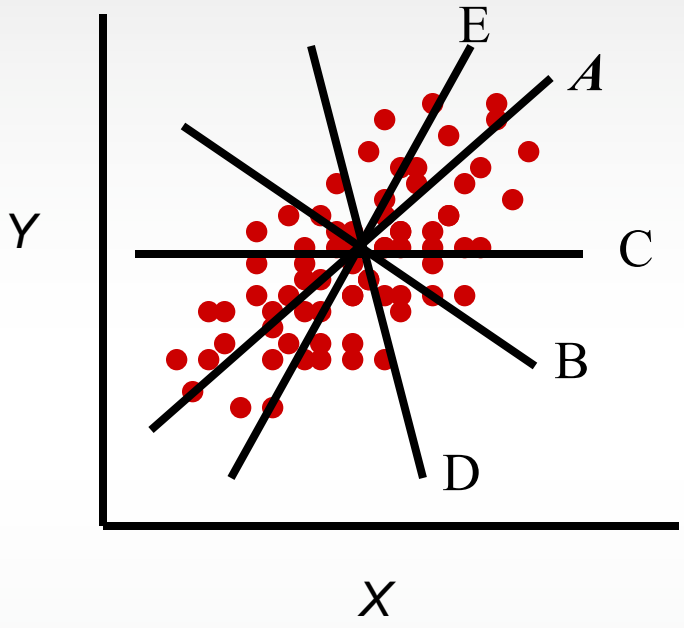


Finding the Line of Best Fit

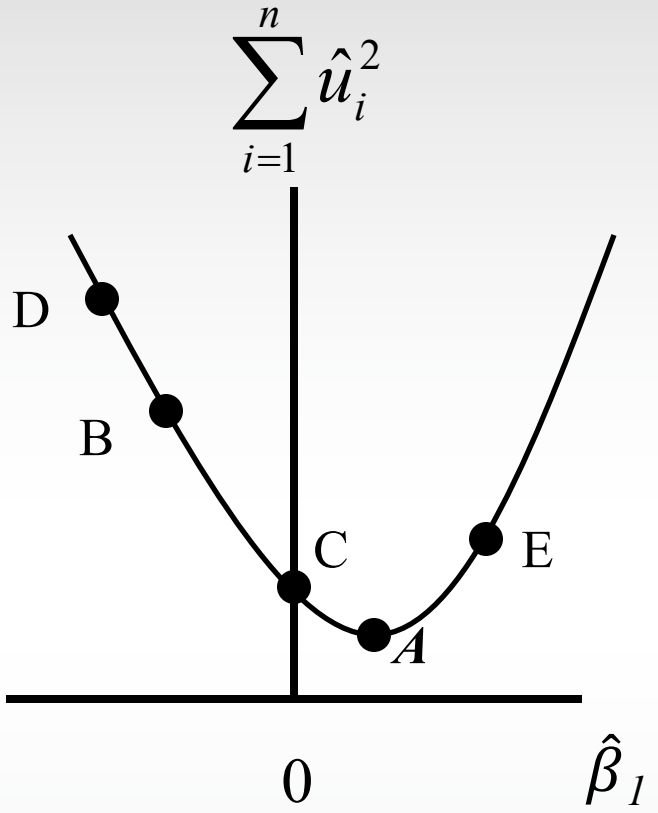
- Mission: Find the line, out of all possible candidates, that *minimizes* the degree of error.
- How?
  - Sum of residuals? Nonsense, because large negative residuals could offset positive ones.
  - Sum of *squared* residuals? Yes, gets around issue discussed above.
- Key criterion underlying least squares regression:

  Find the regression line that minimizes the sum of squared residuals.
- Notation: $\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = SSR$

Finding the Line of Best Fit



5 candidates for "line of best fit" that regress Y on X



Sum of squared errors, or residual sum of squares (SSR), as a function of the slopes from 5 candidate lines. There is **ONE** line of best fit that minimizes SSR (***LINE A***)

Finding the Line of Best Fit

- Next mission: Find parameter estimates for β_0 and β_1 that minimize the residual sum of squares.
  - Previous graph: this is a differential calculus problem.
  - Partial derivatives of residual sum of squares with respect to β_0 and β_1.
- See *Handout 1* (I'll post on Blackboard)

OLS Estimates for Simple Regression

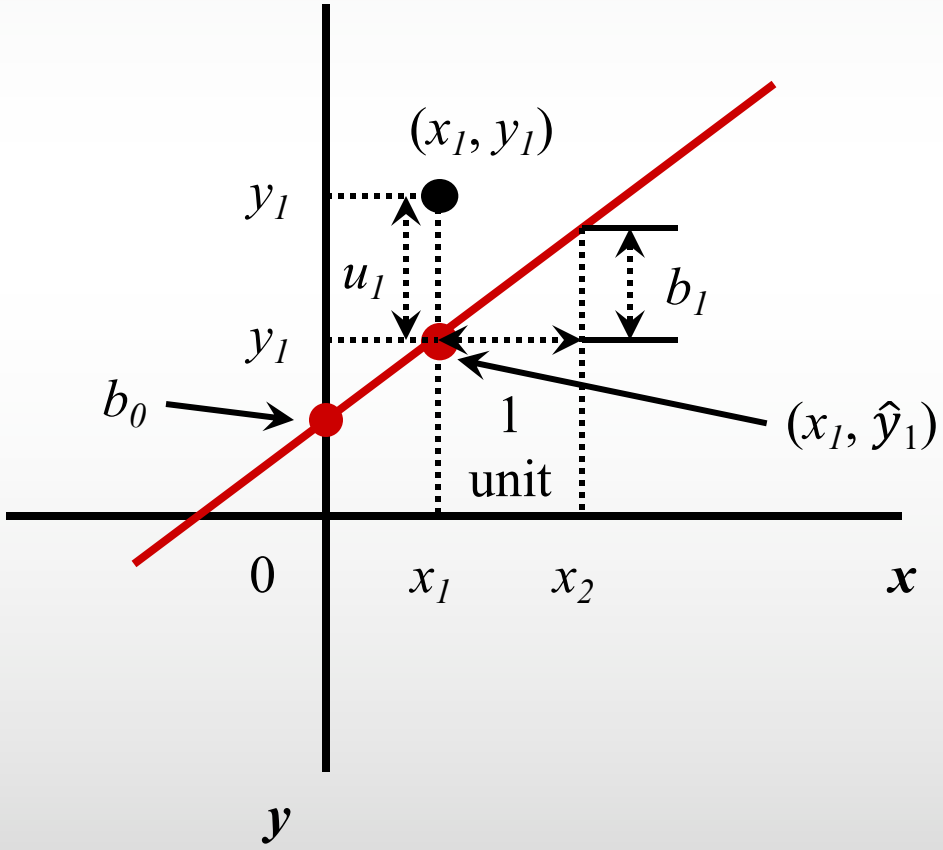
$$\hat{\beta}_1 = \frac{\sum (X_{1i} - \bar{X})(Y_i - \bar{Y})}{\sum (X_{1i} - \bar{X})^2} = \frac{Cov(X_1, Y)}{Var(X_1)} = r_{Y,X_1} \frac{s_Y}{s_{X_1}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- What makes $\hat{\beta}_1$ large or small?

Meaning of Parameters

- Interpreting b_1: *For a 1-unit increase in x, Y increases or decreases by b_1 units.*
- Interpreting b_0: $\hat{y}$ *(prediction of y) when X=0.*



Exercise

- Think of how you would estimate an OLS regression model with pencil and paper.

ID	Y_i	X_i	$\bar{Y}$	$\bar{X}$	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^* (Y_i - \bar{Y})$
1	10	2						
2	4	3						
3	5	0						
4	12	10						
5	20	8						
						$\sum_{i=1}^{n}$		

Exercise

- Let's calculate the *standard error of the regression* (SER)
- What other information do we need?

ID	Y_i	X_i	$\bar{Y}$	$\bar{X}$	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^* (Y_i - \bar{Y})$
1	10	2	*10.2*	*4.6*	*-2.6*	*-0.2*	*6.76*	*0.52*
2	4	3	*10.2*	*4.6*	*-1.6*	*-6.2*	*2.56*	*9.92*
3	5	0	*10.2*	*4.6*	*-4.6*	*-5.2*	*21.16*	*23.92*
4	12	10	*10.2*	*4.6*	*5.4*	*1.8*	*29.16*	*9.72*
5	20	8	*10.2*	*4.6*	*3.4*	*9.8*	*11.56*	*33.32*
						$\sum_{i=1}^{n}$	***71.2***	***77.4***

$B_1 = 77.4/71.2 = 1.087$

$B_0 = 10.2 - 1.087*4.6 = 5.20$

Goodness of Fit: R^2

- **Measures of Variation**

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

Total sum of squares, represents total variation in dependent variable

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

Explained sum of squares, represents variation explained by regression

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2$$

Residual sum of squares, represents variation not explained by regression

R^2

- **Decomposition of total variation**

$$SST = SSE + SSR$$

Total variation — Explained part — Unexplained part

- **Goodness-of-fit measure (R-squared)**

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

R-squared measures the fraction of the total variation that is explained by the regression

R^2

- SST serves as baseline error. R^2 tells us the proportional reduction in error, from a baseline (SST) to our specified model (SSE).

- Relation to correlation:
  - If b_1 is positive: $r = +\sqrt{R^2}$
  - If b_2 is negative: $r = -\sqrt{R^2}$

$$r = \frac{Cov(x, y)}{s_x s_y}$$

Setup for Regression Assumptions

- **Expected values and variances of the OLS estimators**
- **The estimated regression coefficients are estimates from *one* random sample**

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Data is random and depends on particular sample that has been drawn

- **The question is what the estimators will estimate *on average* (given repeated sampling) and how large their variability in repeated samples is**

$$E(\hat{\beta}_0) = ?, \; E(\hat{\beta}_1) = ?$$

Bias or Unbiasedness

$$Var(\hat{\beta}_0) = ?, \; Var(\hat{\beta}_1) = ?$$

Efficiency; random sampling error

Regression Assumptions (Gauss-Markov Assumptions)

- **Standard assumptions for the linear regression model**
- **Assumption SLR.1 (Linear in parameters)**

$$y = \beta_0 + \beta_1 x + u$$

In the population, the relationship between y and x is linear

- **Assumption SLR.2 (Random sampling)**

$$\{(x_i, y_i) : \ i = 1, \ldots n\}$$

The data is a random sample drawn from the population

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

Each data point therefore follows the population equation

Regression Assumptions

- **Assumption SLR.3 (Sample variation in explanatory variable)**

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$$

The values of the explanatory variables are not all the same (otherwise it would be impossible to study how different values of the explanatory variable lead to different values of the dependent variable)

- **Assumption SLR.4 (Zero conditional mean)**

$$E(u_i|x_i) = 0$$

The value of the explanatory variable must contain no information about the mean of the unobserved factors

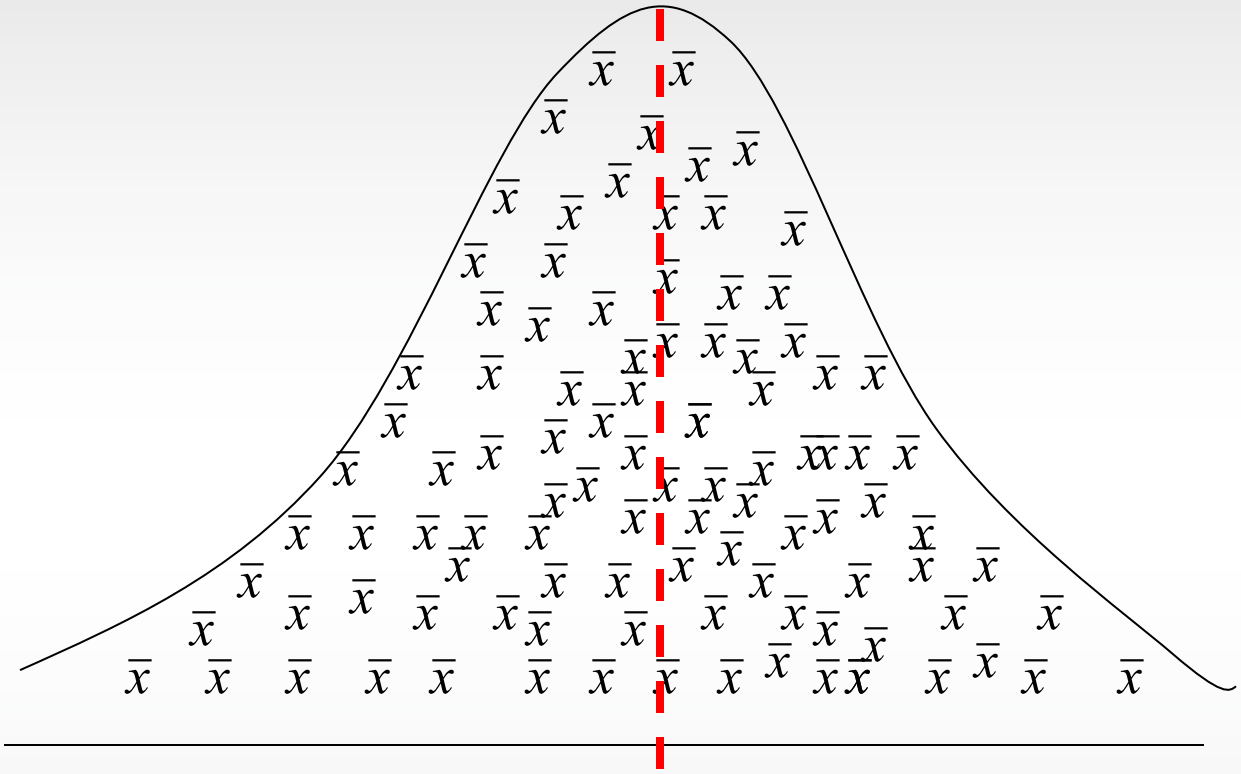
OLS Properties: Unbiasedness

- **Theorem 2.1 (Unbiasedness of OLS)**

$$SLR.1 - SLR.4 \quad \Rightarrow \quad E(\hat{\beta}_0) = \beta_0, \; E(\hat{\beta}_1) = \beta_1$$

- **Interpretation of unbiasedness**
  - The estimated coefficients may be smaller or larger, depending on the sample that is the result of a random draw
  - However, on average, they will be equal to the values that characterize the true relationship between y and x in the population
  - *On average* means if sampling was repeated, i.e. if drawing the random sample und doing the estimation was repeated many times
  - In a given sample, estimates may differ considerably from true values

Illustrating Unbiasedness



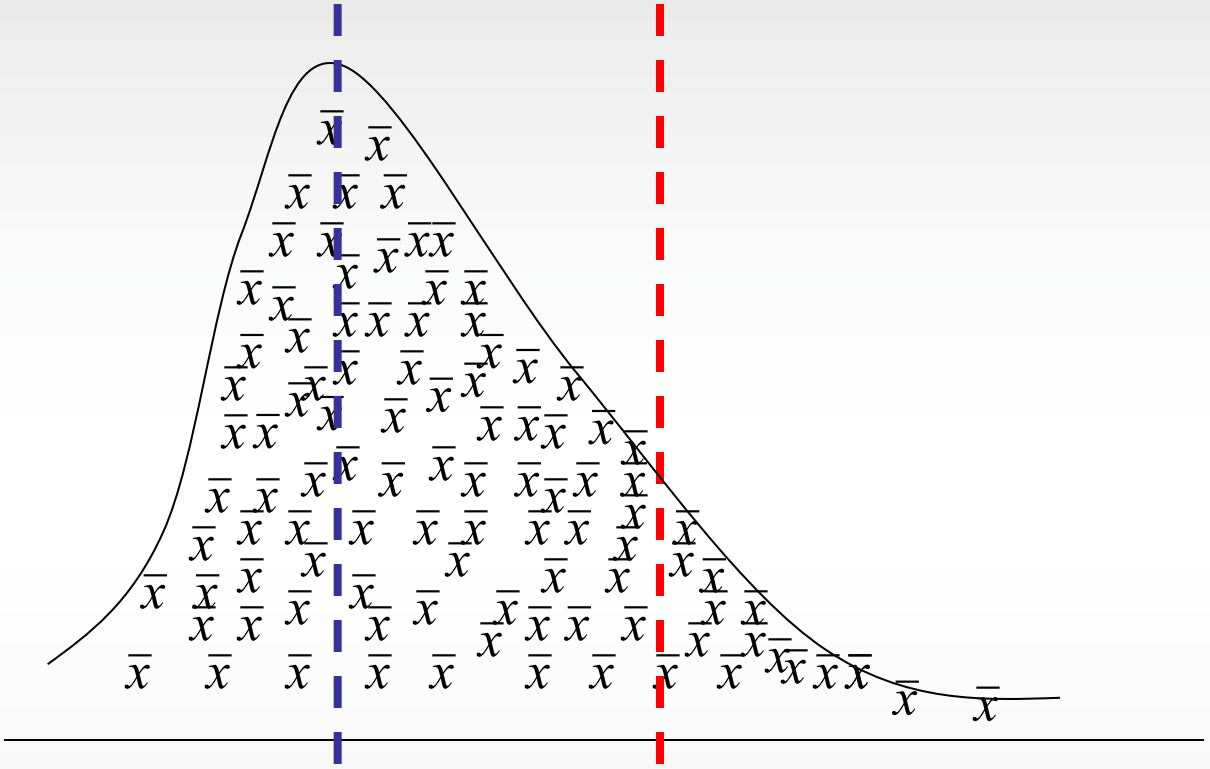
μ = true population mean

Central limit theorem

Unbiasedness*: mean of hypothetical distribution of sample means is the "true" mean.*

•What does dispersion mean in this distribution?

Illustrating Bias



μ_B = *biased* average of sample means

μ = true population mean

Bias*: mean of hypothetical distribution of sample means DEVIATES from the "true" mean.*

- *Systematic* factors inducing bias.

Variance of OLS Estimator

- **Variances of the OLS estimators**
  - Depending on the sample, the estimates will be nearer or farther away from the true population values
  - How far can we expect our estimates to be away from the true population values on average (= sampling variability)?
  - Sampling variability is measured by the estimator‘s variances

$$Var(\hat{\beta}_0), \;\; Var(\hat{\beta}_1)$$

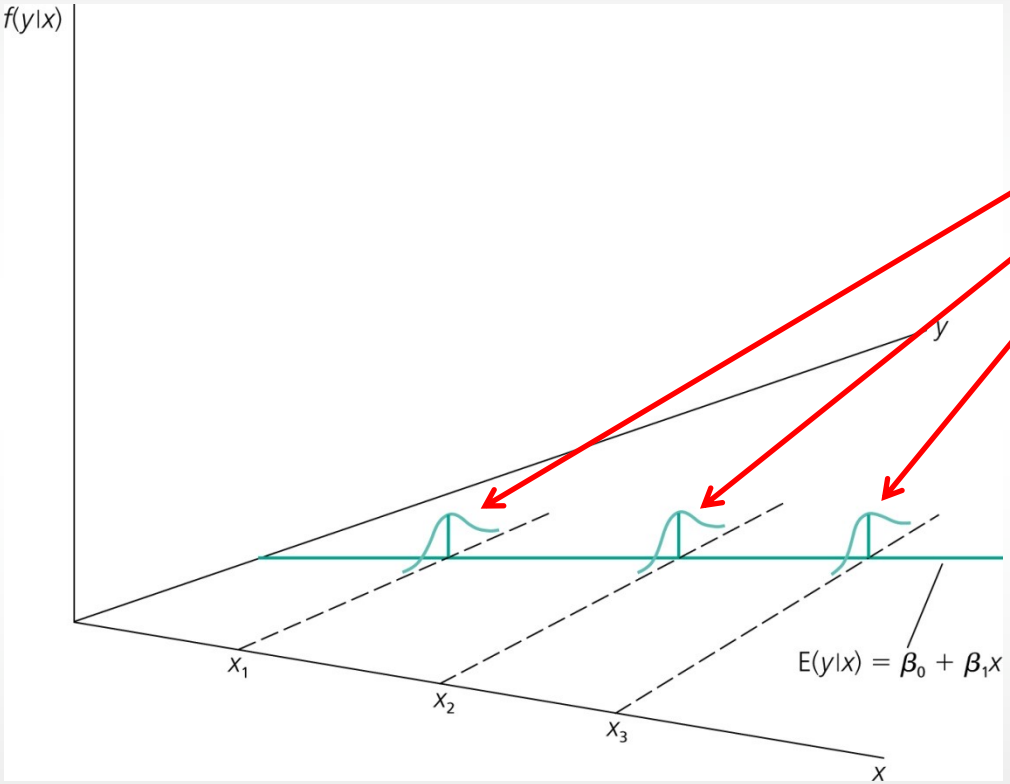
- **Assumption SLR.5 (Homoskedasticity)**

$$Var(u_i|x_i) = \sigma^2$$

The value of the explanatory variable must contain no information about the variability of the unobserved factors

Homoskedasticity

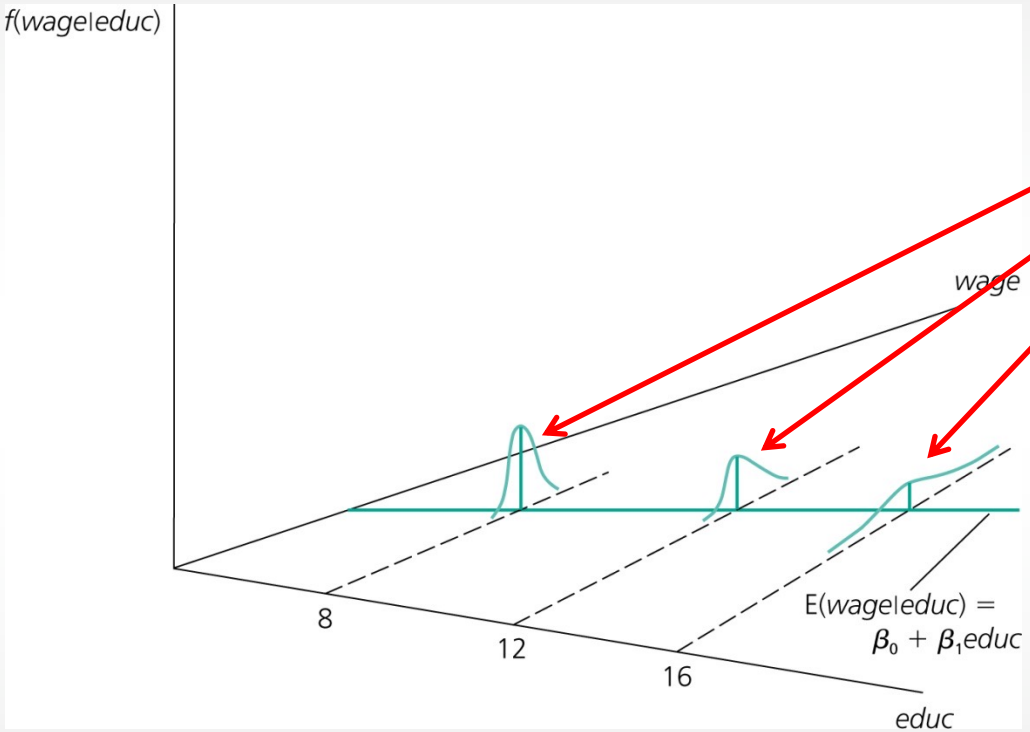
- **Graphical illustration of homoskedasticity**



The variability of the unobserved influences does not dependent on the value of the explanatory variable

Heteroskedasticity

- **An example for heteroskedasticity: Wage and education**



The variance of the unobserved influences of wages increases with the level of education

Sampling Variance

- **Theorem 2.2 (Variances of OLS estimators)**

  Under assumptions SLR.1 – SLR.5:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}$$

$$Var(\hat{\beta}_0) = \frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{SST_x}$$

- **Conclusion:**
  - The sampling variability of the estimated regression coefficients will be higher the larger the variability of the unobserved factors; and lower the higher the variation in the explanatory variable

Error Variance

- **Estimating the error variance**

$$Var(u_i|x_i) = \sigma^2 = Var(u_i)$$

The variance of u does not depend on x, i.e. is equal to the unconditional variance

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2$$

One could estimate the variance of the errors by calculating the variance of the residuals in the sample; unfortunately this estimate would be **biased**

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$$

An **unbiased estimate** of the error variance can be obtained by substracting the number of estimated regression coefficients from the number of observations

Standard Error of the Regression

- Simply the square root of the *error variance*.
- Calculated as (k = number of parameters in model, *including the intercept*):

$$SER = \sqrt{\frac{\sum_{i=1}^{n} u_i^2}{n-k}} = \sqrt{\frac{SSR}{n-k}}$$

- Using rules of standard deviation and the normal distribution: 2/3 of the residuals are $\pm$ 1 S_E from zero. 95% are $\pm$ 2 S_E's from zero. [Note: the mean of u_i is zero.]

Calculating Error Variance and SER

- What other information do we need?

ID	Y_i	X_i	$\bar{Y}$	$\bar{X}$	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(X_i - \bar{X})^* (Y_i - \bar{Y})$
1	10	2	*10.2*	*4.6*	*-2.6*	*-0.2*	*6.76*	*0.52*
2	4	3	*10.2*	*4.6*	*-1.6*	*-6.2*	*2.56*	*9.92*
3	5	0	*10.2*	*4.6*	*-4.6*	*-5.2*	*21.16*	*23.92*
4	12	10	*10.2*	*4.6*	*5.4*	*1.8*	*29.16*	*9.72*
5	20	8	*10.2*	*4.6*	*3.4*	*9.8*	*11.56*	*33.32*
						$\sum_{i=1}^{n}$	***71.2***	***77.4***

$B_1 = 77.4/71.2 = 1.087$

$B_0 = 10.2 - 1.087*4.6 = 5.20$

Standard Errors for OLS Estimates

- **Theorem 2.3 (Unbiasedness of the error variance)**

$$SLR.1 - SLR.5 \quad \Rightarrow \quad E(\hat{\sigma}^2) = \sigma^2$$

- **Calculation of standard errors for regression coefficients**

$$se(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)} = \sqrt{\hat{\sigma}^2 / SST_x}$$

$$se(\hat{\beta}_0) = \sqrt{\widehat{Var}(\hat{\beta}_0)} = \sqrt{\hat{\sigma}^2 n^{-1} \sum_{i=1}^{n} x_i^2 / SST_x}$$

Plug in $\hat{\sigma}^2$ for the unknown σ^2

The estimated standard deviations of the regression coefficients are called „standard errors“. They measure how precisely the regression coefficients are estimated.