L14

Basic Statistics for Economists

Spring 2020

Department of Statistics

Stockholm University

Summary from last time

- A parameter or the difference between two parameters
  - e.g. μ or P e.g. $\mu_A - \mu_B$ or $P_A - P_B$
- Two opposite claims, hypotheses : H_0 and H_1
- Start with what is given and what has to be assumed
  - e.g. normal distribution, known variance, sample size, indep.
- Identify the correct test variable, Z or t_{n-1}
- If not given, decide significance level α
- Find the critical values
  - z_α or $-z_\alpha$ (1-sided) or $\pm z_{\alpha/2}$ (2-sided); or the equivalent t-test
- Calculate and make decision: reject or not reject. **Interpret!**

Stockholm University

One parameter: μ_x

- When should we use z and when should we use t?

Sample size n	Distribution	Variance σ_X^2 known	Variance σ_X^2 unknown, use s_x^2
large, ≥ 30	normal	Z	**approx** $\Rightarrow Z$
	not normal	**CLT** $\Rightarrow Z$	**CLT + approx** $\Rightarrow Z$
small, < 30	normal	Z	t
	not normal	**not included**	**not included**

- In the cases marker blue above, where the sample is large and the variance unknown, one could argue that t should be used instead of z. This would give more conservative results, i.e. slightly smaller risk α; using t adjusts for the slightly larger uncertainty that follows from estimating of the variance.

Two parameters: $\mu_x - \mu_y$

- When should we use z and when should we use t?

Sample sizes n_x, n_y	Distribution	Variances σ_X^2 and σ_Y^2		
		known	unknown: s_x^2, s_y^2	
large, ≥ 30	normal	Z	**approx** $\Rightarrow Z$	
	not normal	**CLT** $\Rightarrow Z$	**CLT + approx** $\Rightarrow Z$	
small, < 30	normal	Z	$\sigma_X^2 = \sigma_Y^2$: t	$\sigma_X^2 \neq \sigma_Y^2$: **not incl.**
	not normal	**not included**	**not included**	

- In the cases marked blue above, where the sample is large and the variances unknown, one could argue that t should be used instead of z. This would give more conservative results, i.e. slightly smaller risk α; using t adjusts for the slightly larger uncertainty that follows from estimating of the variances.

Significance and power, error types

Consequence & probability	Actual situation	
Decision	H_0 true	H_0 false
Acceptera H_0	Correct decision $(1-\alpha)$	Type II error (β)
Reject H_0	Type I error (α)	Correct decision $(1-\beta)$

- Probability α = **significance level** (cf. *p*-value)
- Probability $1-\beta$ = **power** (sv. *testets styrka*)

Stockholm University

Using *p*-values

If the ***p*-value** is **small**, e.g. < 5%

⇒ the null hypothesis is **rejected** on 5% significance level

If the ***p*-value** is **large**, e.g. > 5%

⇒ the null hypothesis **cannot be rejected** on the 5% level

Significance level

- "**statistically significant**" (sv. "*statistiskt säkerställt*"), is synonymous with a **low *p*-value**, typically less than 5%.

Stockholm University

Compare μ_0 against a confidence interval

- Suppose you are going to test a null hypothesis against a **double-sided** alternative at significance level $\boldsymbol{\alpha = 0,05}$

$$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu \neq \mu_0$$

- Your friend has already calculated a $\mathbf{100(1-\alpha) = 95\%}$ CI for $\boldsymbol{\mu}$ based on the same data that you are going to use.

If your μ_0 **lies outside the interval**

$\Rightarrow$ **reject** the null hypothesis

If your μ_0 **lies within the interval**

$\Rightarrow$ **do not reject** null hypothesis

Proportions are an exception!

- When we test $H_0: P = P_0$ we get (according to CLT):

Test variable for P: $$Z = \frac{\hat{p} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} \rightarrow N(0,1)$$

standard error under H_0

- But when we estimate a CI of P:

$100(1-\alpha)\ \%$ KI for P: $$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

estimated standard error

- Since we calculate SE differently, based on P_0 in the test case, or based on $\hat{p}$ in the CI case, the results may differ!

Stockholm University

DIY Example

Sample with $n = 30$ and $\hat{p} = 0.10$, test if $P = 0.25$

- 5% test $H_0: P = 0.25$ v. $H_0: P \neq 0.25$, reject H_0 if $|z_{obs}| > 1.96$

  $$|z_{obs}| = \left|\frac{\hat{p}-P_0}{\sqrt{P_0(1-P_0)/n}}\right| = \left|\frac{0.1-0.25}{\sqrt{0.25 \cdot 0.75/30}}\right| = 1.6444 < 1.96 \;\Rightarrow \text{do not reject } H_0$$

- 95% CI for P:

  $$\hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.10 \pm 1.96 \cdot \sqrt{\frac{0.1 \cdot 0.9}{30}} \;\Rightarrow\; (0.0037, 0.2363)$$

  Since $P = 0.25$ is not covered by the CI we might wrongly conclude that H_0 should be rejected

Today

Introduction to **Linear Regression**

(sections **11.1 – 11.3 in NCT)**

- Intro
- The model
- Estimation of the model parameters
- Predictions
- Residuals and residual variance

What are regression models?

Example:

- Children are born. Can we predict $\boldsymbol{Y}$ = body length as an adult?
  - Empirical rule
  - $\boldsymbol{Y \sim N(\mu_Y, \sigma_Y^2)}$; prediction $\boldsymbol{\mu_Y \pm 2\sigma_Y}$ captures ca 95 %

- The body length at birth varies between children. Could length at birth "explain" the variation in adult body lengths?

- Can other factors affect (explain) body length as an adult?
  - $\boldsymbol{Y = \text{function}(\text{BLNewborn}, \text{Sex}, \text{MomLng}, \text{DadLng}, \ldots)}$

Regression modell!

Stockholm University

A normal distributed variable Y

- Suppose that $\boldsymbol{Y}$ is a normal distributed random variable:

$$\boldsymbol{Y \sim N(\mu_Y, \sigma_Y^2)}$$

- This **model** can also be written like this:

$\boxed{\boldsymbol{Y = \mu_Y + \delta}}$ where $\boxed{\boldsymbol{\delta \sim N(0, \sigma_\delta^2)}}$ NOTE! $\boldsymbol{\mu_Y}$ is a ***constant***

- The random variable $\boldsymbol{\delta}$ is the random distance to $\boldsymbol{\mu_Y}$ or the ***deviation*** or ***error term*** (sv. *felterm*; Greek delta $\boldsymbol{\delta}$)

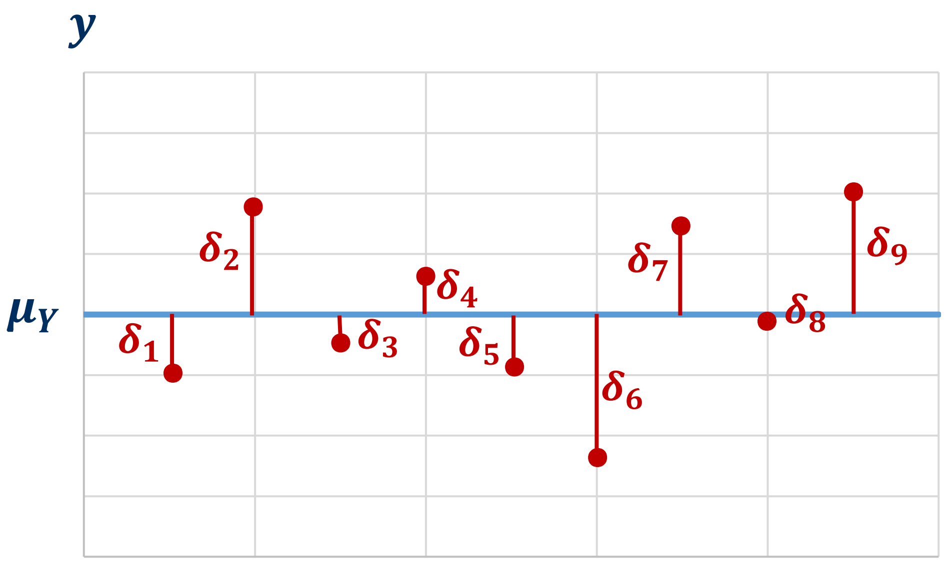
Expectation: $E(Y) = E(\boldsymbol{\mu_Y} + \delta) = \boldsymbol{\mu_Y} + \cancel{E(\delta)} = \boxed{\boldsymbol{\mu_Y}}$ $(E(\delta) = 0)$

Variance: $Var(Y) = Var(\cancel{\boldsymbol{\mu_Y}} + \delta) = Var(\delta) = \boxed{\boldsymbol{\sigma_\delta^2 = \sigma_Y^2}}$

Stockholm University

Illustration

- Each red dot represents an observed Y value



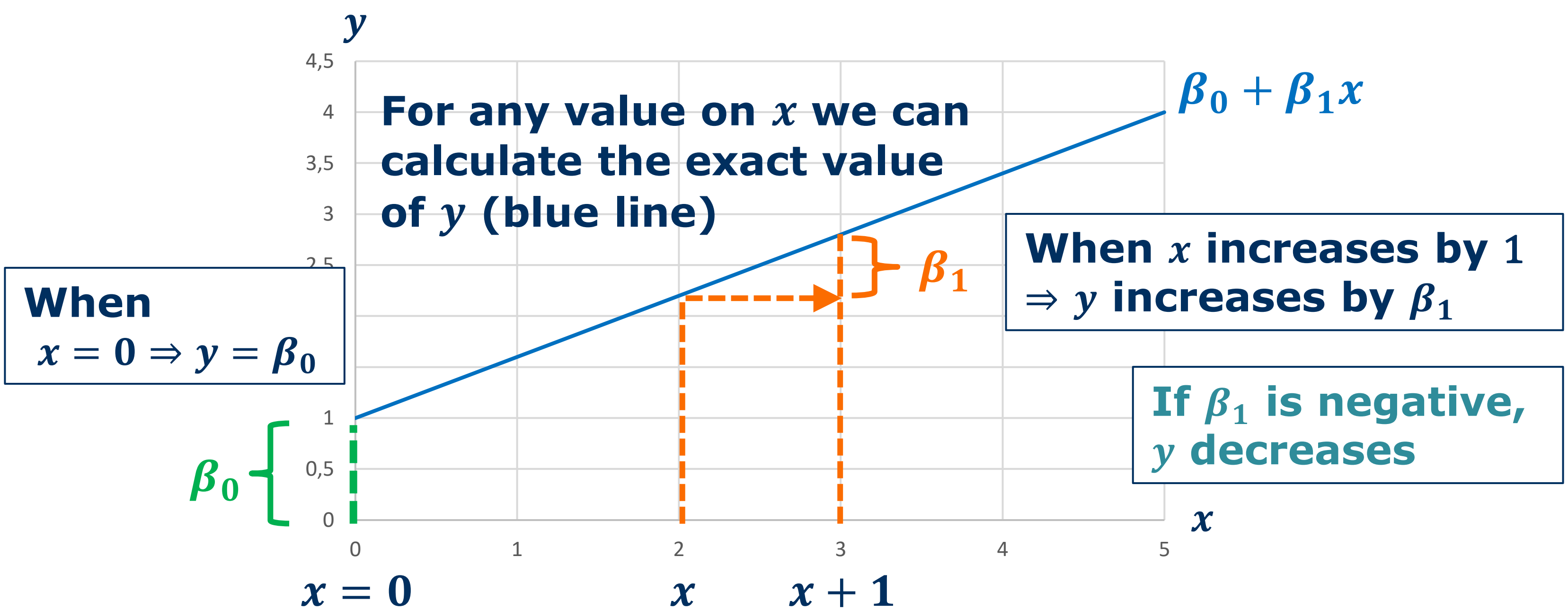
δ_i = distance from the different Y values to μ_Y

$$\boxed{\sigma_\delta^2 = \sigma_Y^2}$$

Stockholm University

Linear function

- $y = f(x) = \beta_0 + \beta_1 x$, the graph of $f(x)$ is a straight line



Stockholm University

Linear function - stochastic version

- $\underbrace{(Y|X = x)}_{\text{Fix } x} = f(x) + \varepsilon = \underbrace{\beta_0 + \beta_1 x}_{\text{Deterministic, linear}} + \underbrace{\varepsilon \quad \text{where} \quad \varepsilon \sim N(0, \sigma_\varepsilon^2)}_{\text{Stochastic, random error } \varepsilon}$

- ***Conditioning*** on the event $X = x \quad \Rightarrow \quad \beta_0 + \beta_1 x$ is a ***constant***

  Expectation: $\mu_{Y|X=x} = E(\beta_0 + \beta_1 x + \varepsilon) = \boxed{\beta_0 + \beta_1 x} + \cancel{E(\varepsilon)}\ (= 0)$

  ***Different** conditional expected values of Y, depends on x*

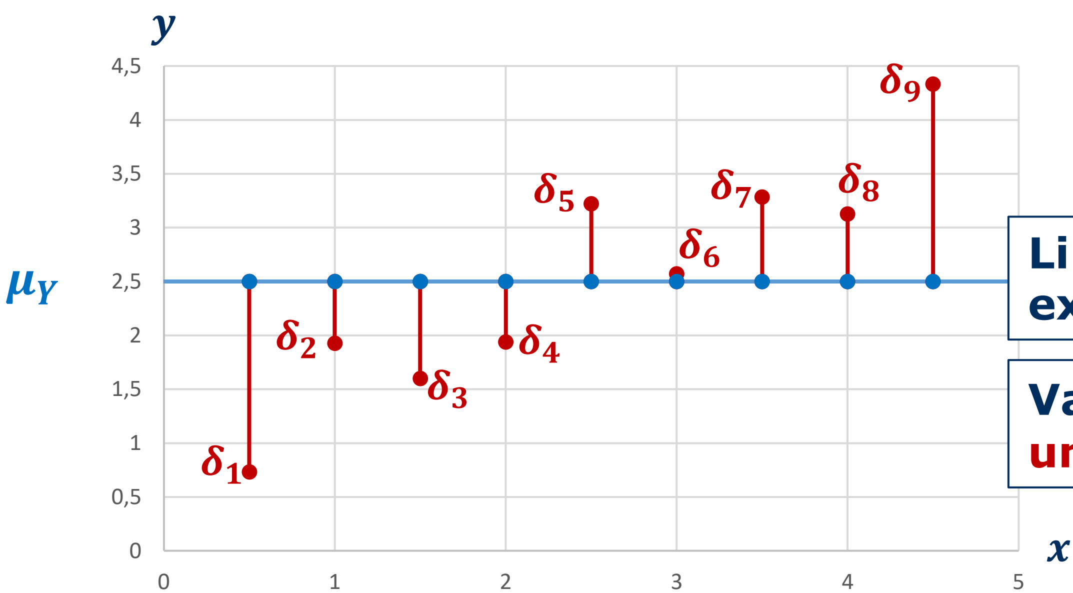
  Variance: $\sigma^2_{Y|X=x} = Var(\cancel{\beta_0 + \beta_1 x} + \varepsilon) = Var(\varepsilon) = \boxed{\sigma_\varepsilon^2 \neq \sigma_Y^2}$

  ***Same** conditional variance of Y regardless of the value of x*

Stockholm University

No linear function, unconditional on x

- The simple model $Y = \mu_Y + \delta$ where $\delta \sim N(0, \sigma_\delta^2)$ $\boxed{\sigma_\delta^2 = \sigma_Y^2}$



Line = μ_Y = **unconditional expectation of** Y

Variance = σ_Y^2 = σ_δ^2 = **unconditional variance of** Y

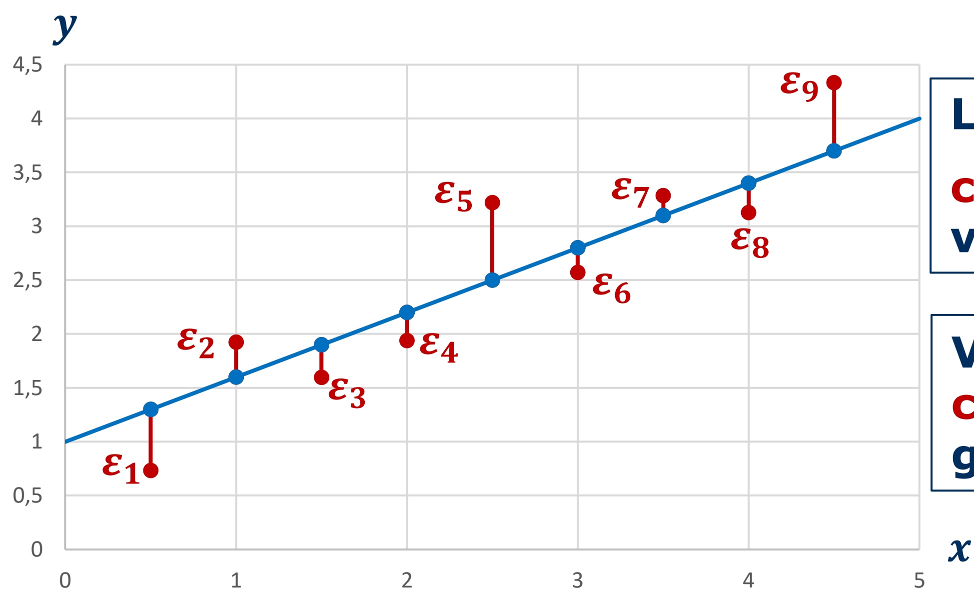
Distance from the dot to the to the line = error term $= \delta$

Stockholm University

Linear function - stochastic version, cont.

- $(Y|X=x) = \beta_0 + \beta_1 x + \varepsilon$ where $\varepsilon \sim N(0, \sigma_\varepsilon^2)$

$\boxed{\sigma_\varepsilon^2 \neq \sigma_Y^2}$



Line $= \beta_0 + \beta_1 x = \mu_{Y|X=x} =$ conditional expected value of Y given $X = x$

Variance $= \sigma_\varepsilon^2 =$ conditional variance of Y given $X = x$

Distance from the dot to the to the line = error term $= \varepsilon$

Stockholm University

Why do we do this?

- **Analysis of relationships**, study how $\boldsymbol{X}$ and $\boldsymbol{Y}$ vary together

- In general (from L1):
  - Descriptive ("*this is the observed relationship between X and Y*")
  - Explanatory, causality ("*the value of Y is this because X is this*")
  - Predictions ("*what happens with Y if X takes this value*")
  - Normative, prescriptive ("*do this to get this result*")

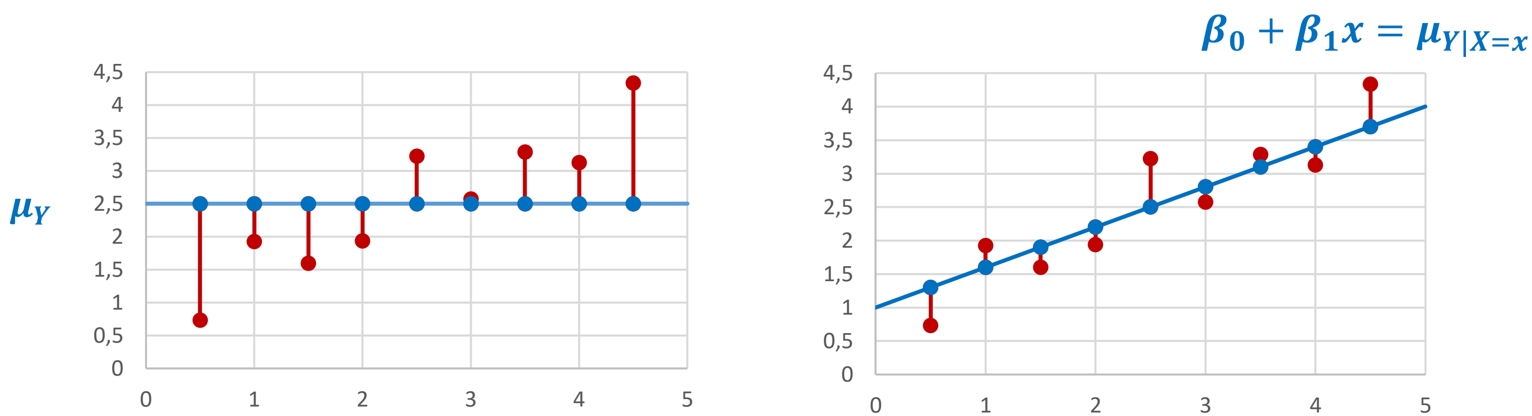
> By using information about a variable X and how it co-varies with another variable Y we can, by conditioning on a particular value $X = x$, make a better guess about the value of Y.

- **Better guess = increased precision = less variance**

Stockholm University

Gain in terms of decreased variance

- Var. of error terms, **unconditional** vs. **conditional** on $X = x$:



$$\boxed{\sigma_Y^2 \;>\; \sigma_{Y|X=x}^2 = \sigma_\varepsilon^2}$$

- Variance decreases = better precision = better predictions

Stockholm University

Interpretation of parameters

- $\boldsymbol{\beta_0}$ = **intercept**, where the regression line cuts the y axis
  - the **average** value of $\boldsymbol{Y}$ when $\boldsymbol{X} = \boldsymbol{0}$
  - estimated with $\boldsymbol{b_0}$

- $\boldsymbol{\beta_1}$ = the **slope** of the line, **the regression coefficient**
  - the **average increase** of $\boldsymbol{Y}$ when $\boldsymbol{X}$ increases 1 unit
  - estimated with $\boldsymbol{b_1}$

- $\boldsymbol{\mu_{Y|X=x}}$ = **conditional expectation** of $\boldsymbol{Y}$, conditional on $\boldsymbol{X = x}$
  - the **average** value of $\boldsymbol{Y}$ when $\boldsymbol{X = x}$
  - according to the model $\boldsymbol{\mu_{Y|X=x} = \beta_0 + \beta_1 x}$
  - estimated with $\boldsymbol{\hat{\mu}_{Y|X=x} = b_0 + b_1 x}$

Stockholm University

Interpretation of parameters, cont.

- $\boldsymbol{\varepsilon}$ = **error**, the distance from $\boldsymbol{Y}$ to the line $\boldsymbol{\mu_{Y|X=x} = \beta_0 + \beta_1 x}$
  - the difference between the actual value of $\boldsymbol{Y}$ and the conditional expected value of $\boldsymbol{Y}$
  - $\boldsymbol{\varepsilon = Y - \mu_{Y|X=x} = Y - \beta_0 - \beta_1 x}$
  - since we typically don't know the values of $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ we will never know the real values of the errors

- $\boldsymbol{\sigma_\varepsilon^2 = Var(\varepsilon)}$ = the **error variance**
  - the conditional variance of $\boldsymbol{Y}$ given $\boldsymbol{X = x}$
  - compare $\boldsymbol{\sigma_\varepsilon^2}$ to $\boldsymbol{\sigma_Y^2}$, the unconditional variance of $\boldsymbol{Y}$
  - we hope that $\boldsymbol{\sigma_\varepsilon^2 < \sigma_Y^2}$
  - estimated with $\boldsymbol{s_e^2}$

Stockholm University

Assumptions for linear regression model

1. $\boldsymbol{Y}$ is a ***linear*** function of $\boldsymbol{x}$ plus an error term $\boldsymbol{\varepsilon}$

$$\boldsymbol{Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i}$$

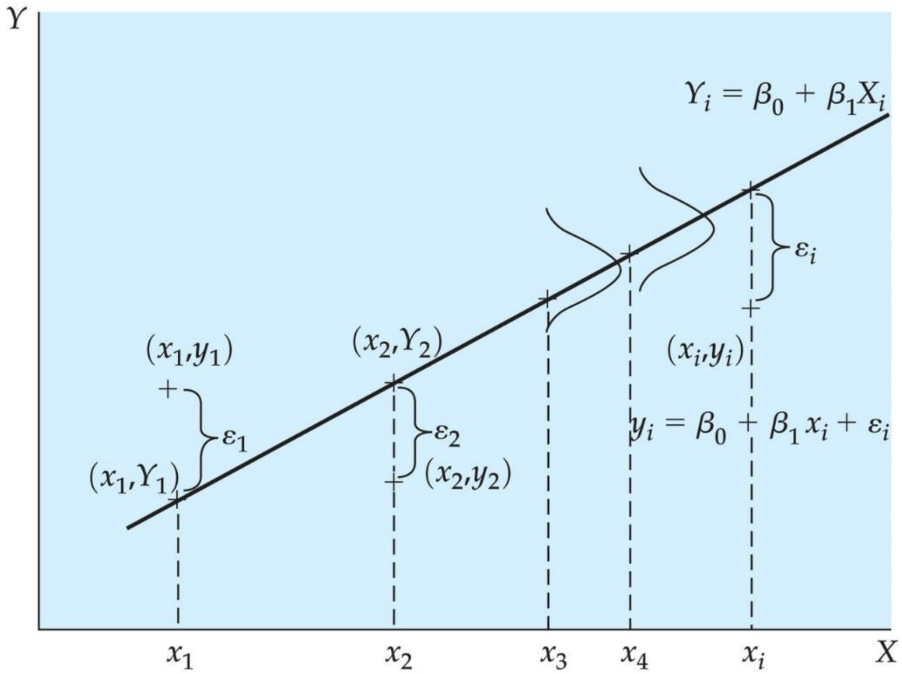
2. The $\boldsymbol{x_i}$-values are either constant (pre-determined) or they are realizations of a r.v. $\boldsymbol{X}$ that are ***independent*** of the errors $\boldsymbol{\varepsilon_i}$

3. The error terms $\boldsymbol{\varepsilon_i}$ are ***independent*** of each other

4. The error terms $\boldsymbol{\varepsilon_i}$ are ***normal distributed*** r.v.'s $\boldsymbol{\varepsilon_i \sim N(0, \sigma_\varepsilon^2)}$ expectation $\boldsymbol{E(\varepsilon_i) = 0}$ and **constant variance** $\boldsymbol{\sigma_\varepsilon^2}$
   - variance of $\boldsymbol{Y}$ is the same regardless of $\boldsymbol{x}$, this is called ***homoscedasticity***

Stockholm University

The conditional expectation – illustration



- Condition on $\boldsymbol{X = x}$

  fix the value of $\boldsymbol{x}$ and study the behavior of $\boldsymbol{Y}$ at that particular $\boldsymbol{x}$ value

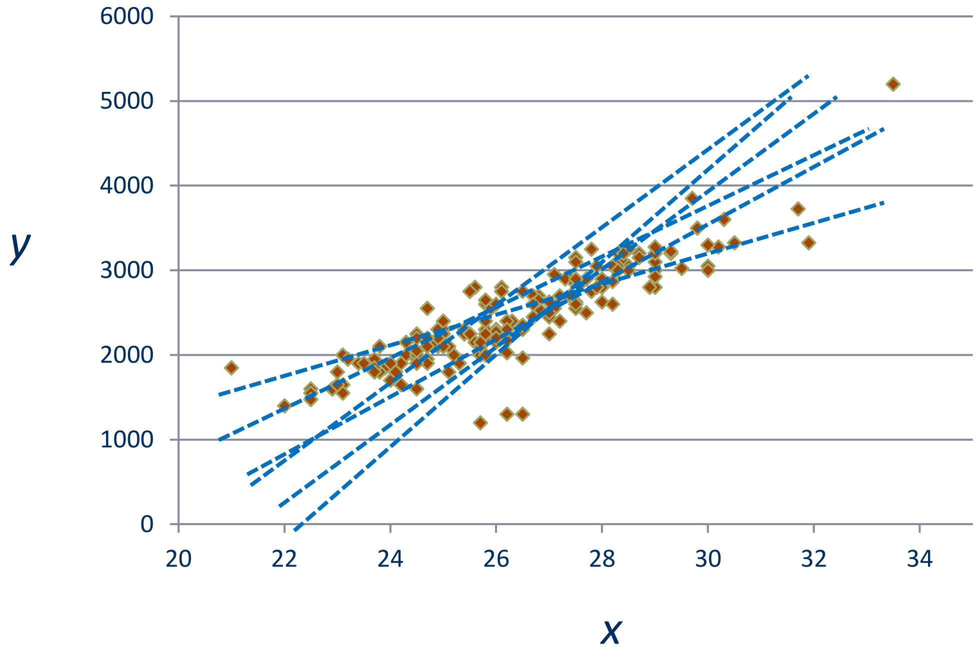
- $\boldsymbol{Y}$ can be above, below, or on the line $\boldsymbol{\beta_0 + \beta_1 x}$

- The distance $\boldsymbol{\varepsilon}$ is a random variable, $\boldsymbol{\varepsilon \sim N(0, \sigma_\varepsilon^2)}$

$$\boxed{\boldsymbol{Y|X = x \sim N(\beta_0 + \beta_1 x\,;\ \sigma_\varepsilon^2)}}$$

Scatterplots – real data

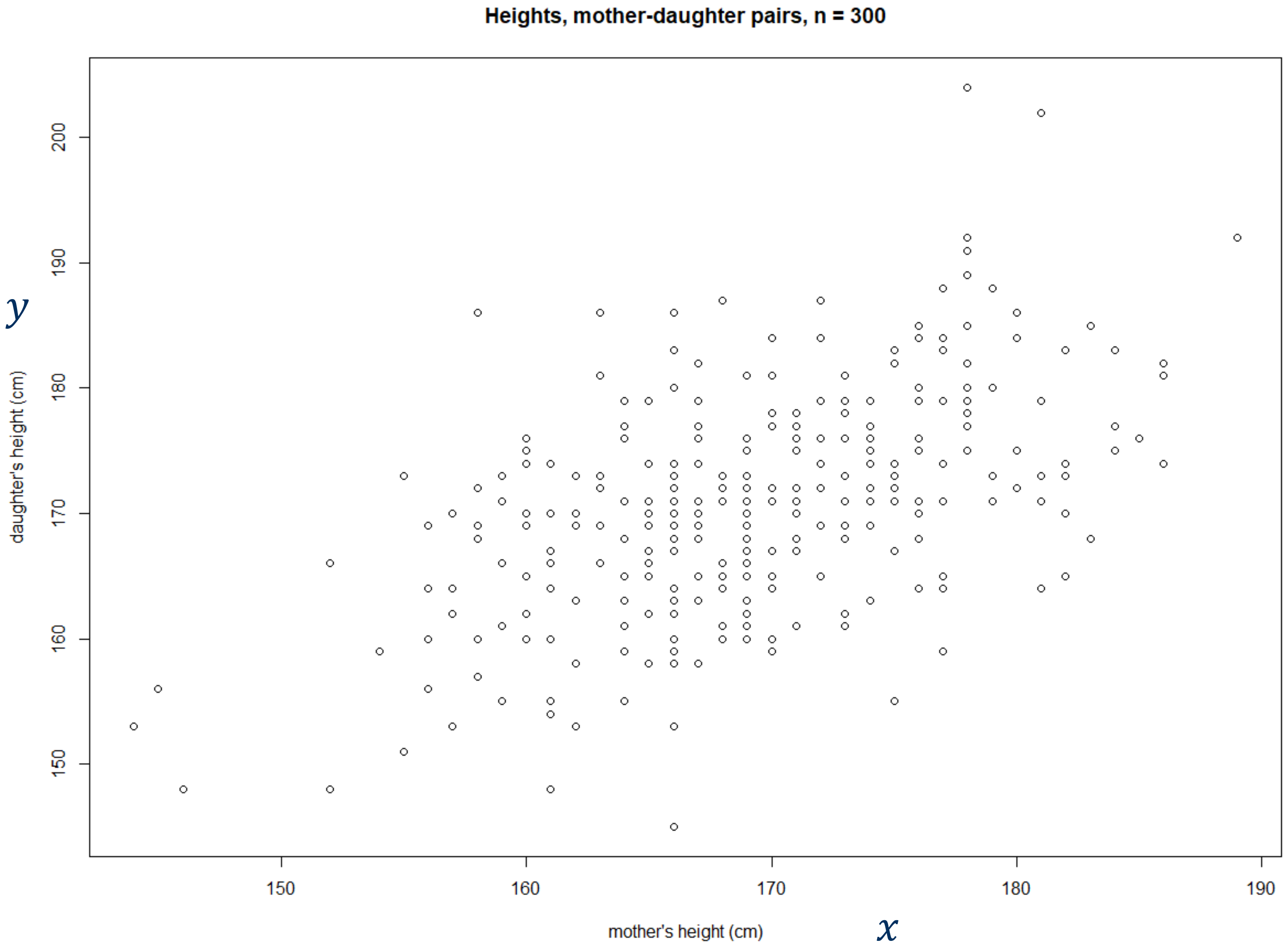
- Assume two numerical, continuous variables: $\boldsymbol{x}$ and $\boldsymbol{y}$
- Every pair $(\boldsymbol{x_i}, \boldsymbol{y_i})$ is represented by a dot



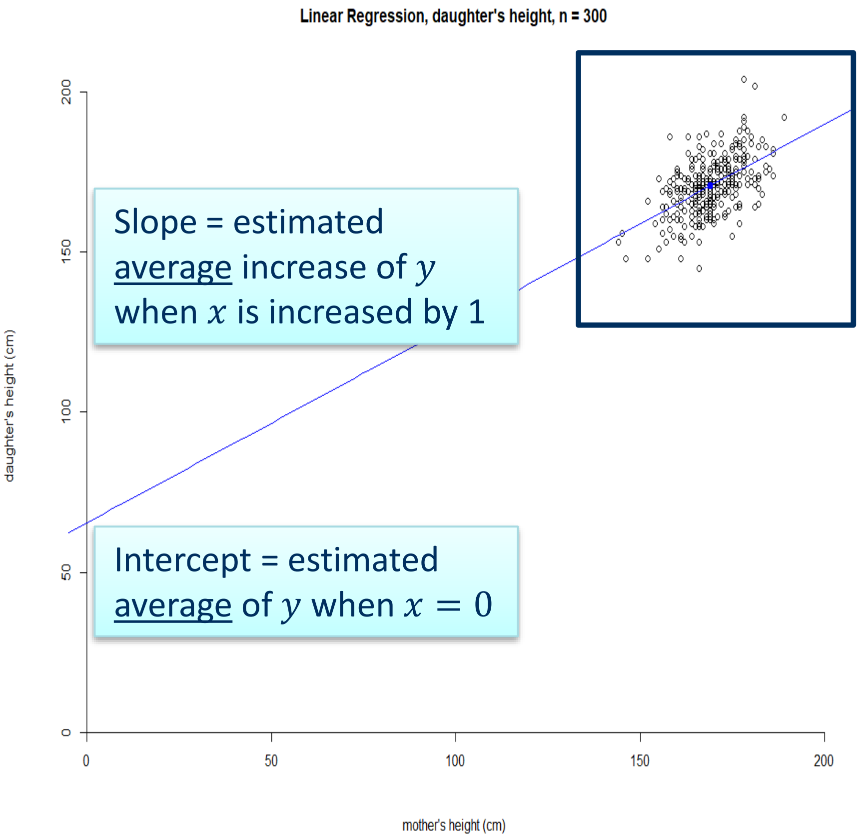
Which line $\boldsymbol{y = b_0 + b_1 x}$ has the best fit to the points?

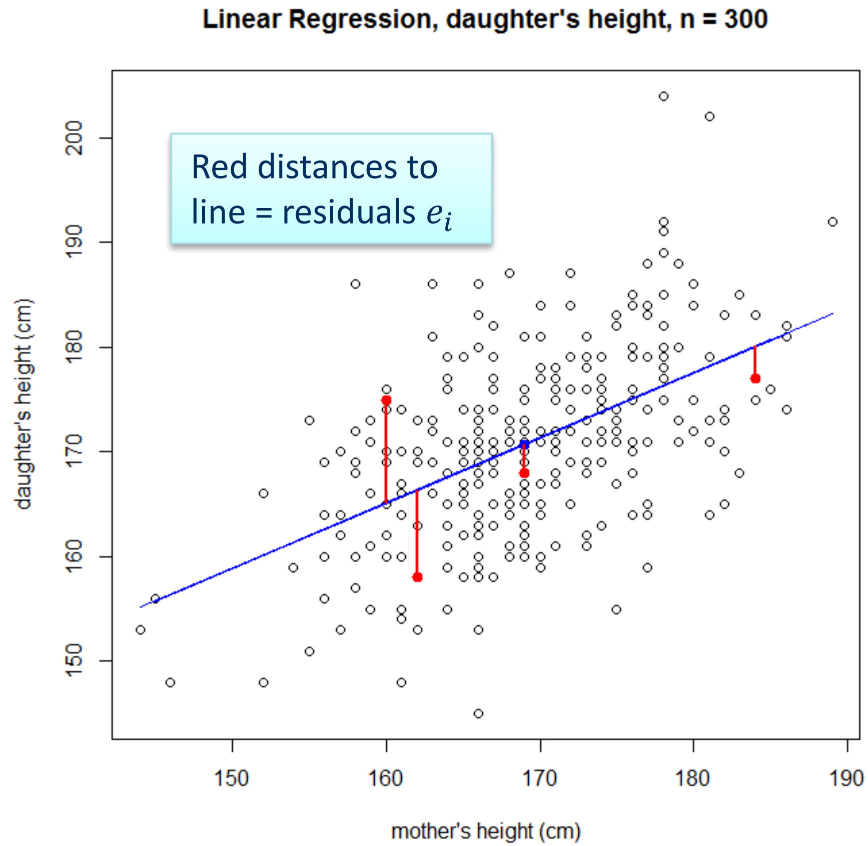
And how do we define best fit?

Example: Height of female adults



Ex. cont. Daughter's height explained by mother's height: $Y_i = \beta_0 + \beta_1 \cdot X_i + \varepsilon_i$





Stockholm University

Exercise

- Suppose that we have a sample of $\boldsymbol{n = 6}$ pairwise values of two random variables $\boldsymbol{x}$ and $\boldsymbol{y}$:

$\boldsymbol{i}$	$\boldsymbol{x_i}$	$\boldsymbol{y_i}$
1	1	2
2	2	3
3	10	8
4	8	6
5	5	7
6	4	8
Sum	30	34

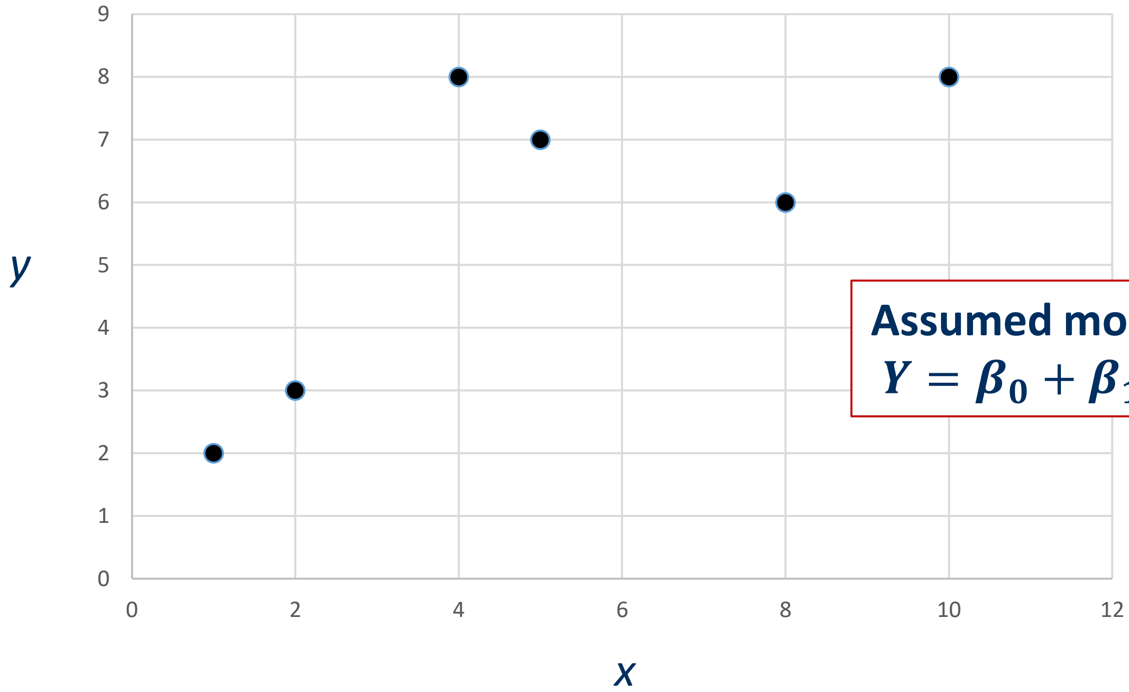
- We wish to ”explain” $\boldsymbol{Y}$ using $\boldsymbol{X}$ by means of a linear regression model

Exercise, cont.

- Plot the observations
- Estimate the model parameters: $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$, and $\boldsymbol{\sigma_\varepsilon^2}$
- Draw the line in the diagram
- Interpret the estimates of the regression parameters
- Make a **prediction** of $\boldsymbol{y}$ when $\boldsymbol{x = 2}$
- Estimate the conditional expectation of $\boldsymbol{y}$ when $\boldsymbol{x = 3}$

Exercise, cont.

- Plot: looks like a positive linear relationship (almost linear?)



Assumed model:

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

Stockholm University

OLS estimation of the model parameters

- Sample (iid) of $\boldsymbol{n}$ pairs $(\boldsymbol{x_i}, \boldsymbol{y_i})$, $\boldsymbol{i = 1, \ldots, n}$
- $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_1}$, and $\boldsymbol{\sigma_\varepsilon^2}$ are estimated with the method of ***ordinary least squares, OLS***

Set $\boldsymbol{b_0}$ and $\boldsymbol{b_1}$ of the line $\boldsymbol{\hat{y} = b_0 + b_1 x}$ such that the sum of the squared distances $\boldsymbol{e_i = y_i - \hat{y}_i}$ is as small as possible.

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

Solved mathematically by taking the derivative of the sum with respect to b_0 and b_1, set the derivatives equal to zero, and then solve for b_0 and b_1.

See formulary

Derived formulas

1. First calculate the estimate of $\boldsymbol{\beta_1}$, one of several ways:

$$\boxed{\boldsymbol{b_1} = r_{xy} \cdot \frac{s_y}{s_x} = \frac{Cov(x,y)}{s_x^2}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \cancel{/(n-1)}}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cancel{/(n-1)}}$$

$$\boxed{= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}}$$

2. Then calculate the estimate of $\boldsymbol{\beta_0}$: $\boxed{\boldsymbol{b_0} = \bar{y} - b_1\bar{x}}$

- To calculate the estimate of the error variance $\boldsymbol{\sigma_\varepsilon^2}$ we first have to calculate the **residuals**.

Stockholm University

Exercise, the estimates

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	1	2	1	4	2
2	2	3	4	9	6
3	10	8	100	64	80
4	8	6	64	36	48
5	5	7	25	49	35
6	4	8	16	64	32
Sum	30	34	210	226	203

$$\bar{x} = \frac{30}{6} = 5$$

$$\bar{y} = \frac{34}{6} = 5{,}6667$$

$$\boldsymbol{b_1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{203 - 6 \cdot 5 \cdot 5{,}6667}{210 - 6 \cdot 5^2} = \mathbf{0,55}$$

$$\boldsymbol{b_0} = \bar{y} - b_1\bar{x} = 5{,}6667 - 0{,}55 \cdot 5 = \mathbf{2,91667}$$

Estimated model:
$\hat{\boldsymbol{y}} = \mathbf{2,92 + 0,55}\boldsymbol{x}$

Stockholm University

Predictions and Residuals

- Once the parameters have been estimated, we calculate the ***predictions*** $\hat{y}_i$ for each pair (x_i, y_i), $i = 1, \dots, n$

$$\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \boldsymbol{b_0} + \boldsymbol{b_1} \boldsymbol{x_i} \quad = \hat{\boldsymbol{\mu}}_{\boldsymbol{Y|X=x_i}}$$

  The predictions *y-hat* are equal to the estimates of the conditional expected values.

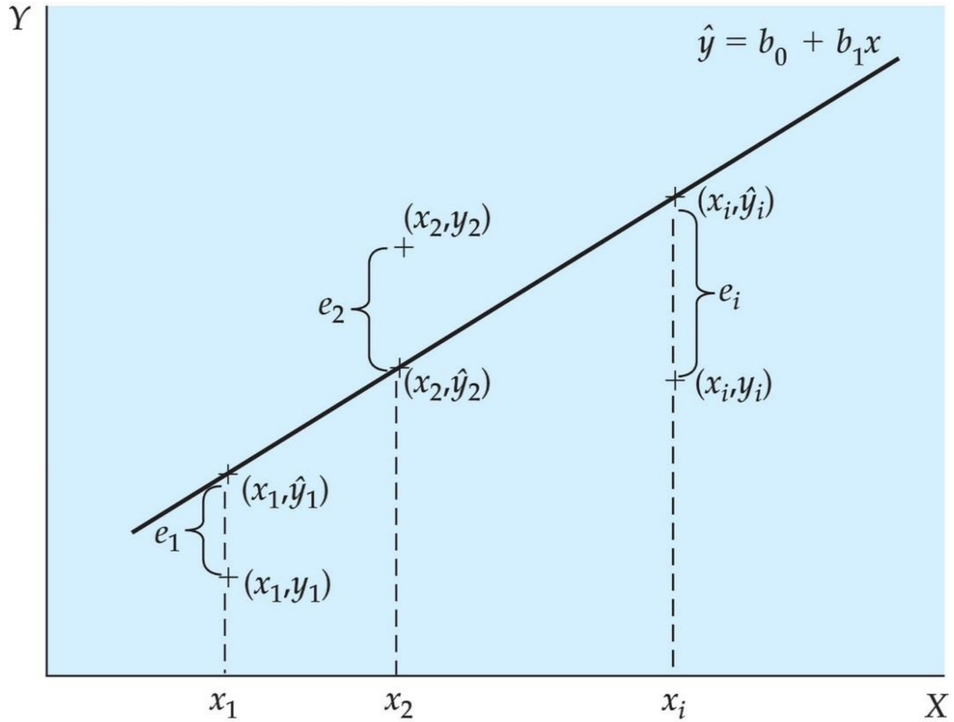
- The **residuals** $\boldsymbol{e_i}$ for $i = 1, \dots, n$ are calculated as

$$\boldsymbol{e_i} = \boldsymbol{y_i} - \hat{\boldsymbol{y}}_{\boldsymbol{i}}$$

  The true parameters $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$ are unknown and hence we <u>cannot</u> observe the true errors $\boldsymbol{\varepsilon_i}$, but we can use $\boldsymbol{e_i}$ as proxies.

Illustration

- Observations (x_i, y_i), predictions $\hat{y}_i$ and residuals e_i:



Observe that the predictions $\hat{y}_i$ all lie on the regression line.

They <u>must</u> lie on the line since

$$\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \boldsymbol{b_0} + \boldsymbol{b_1 x_i}$$

Residual variance

NOTE! The sum of the residuals are always zero! Use this to check your work:

$$\sum_{i=1}^{n} e_i = 0 \;\; \Rightarrow \;\; \bar{e} = 0$$

- Then calculate the square of each residual, $\boldsymbol{e_i^2}$, $i = 1, \ldots n$

$$\boldsymbol{s_e^2} = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{\boldsymbol{n-2}} = \frac{\sum_{i=1}^{n} e_i^2}{\boldsymbol{n-2}}$$

- Two parameters (β_0 and β_1) have to be estimated. This causes the loss of two degrees of freedom; we divide by $\boldsymbol{n-2}$
- The residual variance $\boldsymbol{s_e^2}$ is an estimate of the error variance $\boldsymbol{\sigma_\varepsilon^2}$

Exercise, predictions and residuals

i	x_i	y_i	$\hat{y}_i$	e_i	e_i^2
1	1	2	3.4667	-1.4667	2.1511
2	2	3	4.0167	-1.0167	1.0336
3	10	8	8.4167	-0.4167	0.1736
4	8	6	7.3167	-1.3167	1.7336
5	5	7	5.6667	1.3333	1.7778
6	4	8	5.1167	2.8833	8.3136
Sum	30	34	34.0000	0.0000	**15.1833**

$$\hat{y}_i = b_0 + b_1 x_i$$
$$= 2{,}92 + 0{,}55x$$

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

$$s_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{15.1833}{6-2} = \mathbf{3.7958} \qquad s_e = \sqrt{s_e^2} = \mathbf{1{,}9483}$$

Compare: $s_e^2 < s_y^2 = 6.6667$ $\qquad s_e < s_y = 2.5820$

Stockholm University

Exercise using Excel – interpret output

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0,737902433
R Square	0,5445
Adjusted R Square	0,430625
$\boxed{\text{Standard Error} \quad 1{,}948289848}$	
Observations	6

$$\sqrt{s_e^2} = s_e$$

Note! ”Standard error” is a bit misleading; rather, it should be called ”Residual standard deviation”

ANOVA

	df	SS	MS	F	Significance F
Regression	1	18,15	18,15	4,781558727	0,094040288
Residual	4	15,18333333	3,795833		
Total	5	33,33333333			

	Coefficients	Standard	Upper 95%
Intercept	2,916666667	1,488	7,048102917
X	0,55	0,251	1,248340185

Intercept = intercept: b_0

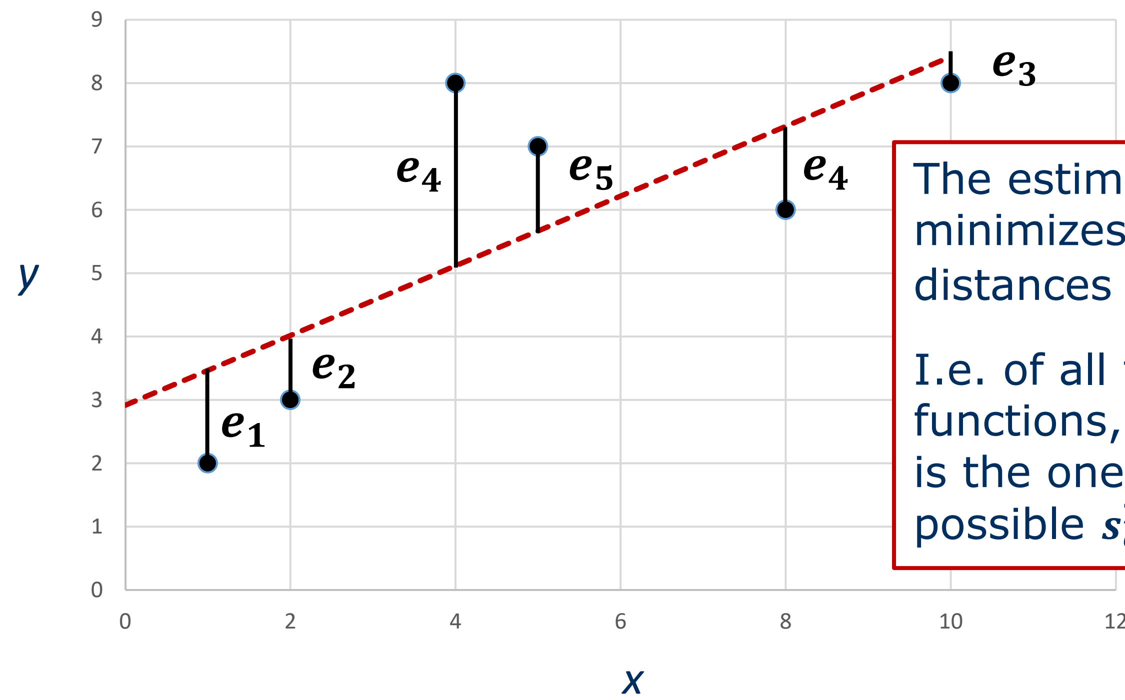
’X’ = name of x-variable: b_1

Stockholm University

Exercise, cont.

To draw the regression line:
- Select to suitable values x_a and x_b
- Calculate $\hat{y}_a$ and $\hat{y}_b$
- Draw a line between the points $(x_a, \hat{y}_a)$ and $(x_b, \hat{y}_b)$

- Scatter plot with estimated regression line



The estimated line is the line that minimizes the sum of the squared distances $\boldsymbol{e_i}$.

I.e. of all the possible linear functions, the line $\hat{\boldsymbol{y}} = \mathbf{2,92 + 0,55}\boldsymbol{x}$ is the one that yields the smallest possible $\boldsymbol{s_e^2}$.

Stockholm University

Exercise, cont.

Interpretation:

- $\boldsymbol{b_0}$ = the intercept = the **estimated average** (expected value) of $\boldsymbol{Y}$ when $\boldsymbol{x = 0}$

- $\boldsymbol{b_1}$ = regression coefficient or slope = the **estimated average** (expected) increase of $\boldsymbol{Y}$ as $\boldsymbol{x}$ increases by 1

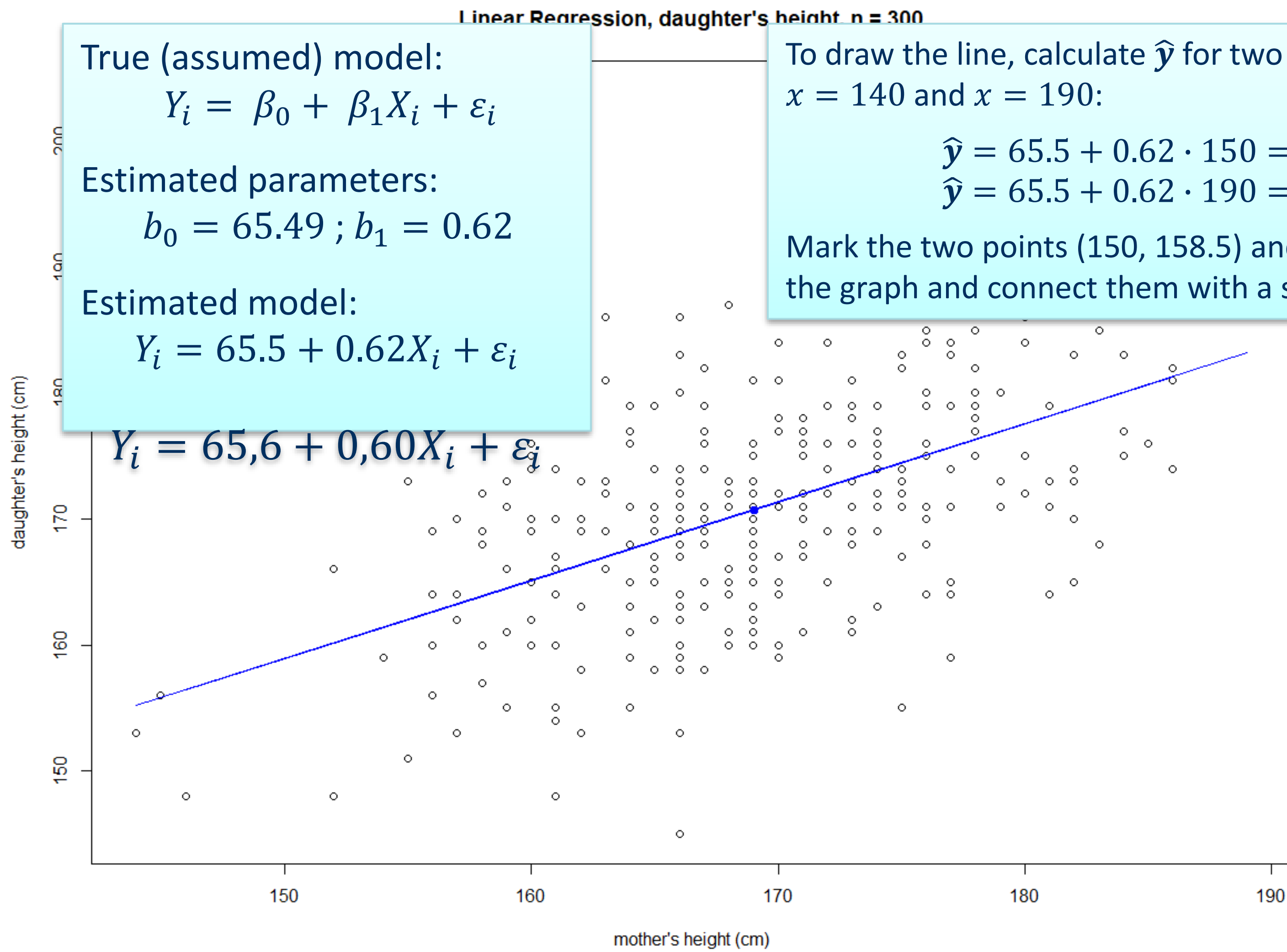
Prediction of $\boldsymbol{Y}$ when $\boldsymbol{x_{n+1} = 2}$:

$$\widehat{\boldsymbol{y}}_{\boldsymbol{n+1}} = 2{,}9167 + 0{,}55 \cdot 2 = \mathbf{4{,}0167}$$

Estimate of conditional mean of $\boldsymbol{Y}$ when $\boldsymbol{x_{n+1} = 3}$:

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{Y|x_{n+1}}} = 2{,}9167 + 0{,}55 \cdot 3 = \mathbf{4{,}5667}$$

Ex. cont. Daughter’s height with mother’s height as independent variable



True (assumed) model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Estimated parameters:

$$b_0 = 65.49 \, ; \, b_1 = 0.62$$

Estimated model:

$$Y_i = 65.5 + 0.62X_i + \varepsilon_i$$

$$Y_i = 65{,}6 + 0{,}60X_i + \varepsilon_i$$

To draw the line, calculate $\hat{\boldsymbol{y}}$ for two values of $\boldsymbol{x}$, say $x = 140$ and $x = 190$:

$$\hat{\boldsymbol{y}} = 65.5 + 0.62 \cdot 150 = 158.5$$
$$\hat{\boldsymbol{y}} = 65.5 + 0.62 \cdot 190 = 183.3$$

Mark the two points (150, 158.5) and (190, 183.3) in the graph and connect them with a straight line.

Next time

Sections **11.4 – 11.6 NCT**

- ANOVA and coefficient of determination $\boldsymbol{R^2}$ and adjusted $\boldsymbol{R^2}$
- Inference for $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$
- Inference for $\boldsymbol{\mu_{Y|x}}$ conditional on a value $\boldsymbol{x}$
- Prediction of $\boldsymbol{y_{n+1}}$ given some (new) value $\boldsymbol{x}$
- Brief discussion of regression and correlation
- A little more on graphic representation, what to look for
  - model assumptions

Stockholm University