Chapter 11
Simple Linear Regression

Introduction

Statistics

Mehmet Güray Güler, PhD

Last updated 28.07.2020

What is SLR?

- Simple linear regression is a statistical method that allows us to summarize and study relationships between two continuous (quantitative) variables:
  - *X*: **predictor**, **explanatory**, or **independent** variable.
  - *Y*: **response**, **outcome**, or **dependent** variable.
- **Types of relationships:**
- <u>Deterministic</u>: the equation *exactly* describes the relationship between the two variable
  - Fahr = 1.8 Cels + 32
  - Circumference = π × diameter
  - Volt = I (current) x R (resistance)

What is SLR?

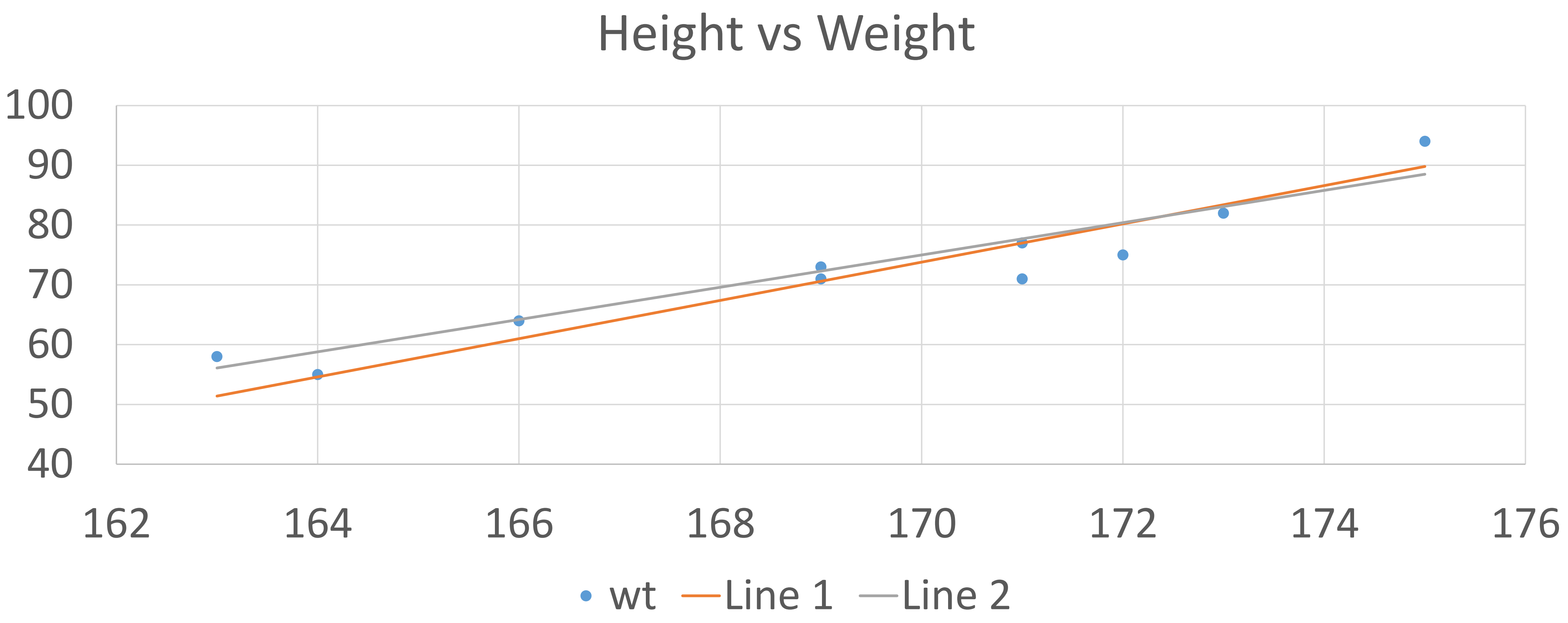
- Instead, we are interested in **statistical relationships**, in which the relationship between the variables is <u>not perfect.</u>
- Here is an example of a statistical relationship.
- The response variable y
  - The weight
- the predictor variable x
  - The height

Sample relations

- Assume we want to find the relation between the weight and the height of the students in this university.
- We have the following data:

Height	**Weight**
163	58
164	55
166	64
169	71
169	73
171	71
171	77
172	75
173	82
175	94

Sample relations



Line 1: Weight = -470.2 + 3.2*Height

Line 2: Weight = -384+2.7*Height

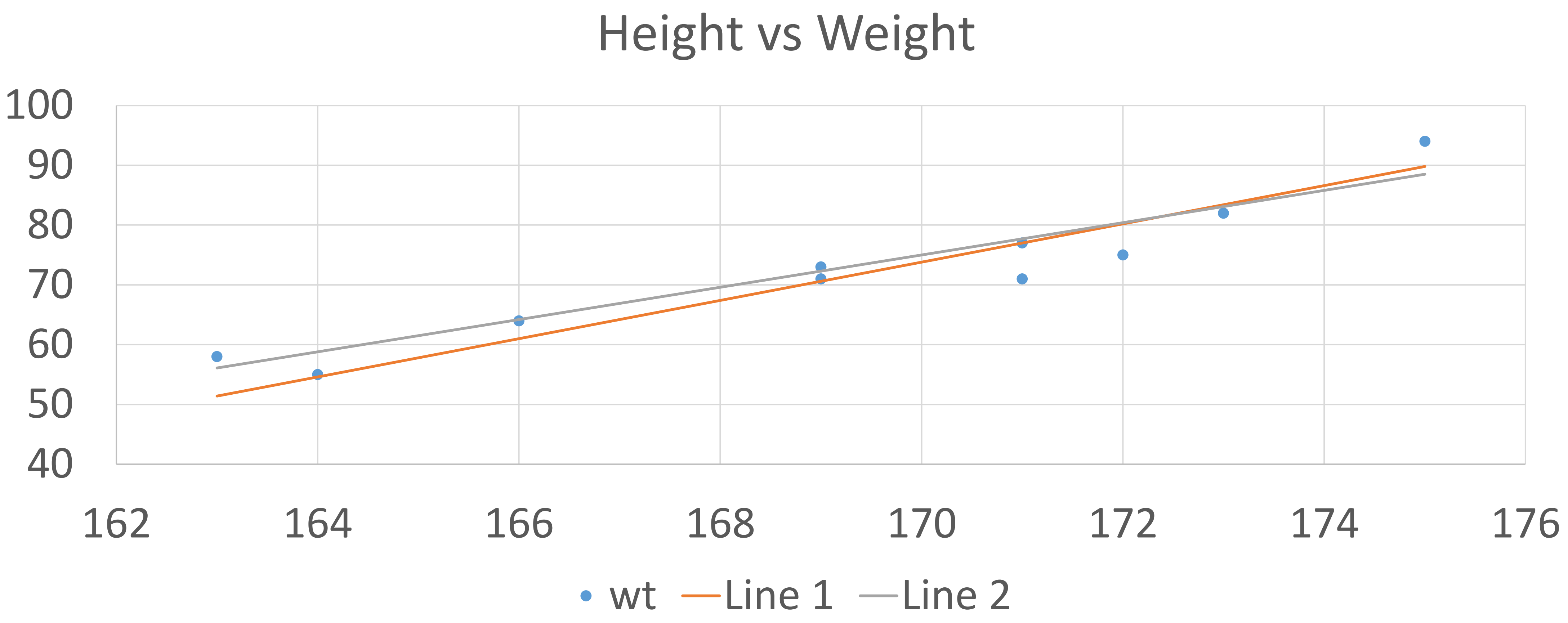
Sample relations

- It appears to be a positive linear relationship between weight and height,
  - but the relationship is not perfect.
- Indeed, the plot exhibits
  - some "**trend**,"
  - but it also exhibits some "**scatter**."
- Therefore, it is a **statistical** relationship, not a deterministic one.
- Examples
  - Size and annual sales - as the size of a shop increases, you'd expect sales to increase, but not perfectly.
  - Anything?

Which line is better?

- Since we are interested in summarizing the trend between two quantitative variables, the natural question arises —
  - **’’which line is better?’’**

Which line is better?



Line 1: Weight = -470.2 + 3.2*Height

Line 2: Weight = -384+2.7*Height

Which line is better?

- In order to examine which of the two lines is a better fit, we first need to introduce some common notation:
  - y_i denotes the observed response for experimental unit i
  - $x_i$denotes the predictor value for experimental unit i
  - $\hat{y}_i$ is the predicted response (or fitted value) for experimental unit i
- Then, the equation for the best fitting line is:
  - $\hat{y}_i = b_0 + b_1 x_i$
- Recall that an "**experimental unit**" is the object or person on which the measurement is made.
  - In our height and weight example, the experimental units are students.

Which line is better?

- First data:
  - $x_1 = 163$ cm, y_1=58 kg.
- If we don't know weight, then
  - using line 1:
    - Weight = -470.2 + 3.2*height
    - Weight = -470.2 + 3.2*163
    - Weight = 51.4
    - $\hat{y}_1 = 51.4$
  - which is the prediction.
- Our prediction wouldn't be perfectly correct — it has some "**prediction error**" (or "**residual error**")
- Here the error is 58-51.4 = 6.6 kg.

Height	Weight
163	58
164	55
166	64
169	71
169	73
171	71
171	77
172	75
173	82
175	94

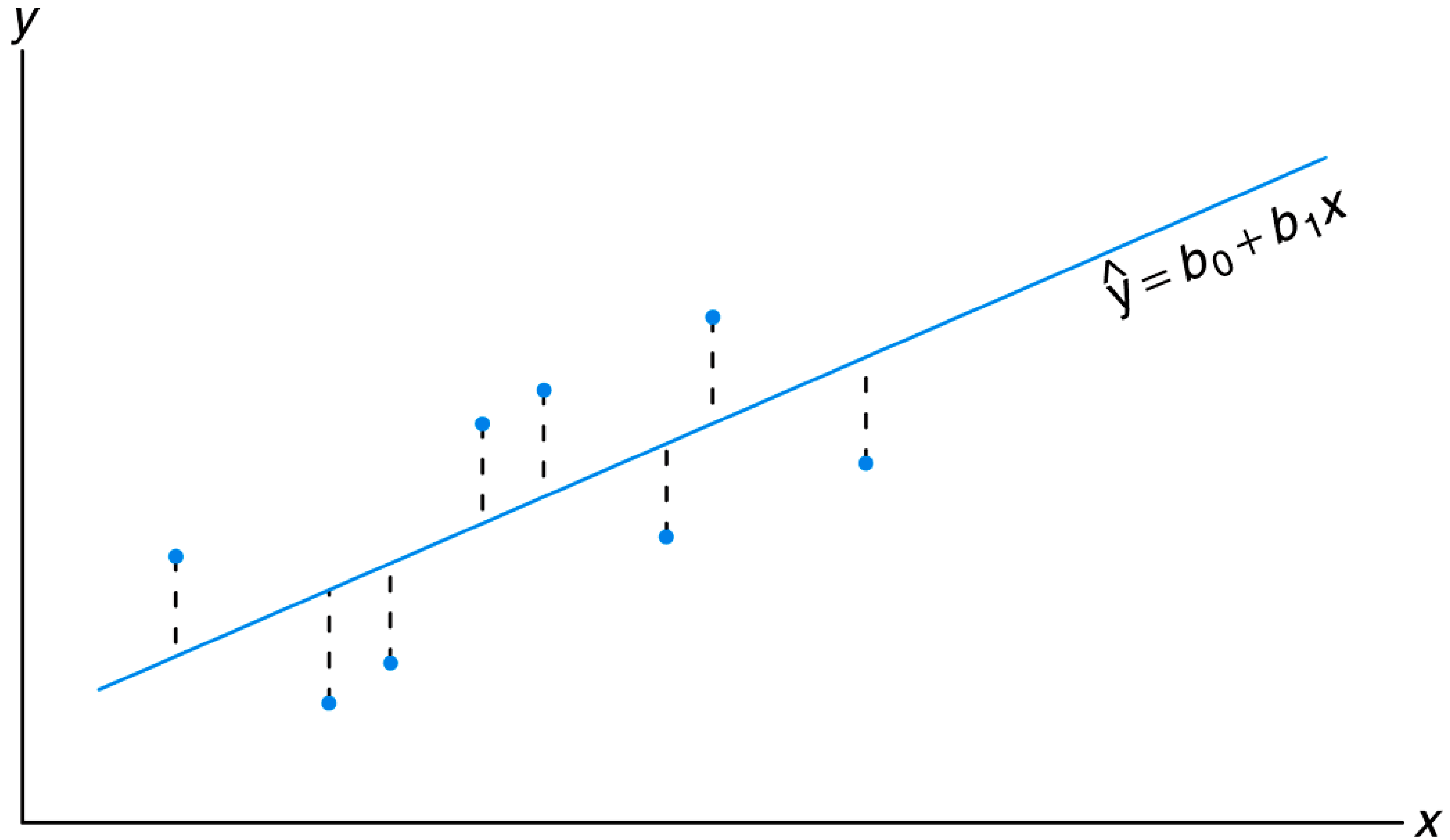
Which line is better?

- If we use $\hat{y}_i = b_0 + b_1 x_i$ to predict y_i, we make a prediction error (or residual error) of size:

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

- e_i value are the **vertical deviations**
  - from points to the predicted regression line.

Which line is better?



Which line is better?

- If we use $\hat{y}_i = b_0 + b_1 x_i$ to predict y_i, we make a prediction error (or residual error) of size:

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

- e_i value are the **vertical deviations** from points to the predicted regression line.
- A line that fits the data "**better**" will be one for which the ***n* prediction errors** — one for each observed data point — **are as small as possible in some overall sense**.
- **S**um of the **S**quared prediction **E**rrors (SSE).

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

Which line is better?

y_i	$\hat{y}$	e_i	e_i^2
58	51.4	6.6	43.56
55	54.6	0.4	0.16
64	61	3	9
71	70.6	0.4	0.16
73	70.6	2.4	5.76
71	77	-6	36
77	77	0	0
75	80.2	-5.2	27.04
82	83.4	-1.4	1.96
94	89.8	4.2	17.64
			141.28

y_i	$\hat{y}$	e_i	e_i^2
58	56.1	1.9	3.61
55	58.8	-3.8	14.44
64	64.2	-0.2	0.04
71	72.3	-1.3	1.69
73	72.3	0.7	0.49
71	77.7	-6.7	44.89
77	77.7	-0.7	0.49
75	80.4	-5.4	29.16
82	83.1	-1.1	1.21
94	88.5	5.5	30.25
			126.27

The Best Fitting Line

- "**best**" line: minimum SSE.
- "**least squares criterion**," which says
- "minimize the **S**um of the **S**quared prediction **E**rrors (SSE) or residual sum of squares.
- Find b_0 and b_1 that minimize:

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

Method of Least Squares

- Differentiating SSE with respect to b_0 and b_1 and setting the resulting equations to zero we get:

$$\frac{\partial(SSE)}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) \qquad \frac{\partial(SSE)}{\partial b_1} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)x_i.$$

- Solving these two equations will yield the computing formulas for b_0 and b_1 as follows:

Method of Least Squares

$$nb_0 + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \qquad b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

$$b_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x}.$$

Method of Least Squares

	x_i	y_i	x_i^2	$x_i y_i$
	163	58	26569	9454
	164	55	26896	9020
	166	64	27556	10624
	169	71	28561	11999
	169	73	28561	12337
	171	71	29241	12141
	171	77	29241	13167
	172	75	29584	12900
	173	82	29929	14186
	175	94	30625	16450
sum	1693	720	286763	122278
avg	169.3	72		

- Using the formula above, our best fit is:
  - Weight = -393.3 + Height *2.7661
- b_0:
  - The intercept,i.e., the value at x=0
  - Here no meaning!
- b_1:
  - The slope
  - The change in y when x increases by 1 unit.

Method of Least Squares – Using MS Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.950328
R Square	0.903124
Adjusted R	0.891014
St Error	3.764064
Observatio	10

ANOVA

	df	*SS*	*MS*	*F*	*ignificance F*
Regression	1	1056.655	1056.655	74.57944	2.51E-05
Residual	8	**113.3454**	14.16818		
Total	9	1170			

	Coefficients	*andard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 95.0%*	*Ipper 95.0%*
Intercept	**-396.303**	54.24025	-7.30643	8.34E-05	-521.381	-271.224	-521.381	-271.224
X Variable 1	**2.766112**	0.320302	8.635939	2.51E-05	2.027493	3.50473	2.027493	3.50473