Class 4

Daniel B. Rowe, Ph.D.

Department of Mathematical and Statistical Sciences

Agenda:

Recap Chapter 2.5, 3.1

Lecture Chapter 3.2, 3.3

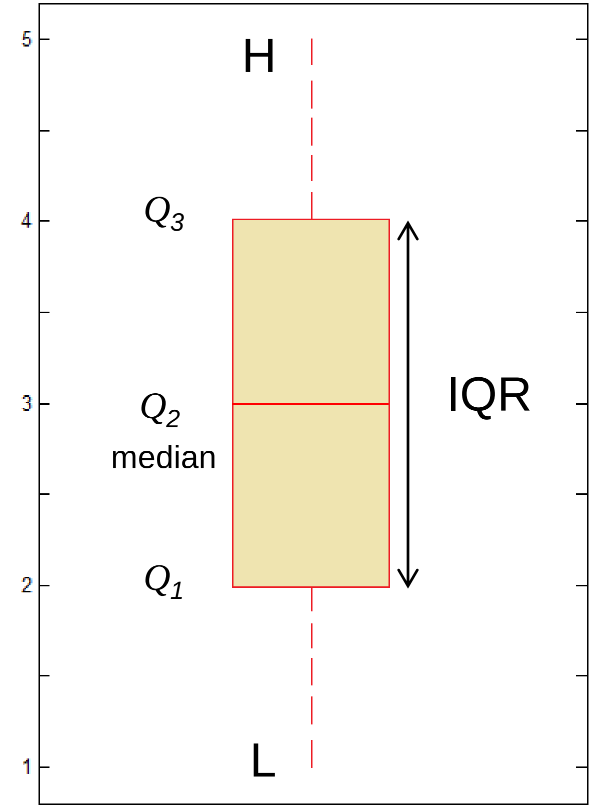
Recap Chapter 2.5

2: Descriptive Analysis and Single Variable Data

2.5 Measures of Position

Measures of Position: Quartiles - ranked data into quarters

L = lowest value
H = highest value
Q_2 = median
Q_1 = 25% smaller
Q_3 = 75% smaller
IQR = Q_3 - Q_1



2: Descriptive Analysis and Single Variable Data

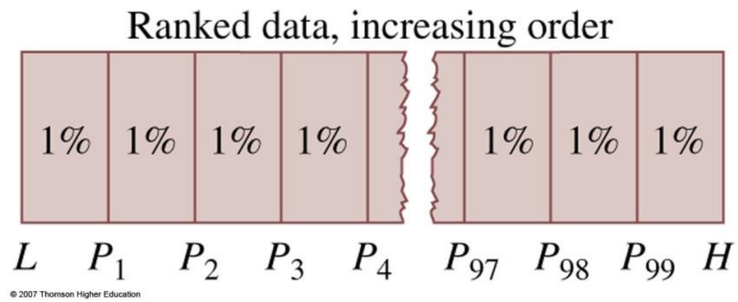
2.5 Measures of Position

Measures of Position: percentiles - rank data into 100$^{\text{ths}}$

L = lowest value

H = highest value

P_k = value where k% are smaller



rank data

$\frac{nk}{100}$ integer, A → p_k halfway between value and next one
average of A^{th} and $(A+1)^{\text{th}}$ values

$\frac{nk}{100}$ not Integer, B → p_k is value in next largest position, B value

Figure from Johnson & Kuby, 2012.

2: Descriptive Analysis and Single Variable Data

2.5 Measures of Position

Standard score, or z-score: The position a particular value of x has relative to the mean, measured in standard deviations.

$$z_i = \frac{i^{\text{th}} \text{ value - mean}}{\text{std. dev.}} = \frac{x_i - \bar{x}}{s}$$

There can be n of these because we have $x_1, x_2, \ldots, x_n$.

2: Descriptive Analysis and Single Variable Data

Questions?

Homework: Read Chapter 2.5-2.7
WebAssign
Chapter 2 # 115, 123c-d, 129, 137

Recap Chapter 3.1

3: Descriptive Analysis and Bivariate Data

3.1 Bivariate Data: two qualitative

Cross-tabulation tables or **contingency tables**

Example:
Construct a 2×3 table.

Name	Gender	Major	Name	Gender	Major	Name	Gender	Major
Adams	M	LA	Feeney	M	T	McGowan	M	BA
Argento	F	BA	Flanigan	M	LA	Mowers	F	BA
Baker	M	LA	Hodge	F	LA	Ornt	M	T
Bennett	F	LA	Holmes	M	T	Palmer	F	LA
Brand	M	T	Jopson	F	T	Pullen	M	T
Brock	M	BA	Kee	M	BA	Rattan	M	BA
Chun	F	LA	Kleeberg	M	LA	Sherman	F	LA
Crain	M	T	Light	M	BA	Small	F	T
Cross	F	BA	Linton	F	LA	Tate	M	BA
Ellis	F	BA	Lopez	M	T	Yamamoto	M	LA

	Major			
Gender	LA	BA	T	Row Total
M	5	6	7	18
F	6	4	2	12
Col. Total	11	10	9	30

M = male
F = female
LA = liberal arts
BA = business admin
T = technology

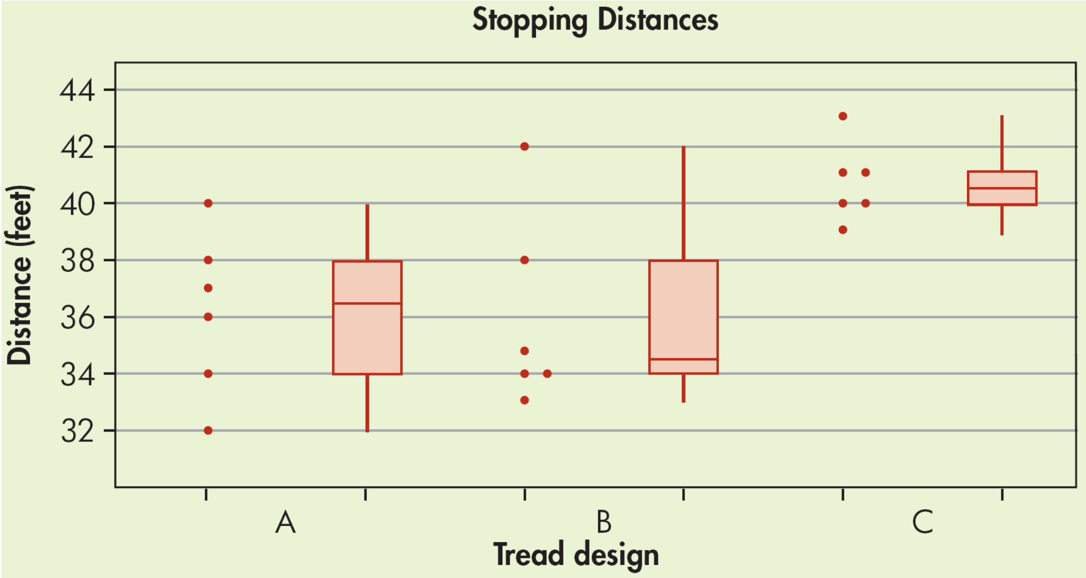
Figures from Johnson & Kuby, 2012.

3: Descriptive Analysis and Bivariate Data

3.1 Bivariate Data: one qualitative and one quantitative

Example:

Design A (n = 6)			Design B (n = 6)			Design C (n = 6)		
37	36	38	33	35	38	40	39	40
34	40	32	34	42	34	41	41	43



Vertical box-and-whiskers

Figures from Johnson & Kuby, 2012.

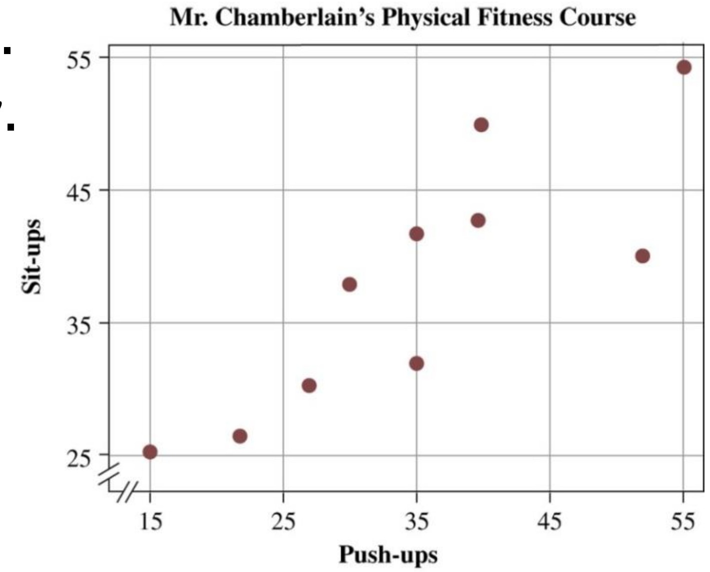
3: Descriptive Analysis and Bivariate Data

3.1 Bivariate Data: two quantitative, Scatter Diagram

Example: Push-ups

Student	1	2	3	4	5	6	7	8	9	10
Push-ups, x	27	22	15	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43

Input variable: independent variable, x.
Output variable: dependent variable, y.



Scatter Diagram: A plot of all the ordered pairs of bivariate data on a coordinate axis system.

(x,y) ordered pairs.

Figures from Johnson & Kuby, 2012.

3: Descriptive Analysis and Bivariate Data

Questions?

Homework: Read Chapter 3
WebAssign
Chapter 3 # 3, 7, 15

3: Descriptive Analysis and Bivariate Data

3.1 Bivariate Data: two quantitative, Scatter Diagram

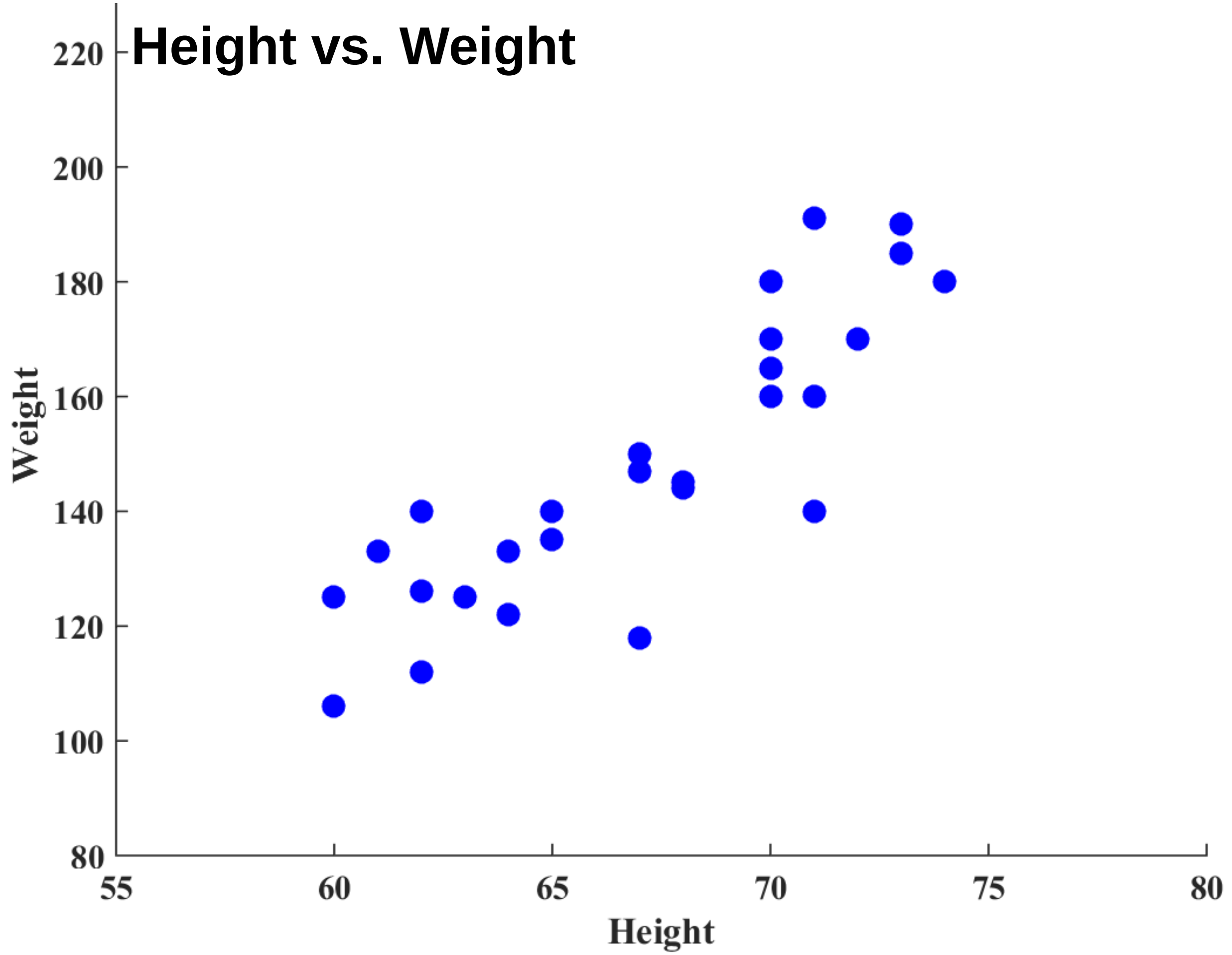
A previous class' data.

Gender, height, weight.

Use height vs. weight (no gender).

3: Descriptive Analysis and Bivariate Data

3.1 Bivariate Data: Scatter Diagram of previous class data.



Chapter 3: Descriptive Analysis and Presentation of Bivariate Data
continued

Daniel B. Rowe, Ph.D.

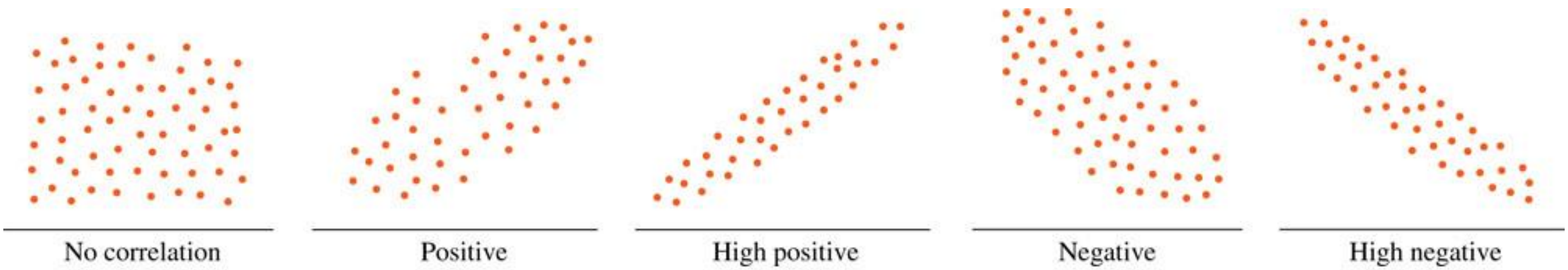
Department of Mathematical and Statistical Sciences

3: Descriptive Analysis and Bivariate Data
3.2 Linear Correlation

Linear Correlation, $\boldsymbol{r}$, is a measure of the strength of a linear relationship between two variables x and y.

$$-1 \le r \le 1$$

Will discuss its computation in a minute.

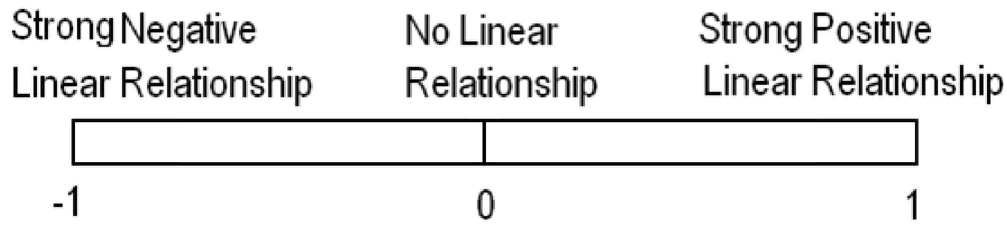


$r \approx 0$ $\quad$ $r \approx 0.5$ $\quad$ $r \approx 0.8$ $\quad$ $r \approx -0.5$ $\quad$ $r \approx -0.8$

Figure from Johnson & Kuby, 2012.

3: Descriptive Analysis and Bivariate Data

3.2 Linear Correlation



Values closer to +1 or -1 mean a stronger relationship.

Values near 0 mean a weak association.

Positive values mean a positive relationship.
positive relationship means as x increases so does y

Negative values mean a negative relationship.
negative relationship means as x increases y decreases

3: Descriptive Analysis and Bivariate Data

3.2 Linear Correlation

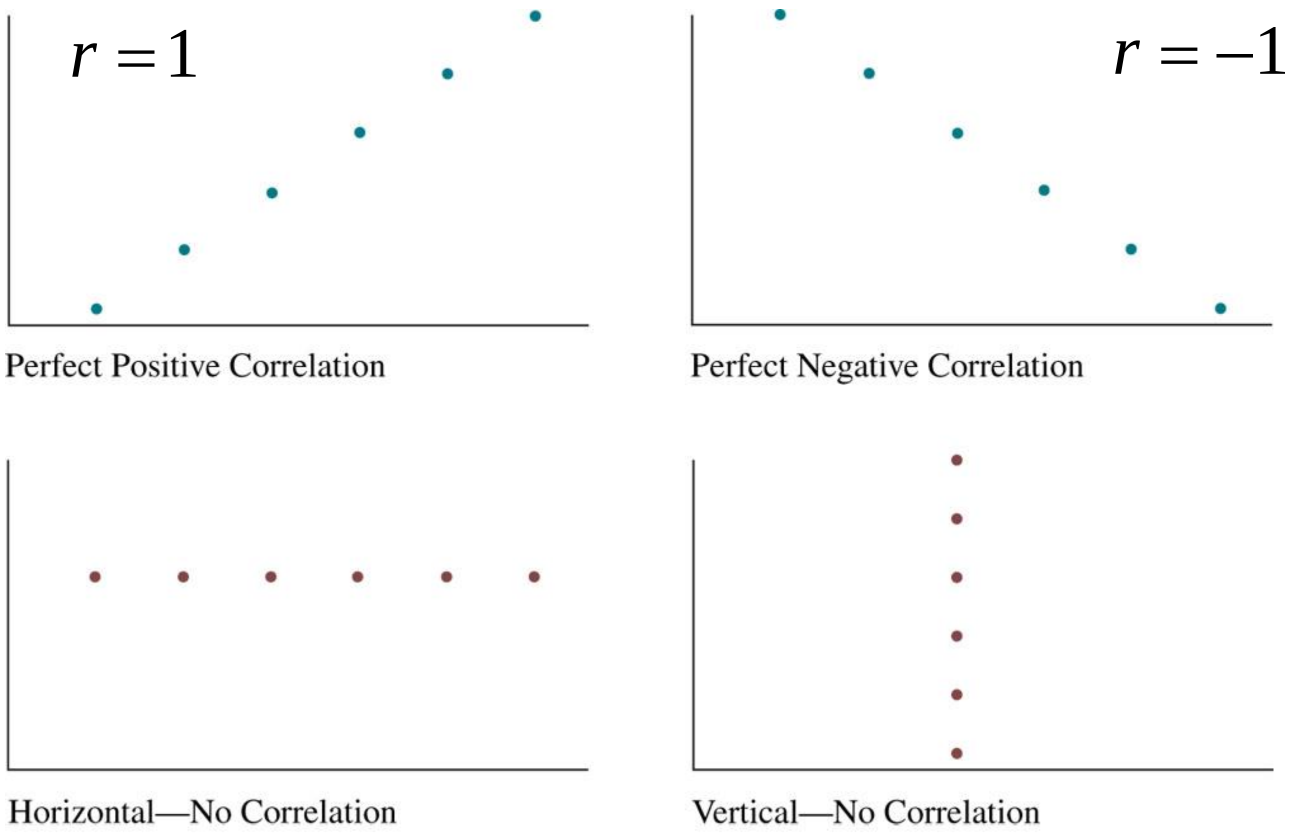


Figure from Johnson & Kuby, 2012.

3: Descriptive Analysis and Bivariate Data
3.2 Linear Correlation

Computing the linear correlation coefficient r.

1.
$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

$$s_x = \text{std } x\text{'s} \qquad s_y = \text{std } y\text{'s}$$

$$SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

2.
$$r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}}$$

$$SS(y) = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

$$SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

1. and **2.** are equivalent.

3: Descriptive Analysis and Bivariate Data

3.2 Linear Correlation

Example:

$$r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}}$$

Student	Push-ups, x	x^2	Sit-ups, y	y^2	xy
1	27	729	30	900	810
2	22	484	26	676	572
3	15	225	25	625	375
4	35	1,225	42	1,764	1,470
5	30	900	38	1,444	1,140
6	52	2,704	40	1,600	2,080
7	35	1,225	32	1,024	1,120
8	55	3,025	54	2,916	2,970
9	40	1,600	50	2,500	2,000
10	40	1,600	43	1,849	1,720
	$\sum x = 351$ sum of x	$\sum x^2 = 13{,}717$ sum of x^2	$\sum y = 380$ sum of y	$\sum y^2 = 15{,}298$ sum of y^2	$\sum xy = 14{,}257$ sum of xy

$$\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i^2$$

$$\sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} y_i^2$$

$$\sum_{i=1}^{n} x_i y_i$$

Figure from Johnson & Kuby, 2012.

3: Descriptive Analysis and Bivariate Data

3.2 Linear Correlation

Example:

Push-ups, x	27	22	15	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43

$$\sum_{i=1}^{n} x_i = 351$$

$$\sum_{i=1}^{n} x_i^2 = 13717$$

$$\sum_{i=1}^{n} y_i = 380$$

$$\sum_{i=1}^{n} y_i^2 = 15298$$

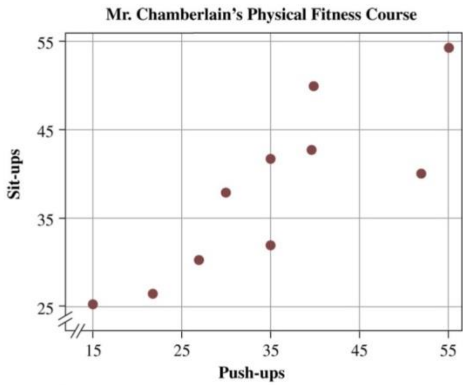
$$\sum_{i=1}^{n} x_i y_i = 14257$$

$$SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 =$$

$$SS(y) = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 =$$

$$SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) =$$

$$r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}} =$$



Figures from Johnson & Kuby, 2012.

3: Descriptive Analysis and Bivariate Data

3.2 Linear Correlation

Example:

Push-ups, x	27	22	15	35	30	52	35	55	40	40
Sit-ups, y	30	26	25	42	38	40	32	54	50	43

$$\sum_{i=1}^{n} x_i = 351$$
$$\sum_{i=1}^{n} x_i^2 = 13717$$
$$\sum_{i=1}^{n} y_i = 380$$
$$\sum_{i=1}^{n} y_i^2 = 15298$$
$$\sum_{i=1}^{n} x_i y_i = 14257$$

$$SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2 = 13717 - \frac{(351)^2}{10} = 1396.9$$

$$SS(y) = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2 = 15298 - \frac{(380)^2}{10} = 858.0$$

$$SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) = 14257 - \frac{(351)(380)}{10} = 919.0$$

$$r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}} = \frac{919.0}{\sqrt{(1396.9)(858.0)}} = 0.84$$

Questions?

Figure from Johnson & Kuby, 2012.

3: Descriptive Analysis and Bivariate Data

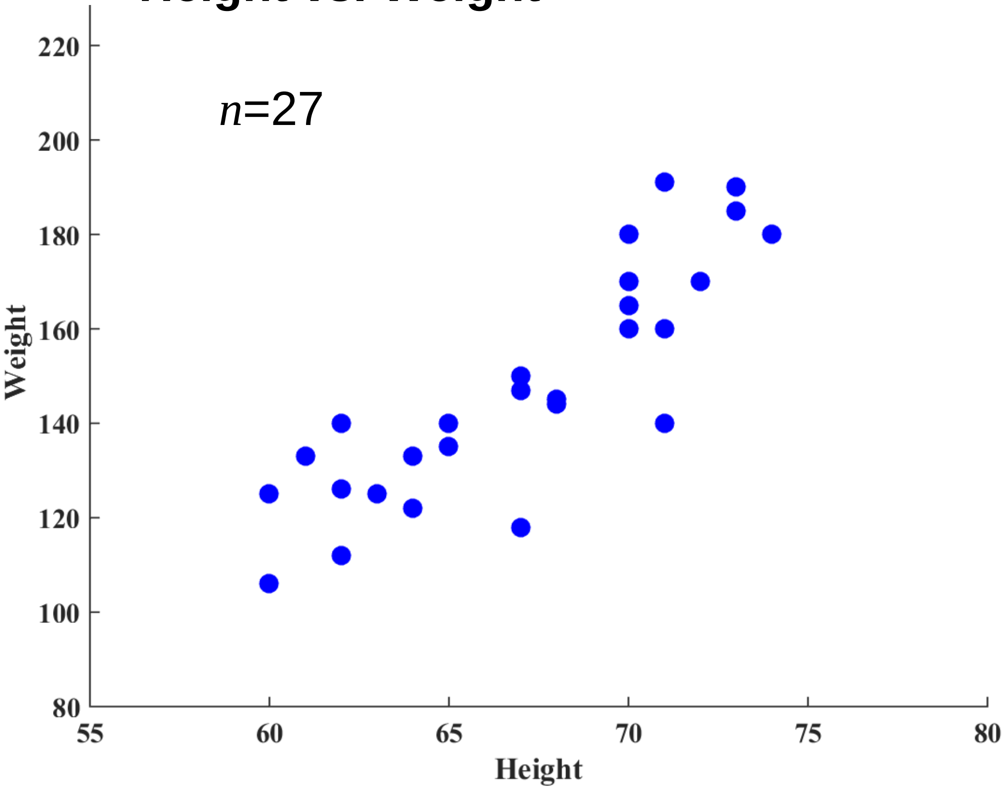
3.2 Linear Correlation

Example: Previous class' data!

$$SS(x)=\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2$$

$$SS(xy)=\sum_{i=1}^{n}x_iy_i-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}y_i\right)$$

Height vs. Weight



$$\sum_{i=1}^{n}x_i=1810$$

$$\sum_{i=1}^{n}x_i^2=121820$$

$$\sum_{i=1}^{n}y_i=3992$$

$$\sum_{i=1}^{n}y_i^2=605818$$

$$\sum_{i=1}^{n}x_iy_i=269982$$

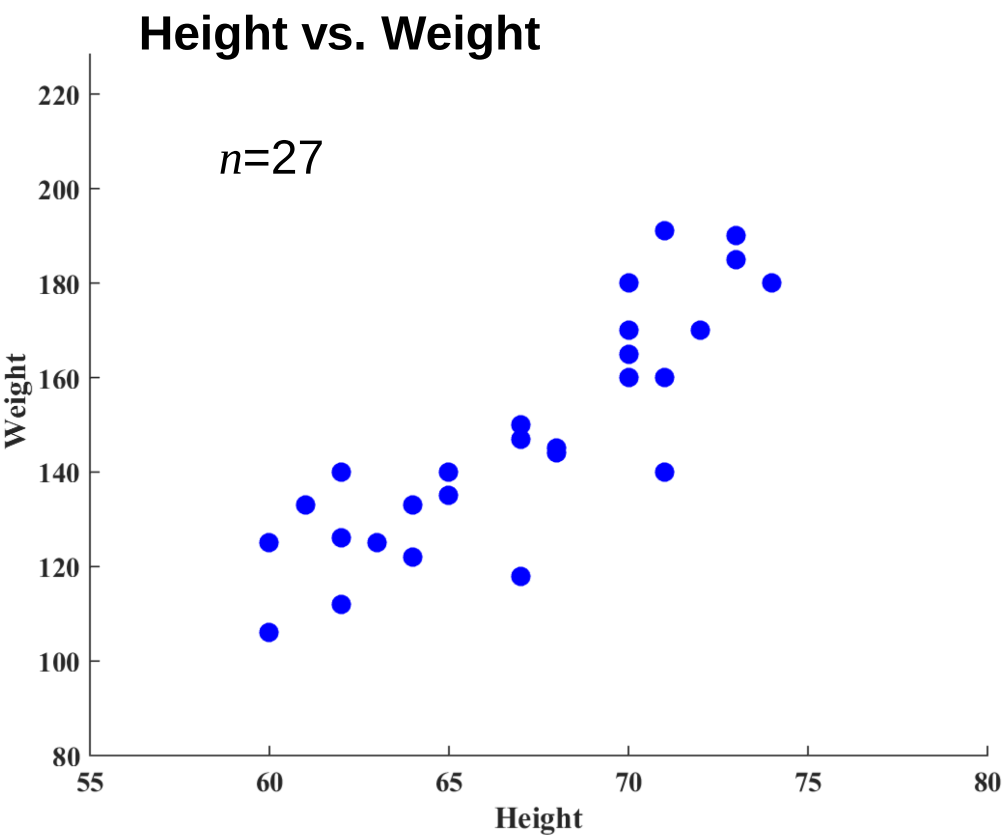
3: Descriptive Analysis and Bivariate Data

3.2 Linear Correlation

Example: Previous class' data!

$$SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$



$$\sum_{i=1}^{n} x_i = 1810 \qquad SS(x) = 483.0$$

$$\sum_{i=1}^{n} x_i^2 = 121820$$

$$\sum_{i=1}^{n} y_i = 3992 \qquad SS(y) = 15593.0$$

$$\sum_{i=1}^{n} y_i^2 = 605818 \qquad SS(xy) = 23701.0$$

$$\sum_{i=1}^{n} x_i y_i = 269982$$

$$r = \frac{SS(xy)}{\sqrt{SS(x)SS(y)}} = 0.86$$

3: Descriptive Analysis and Bivariate Data
3.2 Linear Correlation

Understanding Linear Correlation
Skip for now. Read on own.

Causation and Lurking Variables
Correlation does not necessarily imply causation.
Just because two things are highly related does not mean that one causes the other.

Soda sales go up, flu incidence goes down.
Does soda cause flu to go down?

3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

Regression analysis finds the equation of a line that “best” describes the relationship between the two variables (x and y).

What do we mean by “best?”

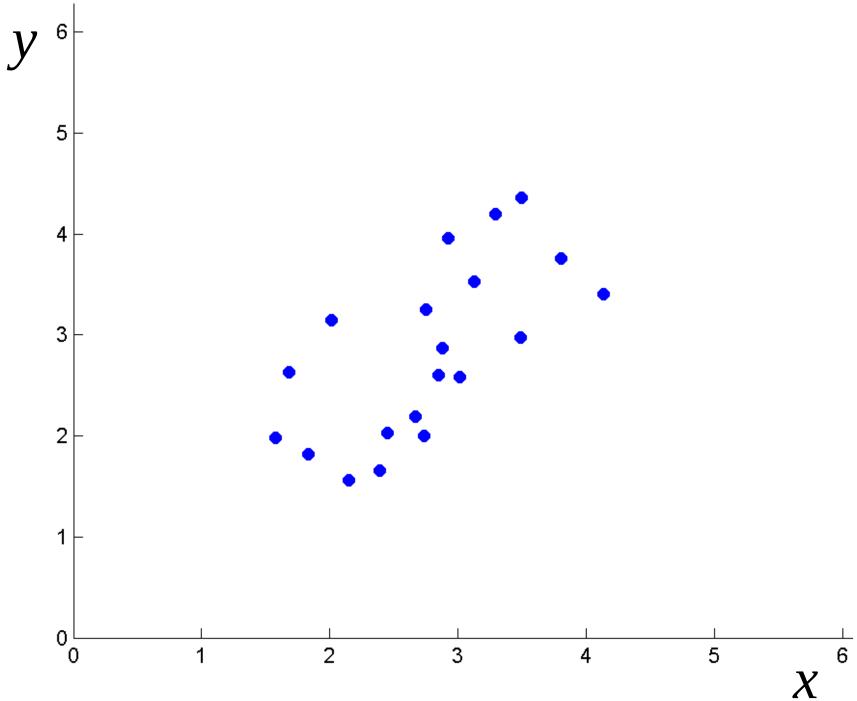
How is “bestness” determined?

Least squares regression.

3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

Let's say that we are given points as in figure.



Imagine that there is an underlying line

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

that the data fits to (or comes from).

β_0 is y-intercept and β_1 is slope.

The points are considered to be

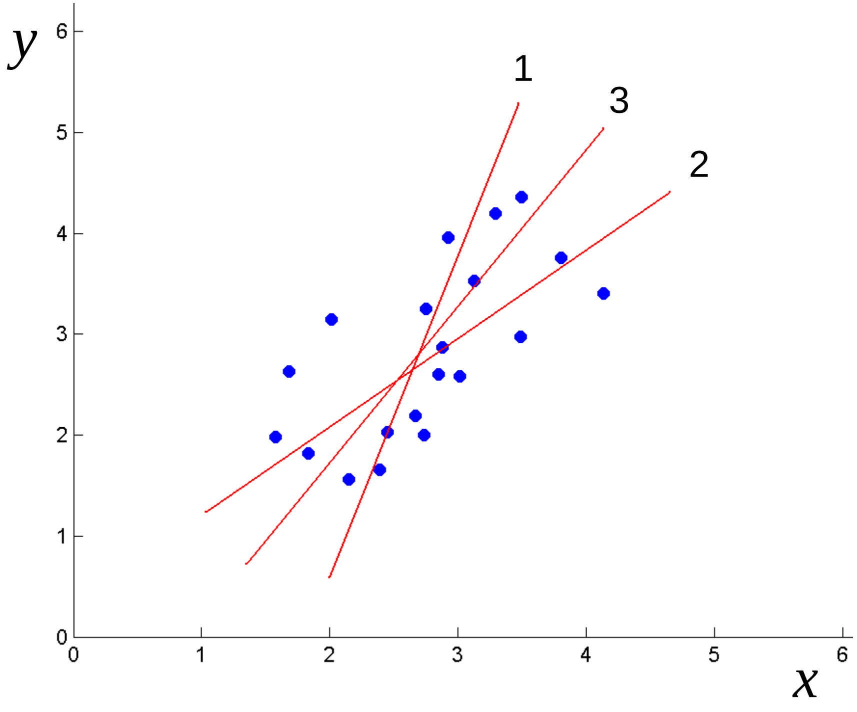
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

random error

We want to find the "best" fit line to the data.

3: Descriptive Analysis and Bivariate Data
3.3 Linear Regression

We can try different lines until we find the "best" one.



Imagine that there is an underlying line

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

that the data fits to (or comes from).

β_0 is y-intercept and β_1 is slope.

The points are considered to be

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \qquad i = 1, \ldots, n$$

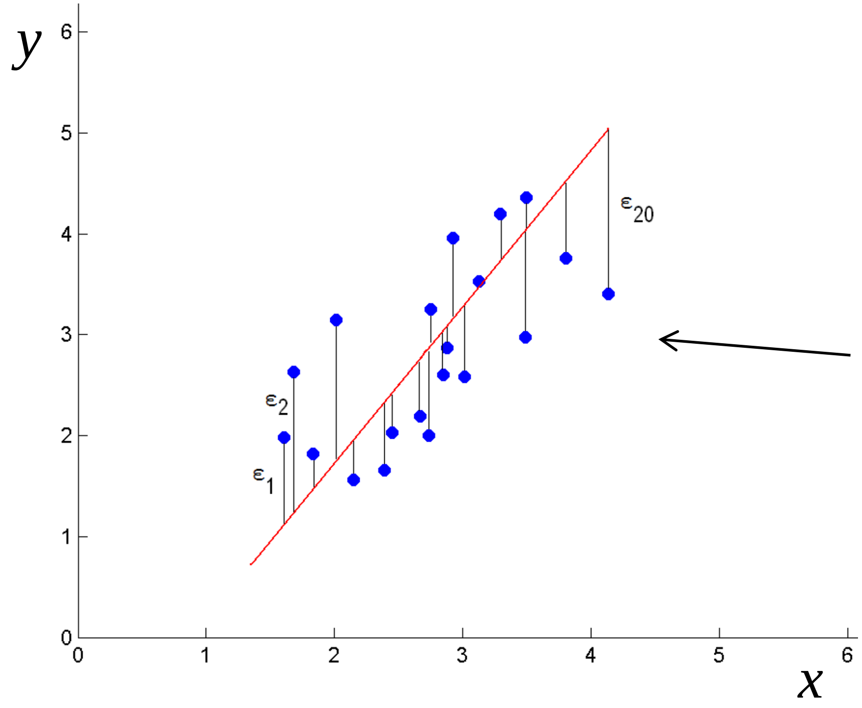
Let's call the "best" one $\hat{y} = b_0 + b_1 x$.

b_0 is estimated y-intercept and b_1 is estimated slope.

3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

What is criteria for bestness? → Sum of squared distances.



The points are considered to be

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

random error

The "best" line value at x_i is

$$\hat{y}_i = b_0 + b_1 x_i$$

These vertical distances ε_i are called residuals.

$$\varepsilon_i = y_i - b_0 - b_1 x_i \qquad i = 1, \ldots, n$$

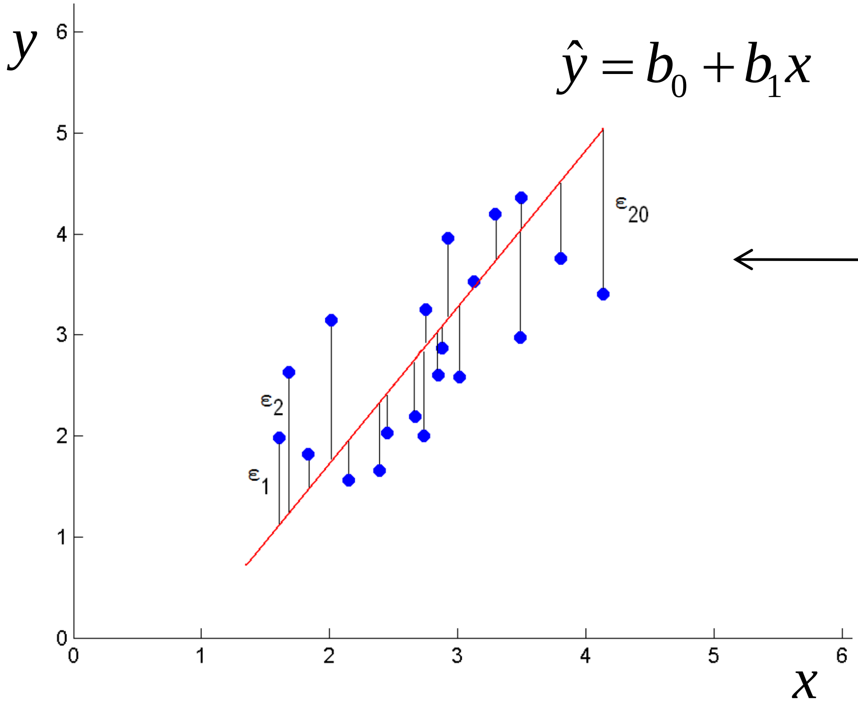
random error
fit error

3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

What is criteria for bestness? → Sum of squared distances.

$$\hat{y} = b_0 + b_1 x$$



We move around the line until the sum of the squared residuals

$$\sum_{i=1}^{n} \varepsilon_i^2$$

is made a minimum.

Least squares line.

This is a measure of misfit and a criterion for the “best“ line.

3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

We don't actually have to move the line around.

We can find the "best" fit line that minimizes the

sum of the squared residuals by using Equations 3.5-3.7a.

$$b_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad \text{or} \quad b_1 = \frac{SS(xy)}{SS(x)}$$

point-slope formula

then $b_0 = \bar{y} - b_1\bar{x}$ because line goes through $(\bar{x},\bar{y})$.

3: Descriptive Analysis and Bivariate Data
3.3 Linear Regression

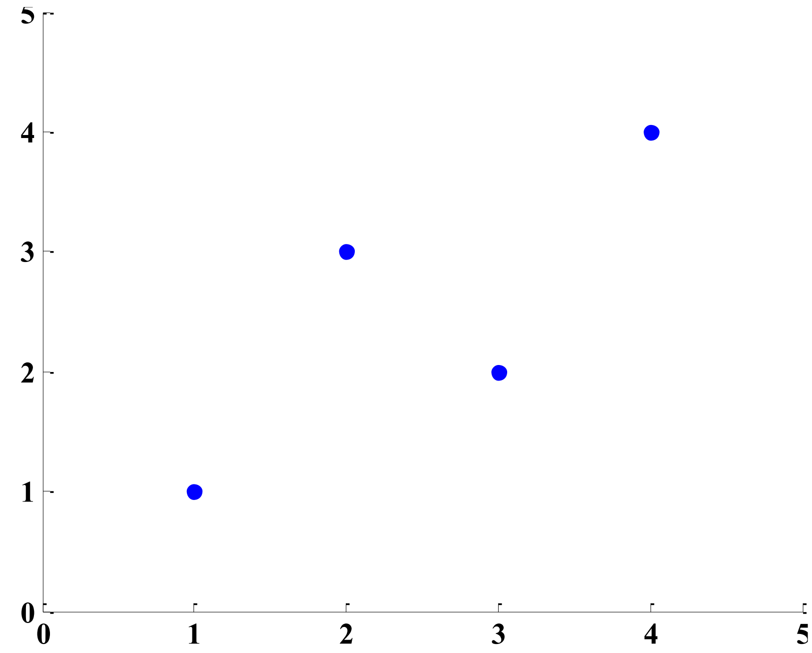
Example: Using (1,1),(3,2),(2,3),(4,4) $\bar{x} = 2.5,\ \ \bar{y} = 2.5$

$$b_1 = \frac{SS(xy)}{SS(x)}$$

$$SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

$$\sum_{i=1}^{n} x_i = \qquad , \sum_{i=1}^{n} y_i = \qquad , \sum_{i=1}^{n} x_i y_i =$$

$$SS(xy) =$$



3: Descriptive Analysis and Bivariate Data
3.3 Linear Regression

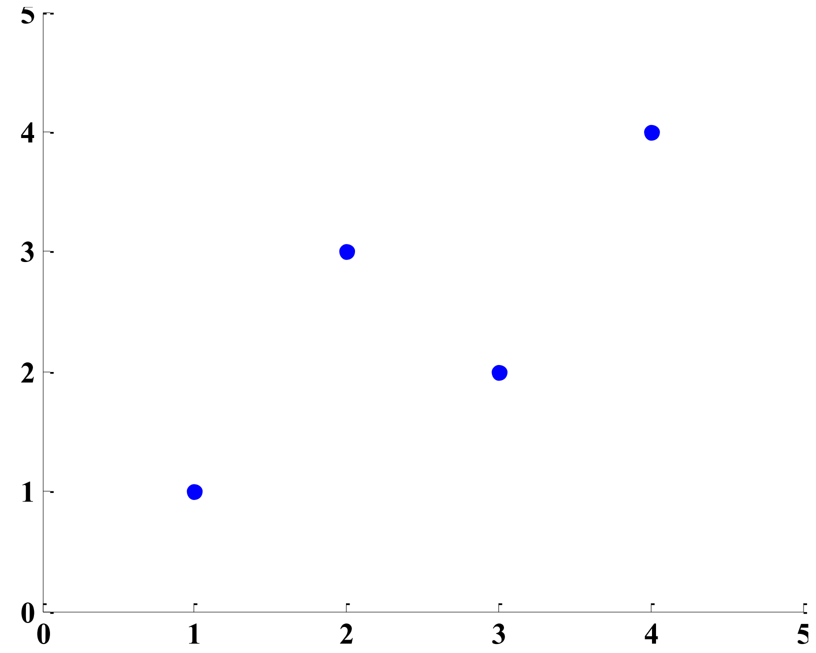
Example: Using (1,1),(3,2),(2,3),(4,4) $\bar{x} = 2.5, \ \bar{y} = 2.5$

$$b_1 = \frac{SS(xy)}{SS(x)}$$

$$SS(xy) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

$$\sum_{i=1}^{n} x_i = 10, \ \sum_{i=1}^{n} y_i = 10, \ \sum_{i=1}^{n} x_i y_i = 29$$

$$SS(xy) = 29 - \frac{1}{4}(10)(10) = 4$$



3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

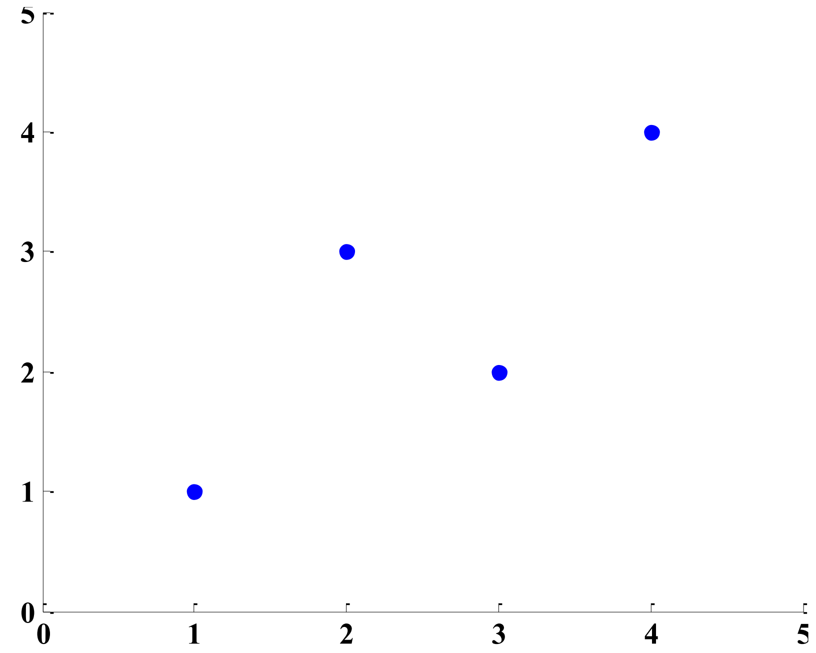
Example: Using (1,1),(3,2),(2,3),(4,4) $\bar{x} = 2.5, \;\; \bar{y} = 2.5$

$$b_1 = \frac{SS(xy)}{SS(x)}$$

$$SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$\sum_{i=1}^{n} x_i = \qquad , \quad \sum_{i=1}^{n} x_i^2 =$$

$$SS(x) =$$



3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

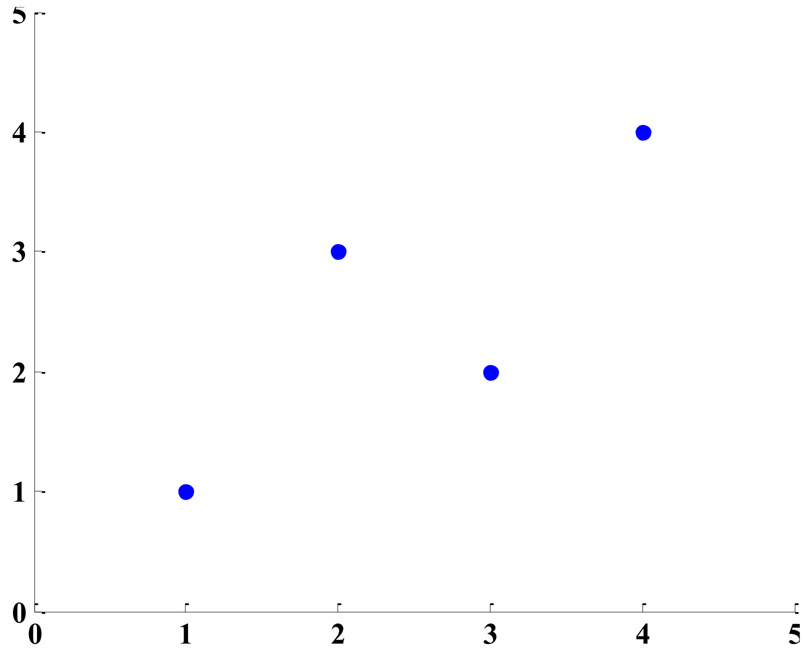
Example: Using (1,1),(3,2),(2,3),(4,4) $\bar{x} = 2.5, \ \bar{y} = 2.5$

$$b_1 = \frac{SS(xy)}{SS(x)}$$

$$SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2$$

$$\sum_{i=1}^{n} x_i = 10 \ , \ \ \sum_{i=1}^{n} x_i^2 = 30$$

$$SS(x) = 30 - \frac{1}{4}(10)^2 = 5$$

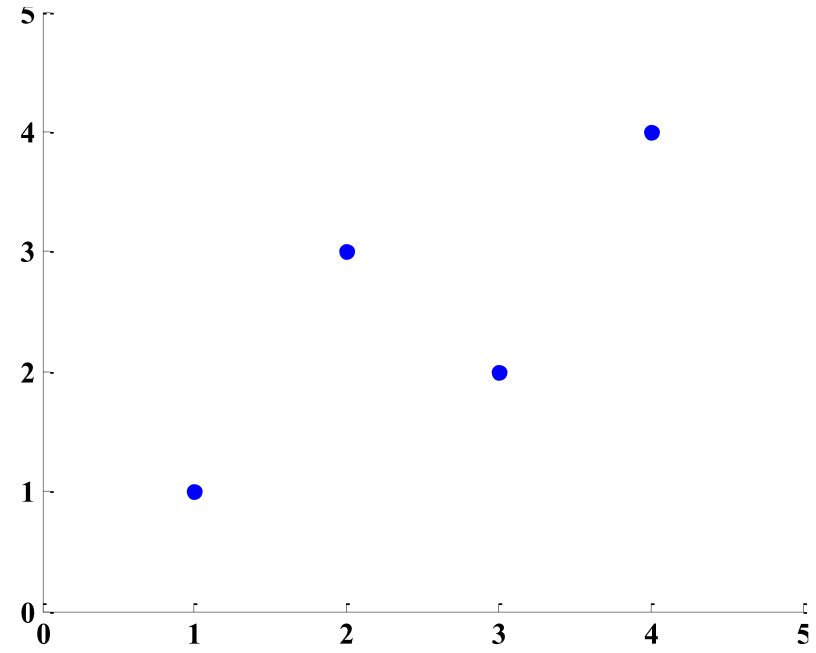


3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

Example: Using (1,1),(3,2),(2,3),(4,4) $\bar{x} = 2.5, \ \bar{y} = 2.5$

$$b_1 = \frac{SS(xy)}{SS(x)}$$



$$b_1 = \frac{SS(xy)}{SS(x)} =$$

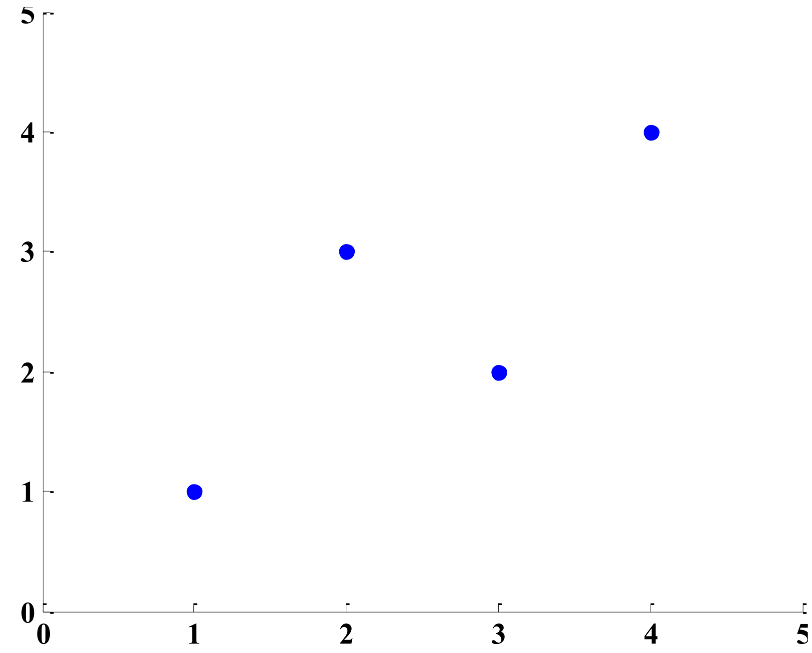
point-slope formula

$$b_0 = \bar{y} - b_1\bar{x} =$$

3: Descriptive Analysis and Bivariate Data
3.3 Linear Regression

Example: Using (1,1),(3,2),(2,3),(4,4) $\bar{x} = 2.5, \ \bar{y} = 2.5$

$$b_1 = \frac{SS(xy)}{SS(x)}$$



$$b_1 = \frac{SS(xy)}{SS(x)} = \frac{4}{5} = 0.8$$

point-slope formula

$$b_0 = \bar{y} - b_1\bar{x} = (2.5) - (0.8)(2.5) = 0.5$$

3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

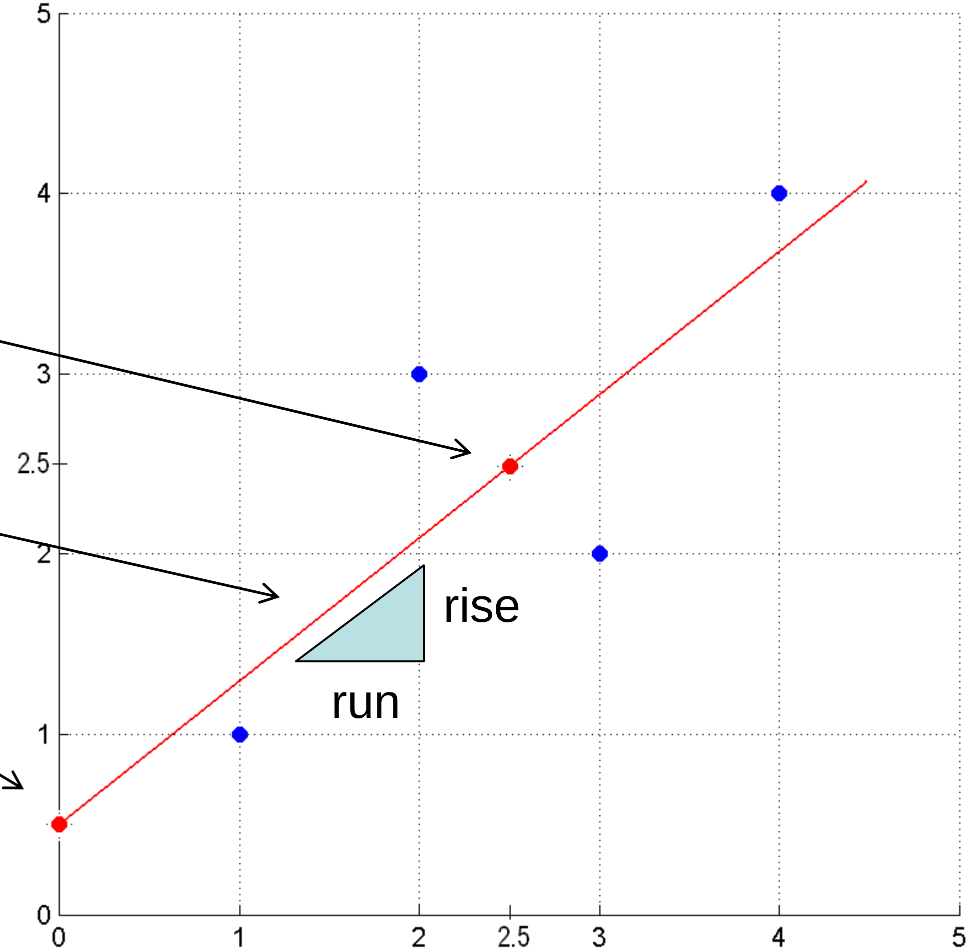
(x,y) pairs: (1,1),(3,2),(2,3),(4,4)

The line goes through $(\bar{x},\bar{y})$.

The slope is b_1=

The y - intercept b_0=

Two points



3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

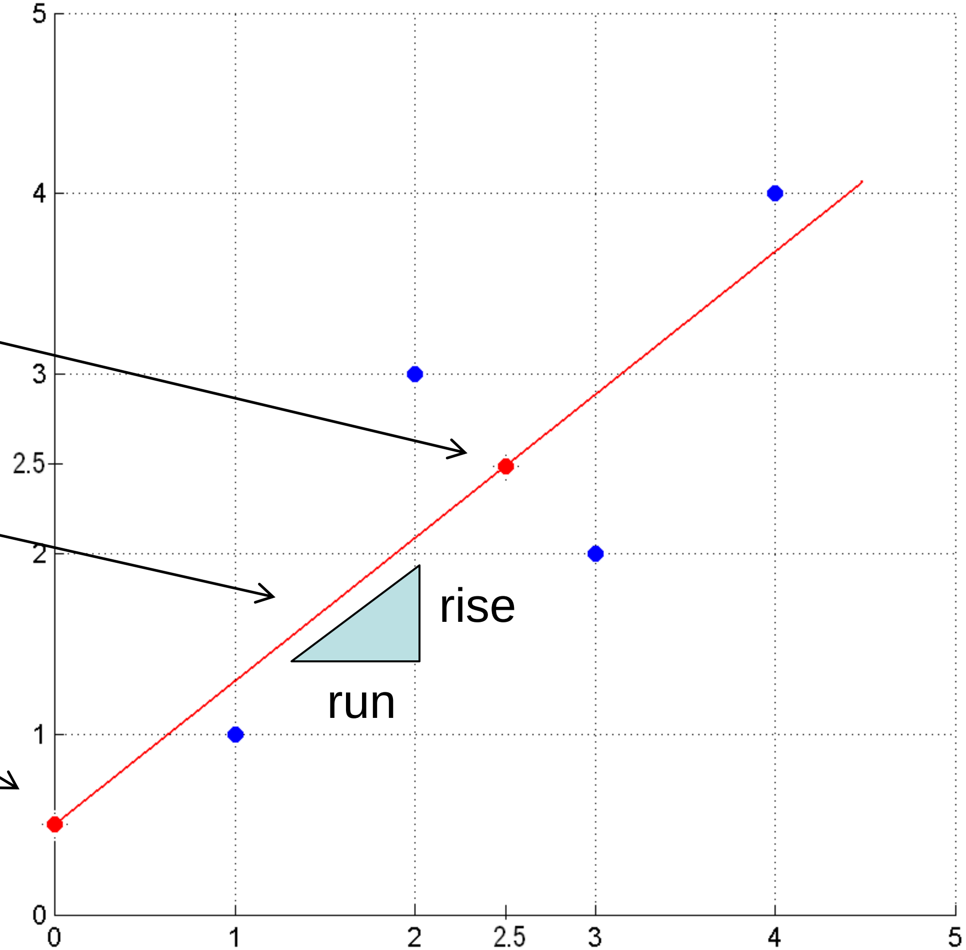
(x,y) pairs: (1,1),(3,2),(2,3),(4,4)

The line goes through $(\bar{x},\bar{y})$.

The slope is b_1=0.8.

The y - intercept b_0=0.5 .

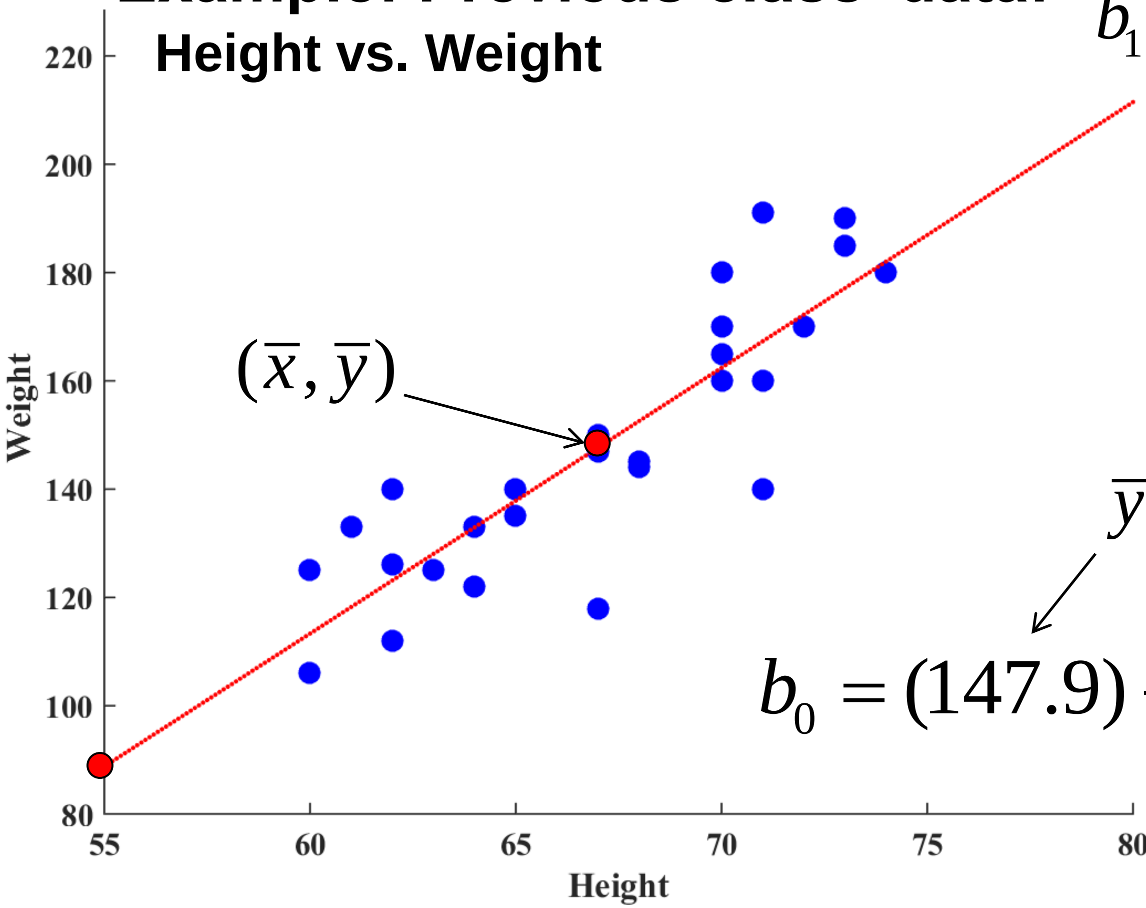
Two points (2.5,2.5) and (0,.5).



3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

Example: Previous class' data!



$$b_1 = \frac{SS(xy)}{SS(x)} = \frac{2370.1}{483.0} = 4.9$$

units of lbs/in

$\bar{y}$ b_1 $\bar{x}$

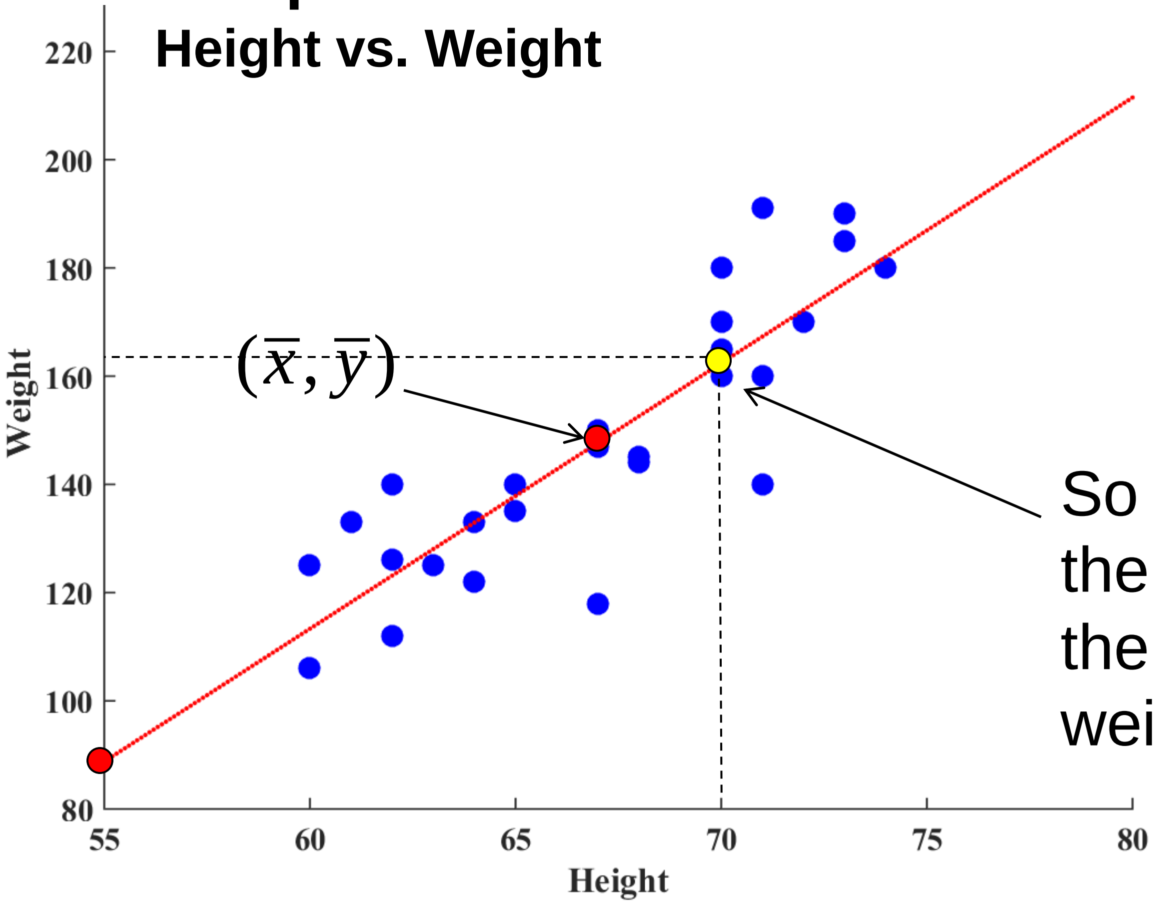
$$b_0 = (147.9) - (4.9)(67.0) = -180.4$$

point-slope formula

3: Descriptive Analysis and Bivariate Data

3.3 Linear Regression

Example: Previous class' data!



$$\hat{y} = -180.4 + 4.9x$$

So if a new student adds the class and is x=70 in tall, the best guess for his/her weight is about y=162.6 lbs.

3: Descriptive Analysis and Bivariate Data

Questions?

Homework: Read Chapter 3.2-3.3
WebAssign
Chapter 3 # 33, 44, 53, 59, 75

Page 169 Problem 3.105

$$r = b_1\sqrt{\frac{SS(x)}{SS(y)}}$$