ECO220Y

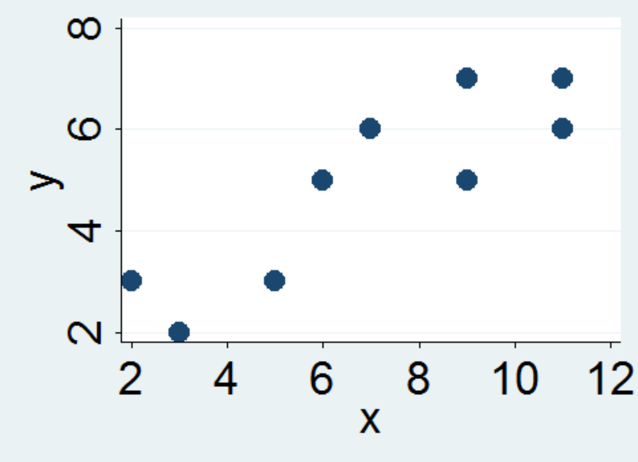
Linear Relationship:
Association, Correlation and Linear Regression
Readings: Chapters 7-8 and Handout

Fall 2011

Lecture 4
Part 2 of 2

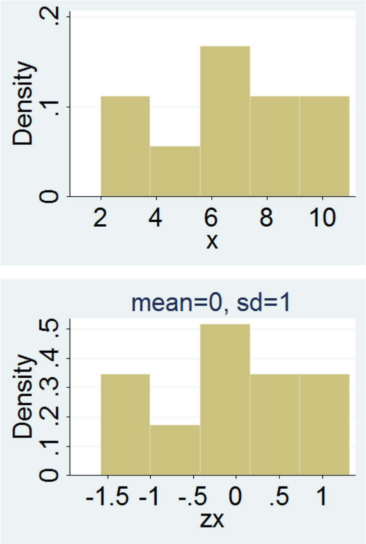
Scatter plot

X	Y
2	3
3	2
5	3
6	5
6	5
7	6
9	7
9	5
11	6
11	7



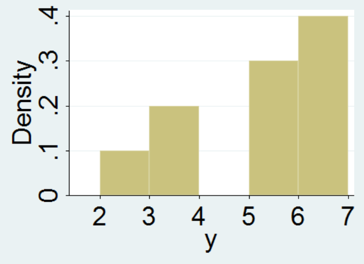
Variable X has mean of 6.9 and standard deviation 3.1

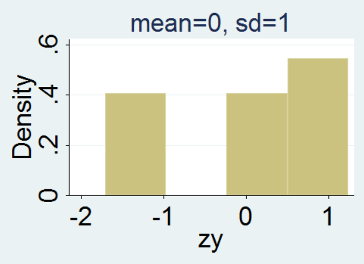
x	$z_x = \frac{x - \bar{x}}{s_x}$
2	-1.58
3	-1.26
5	-0.61
6	-0.29
6	-0.29
7	0.03
9	0.68
9	0.68
11	1.32
11	1.32

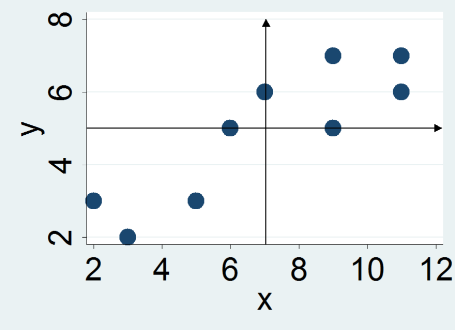


Variable Y has mean of 4.9 and standard deviation 1.7

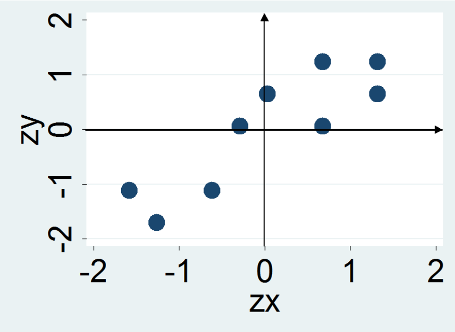
y	$z_y = \frac{y-\bar{y}}{s_y}$
3	-1.12
2	-1.71
3	-1.12
5	0.06
5	0.06
6	0.65
7	1.24
5	0.06
6	0.65
7	1.24







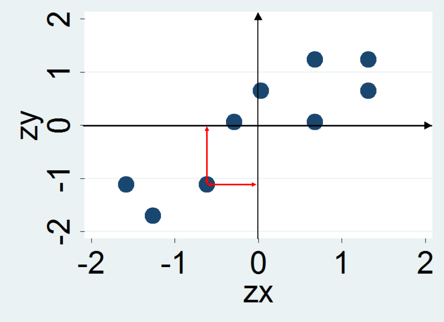
2
4
6
8
y
2
4
6
8
10
12
x



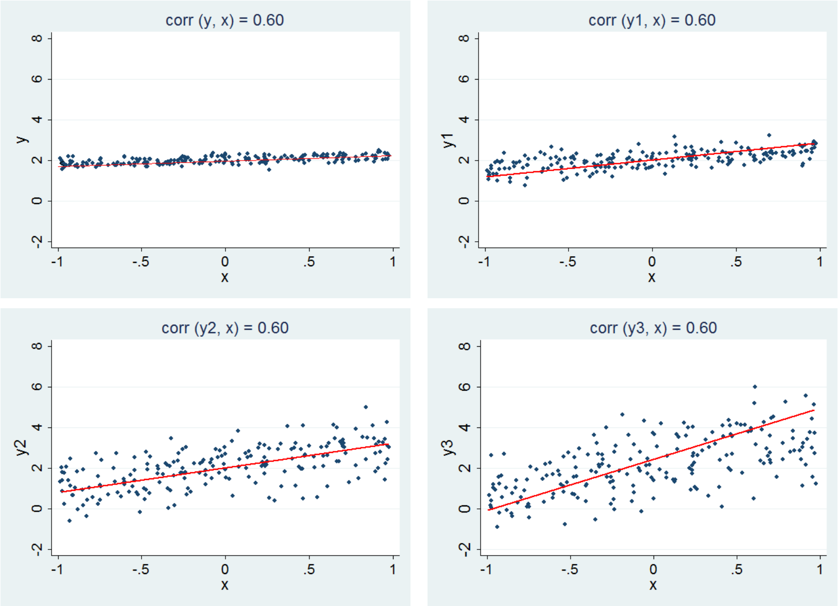
-2
-1
0
1
2
zy
-2
-1
0
1
2
zx

Correlation

z_x	z_y	$z_x z_y$
-1.58	-1.12	1.77
-1.26	-1.71	2.15
-0.61	-1.12	0.69
-0.29	0.06	-0.02
-0.29	0.06	-0.02
0.03	0.65	0.2
0.68	1.24	0.84
0.68	0.06	0.04
1.32	0.65	0.86
1.32	1.24	1.63
$\sum_{i=1}^{N} z_x z_y$		7.95



$$\rho = \frac{\sum_{i=1}^{N} z_x z_y}{n-1} = \frac{7.95}{9} = 0.88$$ (Math Box on page 171)



Slopes: 0.2, 0.6, 1.0, 1.4

The Linear Model

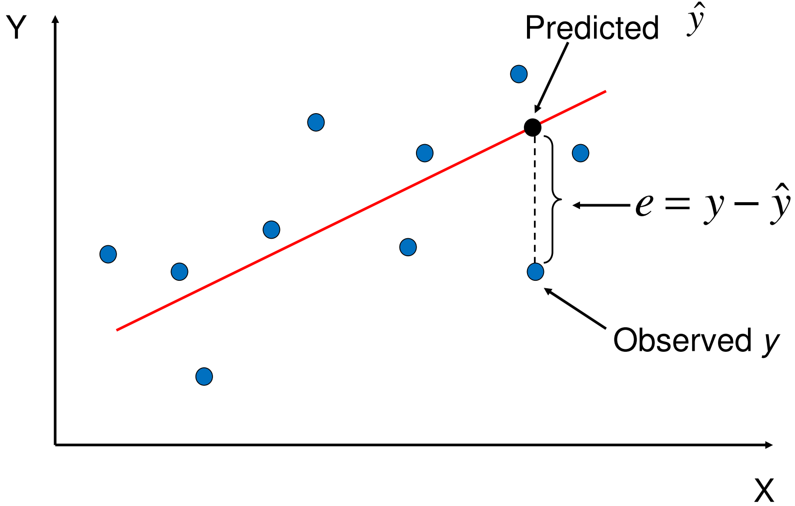
- Can we predict a student's weight from his or her height?
- Can we predict a student's test score on the final from his or her performance on other assessments?
- Can we predict the crop yield from the amounts of rainfall or fertilizer used?

Linear relationship can be described by equation:

$$y = b_0 + b_1 * x$$

where b_0 is called $y - \text{intercept}$
and b_1 is the slope of the line (rise over run)

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



How to find the line of best fit (a.k.a OLS line)?

Minimize sum of squares by solving:

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

or

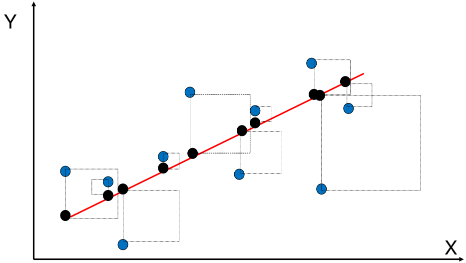
$min \sum_{i=1}^{n} (y_i - b_0 - b_1 * x)^2$

Solution is given by:

$\boxed{b_1 = r \frac{s_y}{s_x}}$

and

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



$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

Note: OLS=Ordinary Least Squares

Math Box

Minimization problem: $min \sum_{i=1}^{n}(y_i - b_0 - b_1 * x)^2$

Take two derivatives: with respect to b_0 and with respect to b_1.

Solve two equations with two unknowns (b_0 and b_1).

Result:

Familiar formula?

$\downarrow$

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

$\uparrow$

Familiar formula?

$$\boxed{b_1 = \frac{s_{xy}}{s_x^2}} \quad + \quad \boxed{s_{xy} = r_{xy} s_x s_y} \quad = \quad \boxed{b_1 = r \frac{s_y}{s_x}}$$

Note: The regression line always passes through point $(\bar{y}, \bar{x})$. Why?

Math Box Cont'd

$$SST = SSE + SSR$$

$$SST = \sum(y_i - \bar{y})^2$$

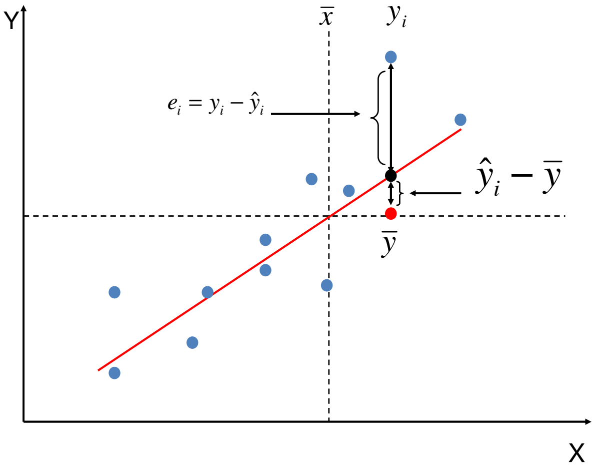
$$SSE = \sum(y_i - \hat{y}_i)^2$$

$$SSR = \sum(\hat{y}_i - \bar{y})^2$$

$$\frac{SST}{SST} = \frac{SSE}{SST} + \frac{SSR}{SST}$$

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

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



R^2 measures how well the line fits the data

Standardized Regression=Regression to the Mean

What is b_1 for the "standardized" regression?

$$\boxed{b_1 = r\frac{s_y}{s_x}, \text{ but } s_{zy} = 1 \text{ and } s_{zx} = 1 \Rightarrow b_1 = r!}$$

"Standardized" Regression Line:

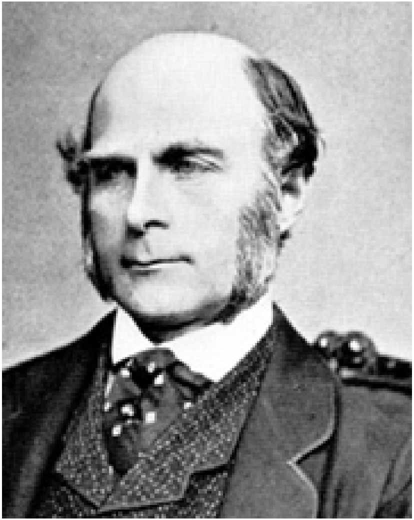
$$\boxed{\hat{z}_y = r z_x}$$

$R^2 = r^2 \Rightarrow 0\% \leq R^2 \leq 100\%$
since R^2 is measured in percentage

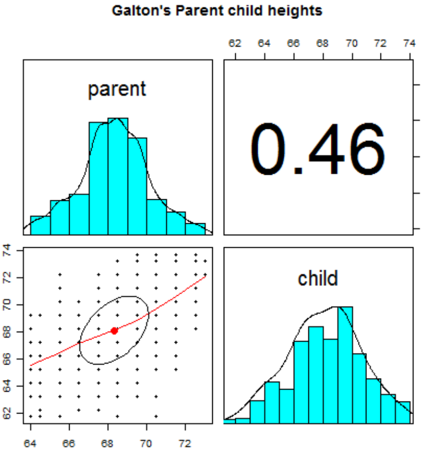
What does R^2 of 100% indicate?

What does R^2 of 0 indicate?

Regression to the Mean

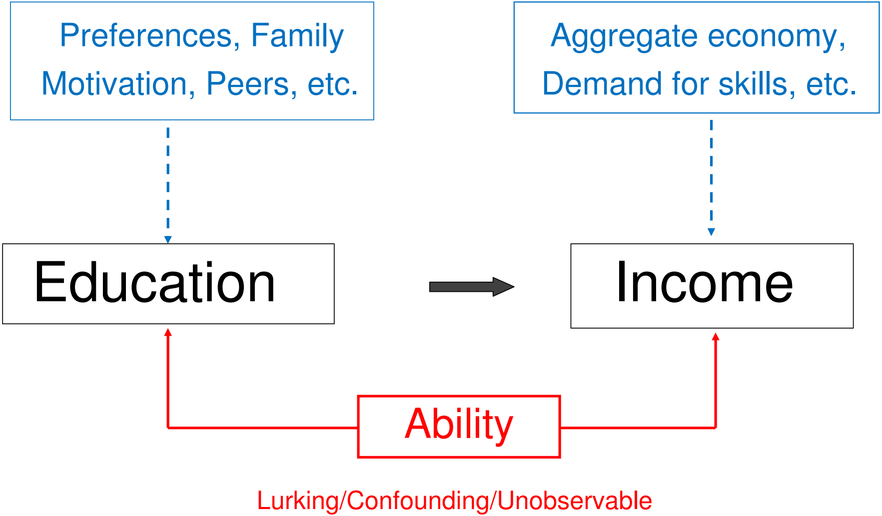


Sir Francis Galton
(1822-1911)



Interpretation of the OLS line

- Intercept has no particular meaning. It is tempting to say that when the independent variable x is 0, dependent variable y is equal to the number represented by the intercept. This is wrong. Often, we do not observe x=0 at all, and the interpretation does not make sense.
- Slope (b_1) measures marginal change in the dependent variable y associated with a change in the independent variable x. Mathematically, $b_1 = \frac{\Delta y}{\Delta x}$.
- Does the existence of correlation (slope) imply the *causal* effect or direct effect of x on y?


Preferences, Family
Motivation, Peers, etc.
Aggregate economy,
Demand for skills, etc.
Education
Income
Ability
Lurking/Confounding/Unobservable

Summary of Data Analysis

One variable		Two variables
	Want to learn about	
↙		↘
Distribution		Relationship
	Graphical Description	
↙		↘
Histogram Bar Chart, Pie Chart Line Graph		Scatter Plot Bar Chart
	Descriptive Statistics	
↙		↘
mean, median, mode IQR, range, percentile standard deviation		covariance correlation OLS line