Lesson 2: Bivariate regression, correlation, and visualization

Oscar Torres-Reyna

1

Univariate analysis → single variable (distribution, central location, and variability).

Bivariate analysis → explores the *nature* and *strength* of the relationship between two variables (correlation coefficients and linear regression).

2

Bivariate analysis → **Association between variables**

Information on one variable tells you something about another variable.

How one variable behaves in the presence of another.

3

Correlation →	**Strength and direction of the association**

Bivariate regression →	**Nature of the association:** **Value of X → Value of Y** **Change in X → Change in Y**

4

Correlation →	Goes both ways: X → Y Y → X

Bivariate regression →	One direction: Explanatory X → Response Y Change in explanatory X → Change in response Y

Total shared variation and Covariance

Total combined variation and sample covariance

The sample covariance, s_{XY}, is defined as the sum of the cross products of the x deviations and y deviations from their respective means divided by *n-1*. In the same way as the variance, the units of the covariance will be in squared units (i.e. squared dollars)

$$s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{S_{XY}}{n-1}$$

Total shared variation (numerator)

If values of *Y* tend to be above their mean when values of *X* are also above their mean, or if values of *Y* tend to be below their mean when values of *X* also are below their mean, then the covariance will be positive indicating that both variables move in the same direction.

If values of *Y* tend to be above their mean when values of *X* are below their mean, or if values of *Y* tend to be below their mean when values of *X* are above their mean, then the covariance will be negative indicating that both variables move in different direction.

Country	Median per capita daily income (MI) (US$)	Inclusive Development Index (IDI) (1 = worst, 7 = best)
Australia	44.4	5.36
Belgium	43.8	5.14
Canada	49.2	5.06
Czech Republic	24.3	5.09
Estonia	22.1	4.74
Iceland	43.4	6.07
Ireland	38.0	5.44
Israel	25.8	4.51
Italy	34.3	4.31
Japan	34.8	4.53
Korea Rep	34.2	5.09
Netherlands	43.3	5.61
Norway	63.8	6.08
Portugal	21.2	3.97
United Kingdom	39.4	4.89

The Inclusive Development Index 2018, World Economic Forum, https://www3.weforum.org/docs/WEF_Forum_IncGrwth_2018.pdf

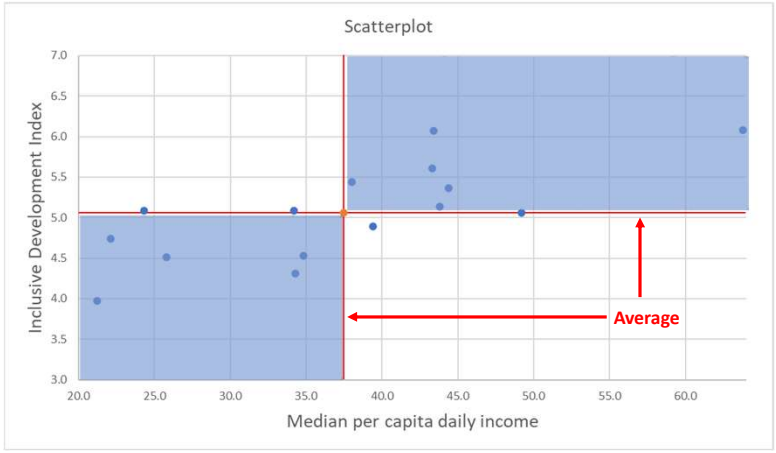
MMC, p. 99

Country	mi (x)	idi (y)	(mi-mean_mi)	(idi-mean_idi)	(idi-mean_idi)(mi-mean_mi)
Australia	44.4	5.4	6.9333	0.3007	2.0846
Belgium	43.8	5.1	6.3333	0.0807	0.5109
Canada	49.2	5.1	11.7333	0.0007	0.0078
Czech Republic	24.3	5.1	-13.1667	0.0307	-0.4038
Estonia	22.1	4.7	-15.3667	-0.3193	4.9071
Iceland	43.4	6.1	5.9333	1.0107	5.9966
Ireland	38.0	5.4	0.5333	0.3807	0.2030
Israel	25.8	4.5	-11.6667	-0.5493	6.4089
Italy	34.3	4.3	-3.1667	-0.7493	2.3729
Japan	34.8	4.5	-2.6667	-0.5293	1.4116
Korea Rep	34.2	5.1	-3.2667	0.0307	-0.1002
Netherlands	43.3	5.6	5.8333	0.5507	3.2122
Norway	63.8	6.1	26.3333	1.0207	26.8776
Portugal	21.2	4.0	-16.2667	-1.0893	17.7198
United Kingdom	39.4	4.9	1.9333	-0.1693	-0.3274
Mean	37.5	5.1			
Sum			0.0000	0.0000	70.8817

Sxy

Total shared variation

$$\textbf{Sample covariance} = s_{XY} = \frac{S_{XY}}{n-1} = \frac{70.9}{15-1} = 5.063$$



Correlation coefficient

Estimation of the correlation coefficient

The correlation coefficient is the proportion of the total shared variation, S_{XY} out of the combined variation between X and Y (i.e. the product of the square root of the total variation of *X* and *Y)*:

$$Correlation\ coefficient = r = \frac{S_{XY}}{\sqrt{S_X}\sqrt{S_Y}}$$

$$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

Sum of the cross products of the *x* deviations with the *y* deviations (total shared variation)

$$S_{XX} = \sum (X_i - \bar{X})^2$$

$$S_{YY} = \sum (Y_i - Y)^2$$

Sum of the squared deviations of the *x* and *y* observations

Total variation

Manual estimation of the correlation coefficient in Excel

$$r = \frac{S_{XY}}{\sqrt{S_X}\sqrt{S_Y}}$$

$$S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

$$S_{XX} = \sum (X_i - \bar{X})^2$$

$$S_{YY} = \sum (Y_i - Y)^2$$

	x	y	(x-x̄)	(x-x̄)²	(y-ȳ)	(y-ȳ)²	(x-x̄)*(y-ȳ)
	2	5	0	0	1	1	0
	1	3	-1	1	-1	1	1
	5	6	3	9	2	4	6
	0	2	-2	4	-2	4	4
Mean	2	4		14		10	11
	↑	↑		↑		↑	↑
	x̄	ȳ		Sxx		Syy	Sxy
Correlation coefficient =					0.9297		

	x	y	(x-x̄)	(x-x̄)²	(y-ȳ)	(y-ȳ)²	(x-x̄)*(y-ȳ)
	2	5	=+B2-B$7	=+D2^2	=+C2-C$7	=+F2^2	=+D2*F2
	1	3	=+B3-B$7	=+D3^2	=+C3-C$7	=+F3^2	=+D3*F3
	5	6	=+B4-B$7	=+D4^2	=+C4-C$7	=+F4^2	=+D4*F4
	0	2	=+B5-B$7	=+D5^2	=+C5-C$7	=+F5^2	=+D5*F5
Mean	=+AVERAGE(B2:B5)	=+AVERAGE(C2:C5)		=SUM(E2:E6)		=SUM(G2:G6)	=+SUM(H2:H5)
	↑	↑		↑		↑	↑
	x̄	ȳ		Sxx		Syy	Sxy
Correlation coefficient =	=+H7/(SQRT(E7)*SQRT(G7))						

Estimation of the correlation coefficient in Excel

Excel can estimate the correlation coefficient automatically by using the Data Analysis ToolPak. Go to Data → Data Analysis. A window will pop-up, select "Correlation" and click OK. The 'Correlation' window will pop-up, fill in as indicated below, then click OK.

	x	y	(x-x̄)	(x-x̄)²	(y-ȳ)	(y-ȳ)²	(x-x̄)*(y-ȳ)
	2	5	0	0	1	1	0
	1	3	-1	1	-1	1	1
	5	6	3	9	2	4	6
	0	2	-2	4	-2	4	4
Mean	2	4		14		10	11
	↑	↑		↑		↑	↑
	x̄	ȳ		Sxx		Syy	Sxy
Correlation coefficient =			0.9297				

Correlation window: Input Range: $B$1:$C$5; Grouped By: Columns; Labels in first row; Output options: Output Range: $G$11; OK; Cancel; Help

	x	y
x	1	
y	0.92966968	1

Excel has also the `correl()` function for direct estimation.

Alternative estimation of the correlation coefficient

The correlation coefficient can also be understood as the sum of the product of the z-score of *X* and the z-score of *Y* divided by *n-1*.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

	x	y	z-score of x (x-x̄) / std. dev.	z-score of y (y-ȳ) / std. dev.	(z-score of x)*(z-score of y) [(x-x̄) / std. dev.]*[(y-ȳ) / std. dev.]
	2	5	0	0.547722558	0
	1	3	-0.46291005	-0.547722558	0.253546276
	5	6	1.38873015	1.095445115	1.521277659
	0	2	-0.9258201	-1.095445115	1.014185106
Mean	2	4			
Standard deviation	2.160246899	1.825741858			
Sum					2.789009041
N	4	4			
Correlation coefficient =	0.92966968				

MMC, p. 92

Alternative estimation of the correlation coefficient

The correlation coefficient can also be understood as the sum of the product of the z-score of *X* and the z-score of *Y* divided by *n-1*.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

	x	y	z-score of x (x-x̄) / std. dev.	z-score of y (y-ȳ) / std. dev.	(z-score of x)*(z-score of y) [(x-x̄) / std. dev.]*[(y-ȳ) / std. dev.]
	2	5	=+(B3-B$8)/B$9	=+(C3-C$8)/C$9	=+D3*E3
	1	3	=+(B4-B$8)/B$9	=+(C4-C$8)/C$9	=+D4*E4
	5	6	=+(B5-B$8)/B$9	=+(C5-C$8)/C$9	=+D5*E5
	0	2	=+(B6-B$8)/B$9	=+(C6-C$8)/C$9	=+D6*E6
Mean	=+AVERAGE(B3:B6)	=+AVERAGE(C3:C6)			
Standard deviation	=+STDEV.S(B3:B6)	=+STDEV.S(C3:C6)			
Sum					=SUM(F3:F9)
N	=+ROWS(B3:B6)	=+ROWS(C3:C6)			
Correlation coefficient =	=+F10/(B11-1)				

MMC, p. 92

Correlation coefficient (*r*)

The correlation coefficient or *Pearson product-moment correlation*:

1. It is an index, it goes from -1 to 1.
2. Measures the **strength** and **direction** of the linear relationship or linear association between two variables.
3. Explores how the value of one variable move when another moves.
4. It is used only with continuous variables and, with some care, with integer variables.
5. It assumes a linear relationship.

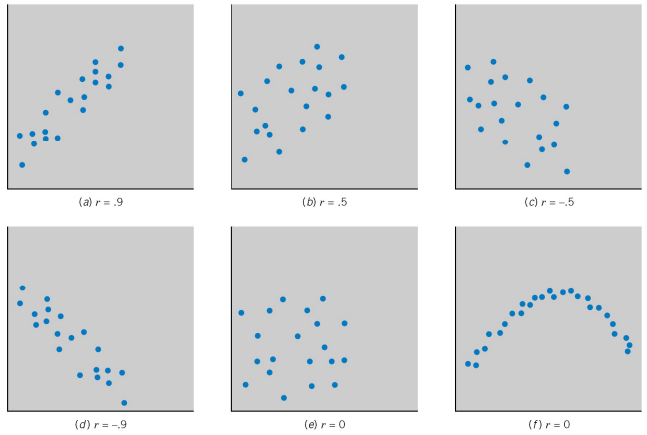
Correlation coefficient (*r*)

Characteristics of the correlation coefficient are:

1. Its magnitude defines the strength of the relationship
   1. *r* closest to 1 means a very strong positive linear relationship.
   2. *r* closest to -1 means a very strong negative linear relationship.
   3. *r* closest to 0 means a weak or no linear relationship.
2. Its sign defines the direction of the relation:
   1. *When positive*, if *x* increases then *y* also increases
   2. *When negative*, if *x* increases then *y* decreases
   3. *When equal to zero*, either no relationship between *x* and *y*, or is not linear in nature.
   4. *When equals to 1*, perfect positive linear relationship.
   5. *When equals to -1*, perfect negative linear relationship

J&B, 96-98

Correlation coefficient (*r*) - Scatterplots



Source: Johnson, Richard A., Gouri K. Bhattacharyya, *Statistics. Principles and Methods*, 7th edition, John Wiley & Sons, 2014, p. 97

Type of relationship given level of correlation coefficient

Level (absolute value)	Economics	Social sciences	Hard sciences
0.80 - 1.0	Very strong	Very strong	Very strong
0.60 – 0.79	Strong	Very strong	Strong
0.40 – 0.59	Moderate	Strong	Moderate
0.20 – 0.39	Weak	Moderate	Weak
0.00 – 0.19	Very weak/none	Weak/very/none	Very weak/none

Main idea from Salkind

Coefficient of determination (r^2)

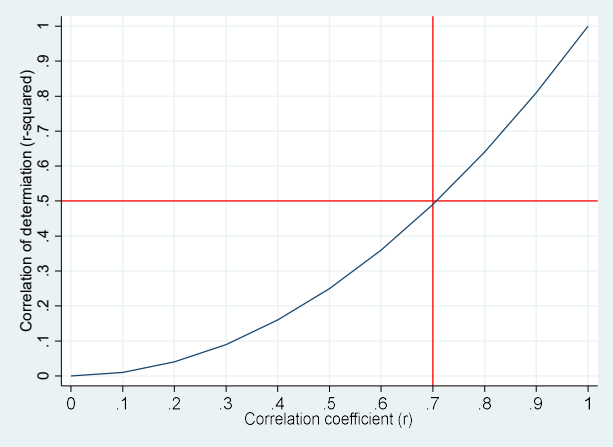
Coefficient of determination (r^2)

The correlation coefficient measures the strength of the relationship between two variables.

The coefficient of determination measures how much the variance observed in variable *x* is shared with variable *y*. This is, how much of the variance in *x* is explained by *y* (and vice versa).

It is estimated by squaring the correlation coefficient.

Correlation coefficient and coefficient of determination



Linear Regression

Correlation →	**Strength and direction of the association**

Linear bivariate regression →	**Nature of the association:** **Value of X → Value of Y** **Change in X → Change in Y**

- Does on-the-job training improve workers' skills?
- Does small class size improve students' learning?
- Do higher cigarette taxes lower smoking habits?
- Do higher tariffs will generate higher GDP?

Linear regression

Linear regression is a method to fit a regression line to predict the value of outcome *y* from a predictor *x*.

The regression line shows how much *y* changes when *x* changes by one unit.

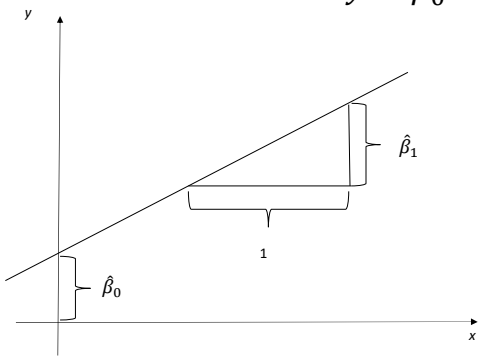
Some conditions apply:

- *Y* must be continuous.
- *X* can be continuous, discrete and/or categorical.

Linear regression

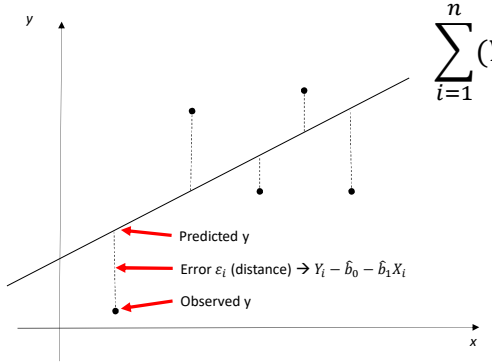
The linear regression is represented by the following equation:

$$\hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x$$



Ordinary Least-Squares (OLS)

The method of Ordinary Least-Squares (OLS) selects the linear fit that has the lowest sum of the error squares for all data points:



Linear regression

The linear regression equation is of the form:

$$y = b_o + b_1 x + \varepsilon_i$$

y : the response or outcome.

b_0 : intercept, indicates where the regression line crosses the y-axis.

b_1: slope, indicates *how much y increases (or decreases) when x increases by one unit*.

x : the predictor, input, explanatory variable(s).

ε: error term, part not explained by the explanatory variable.

Linear regression

The linear equation is $y = b_o + b_1 x$

Slope ➡ $b_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{S_{xy}}{S_{xx}} = r\frac{s_y}{s_x}$

$S_{XY} = \sum(X_i - \bar{X})(Y_i - \bar{Y})$ Sum of the cross products of the *x* deviations with the *y* deviations

$S_{XX} = \sum(X_i - \bar{X})^2$

$S_{YY} = \sum(Y_i - Y)^2$

Sum of the squared deviations of the *x* and *y* observations (total variation)

r → correlation coefficient
s_y → standard deviation of y
s_x → standard deviation of x

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

Country	Inclusive Development Index (IDI) (1 = worst, 7 = best)	Median per capita daily income (MI) (US$)
Australia	5.36	44.4
Belgium	5.14	43.8
Canada	5.06	49.2
Czech Republic	5.09	24.3
Estonia	4.74	22.1
Iceland	6.07	43.4
Ireland	5.44	38.0
Israel	4.51	25.8
Italy	4.31	34.3
Japan	4.53	34.8
Korea Rep	5.09	34.2
Netherlands	5.61	43.3
Norway	6.08	63.8
Portugal	3.97	21.2
United Kingdom	4.89	39.4

The Inclusive Development Index 2018, World Economic Forum, https://www3.weforum.org/docs/WEF_Forum_IncGrwth_2018.pdf

MMC, p. 99

Linear regression (Excel table)

Country	mi (x)	idi (y)	(mi-mean_mi)	(mi-mean_mi)²	(idi-mean_idi)	(idi-mean_idi)²	(idi-mean_idi)(mi-mean_mi)
Australia	44.4	5.4	6.9	48.1	0.3	0.1	2.1
Belgium	43.8	5.1	6.3	40.1	0.1	0.0	0.5
Canada	49.2	5.1	11.7	137.7	0.0	0.0	0.0
Czech Republic	24.3	5.1	-13.2	173.4	0.0	0.0	-0.4
Estonia	22.1	4.7	-15.4	236.1	-0.3	0.1	4.9
Iceland	43.4	6.1	5.9	35.2	1.0	1.0	6.0
Ireland	38.0	5.4	0.5	0.3	0.4	0.1	0.2
Israel	25.8	4.5	-11.7	136.1	-0.5	0.3	6.4
Italy	34.3	4.3	-3.2	10.0	-0.7	0.6	2.4
Japan	34.8	4.5	-2.7	7.1	-0.5	0.3	1.4
Korea Rep	34.2	5.1	-3.3	10.7	0.0	0.0	-0.1
Netherlands	43.3	5.6	5.8	34.0	0.6	0.3	3.2
Norway	63.8	6.1	26.3	693.4	1.0	1.0	26.9
Portugal	21.2	4.0	-16.3	264.6	-1.1	1.2	17.7
United Kingdom	39.4	4.9	1.9	3.7	-0.2	0.0	-0.3
Mean	37.5	5.1					
Sum			0.0	1830.6	0.0	5.1	70.9
	Mean x	Mean y		Sxx		Syy	Sxy

Correlation coefficient = 0.7357 ← $r = \frac{S_{XY}}{\sqrt{S_X}\sqrt{S_Y}}$

Slope = 0.0387 ← $b_1 = \frac{S_{xy}}{S_{xx}}$

Intercept= 3.6086 ← $b_0 = \bar{y} - b_1\bar{x}$

$S_{XY} = \sum(X_i - \bar{X})(Y_i - \bar{Y})$

$S_{XX} = \sum(X_i - \bar{X})^2$

$S_{YY} = \sum(Y_i - Y)^2$

average idi = 3.6086 + 0.0387*mi

Linear regression (Excel table)

	A	B	C	D	E	F	G	H
1	Country	mi (x)	idi (y)	(mi-mean_mi)	(mi-mean_mi)^2	(idi-mean_idi)	(idi-mean_idi)^2	(idi-mean_idi)(mi-mean_mi)
2	Australia	44.4	5.4	=+B2-B$18	=+D2^2	=+C2-C$18	=+F2^2	=+F2*D2
3	Belgium	43.8	5.1	=+B3-B$18	=+D3^2	=+C3-C$18	=+F3^2	=+F3*D3
4	Canada	49.2	5.1	=+B4-B$18	=+D4^2	=+C4-C$18	=+F4^2	=+F4*D4
5	Czech Republic	24.3	5.1	=+B5-B$18	=+D5^2	=+C5-C$18	=+F5^2	=+F5*D5
6	Estonia	22.1	4.7	=+B6-B$18	=+D6^2	=+C6-C$18	=+F6^2	=+F6*D6
7	Iceland	43.4	6.1	=+B7-B$18	=+D7^2	=+C7-C$18	=+F7^2	=+F7*D7
8	Ireland	38.0	5.4	=+B8-B$18	=+D8^2	=+C8-C$18	=+F8^2	=+F8*D8
9	Israel	25.8	4.5	=+B9-B$18	=+D9^2	=+C9-C$18	=+F9^2	=+F9*D9
10	Italy	34.3	4.3	=+B10-B$18	=+D10^2	=+C10-C$18	=+F10^2	=+F10*D10
11	Japan	34.8	4.5	=+B11-B$18	=+D11^2	=+C11-C$18	=+F11^2	=+F11*D11
12	Korea Rep	34.2	5.1	=+B12-B$18	=+D12^2	=+C12-C$18	=+F12^2	=+F12*D12
13	Netherlands	43.3	5.6	=+B13-B$18	=+D13^2	=+C13-C$18	=+F13^2	=+F13*D13
14	Norway	63.8	6.1	=+B14-B$18	=+D14^2	=+C14-C$18	=+F14^2	=+F14*D14
15	Portugal	21.2	4.0	=+B15-B$18	=+D15^2	=+C15-C$18	=+F15^2	=+F15*D15
16	United Kingdom	39.4	4.9	=+B16-B$18	=+D16^2	=+C16-C$18	=+F16^2	=+F16*D16
17								
18	Mean	=+AVERAGE(B2:B16)	=+AVERAGE(C2:C16)					
19	Sum			=SUM(D2:D18)	=SUM(E2:E18)	=SUM(F2:F18)	=SUM(G2:G18)	=SUM(H2:H18)
20		Mean x	Mean y		Sxx		Syy	Sxy
21								
22	Correlation coefficient =		=+H19/(SQRT(E19)*SQRT(G19))					
23	Beta-hat 1 =		=+H19/E19					
24	beta-hat 0 =		=+C18-(B23*B18)					
25	average idi = 3.6086 + 0.0387*mi							

$$r = \frac{S_{XY}}{\sqrt{S_X}\sqrt{S_Y}} \qquad b_1 = \frac{S_{xy}}{S_{xx}} \qquad S_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) \qquad S_{XX} = \sum (X_i - \bar{X})^2$$

$$b_0 = \bar{y} - b_1\bar{x} \qquad S_{YY} = \sum (Y_i - Y)^2$$

Interpretation

The equation is

average idi = 3.6086 + 0.0387 * mi

or random variable (idi) = 3.6086 + 0.0387 * random variable(mi)

Slope = 0.0387, means that when median income increases one dollar, the average Inclusive Development Income (IDI) increases by 0.0387 points. The operating word here is "average" points.

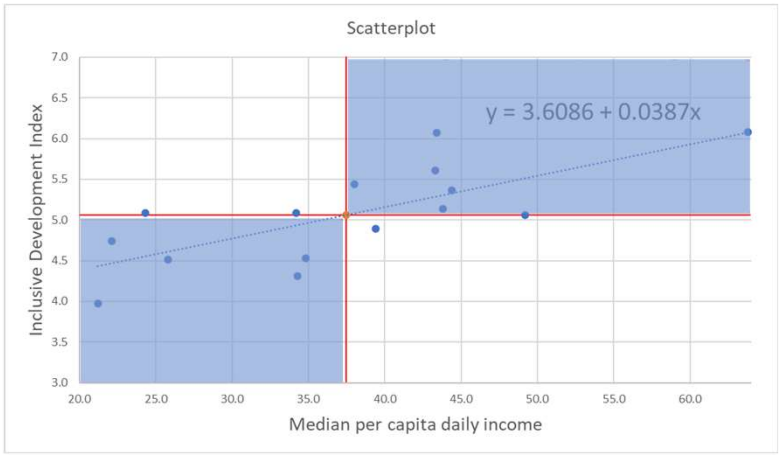
If, for example we want to predict the average IDI when median income is $33, we substitute in the equation:

```
average idi = 3.6086 + 0.0387 * 33 = 4.8857
```

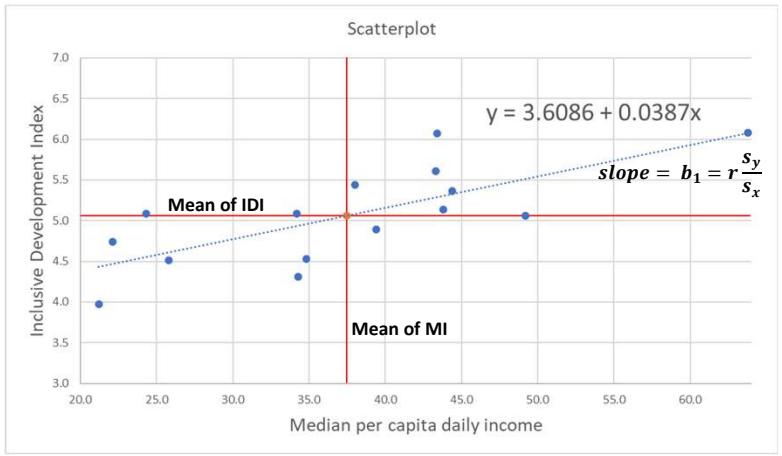
The predicted IDI time with a median income of $33 is 4.88

Intercept = 3.6086, means that if median income is equal to zero, the average IDI is 3.6086. This is the point where the line crosses the y-axis.

Scatterplot – adding linear fit

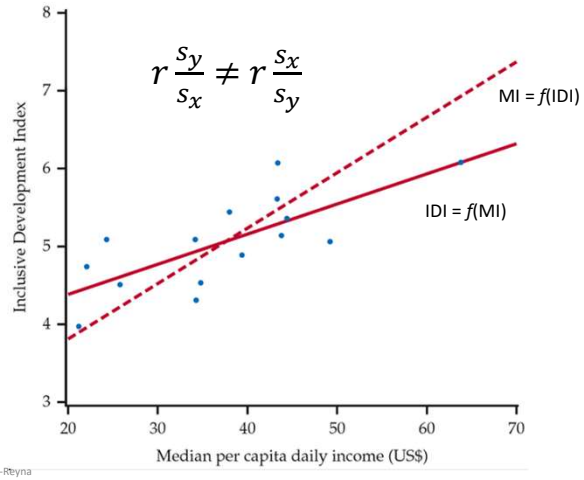


The mean of *y* and mean of *x* are on the regression line



It matters which is the outcome and which is the predictor

$$r\frac{s_y}{s_x} \neq r\frac{s_x}{s_y}$$



MMC, p. 106

Connection between correlation (*r*) and the slope (b_1)

The slope is a weighted value of the ratio between the standard deviation of *y* and standard deviation of *x*.

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

Connection between correlation (*r*) and the slope (b_1)

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

The correlation coefficient shows the strength and direction of the relationship between y and x. It provides the sign of the slope (positive or negative).

The ratio $\frac{s_y}{s_x}$ shows how *y* moves per unit of *x*.

Understanding linear regression as a modified mean

In the absence of any other variable, we can predict a value of *y* by taking its mean.

In the presence of other variables like *x*, we can predict an average value of *y* by using the linear equation

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

No effect of X on Y

		y	x - mean(x)	y - mean(y)	[x - mean(x)]²	[y - mean(y)]²	[x - mean(x)]*[y - mean(y)]
	3	2	-1.333	-1.500	1.778	2.250	2.000
	3	5	-1.333	1.500	1.778	2.250	-2.000
	4	2	-0.333	-1.500	0.111	2.250	0.500
	4	5	-0.333	1.500	0.111	2.250	-0.500
	6	2	1.667	-1.500	2.778	2.250	-2.500
	6	5	1.667	1.500	2.778	2.250	2.500
Sum			0.000	0.000	9.333	13.500	0.000
Mean	4.333	3.500					
Variance	1.867	2.700					
St. Dev. (s)	1.366	1.643					
					Sxx	Syy	Sxy
Correlation (r)	0						
Beta-hat 1	0	0					
Beta-hat 0	3.500	3.500					

$$b_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{S_{xy}}{S_{xx}} = r\frac{s_y}{s_x}$$

y = 3.5 + 0 * x

$$b_0 = \bar{y} - b_1\bar{x} \qquad r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Notice b_0 is equal to the mean of *y*

Values of *X* may affect average of *Y*

	x	y	x - mean(x)	y - mean(y)	[x - mean(x)]²	[y - mean(y)]²	[x - mean(x)]*[y - mean(y)]
	3	13	-0.667	-4.167	0.444	17.361	2.778
	4	33	0.333	15.833	0.111	250.694	5.278
	8	21	4.333	3.833	18.778	14.694	16.611
	2	5	-1.667	-12.167	2.778	148.028	20.278
	4	3	0.333	-14.167	0.111	200.694	-4.722
	1	28	-2.667	10.833	7.111	117.361	-28.889
Sum			0.000	0.000	29.333	748.833	11.333
Mean	3.667	17.167					
Variance	5.867	149.767					
St. Dev. (s)	2.422	12.238					
					Sxx	Syy	Sxy
Correlation (r)	0.076						
Beta-hat 1	0.386	0.386					
Beta-hat 0	15.750	15.750					

$$b_1 = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{S_{xy}}{S_{xx}} = r\frac{s_y}{s_x}$$

y = 15.750 + 0.386 * x

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

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

The r^2 in linear regression

In the absence of any other variable, the total variation of *y* is estimated as

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

In the presence of *x*, the total variation of the estimate $\hat{y}$ is calculated as (see next slide):

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

Where $\hat{y} = b_o + b_1 x$

Having *x*, moves the deviation or distance from:

$$(y - \bar{y}) \rightarrow (y - \hat{y})$$

If we take the difference:

$$(y - \bar{y}) - (y - \hat{y})$$

$$y - \bar{y} - y + \hat{y}$$

$$\hat{y} - \bar{y}$$

Therefore, the total variation of $\hat{y}$ is estimated as:

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

The ratio of the total variation due to the regression and the total variation in the absence of regression gives the proportion of variation explained by the linear regression:

$$\frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2} = r^2$$

	x	y	y-hat	y - mean(y)	[y - mean(y)]²	[y-hat - mean(y)]	[y-hat - mean(y)]²
	3	13	17	-4.167	17.361	-0.258	0.066
	4	33	17	15.833	250.694	0.129	0.017
	8	21	19	3.833	14.694	1.674	2.803
	2	5	17	-12.167	148.028	-0.644	0.415
	4	3	17	-14.167	200.694	0.129	0.017
	1	28	16	10.833	117.361	-1.030	1.062
Sum				0.000	748.833	0.000	4.379
Mean	3.667	17.167					
					Syy		Syy-hat
Correlation (*r*)	0.076						
Beta-hat 1	0.386			***y = 15.750 + 0.386*x***			
Beta-hat 0	15.750						
R-squared (r^2)	0.005847						0.005847

$$r^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

	x	y	y-hat	y - mean(y)	$[y - mean(y)]^2$	[y-hat - mean(y)]	$[y\text{-hat} - mean(y)]^2$
	3	13	=+$B$17+($B$16*B2)	=+C2-C$10	=+E2^2	=+D2-$C$10	=+G2^2
	4	33	=+$B$17+($B$16*B3)	=+C3-C$10	=+E3^2	=+D3-$C$10	=+G3^2
	8	21	=+$B$17+($B$16*B4)	=+C4-C$10	=+E4^2	=+D4-$C$10	=+G4^2
	2	5	=+$B$17+($B$16*B5)	=+C5-C$10	=+E5^2	=+D5-$C$10	=+G5^2
	4	3	=+$B$17+($B$16*B6)	=+C6-C$10	=+E6^2	=+D6-$C$10	=+G6^2
	1	28	=+$B$17+($B$16*B7)	=+C7-C$10	=+E7^2	=+D7-$C$10	=+G7^2
Sum				=SUM(E2:E8)	=SUM(F2:F8)	=SUM(G2:G8)	=SUM(H2:H8)
Mean	=+AVERAGE(B2:B7)	=+AVERAGE(C2:C7)					
Variance	=+VAR.S(B2:B7)	=+VAR.S(C2:C7)					
St. Dev. (s)	=+STDEV.S(B2:B7)	=+STDEV.S(C2:C7)					
					Syy		Syy-hat
Correlation (r)	0.076						
Beta-hat 1	0.386						
Beta-hat 0	15.750						
R-squared	=+B15^2						=+H9/F9

y = 15.750 + 0.386*x

$$r^2 = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

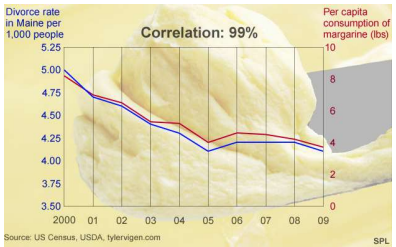
Assumptions of linear regression

1. The relationship between Y and X is linear in the parameter b_0 and b_1.
2. The expected value of the error term is 0; $E(\varepsilon)=0$.
3. The variance of ε is the same for all observations, i.e. heteroskedastic, $E(\varepsilon^2)=\sigma_\epsilon^2$.
4. $E(\epsilon_i \epsilon_j) = 0, for\ i \neq j$. The error term is uncorrelated across all observations.
5. The error term is normally distributed.

Lurking variable

Lurking variable, correlation ≠ causation

High correlation between *x* and *y* does not necessary mean that *x* cause *y*. In some cases, that correlation may be mediated by a third variable. This third variable is called a **lurking** variable, and the false correlation it generates is called a **spurious correlation**.



"Margarine consumption is linked to divorce"

Shall we ban the sale of margarine to stop people from getting divorce?

Is there a third variable in place?

How about economic crisis

Source: http://www.bbc.com/news/magazine-27537142

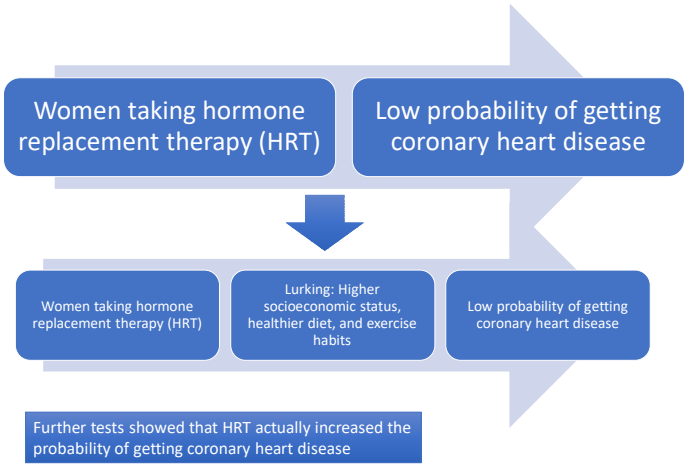
Lurking variable, correlation ≠ causation

A correlation of 0.77 was found between the number of personal fouls and the number of points scored for each member of the Los Angeles Lakers basketball team for the 2009-2010 season (source: *Statistical Reasoning in Sports*, by Josh Tabor and Chris Franklin, p. 357).

Does this mean that, in order to score more points, players should commit more fouls?

Not necessarily, apparently, the lurking variable here was play time. The more the players play, the more the chances of scoring points, but also the more the chance to commit personal fouls.

Lurking variable, correlation ≠ causation



Source: http://www.bbc.com/news/magazine-27537142

Lurking variable, correlation ≠ causation

Lurking Variables

The Insurance Institute for Highway Safety 2011 report announced the safest and unsafest 2005–2008 car models for the period 2006 to 2009 in terms of fewest fatalities per one million registered vehicle years. The death rates, shown in parentheses, are given in terms of one million cars that are registered for the year.

Lowest Fatality Rates		Highest Fatality Rates	
Mercedes E-class 4-door AWD	(0)	Nissan 350Z	(143)
Mercedes E-class 4-door	(12)	Chevrolet Aveo 4-door	(119)
Saab 9–3 4-door	(16)	Chevrolet Cobalt 4-door	(117)
Honda Accord 4-door	(19)	Kia Spectra	(102)
Accura 3.2 TL	(21)	Chevrolet Malibu Classic 4-door	(99)
Accura RL	(21)	Hyundai Tiburon 2-door	(96)
		Nissan Versa 4-door	(96)

Although it must be acknowledged there is truth in the statement that large cars are generally safer than small cars, there is a big lurking variable here—the driver. How often does the teenager cruise in the luxury car? There is a strong correlation between the age of the driver and the type of car driven and also between the age of the driver and driver behavior.

J&B, p. 103

Lurking variable, correlation ≠ causation

The lesson is that just because two variables move in the same direction (or in the exact opposite one), does not mean that one cause the other.

For more examples of spurious correlations see the following site:

http://www.tylervigen.com/

Book references

- CFA Institute, *Quantitative Investment Analysis*, 4th edition, 2015. [QIA].
- Johnson, Richard A., Gouri K. Bhattacharyya, *Principles and Methods*, 7th edition, John Wiley & Sons, 2014 [JB].
- Moore, David, George McCabe, and Bruce Craig, *Introduction to the Practice of Statistics*, 10th edition, W. H. Freeman. Any other edition works as well. [MMC].