Data Analysis Methods: Correlation and Regression.

Montevideo. March 2015.

Lily House-Peters

Sebastian Bonelli

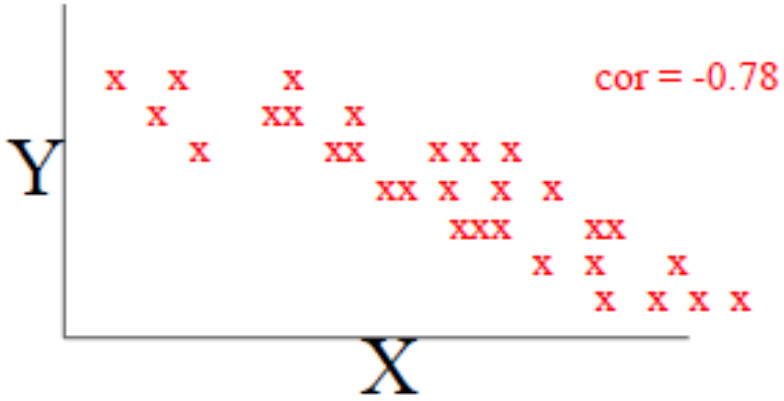
Correlation

Correlation is a systematic relationship between two variables.

These two variables may increase/decrease together. Correlation measures this trend. Need to have corresponding pairs of cases of x, y.

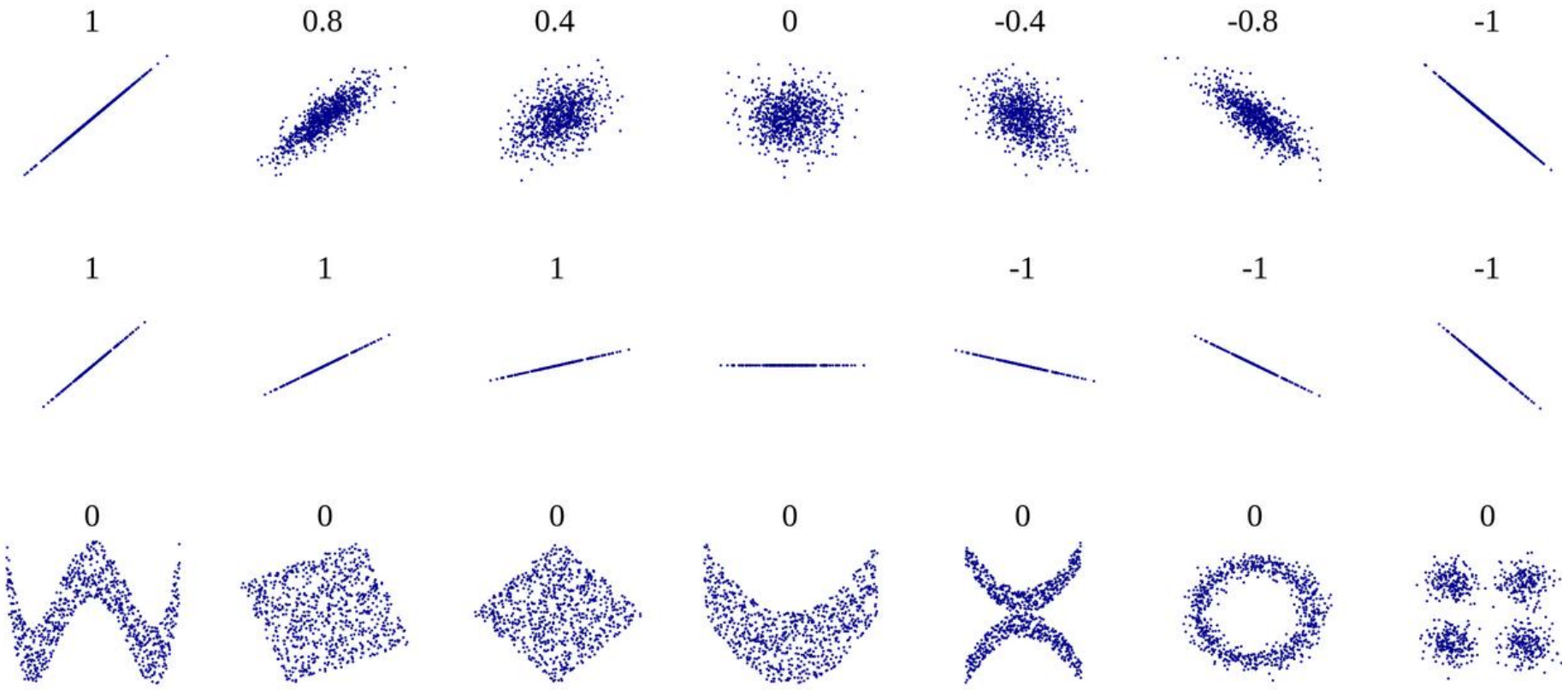
Correlation

- Perfect positive correlation is -1
- Perfect negative is -1
- Correlation can be anywhere between -1 and+ 1



- Relation may be casual
- If not casual, can be controlled by a third factor
- Pearson product-moment describes linear relation between x and y.

Correlation



Wikipedia, 2015. http://en.wikipedia.org/wiki/Correlation_and_dependence

How is measured?

Index	X	Y	zX	zY
1	1	417	-1.56	-1.87
2	6	492	-1.10	-0.96
3	8	510	-0.92	-0.74
4	9	531	-0.82	-0.49
5	16	537	-0.18	-0.42
6	17	553	-0.08	-0.22
7	17	590	-0.08	0.23
8	20	598	0.19	0.32
9	27	600	0.84	0.35
10	28	643	0.93	0.87
11	29	667	1.02	1.16
12	37	719	1.76	1.79
MEAN	**17.92**	**571.42**		
ST DEV	**10.82**	**82.57**		

the root

$$SD = s = \sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}}$$

$$z_x = \frac{x-\overline{x}}{SD_x}$$

$$z_y = \frac{y-\overline{y}}{SD_y}$$

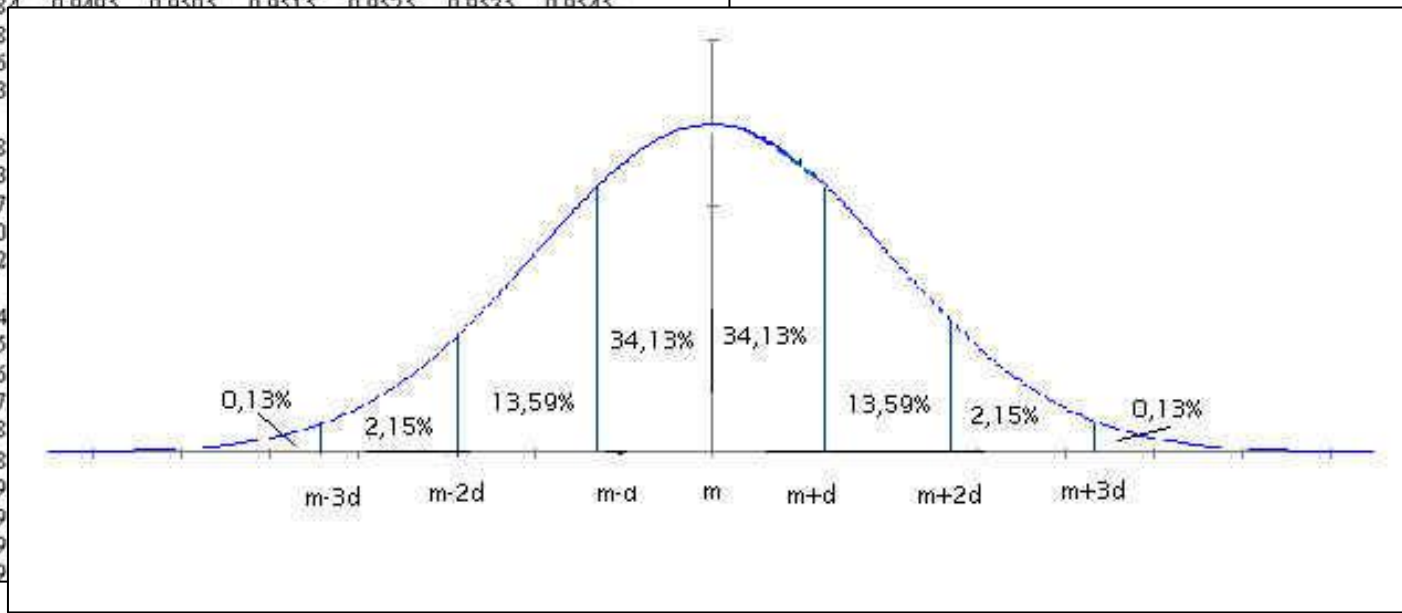
$$Correlation = \frac{1}{n}\sum_{i=1}^{n} z(x_i)z(y_i)$$

Normal curve

Tabla de la distribución normal

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.958						
1.8	0.9641	0.9649	0.9656	0.966						
1.9	0.9713	0.9719	0.9726	0.973						
2.0	0.9772	0.9778	0.9783	0.978						
2.1	0.9821	0.9826	0.9830	0.983						
2.2	0.9861	0.9864	0.9868	0.987						
2.3	0.9893	0.9896	0.9898	0.990						
2.4	0.9918	0.9920	0.9922	0.992						
2.5	0.9938	0.9940	0.9941	0.994						
2.6	0.9953	0.9955	0.9956	0.995						
2.7	0.9965	0.9966	0.9967	0.996						
2.8	0.9974	0.9975	0.9976	0.997						
2.9	0.9981	0.9982	0.9982	0.998						
3.0	0.9987	0.9987	0.9987	0.998						
3.1	0.9990	0.9991	0.9991	0.999						
3.2	0.9993	0.9993	0.9994	0.999						
3.3	0.9995	0.9995	0.9995	0.999						
3.4	0.9997	0.9997	0.9997	0.999						

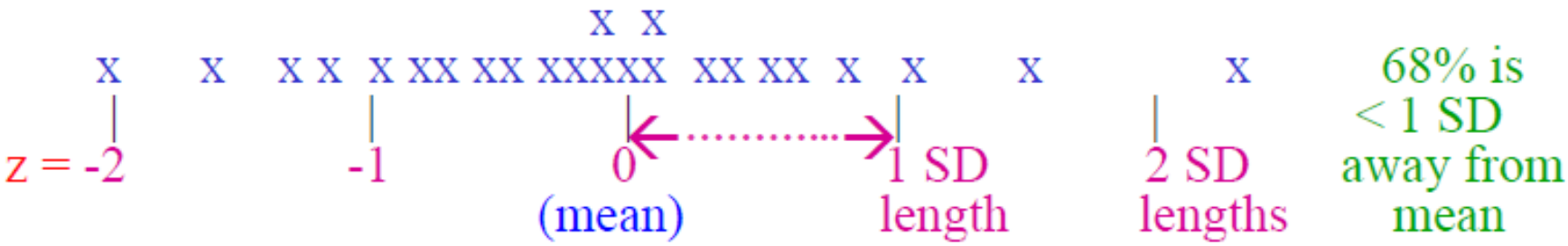
$$z_x = \frac{x - \overline{x}}{SD_x}$$



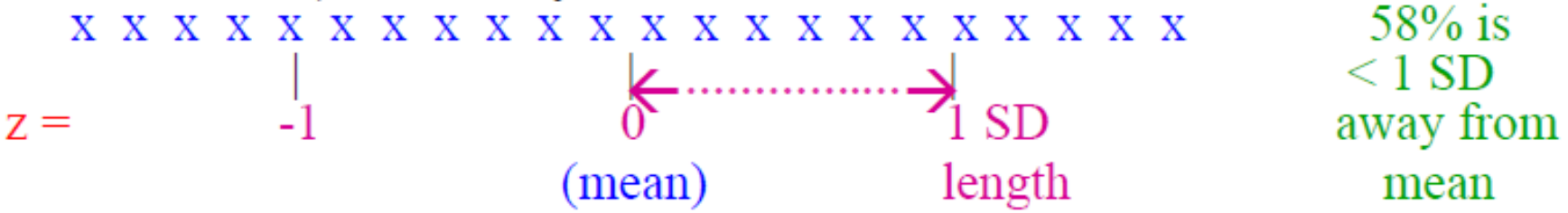
Size of Standard Deviation Relative to the Data Distribution

define z = number of SD lengths above or below the mean

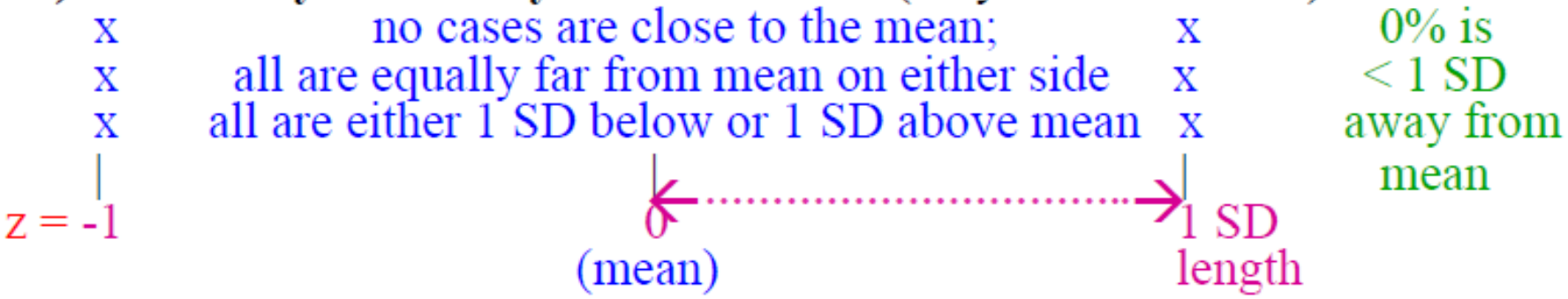
1) ~**Normally** distributed data:



2) **Uniformly** Distributed Data:



3) Maximally Bimodally distributed data (only 2 values occur):



IRI, Barnston A. 2015. Lecture slides. https://courseworks.columbia.edu/welcome/

Regression

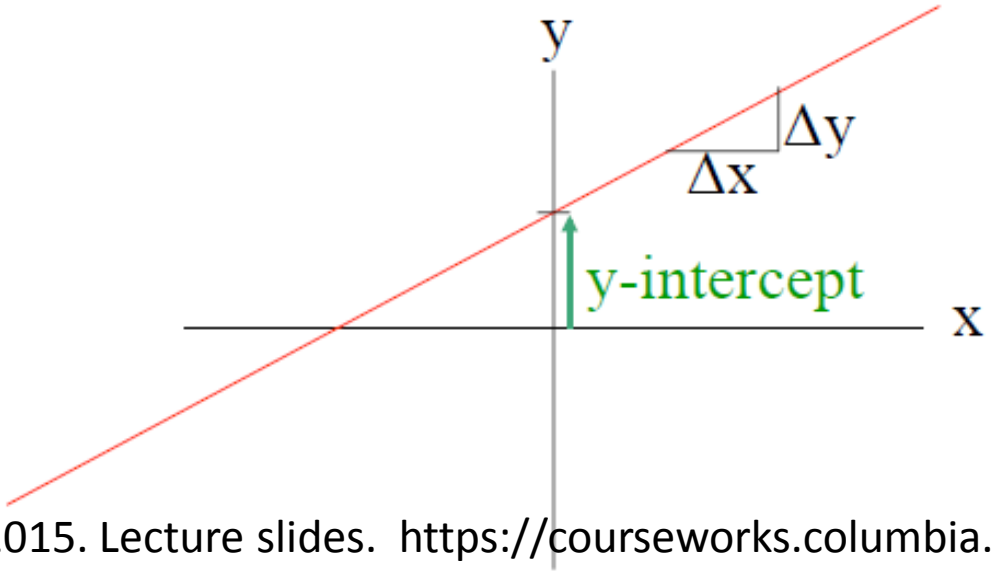
A line in the x vs. y coordinate system has the form

$y = a + bx$ a is y-intercept b is slope

x is the predictor, y is what is being predicted from x

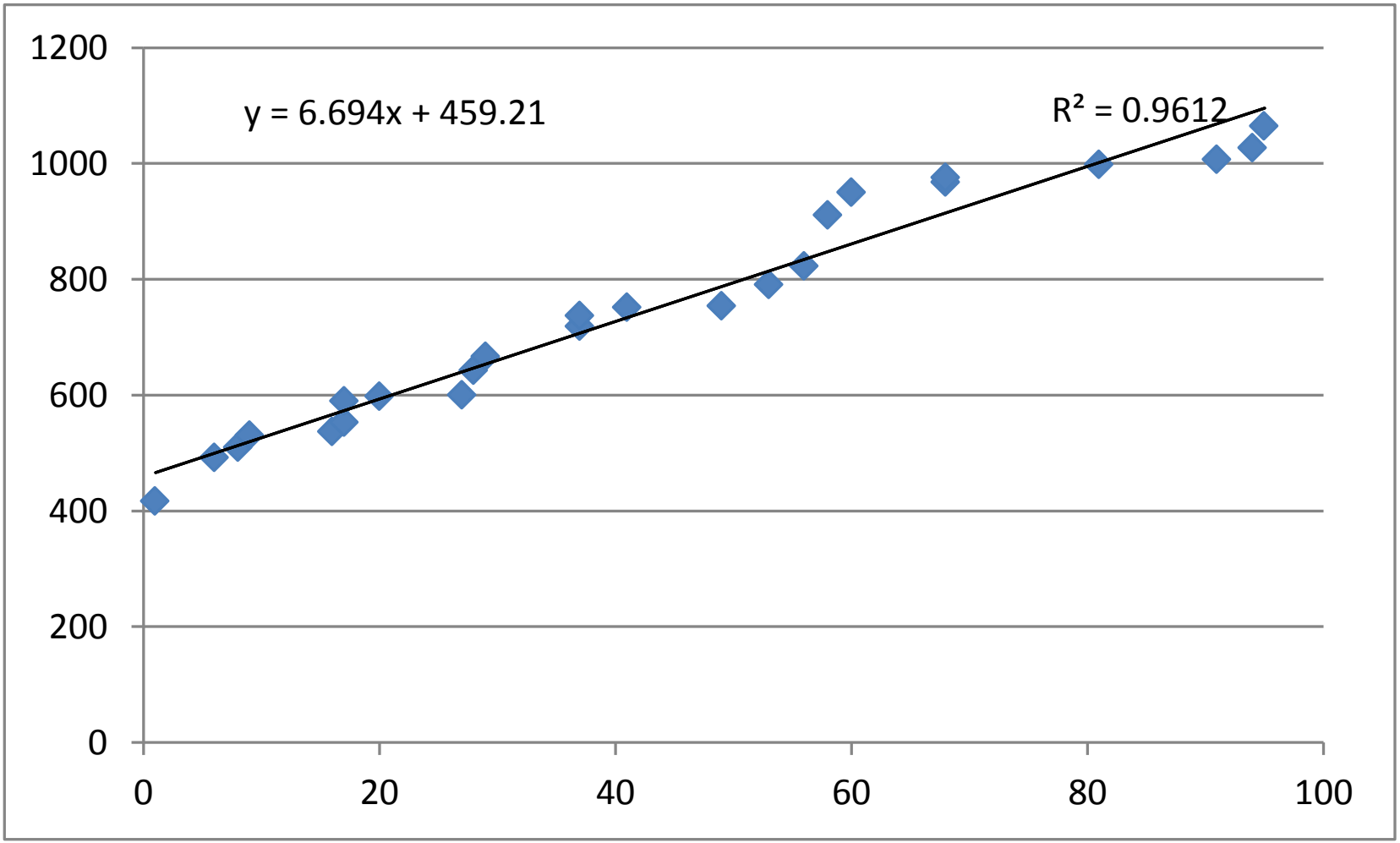
The slope is defined as: $\frac{\text{change in y}}{\text{change in x}}$ or $\frac{\Delta y}{\Delta x}$

The y-intercept is the y value that would occur when x is equal to 0.

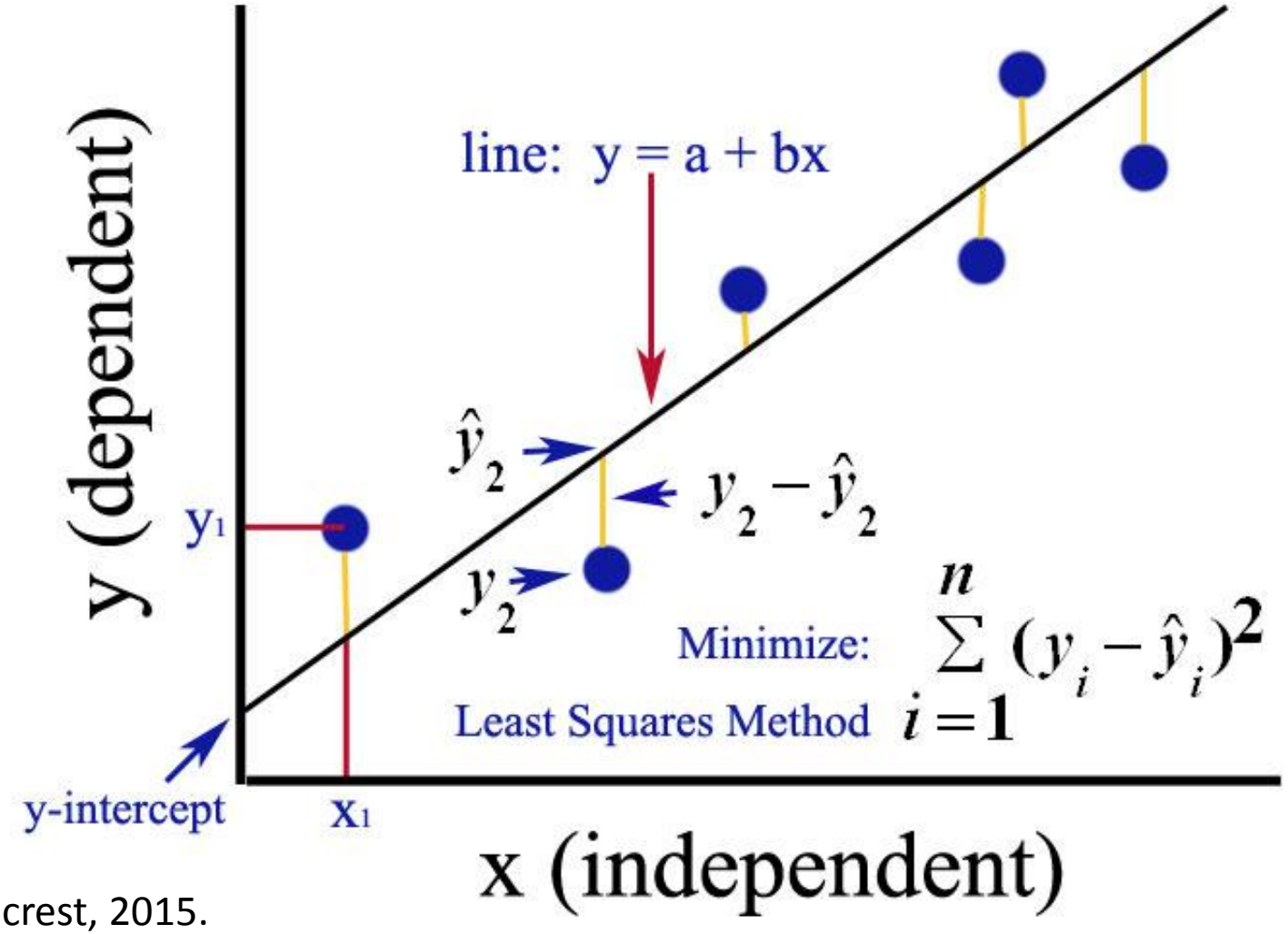


IRI, Barnston A. 2015. Lecture slides. https://courseworks.columbia.edu/welcome/

Regression


y = 6.694x + 459.21
R² = 0.9612
1200
1000
800
600
400
200
0
0
20
40
60
80
100

Regression line is defined such as the sum of squares of the errors (predictes y vs the true y) is minimized



Cedacrest, 2015.
http://www2.cedarcrest.edu/academic/bio/hale/biostat/session24links/linegraph.jpg

Regression

- Such a line predicts y from x such that in standarized (z) units for x and y:

$$z_y = cor_{xy} z_x$$

- If cor_{xy}=0.5, then y will be predicted to be half as many SDs away from its mean as x.