Chapter 14: Linear Regression

Note Title 11/30/2016

14.1: Linear Equations

Slope-intercept form: $y = mx + b$

$m = \text{slope} = \dfrac{y_2 - y_1}{x_2 - x_1} = \dfrac{\text{"rise"}}{\text{run}}$

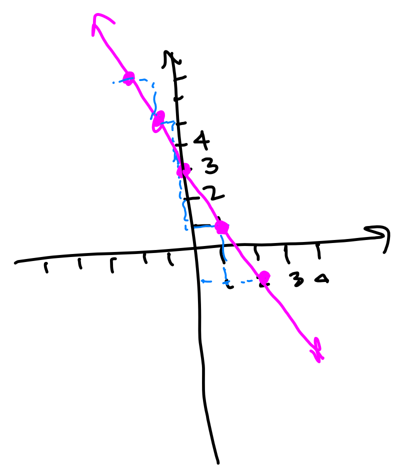
$b = \text{y-intercept}$

Ex: Graph the line

$y = -2x + 3$

Slope: $m = -2 = -\dfrac{2}{1}$

y-intercept: $b = 3$ $(0,3)$



positive slopes: uphill from left to right

Negative slopes: downhill from left to right

Scatter Plot graph of observed points (x,y)

In statistics, we use b_1 and b_0 instead of m and b.

$$y = b_1 x + b_0$$

Slope: b_1

y-intercept: b_0

x is the predictor (independent) variable.

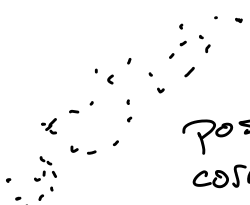
y is the response (dependent) variable

Why? Linear regression can be done with multiple predictor (independent) variable.

Then you have

$$y = b_3 x_3 + b_2 x_2 + b_1 x_1 + b_0$$

If there is a discernible pattern in the scatterplot, x can be used to predict y.



positive linear correlation

As $x\uparrow$, $y\uparrow$

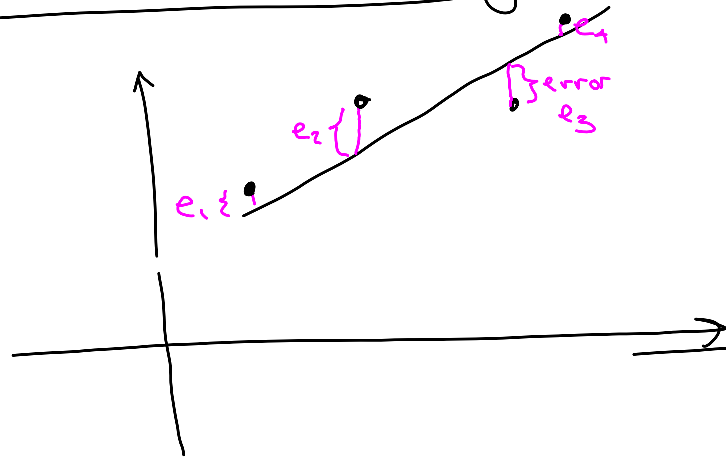
negative linear correlation

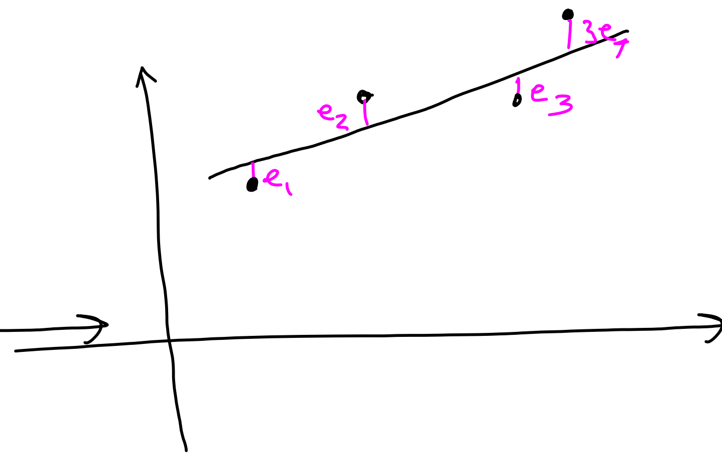
as $x\uparrow$, $y\downarrow$

no correlation

non linear correlation

14.2: The Regression Equation





e: error = the signed vertical distance between the data point and the line.

Which line is "better"? We choose the line that minimizes the squares of the errors.

The best line is called the <u>Least Squares Regression Line</u>.

$\hat{y}$ = predicted value of y.

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

Actual y: $y = b_1 x + b_0 + e$ ← error term

<u>Example</u>: suppose x = # hours studied, y = grade on exam; for this data set:

x	y
10	96
3	64
6	80
7	76
8	92
4	60

Next page ...

Ex 1 Continued:

x	y	xy	x^2	y^2
10	96	960	100	9216
3	64	192	9	4096
6	80	480	36	6400
7	76	532	49	5776
8	92	736	64	8464
4	60	240	16	3600
Sums 38	468	3140	274	37552

Σ means "sum" (sigma notation)

We must calculate S_{xy}, S_{xx}, S_{yy}

Conceptual formulas:

$$S_{xy} = \Sigma (x_i - \bar{x})(y_i - \bar{y})$$

$$S_{xx} = \Sigma (x_i - \bar{x})^2$$

$$S_{yy} = \Sigma (y_i - \bar{y})^2$$

Computational Formulas

$$S_{xy} = \Sigma (x_i y_i) - \frac{\Sigma x_i \Sigma y_i}{n}$$

$$S_{xx} = \Sigma (x_i^2) - \frac{(\Sigma x_i)^2}{n}$$

$$S_{yy} = \Sigma (y_i^2) - \frac{(\Sigma y_i)^2}{n}$$

Slope of regression line: $b_1 = \frac{S_{xy}}{S_{xx}}$

Regression line is: $\bar{y} = b_1\bar{x} + b_0$. Use this to get b_0

Correlation Coefficient: $r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$
(section 14.4)

The correlation coefficient r (sometimes just called the correlation) is always between -1 and 1. It tells us how well the x's predict the y's, and the direction of the relationship.

:

x	y	xy	x^2	y^2
10	96	960	100	9216
3	64	192	9	4096
6	80	480	36	6400
7	76	532	49	5776
8	92	736	64	8464
4	60	240	16	3600
Sums 38	468	3140	274	37552
Σx	Σy	Σxy	Σx^2	Σy^2

Computational Formulas

$$S_{xy} = \sum (x_i y_i) - \frac{\Sigma x_i \Sigma y_i}{n}$$

$$S_{xx} = \Sigma (x_i^2) - \frac{(\Sigma x_i)^2}{n}$$

$$S_{yy} = \Sigma (y_i^2) - \frac{(\Sigma y_i)^2}{n}$$

$$S_{xy} = \Sigma(xy) - \frac{\Sigma x \Sigma y}{n} = 3140 - \frac{38(468)}{6} = 176$$

$$S_{xx} = \Sigma(x^2) - \frac{(\Sigma x)^2}{n} = 274 - \frac{(38)^2}{6} = 33.3333$$

$$S_{yy} = \Sigma(y^2) - \frac{(\Sigma y)^2}{n} = 37552 - \frac{(468)^2}{6} = 1048$$

Slope of regression line:

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{176}{33.3333} = 5.28$$

$$\bar{x} = \frac{\Sigma x}{n} = \frac{38}{6} = 6.3333 \quad , \quad \bar{y} = \frac{\Sigma y}{n} = \frac{468}{6} = 78$$

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

$$78 = 5.28(6.3333) + b_0$$

$$b_0 = 78 - 5.28(6.333) = 44.56$$

Regression Line: $y = b_1 x + b_0$, so $\boxed{y = 5.28x + 44.56}$

Use the regression line to predict values of y for x's within the boundaries of the original x-values.

For an x-value of 4, the predicted y-value is

$$\hat{y} = 5.28(4) + 44.56 = 65.68.$$

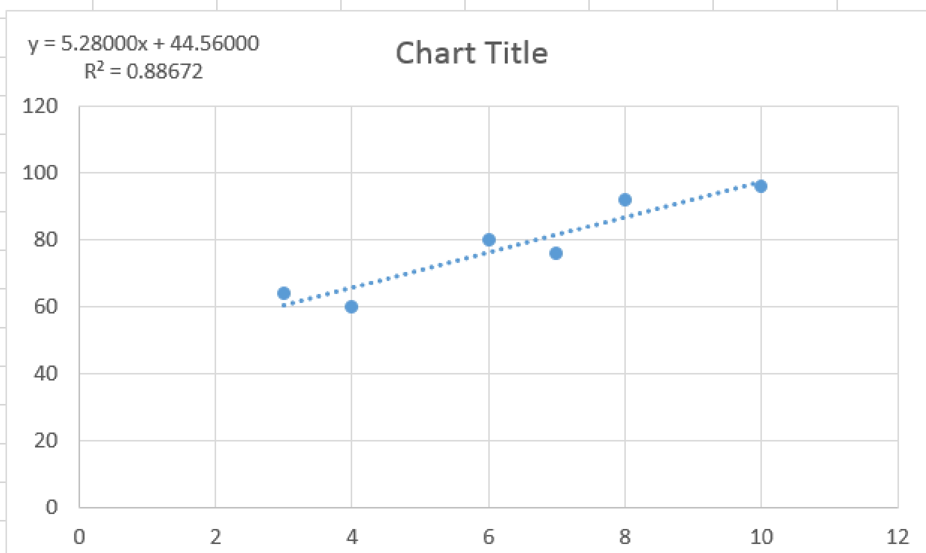
Calculate correlation coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = \frac{176}{\sqrt{33.3333(1048)}}$$

$$= \boxed{0.942}$$ strong positive correlation (close to 1)

Also we can use Excel to find r and the regression line

	A	B
1	x	y
2	10	96
3	3	64
4	6	80
5	7	76
6	8	92
7	4	60
8		
9		



The R^2 is the coefficient of determination... this means 0.88672 (or 88.7%) of the variance in values of the response variable can be explained by the predictor value.

To find the correlation r, we take the square root of R^2.

$$r = \sqrt{R^2} = \sqrt{0.88672} = 0.942.$$ (same as before)