Chapter 14: Linear Regression

Note Title 4/24/2018

14.1: Linear Equations

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

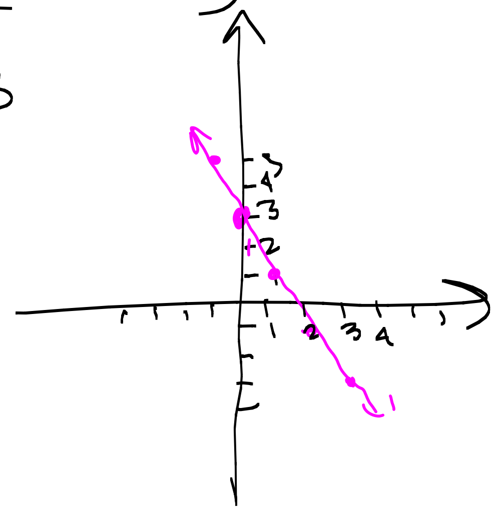
$$m = \text{slope} = \frac{y_2 - y_1}{x_2 - x_1} = \text{“}\frac{\text{rise}}{\text{run}}\text{”}$$

b = y-intercept (or $(0, b)$)

Ex: Graph the line $y = -2x + 3$

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

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



Positive slopes: uphill left to right

Negative slopes: downhill 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 b_1 and b_0?

Linear regression can be done with multiple predictor (independent) variables. 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.

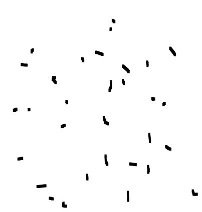
positive linear correlation

as $x\uparrow$, $y\uparrow$ also

negative linear correlation

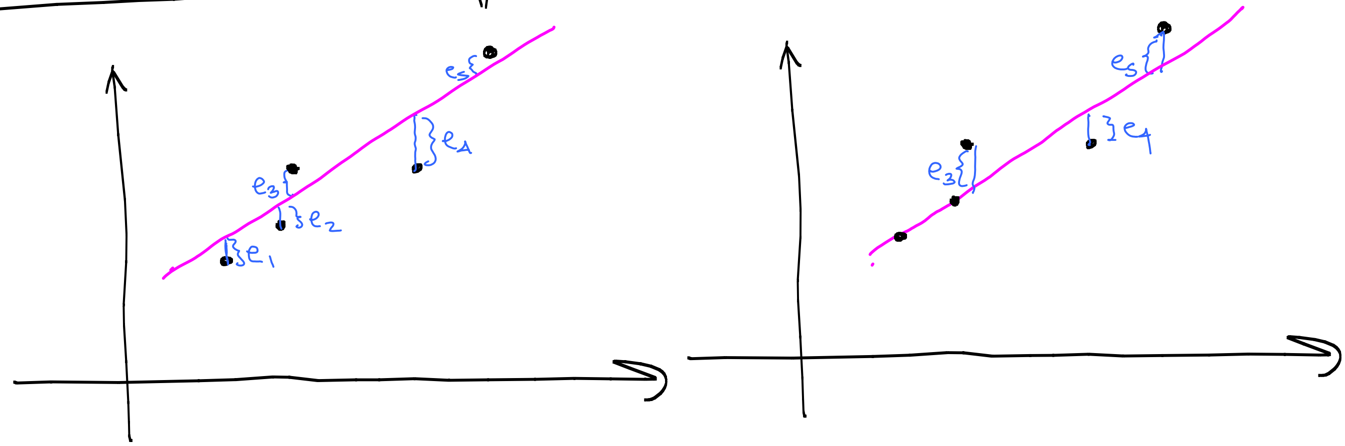
as $x\uparrow$, $y\downarrow$

nonlinear correlation



no correlation

14.2: The Regression equation



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

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

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

y-hat

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

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

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

Calculating the Least Squares Regression Line

Σ means "sum" (sigma notation)

we must calculate S_{xy}, S_{xx}, S_{yy} (sometimes written as SS_{xy}, SS_{xx}, S_{yy}) stands for Sum of Squares

Conceptual Formulas

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

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

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

Computational Formulas

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

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

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

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

Regression line (least squares line): $\bar{y} = b_1 \bar{x} + b_0$. Use this to get b_0.

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

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

(r close to 1 or close to -1 $\Rightarrow$ very strong relationship

r close to 0 $\Rightarrow$ weak relationship)

Example: Suppose x = # hours studied
y = grade on exam

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

	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 Σx_i	Σy	Σxy	Σx^2	Σy^2

Computational Formulas

$$S_{xy} = \Sigma(xy) - \frac{\Sigma x \Sigma y}{n}$$

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

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

Note: $n = 6$ (number of ordered pairs)

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

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

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

Slope of regression line: $b_1 = \frac{S_{xy}}{S_{xx}} = \frac{176}{33.3333} = 5.28$

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

The pair $(\bar{x}, \bar{y})$ is always on the line.

next page

Previous example cont'd:

$y = b_1 x + b_0$ is line

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

$\left.\begin{array}{l}\bar{x} = 6.3333, \\ \bar{y} = 78, \\ b_1 = 5.28\end{array}\right\} \Longrightarrow$ $78 = 5.28(6.3333) + b_0$

$b_0 = 78 - 528(6.3333) = 44.56$

Regression line: $\boxed{y = 5.28x + 44.56}$

Use the regression line to predict y-values for x's within the range of the x's in the data set.

For an x-value of 4, the predicted value of y is $\hat{y} = 5.28(4) + 44.56 = \boxed{65.68}$

Calculate the correlation coefficient:

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

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