Chapter 14: Linear Regression

Note Title 7/5/2017

14.1: Linear Equations

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

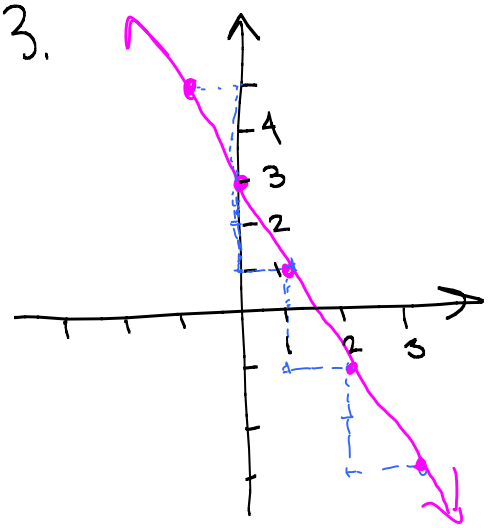
m = slope $= \frac{y_2 - y_1}{x_2 - x_1} =$ "$\frac{\text{rise}}{\text{run}}$"

b = y-intercept

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

Slope: $m = -2 = -\frac{2}{1}$ ↘

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



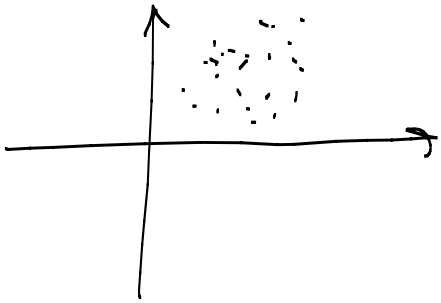
Positive slopes: uphill from left to right ↗

Negative slopes: downhill from left to right ↘

Zero slope: horizontal

undefined: vertical

Scatterplot: graph of observed points (x, y)



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

positive linear correlation

as x increases, y increases also

Negative linear correlation

as x increases, y decreases

no correlation

nonlinear correlation

We will learn how to create a best-fit line for a data set with a positive or negative linear correlation.

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

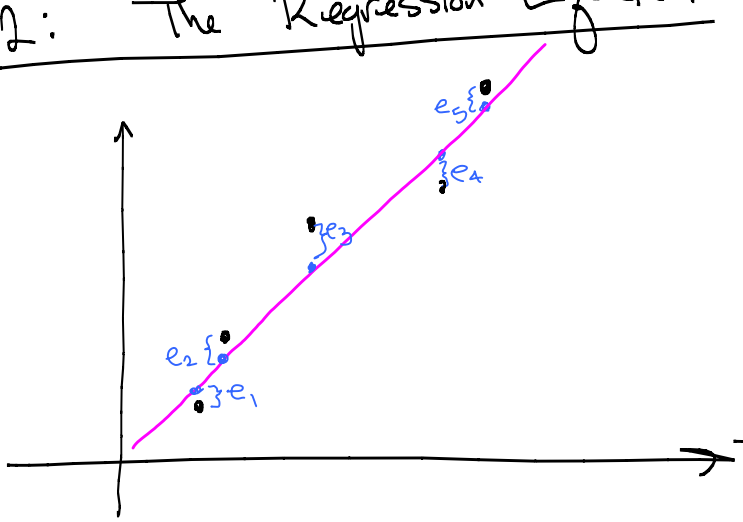
(sometimes called the outcome variable)

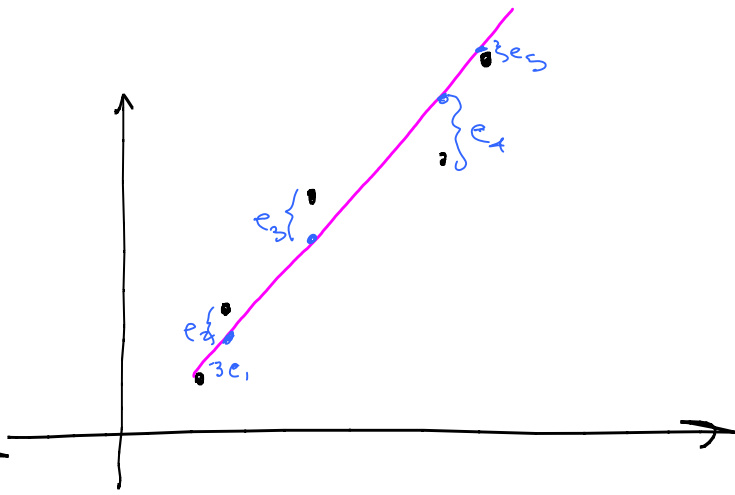
Why use b_1 and b_0?

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

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

4.2: 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 total squares of the errors.

This line is called the Least Squares Regression Line.

$\hat{y}$ = predicted value of y (using the equation of the line)

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

Actual y: $y = \hat{y} + e$ ← error term

$y = b_1 x + b_0 + e$

Formulas for the regression line

Conceptual Formulas

$$SS_{xy} \text{ or } S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$SS_{xx} \text{ or } S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$SS_{yy} \text{ or } S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

(SS stands for Sum of Squares)

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: $b_1 = \frac{S_{xy}}{S_{xx}}$

Regression line: $\hat{y} = b_1 \bar{x} + b_0$ (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 a number between -1 and 1.

It tells us how well the x's predict the ys, and the direction of the relationship.

$r = 1 \Rightarrow$ all points exactly on line, positive slope

$r = 0.9 \Rightarrow$ points very close to line, positive slope

$r = -1 \Rightarrow$ all points exactly on line, negative slope

$r = -0.9 \Rightarrow$ points very close to line, negative slope

Example 1: 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

Create table:

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	$\Sigma(xy)$	$\Sigma(x^2)$	$\Sigma(y^2)$

$n = 6$

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

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

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

Find slope of regression line:

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

Find equation of regression line:

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

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

See next page

Ex 1 cont'd:

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

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

$$b_0 = 78 - 5.28(6.3333) = 44.56$$

Eqn of regression line: $y = b_1x + b_0$

$$\hat{y} = 5.28x + 44.56$$ eqn of regression line

Use the regression line to predict values of y within the range of the original x-values.

Here, we can only use the regression line for x-values that are between 3 and 10.

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

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

(so the point $(4, 65.68)$ should be exactly on the line)

Calculate the correlation coefficient

$$r = \frac{S_{xy}}{\sqrt{S_{xx}\,S_{yy}}} = \frac{176}{\sqrt{33.3333(1048)}} = 0.942$$

strong positive correlation

The R^2 Excel gives us is the coefficient of determination.

For $R^2 = 0.88672$, this means 88.7% of the variance in values of the response variable (y) can be explained by the predictor variable (x).

The correlation coefficient r is the square root of the R^2.

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