Chapter 12

Linear Regression & Correlation

$$f(x) = y = mx + b \quad \text{(Algebra)}$$

$$\hat{y} = b_0 + b_1 x \quad \text{(Statistics)}$$

$$r = \left\{ \begin{array}{l} \text{strength \&} \\ \text{Direction of line} \end{array} \right\} \quad \text{(chapter 3)}$$

$$\left\{ \begin{array}{l} \text{dependent variable} \\ \text{predicted variable} \end{array} \right\} = y = \hat{y} = f(x)$$

$$\left\{ \begin{array}{l} \text{independent variable} \\ \text{predictor variable} \end{array} \right\} = x$$

$$\{\text{y-intercept}\} = b = b_0 = \begin{array}{l} \text{INTERCEPT} \\ \text{Excel function} \end{array}$$

$$\{\text{slope}\} = \left\{ \begin{array}{l} \text{How much y moves} \\ \text{for every 1 unit} \\ \text{of x} \end{array} \right\} = m = b_1 = \begin{array}{l} \text{SLOPE} \\ \text{Excel Function} \end{array}$$

$$r = \left\{ \begin{array}{l} \text{strength \&} \\ \text{Direction} \\ \text{of Linear} \\ \text{equation} \end{array} \right\} = r = \begin{array}{l} \text{PEARSON} \\ \text{Excel function} \end{array}$$

$$r^2 = \left\{ \begin{array}{l} \text{coefficient of} \\ \text{Determination} \\ \text{"Goodness of fit} \\ \text{of equation"} \end{array} \right\} = r^2 = \begin{array}{l} \text{RSQ} \\ \text{Excel function} \end{array}$$

chapter 12 : Simple Linear Regression

What we already know about "Relationship between 2 quantitative variables" from math & chapter 3:

① **Independent variable (X)**

predictor variable.

② **Dependent variable** (y or $f(x)$ or $E(y)$ or $\hat{y}$)

Variable that is predicted or estimated

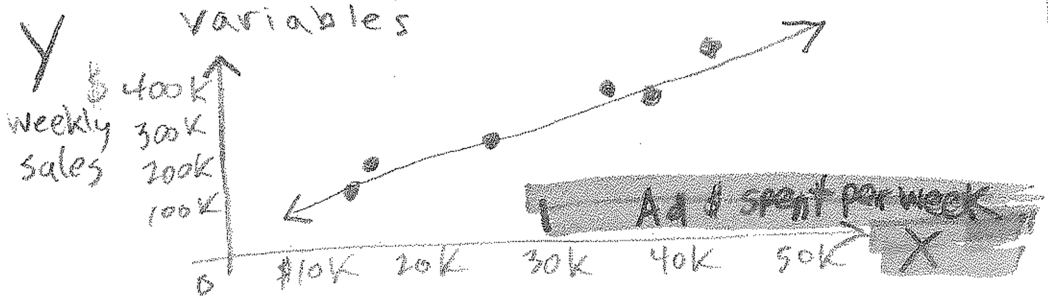
Example:

Collected data from past:

X Ad $ spent per week	Y weekly sales
14,000	97,000
27,000	185,000
40,000	260,000
17,000	143,000
34,000	270,000
43,000	398,000

③ **Scatter Diagram/chart/Plot**

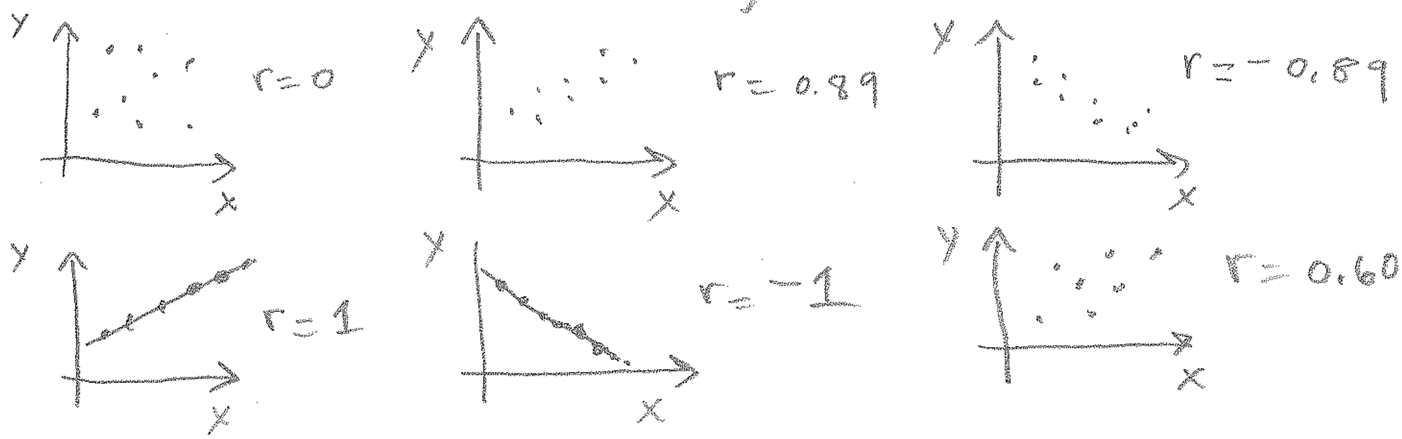
Graphical technique to show relationship between 2 quantitative variables



4 Coefficient of Correlation (r) (Interval or Ratio Level Data)

Coefficient of correlation = r

1. Measure of the strength & direction of the linear relationship (Between -1 and 1).
2. 0 = No correlation.
3. 0.5 or -0.5 = Moderate correlation.
4. Near -1 or 1 = Strong correlation.



5 Coefficient of determination (r^2)

A measure of the goodness of fit of the estimated regression equation. The proportion of the total variation in the dependent variable, y, that is explained, or accounted for, by the variation in the idependent variable, X. It does NOT say anything about causation, that is, it does not say that X causes y.

{Coefficient of Correlation} $= r = \dfrac{\sum (X-\bar{X}) * (Y-\bar{Y})}{(n-1) * S_x * S_y}$

only 1 independent variable

How to calculate coefficient of correlation Example 1

$n = 6$
$n-1 = 5$

$\bar{X} = 29150$

$\bar{Y} = 225{,}500$

$S_x = 11{,}938.80$

$S_y = 107{,}596.93$

X = particular value
$\bar{X}$ = mean of Xs
Y = particular value
$\bar{Y}$ = mean of Ys
n = count of observed pairs
S_x = standard deviation of Xs
S_y = standard deviations of Ys

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})*(Y-\bar{Y})$
14,000	97,000	14000 − 29150 = −15150	97000 − 225500 = −128500	+ 1,946,775,000
17,000	143,000	17000 − 29150 = −12150	143000 − 225500 = −82500	+ 1002 375000
27,000	185,000	27000 − 29150 = −2150	185000 − 225500 = −40500	+ 87 075 000
34,000	270,000	34000 − 29150 = 4850	270000 − 225500 = 44500	+ 215 825 000
39,900	260,000	39900 − 29150 = 10750	260000 − 225500 = 34500	+ 370 875 000
43,000	398,000	43000 − 29150 = 13850	398000 − 225500 = 172500	+ 2,389 125000
		$\Sigma = 0$	$\Sigma = 0$	$\Sigma = 6{,}012{,}050{,}000$

{Strength & direction} $= r = \dfrac{6{,}012{,}050{,}000}{5 * 11938.8 * 107{,}596.93} = .936034653$

{what does .936 mean?}
→ strength = Strong
→ direction = direct, as X increase, Y increase.

In Excel: → PEARSON or CORREL function!
→ =PEARSON(Yrange, Xrange)

$\left\{\begin{array}{c}\text{coefficient}\\ \text{of}\\ \text{correlation}\end{array}\right\} = r = .936034653$

Ad dollars spent & sales earned Example 1:

.936034653 indicates that the relationship between Ad dollars spent & sales earned is very strong. In addition the relationship is direct, which means that as Ad dollars increase Sales increase. However "strong" is not numerically precise. But coefficient of Determination is numerically precise.

$\left\{\begin{array}{c}\text{coefficient}\\ \text{of}\\ \text{Determination}\end{array}\right\} = r^2 = \left\{\begin{array}{c}\text{Influence } X\\ \text{has on } y,\\ \text{(Not causation)}\end{array}\right\} = .936034^2 = .876$

Goodness of Fit

Ad dollars spent & sales earned Example 1:

We can say that 87.6% of the variation in y $(f(x))$ can be explained by the variation in X (not causation). Correlation does not mean causation!! (from textbook: As population of donkeys decreases, number of doctorial degrees increases.) This is called "spurious correlations." Conclusion: correlation is great for building models we can use to make predictions, but correlation does not mean causation.

6 Estimated Simple Linear Regression Equation

An equation that expresses the linear relationship between 2 variables

$y = mx + b$ (Algebra)

$\hat{y} = b_0 + b_1 x$ (Statistics)

$y = \hat{y}$ = predicted variable = dependent variable

$x = x$ = predictor variable = Independent variable

$m = b_1$ = slope = How much y moves for 1 unit of x

$b = b_0$ = y-intercept

x_i = particular x

y_i = particular y

$\bar{x}$ = xbar = sample mean

$\bar{Y}$ = Ybar = sample mean

$$b_1 = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

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

$\bar{X} = 29150$ $n = 6$

$n - 1 = 5$

$\bar{Y} = 225,500$

$S_x = 11,938.80$

$S_y = 107,596.93$

From bottom:

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

$$\boxed{\hat{y} = -20,406 + 8.44X}$$

Can use to estimate!!

X	Y	(X−Xbar)*(Y−Ybar)	$(X-Xbar)^2$
14,000	97,000	(14000 − 29150) * (97000 − 225500) = 1,946,775,000	$(14000-29150)^2$ = 229,522,500
27,000	185,000	(27,000 − 29150) * (185,000 − 225500) = 87,075,000	$(27,000-29150)^2$ = 4,622,500
39,900	260,000	(39,900 − 29150) * (260,000 − 225500) = 370,875,000	$(39900-29150)^2$ = 115,562,500
17,000	143,000	(17,000 − 29150) * (143,000 − 225,500) = 1,002,375,000	$(17,000-29,150)^2$ = 147,622,500
34,000	270,000	(34,000 − 29150) * (270,000 − 225500) = 215,825,000	$(34,000-29,150)^2$ = 23,522,500
43,000	398,000	(43000 − 29150) * (398,000 − 225500) = 2,389,125,000	$(43000-29150)^2$ = 191,822,500
Total:		6,012,050,000	712,675,000

Slope $= m = b_1 = \dfrac{\Sigma(x-xbar)*(y-ybar)}{\Sigma(x-xbar)^2} = \dfrac{6,012,050,000}{712,675,000} = 8.44$

Intercept $= b = b_0 =$ Ybar $- b_1 *$ Xbar $= 225,500 - 8.44 * 29150 = -20,406.28$

$$\underbrace{\hat{y} = -20{,}406.28 + 8.44X}_{\text{use to make predictions}}$$

If we spend \$55,000 on weekly ads, we can expect the sales Revenue to be:

$$f(x)\ \hat{y} = -20{,}406.28 + 8.44 * 55{,}000$$

$$f(x) = \hat{y} = 443{,}567.83 = f(55{,}000)$$