# Linear Regression & Correlation

Excel:

1. For slope of Line use:

   = SLOPE( y values, x values)

2. For Intercept of Line use:

   = INTERCEPT( y values, x values)

3. For Correlation coefficient use:

   = PEARSON ( x values, y values)

4. For coefficient of Determination use:

   = RSQ ( y values, x values)

## Scatter Diagram

graphical technique to show relationship between 2 variables

**Independent Variable (X)**

predictor variable

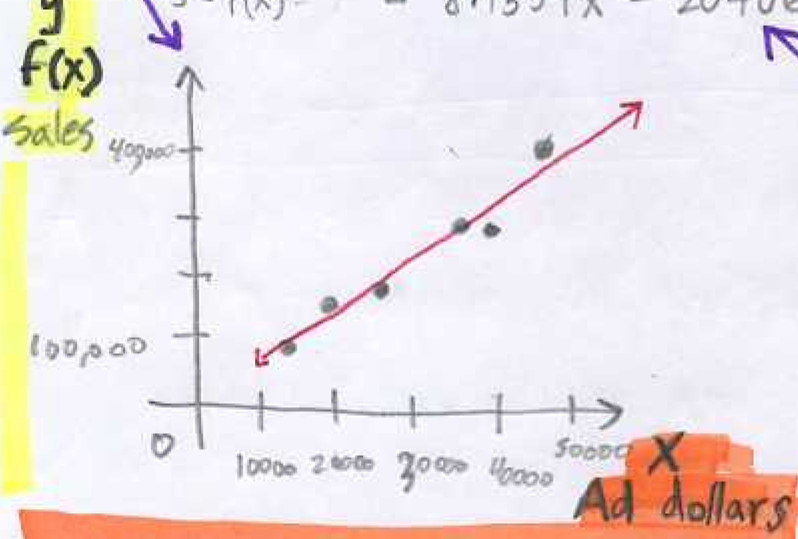
**Dependent Variable f(x) or y**

Variable that is predicted or estimated

"we want to estimate what sales will be based on what Advertising dollars have been spent"

↓ look at past Data

$y = f(x) = 8.4359X - 20406$




EXAMPLE 1

| X Ad dollars | y f(x) sales |
|---|---|
| 14,000 | 97,000 |
| 17,000 | 143,000 |
| 27,000 | 185,000 |
| 34,000 | 270,000 |
| 39,900 | 260,000 |
| 43,000 | 398,000 |

## Regression Equation

An Equation that Expresses the linear relationship between 2 variables

$E(y) = \beta_0 + \beta_1 X$ (Statistics)

$y = f(x) = mx + b$ (Algebra)

$\beta_0 = b$ = y intercept

$X = x$ = predictor variable

$\beta_1 = m$ = slope = for every 1 unit of x, how much you go up/down y-axis

$E(y)$ $y = f(x)$ = predicted variable

## Coefficient of correlation = r

Strength & direction of Relationship (not causation)

$r = .936$

## Coefficient of Determination = $r^2$

Amount of influence X has on f(x) (correlation not causation)

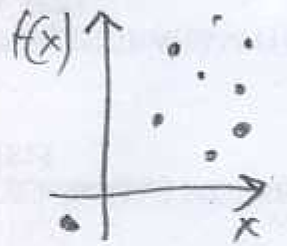
$r^2 = .876$

## coefficient of correlation = { strength and direction of relationship (between −1 and 1)

0 = NO correlation
.5 or −.5 = moderate
near 1 or −1 = strong }

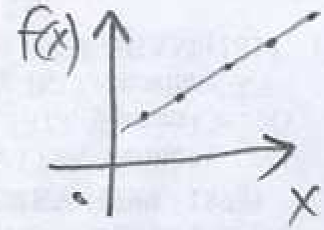
* Not causation

* Interval or Ratio Level

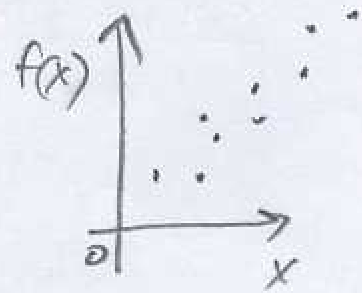


$r = 0$

Direct Relationship

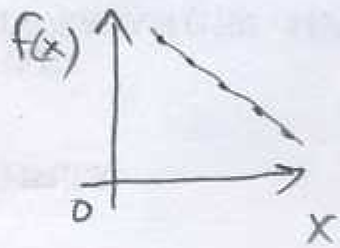


$r = 1$

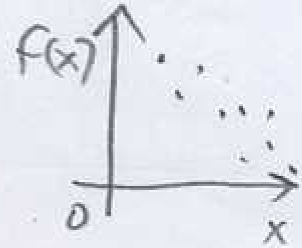


$r = .89$

Negative or Indirect or Inverse Relationship



$r = -1$



$r = -.89$

## coefficient of determination = { Amount of influence X has on f(x) → Not causation. Even though we may be able to use our model to predict, this does not mean causation }

$r^2 = r^2$

In General:
- .25 → weak influence
- .50 → moderate influence
- .75 → strong influence

"The proportion of the total variation in the dependent variable Y that is explained, or accounted for, by the variation in the independent variable X."

---

If we spend \$50,000 on Ads, what is an estimate for Sales?

$$f(50{,}000) = 8.4359(50{,}000) - 26406 = 401{,}388.37$$

{Coefficient of Correlation} $= r = .936034653$

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.

{Coefficient of Determination} $= r^2 =$ {Influence x has on y, (Not causation)} $= .936034^2 = .876$

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.