

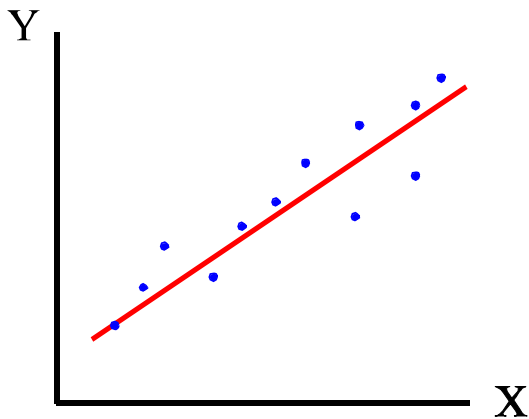
Correlation and Regression

Linear correlation:

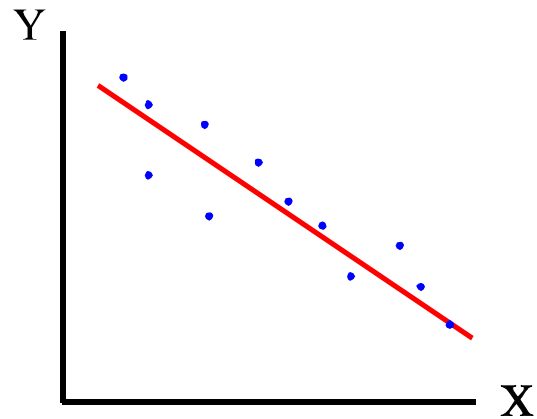
- Does one variable increase or decrease linearly with another?
- Is there a linear relationship between two or more variables?

Types of linear relationships:

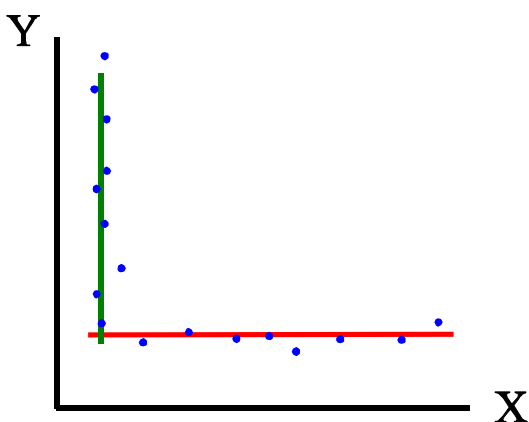
Positive linear



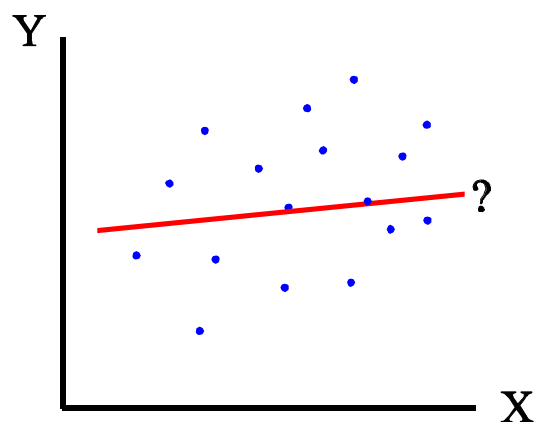
Negative linear



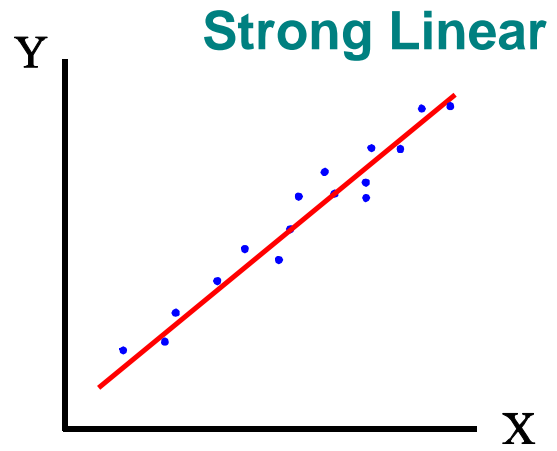
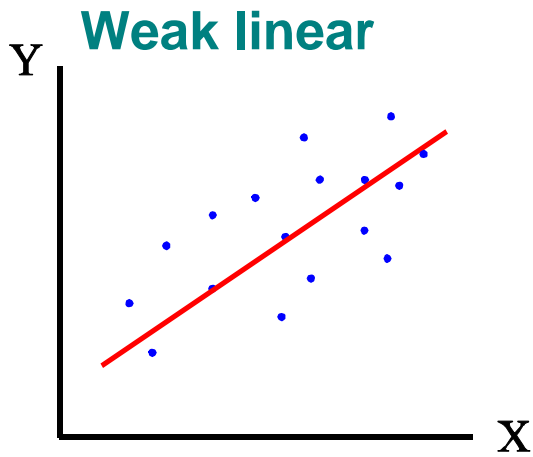
No relationship



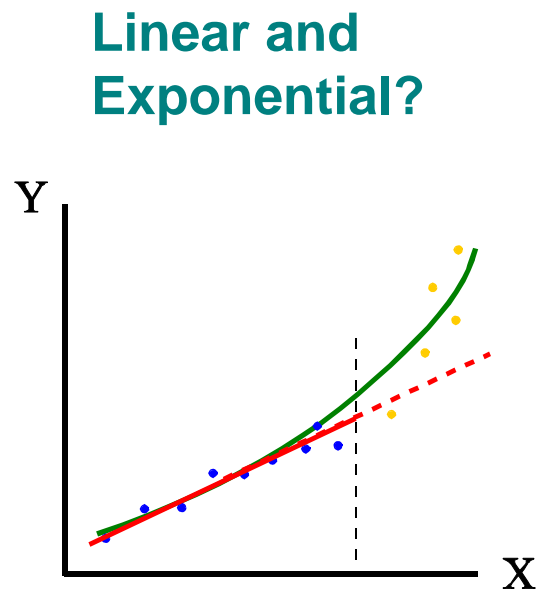
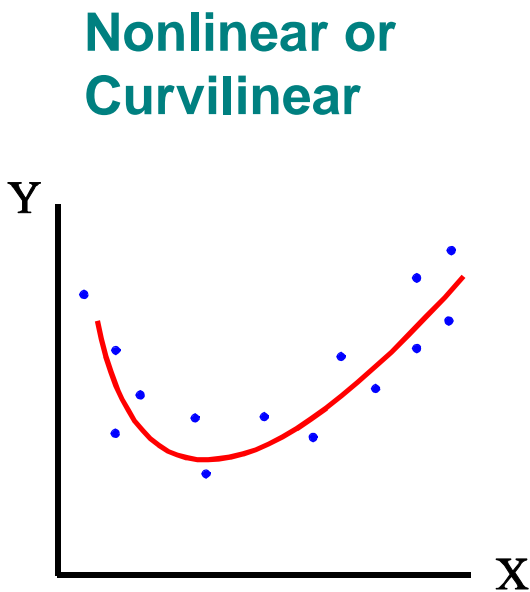
None or weak



Scattergrams



Other relationships:



Correlation

Pearson Product Moment Correlation

Coefficient:

- Simply called correlation coefficient, PPMC or r -value
- Linear correlation between two variables

Examples:

Weight increases with height.

IQ with brain size?!

Used for calibration of instruments, force transducers, spring scales, electrogoniometers (measure joint angles).

$$r = \frac{n(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[n(\sum X^2) - (\sum X)^2][n(\sum Y^2) - (\sum Y)^2]}}$$

Multiple Correlation:

- Used when several independent variables influence a dependent variable
- R-value

Defined as: $Y = A + B_1 X_1 + B_2 X_2 + B_3 X_3 + \dots + B_n X_n$

Examples:

Heart disease is affected by family history, obesity, smoking, diet etc.

Academic performance is affected by intelligence, economics, experience, memory etc.

Lean body mass is predicted by a combination of body mass, thigh, triceps and abdominal skinfold measures.

Significance of Correlation Coefficient

Method 1

Step 1: $H_0: \rho = 0$; $H_1: \rho \neq 0$

Step 2: Look up r_{crit} for $n-2$ degrees of freedom
(Table I)

Step 3: Compute sample r (as above)

Step 4: Sample r is significant if it is greater than r_{crit}

Step 5: If significance occurs data are linearly correlated otherwise they are not.

If table of significant correlation coefficients is not available or significance level (α) is not 0.05 or 0.01 use Method 2.

Method 2

Step 1: $H_0: \rho = 0$; $H_1: \rho \neq 0$

Step 2: Look up t_{crit} for $n-2$ degrees of freedom

Step 3: Compute sample r then t

$$t = \sqrt{\frac{n-2}{1-r^2}}$$

Step 4: Sample t is significant if it is greater than t_{crit}

Step 5: If significance occurs data are linearly correlated otherwise they are not.

Regression

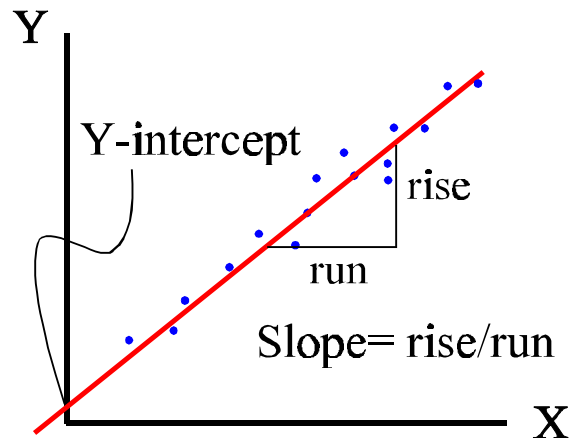
Regression:

- **Can only be done if a significant correlation exists.**
- Equation of line or curve which defines the relationship between variables.
- The “line of best fit.”
- Mathematical technique is called “least squares” method. This technique computes the line that minimizes the squares of the deviations of the data from the line.

$$\hat{Y} = mX + b$$

therefore

$$\hat{X} = \frac{1}{m} Y - \frac{b}{m}$$



$$\text{slope} = m = \frac{n(\sum XY) - (\sum X)(\sum Y)}{n(\sum X^2) - (\sum X)^2}$$

$$\text{Y - intercept} = b = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{n(\sum X^2) - (\sum X)^2}$$

Coefficient of Determination and Standard Error of Estimate

Coefficient of Determination

- Measures the strength of the relationship between the two variables.
- Equal to the explained variation divided by the total variation = r^2
- Usually given as a percentage, i.e.,

$$\text{coefficient of determination} = r^2 \times 100\%$$

For example, an r of 0.90 has 81% of total variation explained but an r of 0.60 has only 36% of its variation. A correlation may be significant but explain very little.

Standard Error Of Estimate

- Measure of the variability of the observed values about the regression line
- Can be used to compute a confidence interval for a predicted value

standard error of estimate:

$$s_{est} = \sqrt{\frac{\Sigma(Y - \hat{Y})^2}{n - 2}}$$
$$= \sqrt{\frac{\Sigma Y^2 - b\Sigma Y - m\Sigma(XY)}{n - 2}}$$

Possible Reasons for a Significant Correlation

1. There is a **direct cause-and-effect relationship** between the variables. That is, x causes y. For example, positive reinforcement improves learning, smoking causes lung cancer and heat causes ice to melt.
2. There is a **reverse cause-and-effect relationship** between the variables. That is, y causes x. For example, suppose a researcher believes excessive coffee consumption causes nervousness, but the researcher fails to consider that the reverse situation may occur. That is, it may be that an nervous people crave coffee.
3. The relationship between the variables may be **caused by a third variable**. For example, if a statistician correlated the number of deaths due to drowning and the number of cans of soft drinks consumed during the summer, he or she would probably find a significant relationship. However, the soft drink is not necessarily responsible for the deaths, since both variables may be related to heat and humidity.
4. There may be a **complexity of interrelationships** among many variables. For example, a researcher may find a significant relationship between students' high school grades and college grades. But there probably are many other variables involved, such as IQ, hours of study, influence of parents, motivation, age and instructors.
5. The relationship may be **coincidental**. For example, a researcher may be able to find a significant relationship between the increase in the number of people who are exercising and the increase in the number of people who are committing crimes. But common sense dictates that any relationship between these two variables must be due to coincidence.