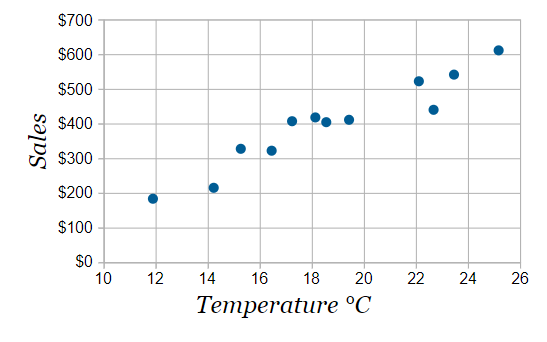
**STATISTICS**

**Bivariate Data:**

**Bivariate data** deals with two variables that can change and are compared to find relationships. If one variable is influencing another variable, then you will have bivariate data that has an independent and a dependent variable. This is because one variable depends on the other for change. An **independent variable** is a condition or piece of data in an experiment that can be controlled or changed. A **dependent variable** is a condition or piece of data in an experiment that is controlled or influenced by an outside factor, most often the independent variable.

**Example:** An ice cream shop keeps track of how much ice cream they sell versus the temperature on that day. The two variables are **Ice Cream Sales** and **Temperature**

Now we can easily see that **warmer weather** and **more ice cream sales** are linked, but the relationship is not perfect.

So with bivariate data we are interested in **comparing** the two sets of data and finding any **relationships**.

We can use Tables, Scatter Plots, Mean, Covariance, Correlation, Line of Best Fit, and plain old common sense.

**Bivariate Normal Distribution**

The bivariate normal distribution is the statistical distribution with probability density function

|  |  |
| --- | --- |
| P(x_1,x_2)=1/(2pisigma_1sigma_2sqrt(1-rho^2))exp[-z/(2(1-rho^2))], | (1) |

where

|  |  |
| --- | --- |
| z=((x_1-mu_1)^2)/(sigma_1^2)-(2rho(x_1-mu_1)(x_2-mu_2))/(sigma_1sigma_2)+((x_2-mu_2)^2)/(sigma_2^2), | (2) |

and

|  |  |
| --- | --- |
| rho=cor(x_1,x_2)=(V_(12))/(sigma_1sigma_2) | (3) |

is the correlation of x_1 and x_2 and V_(12) is the covariance.

The probability density function of the bivariate normal distribution is implemented as MultinormalDistribution[{*mu*1, *mu*2}, {{*sigma*11, *sigma*12}, {*sigma*12, *sigma*22}}] in the Wolfram Language package MultivariateStatistics` .

The marginal probabilities are then

|  |  |  |  |
| --- | --- | --- | --- |
| P(x_1) | = | int_(-infty)^inftyP(x_1,x_2)dx_2 | (4) |
| http://mathworld.wolfram.com/images/equations/BivariateNormalDistribution/Inline15.gif | = | 1/(sigma_1sqrt(2pi))e^(-(x_1-mu_1)^2/(2sigma_1^2)) | (5) |

and

|  |  |  |  |
| --- | --- | --- | --- |
| P(x_2) | = | int_(-infty)^inftyP(x_1,x_2)dx_1 | (6) |
| http://mathworld.wolfram.com/images/equations/BivariateNormalDistribution/Inline21.gif | = | 1/(sigma_2sqrt(2pi))e^(-(x_2-mu_2)^2/(2sigma_2^2)) | (7) |

**Covariance:**

Covariance provides a measure of the strength of the correlation between two or more sets of random variates. The covariance for two random variates Xand Y, each with sample size N, is defined by the expectation value

|  |  |  |  |
| --- | --- | --- | --- |
| cov(X,Y) | = | <(X-mu_X)(Y-mu_Y)> | (1) |
| http://mathworld.wolfram.com/images/equations/Covariance/Inline7.gif | = | <XY>-mu_Xmu_y | (2) |

where mu_x=<X> and mu_y=<Y> are the respective means, which can be written out explicitly as

|  |  |
| --- | --- |
| cov(X,Y)=sum_(i=1)^N((x_i-x^_)(y_i-y^_))/N. | (3) |

**For uncorrelated variates,**

|  |  |
| --- | --- |
| cov(X,Y)=<XY>-mu_Xmu_Y=<X><Y>-mu_Xmu_Y=0, | (4) |

so the covariance is zero. However, if the variables are correlated in some way, then their covariance will be nonzero. In fact, if cov(X,Y)>0, then Ytends to increase as X increases, and if cov(X,Y)<0,   
then Y tends to decrease as X increases.

Note that while statistically independent variables are always uncorrelated, the converse is not necessarily true.

In the special case of Y=X,

|  |  |  |  |
| --- | --- | --- | --- |
| cov(X,X) | = | <X^2>-<X>^2 | (5) |
| http://mathworld.wolfram.com/images/equations/Covariance/Inline22.gif | = | sigma_X^2, | (6) |

so the covariance reduces to the usual variance sigma_X^2=var(X). This motivates the use of the symbol sigma_(XY)=cov(X,Y), which then provides a consistent way of denoting the variance as sigma_(XX)=sigma_X^2, where sigma_X is the standard deviation.

**Properties of variance and covariance**

(a) If $ X$ and $ Y$ are independent, then $ {\rm Cov}(X,Y) = 0$ by observing that $ E[XY] = E[X] \cdot E[Y]$.

(b) In contrast to the expectation, the variance is *not* a linear operator. For two random variables $ X$ and $ Y$, we have

|  |  |
| --- | --- |
| $\displaystyle {\rm Var}(X + Y) = {\rm Var}(X) + {\rm Var}(Y) + 2 {\rm Cov}(X,Y).$ | (3) |

However, if $ X$ and $ Y$ are independent, by observing that $ {\rm Cov}(X,Y) = 0$ in, we have

|  |  |
| --- | --- |
| $\displaystyle {\rm Var}(X + Y) = {\rm Var}(X) + {\rm Var}(Y).$ | (4) |

In general, when we have a sequence of independent random variables $ X_1,\ldots,X_n$, the property  is extended to

$\displaystyle {\rm Var}(X_1 + \cdots + X_n)
= {\rm Var}(X_1) + \cdots + {\rm Var}(X_n).
$

Variance and covariance under linear transformation. Let $ a$ and $ b$ be scalars (that is, real-valued constants), and let $ X$ be a random variable. Then the variance of $ aX + b$ is given by

$\displaystyle {\rm Var}(aX + b) = a^2 {\rm Var}(X).
$

Now let $ a_1$, $ a_2$, $ b_1$ and $ b_2$ be scalars, and let $ X$ and $ Y$ be random variables. Then similarly the covariance of $ a_1 X + b_1$ and $ a_2 Y + b_2$ can be given by

$\displaystyle {\rm Cov}(a_1 X + b_1, a_2 Y + b_2) = a_1 a_2 {\rm Cov}(X,Y)
$

**Analysis of Bivariate Data:** when data regarding two or more variables are available, we may have to study the relative variation between the variables. This process is called analysis of bivariate data

**Two Questions of Bivariate Data Analysis**

1. What is the degree of linearity? is there a line?
2. What is the degree of association? how strong is the line?

**Plot the Data in Scatter Plot and observe**

* Shape: Does relationship look linear?
* Outliers: Are there any unusual points?
* Direction: Is linear relationship positive or negative?
* Strength: Is the line strong?

**Correlation:** A statistical device that helps in studying the covariance of two or more variables. a study of interdependence. r is the symbol for correlation r takes on values between -1 and 1

**Categories of correlation:**

1. **Simple Correlation:** Covariance between two related variables
2. **Multiple Correlation: R**elated variations between three or more variables. all the variables are studied simultaneously.
3. **Partial Correlation:** Related variation between two or more variables. only consider two variables to be influencing each other and the effect of other influencing variables kept constant

if x and y are correlated, the three types of relationships are

1. x is the cause and y is the effect
2. x is the effect and y is the cause
3. x and y are the effects of some other causes

**Significance of study of correlation:** measures the degree of relationship between variables can estimate the value of one variable with the help of another value when the two are related helps in progressive development in the methods of science and philosophy. Reduces uncertainty in any field correlation analysis is useful in economics and business

**Types of correlation**

1. **Positive Correlation:** if the variables vary in the same direction. also called direct correlation
2. **Negative Correlation:** if the variables vary in the opposite direction. also called inverse correlation
3. **Non Correlation :** if the two variables show no associated variation
4. **Perfect Correlation:** if the variables vary in the same proportion, i.e. the variables show exact linear relationship, then the correlation is said to be perfect. May be positive or negative.

**Correlation ≠ Causation:**

Just because we have a high r value does not mean that x causes y

There could be a “lurking variable” that actually is the cause.

Example

**Minutes Late to Work**

1. Direct Causation: late to work and rain
2. Reverse Cause and Effect: late to work and poor relationship with boss
3. 3rd variable: late to work and rain and # of children at home
4. Coincidence: late to work and height

**Regression:** The property of the tendency of the actual value to lie close to the estimated value is called regression. It is the theory of estimation of unknown value of a variable with the help of known values of the other variables.

y = mx + c 🡺 represents a straight line graph on the xy plane with slope m and intercept c.

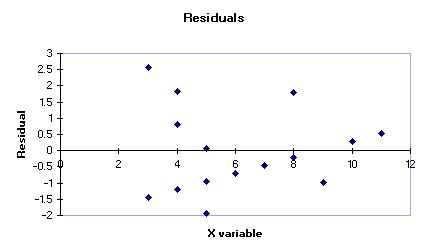
**Centroid:** Regression line always passes through the centroid point of data, (x̄,ȳ)

**Curve of regression and line of regression:** if the variables in a bivariate data are correlated we may find that the points in the scatter diagram will cluster in the form of a curve----> curve of regression or in the form of a straight line ------> line of regression.

**Line of Least Squares:** The regression line is sometimes called the line of least squares. The “best fit” minimizes the squares of the residuals of each point from the regression line.

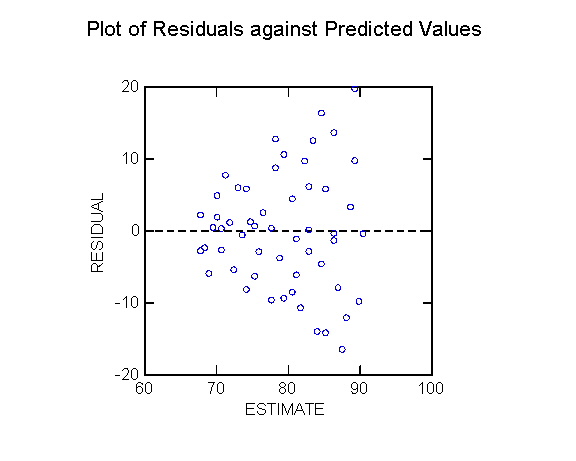
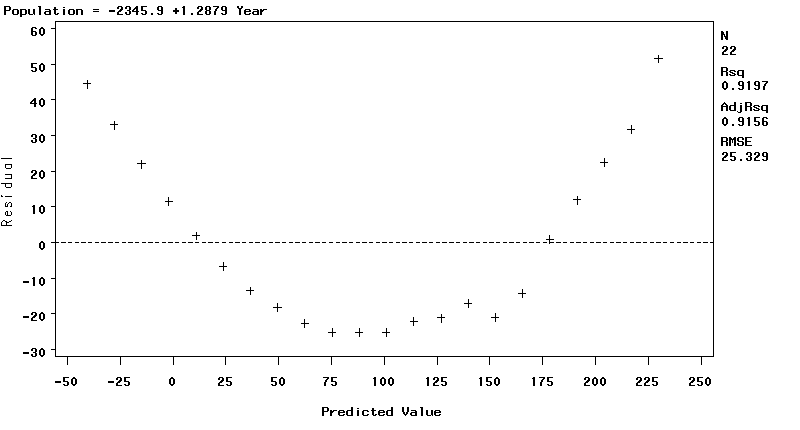
* ***r is the correlation coefficient***
  + *Measures strength of linear relationship*
  + *Can be positive or negative*
  + *Is always a number between -1 and 1*
* ***r2 is the coefficient of determination***
  + *Measures what % of the variation in y is explained by (or determined by) the variation in x*
  + *Will be given as a %*
* ***1 – r2 is the coefficient of non-determination***
  + *Measures what % of the variation in y which is explained by chance and other factors*
  + *% of variation in y NOT explained by variation in x*

**Is Linear Model Appropriate?**

* *Check SCATTERPLOT and r value for strong linearity*
* *Look at RESIDUAL PLOT—it should be random* 
  + *A graph of the residual values compared to the x values.*
  + *We do not want to see a pattern of residuals increasing or decreasing or fitting any noticeable curve.*

**Good Residual = Random 🡺**

**Bad Residual = Pattern**

******

**Correlation and Coefficient Strength:**

Threshold = 1.96 /  √n 🡺 n = sample size

* If r > threshold correlation and coefficient is significant
* If r < threshold correlation and coefficient is may be or may not be significant
  + Will lead to more analysis of data to ascertain correlation and coefficient is significant or not

**Good Model = Causation??**

If a linear regression model is good (strong r, linear scatterplot, random residuals), that DOES NOT mean that x causes y. For a researcher to assert that x causes y, the data must have come from a PLANNED, CONTROLLED EXPERIMENT.

**Re-expressing data=transformation**

We transform data if a linear model is not appropriate. Transforming data means re-expressing the numbers using algebraic operations that take the “curve” out of the original data.

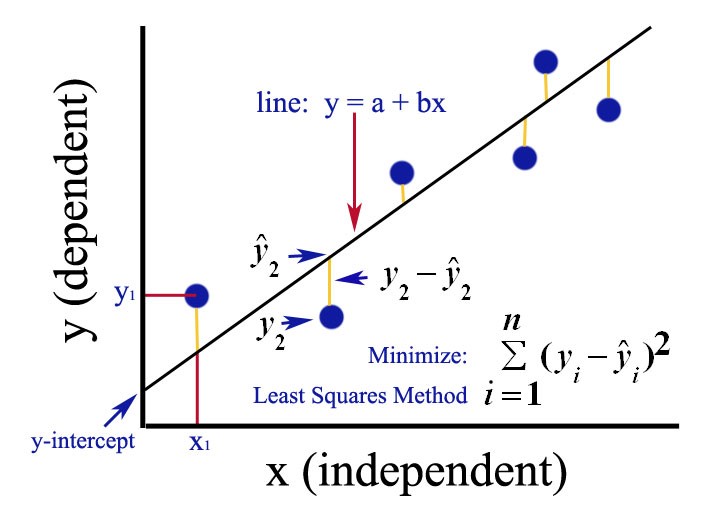
**Examples:**  take square root, take log or ln, use reciprocals

**Log Transformations: How to Handle Negative Data Values?**

The log transformation is one of the most useful transformations in data analysis. It is used as a transformation to normality and as a variance stabilizing transformation.

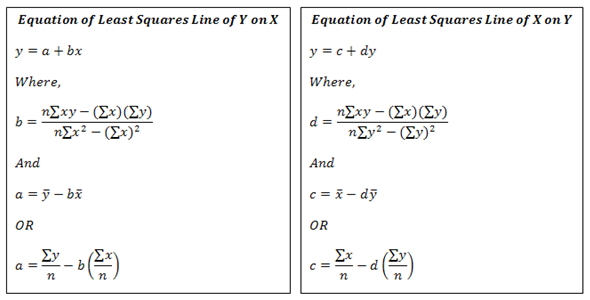
A log transformation is often used as part of exploratory data analysis in order to visualize (and later model) data that ranges over several orders of magnitude. Common examples include data on income, revenue, populations of cities, sizes of things, weights of things, and so forth.

**Translate, then Transform:** A common technique for handling negative values is to add a constant value to the data prior to applying the log transform. The transformation is therefore log(*Y+a*) where *a* is the constant. Some people like to choose *a* so that min(*Y+a*) is a very small positive number (like 0.001). Others choose *a* so that min(*Y+a*) = 1. For the latter choice, you can show that *a = b* – min(*Y*), where *b* is either a small number or is 1.

**Least-Squares Regression Line**:

The linear fit that matches the pattern of a set of paired data as closely as possible. Out of all possible linear fits, the least-squares regression line is the one that has the smallest possible value for the sum of the squares of the residuals.

**Formula:**



**Generic Regression Function:**

y = f(x, β) + ∈ (1)

y = Dependent Variable x = Independent Variable ∈ = Noise/error  
**Why there is a noise?**

The errors represent everything that the model does not have into account. And why is that? Because it would be extremely unlikely for a model to perfectly predict a variable, as it is impossible to control every possible condition that may interfere with the response variable. The errors may also include reading or measuring inaccuracies. Considering the regression line of best fit, the errors are based on the distance from each point to that line.

**Assumptions:**

* Normally distributed
* 0 mean
* Un correlated
* ∈ ~ n (o, σ2) 🡺 o = mean | σ = variance n = sample
* Covariance between errors should be 0 🡺 Cov (Et, Et+1) = 0
* Independent

**Example1**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **n=5** | **x** | **y** | **x–x̅** | **y–y̅** | **(x–x̅)²** | **(x–x̅)(y–y̅)** |
|  | 0 | 6 | –3.2 | 10.8 | 10.24 | –34.56 |
|  | 2 | –1 | –1.2 | 3.8 | 1.44 | –4.56 |
|  | 3 | –3 | –0.2 | 1.8 | –0.04 | –0.36 |
|  | 5 | –10 | 1.8 | –5.2 | 3.24 | –9.36 |
|  | 6 | –16 | 2.8 | –11.2 | 7.84 | –31.36 |
| ∑ | 16 | –24 | 0 | 0 | 22.8 | –80.2 |
| mean | 3.2 | –4.8 |  |  |  |  |

Now compute m and b:

m = −80.2 / 22.8 = −3.5175

b = −4.8 − (−3.5175)(3.2) = 6.456

**Example2**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **n=5** | **x** | **y** | **x²** | **xy** |
|  | 0 | 6 | 0 | 0 |
|  | 2 | –1 | 4 | –2 |
|  | 3 | –3 | 9 | –9 |
|  | 5 | –10 | 25 | –50 |
|  | 6 | –16 | 36 | –96 |
| ∑ | 16 | –24 | 74 | –157 |

Now compute m and b:

m = 5(−157) − (16)(−24) / 5(74) − 16²

m = −401 / 114

m = −3.5175

b = −24 − (−3.5175)(16) / 5

b = 32.28 / 5 = 6.456

**Transformations of functions**

**Vertical and horizontal Shifts:**

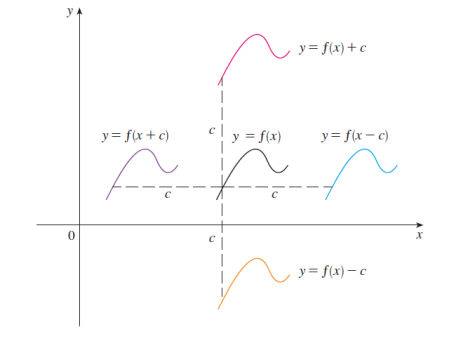
Suppose c > 0. To obtain the graph of

y = f(x) + c, shift the graph of y = f(x) a distance c units upward

y = f(x) − c, shift the graph of y = f(x) a distance c units downward

y = f(x + c), shift the graph of y = f(x) a distance c units to the left

y = f(x − c), shift the graph of y = f(x) a distance c units to the right



**Vertical and horizontal Stretching and Reflecting:**

Suppose c > 1. To obtain the graph of

y = cf(x), stretch the graph of y = f(x) vertically by a factor of c

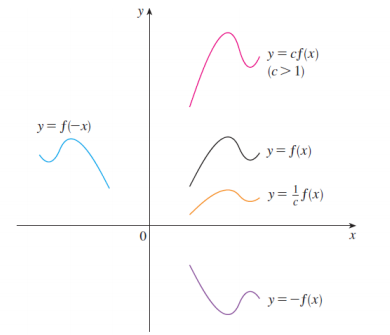
y = (1/c)f(x), shrink the graph of y = f(x) vertically by a factor of c

y = f(cx), shrink the graph of y = f(x) horizontally by a factor of c

y = f(x/c), shrink the graph of y = f(x) horizontally by a factor of c

y = −f(x), reflect the graph of y = f(x) about the x-axis

y = f(−x), reflect the graph of y = f(x) about the y-axis



To shift the graph y0 units upwards and x0 units to the right then the equation becomes

y+y0=m(x−x0)+b

For downward shifting and shifting to the left you have to change the signs.

**Numerical example**

Suppose the initial function is y=2x+2. And now we want to shift it 1 unit upward and 3 units to the right. The equation becomes

y+1=2(x−3)+2

Multiplying out the brackets

y+1=2x−6+2

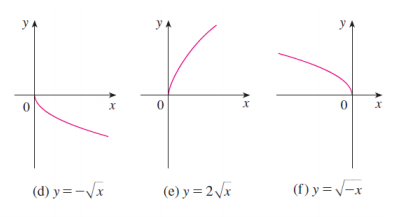
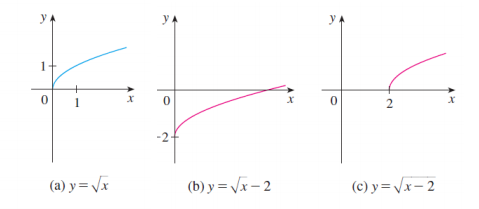
Subtracting 1 on both sides

y=2x−6+2−1

y=2x−5

**Graphical example**

Given the graph of y = √ x, use transformations to graph y = √ x − 2, y = √ x − 2, y = − √ x, y = 2√ x, y = √ −x

****