VBM683
Machine Learning

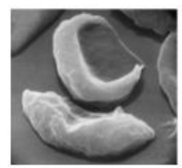
Pinar Duygulu

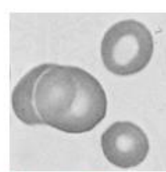
Slides are adapted from
Dhruv Batra,
Aarti Singh, Barnabas Poczos,
Wenjiang Fu
Aykut Erdem

Classification

- Input: X
  - Real valued, vectors over real.
  - Discrete values (0,1,2,…)
  - Other structures (e.g., strings, graphs, etc.)
- Output: Y
  - Discrete (0,1,2,...)

Sports
Science
News





Anemic cell
Healthy cell

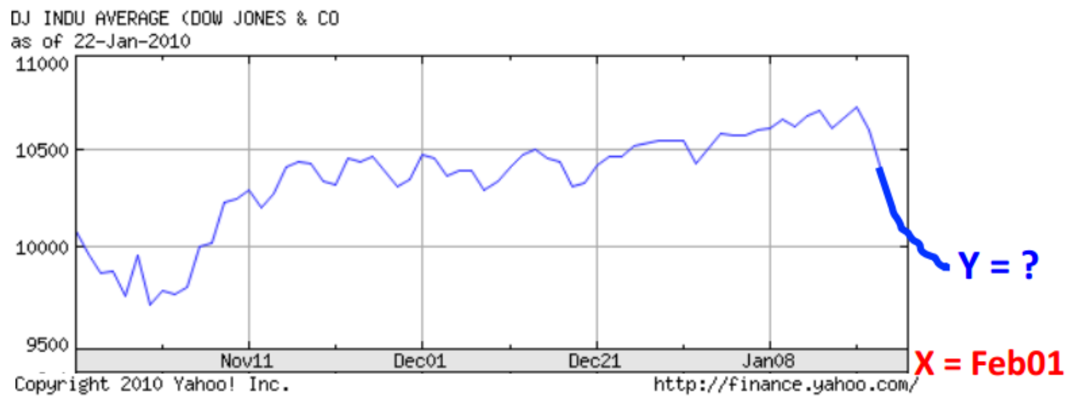
X = Document **Y = Topic** **X = Cell Image** **Y = Diagnosis**

slide by Aarti Singh and Barnabas Poczos

Regression

- Input: X
  - Real valued, vectors over real.
  - Discrete values (0,1,2,…)
  - Other structures (e.g., strings, graphs, etc.)
- Output: Y
  - Real valued, vectors over real.

Stock Market Prediction



slide by Aarti Singh and Barnabas Poczos

What should I watch tonight?

slide by Sanja Fidler

1-NN for Regression

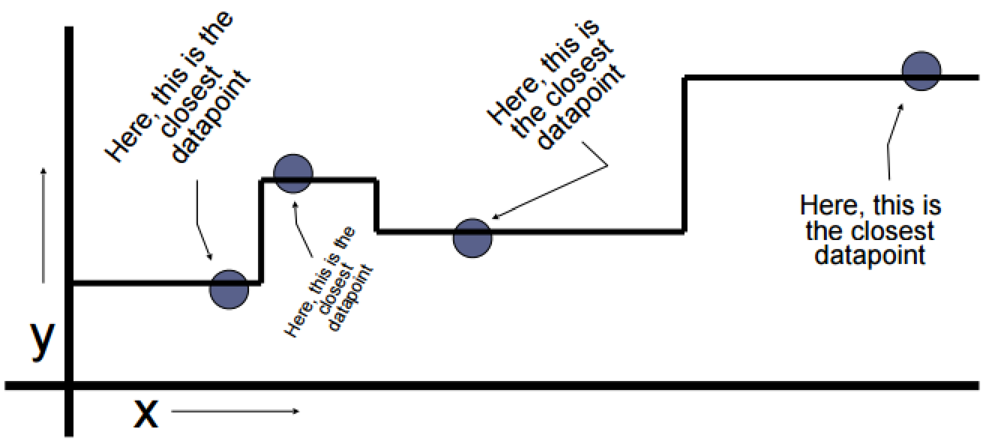
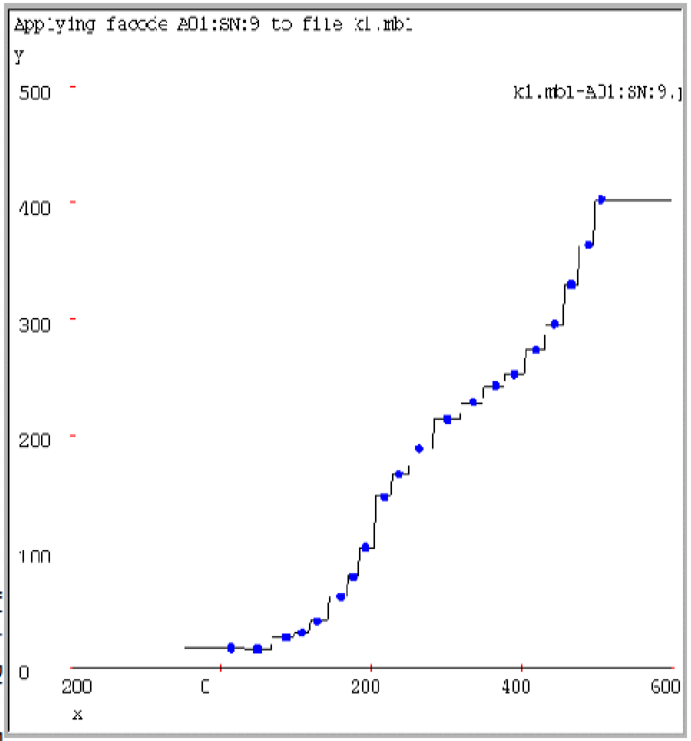


Figure Credit: Carlos Guestrin

slide by Dhruv Batra

1-NN for Regression

- Often bumpy (overfits)



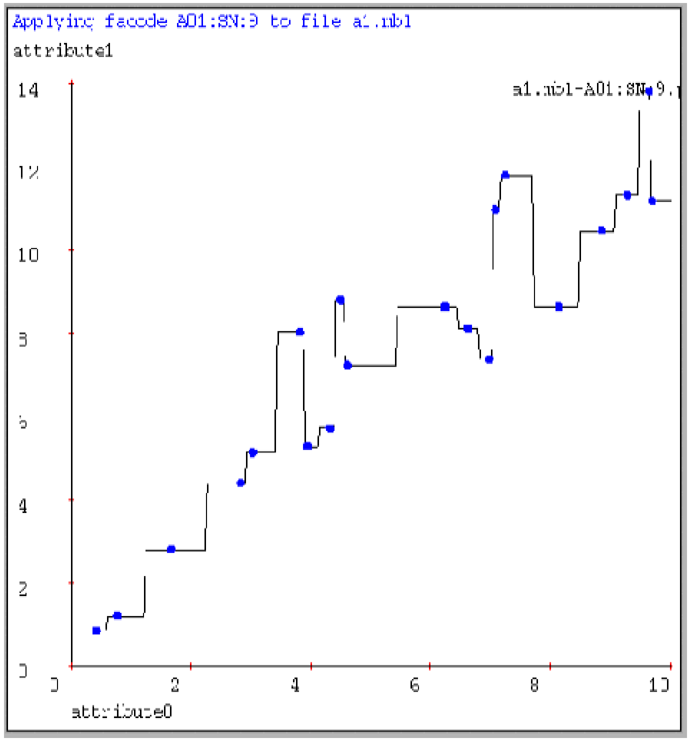
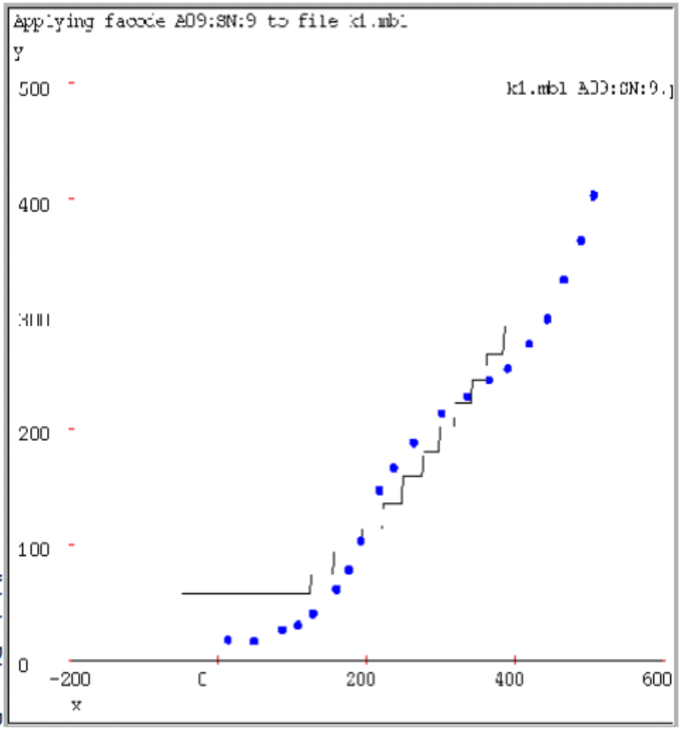


Figure Credit: Andrew Moore

slide by Dhruv Batra

9-NN for Regression

- Often bumpy (overfits)



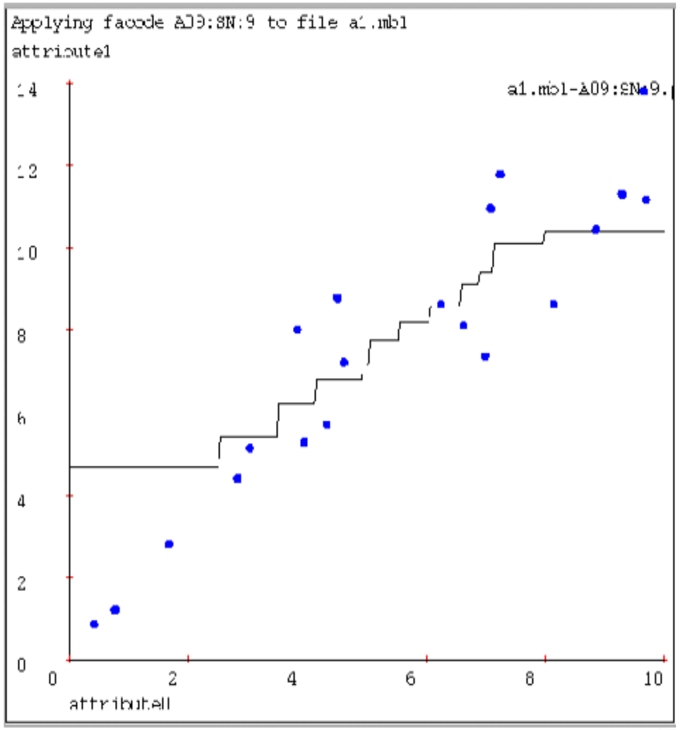
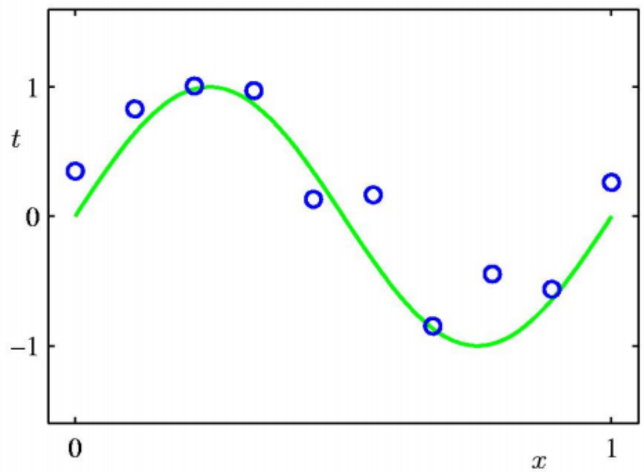


Figure Credit: Andrew Moore

slide by Dhruv Batra

Simple 1-D Regression



- Circles are data points (i.e., training examples) that are given to us
- The data points are uniform in x, but may be displaced in y

$$t(x) = f(x) + \varepsilon$$

  with ε some noise
- In green is the “true” curve that we don’t know

slide by Sanja Fidler

What is a Model?

1. Often Describe Relationship between Variables

2. Types
   - Deterministic Models (no randomness)

   - Probabilistic Models (with randomness)

Deterministic Models

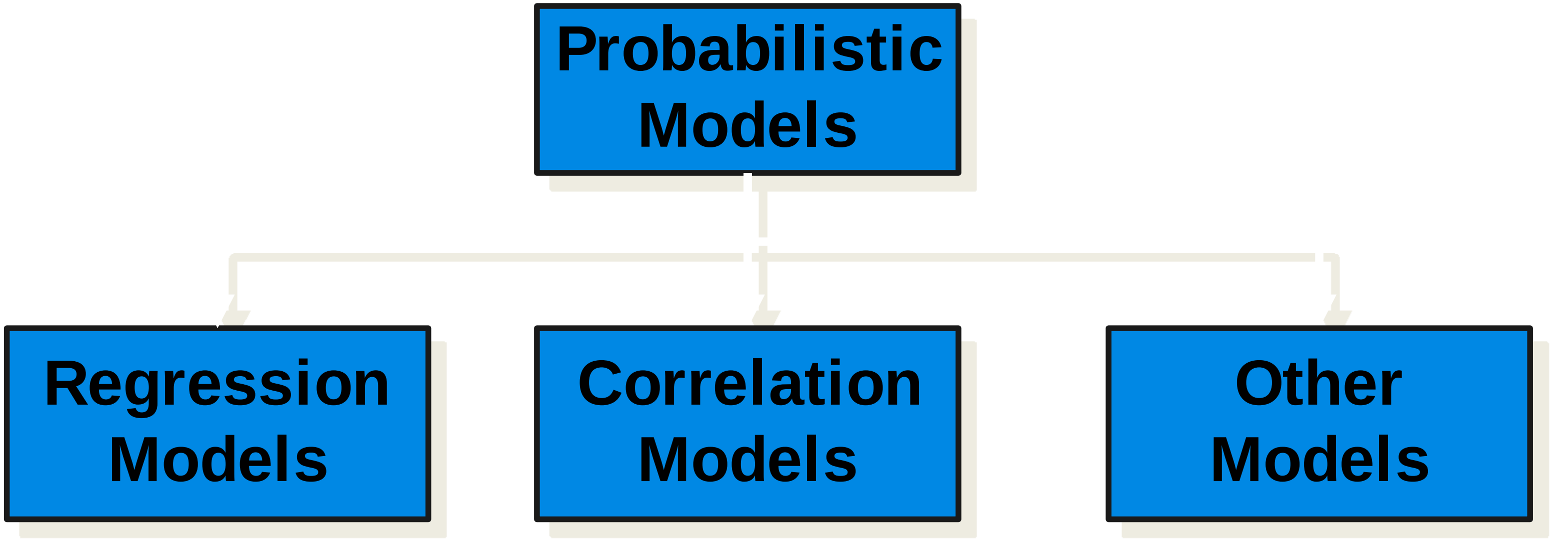
1. Hypothesize Exact Relationships
2. Suitable When Prediction Error is Negligible
3. Example: Body mass index (BMI) is measure of body fat based

   – $$BMI = \frac{\text{Weight in Kilograms}}{(\text{Height in Meters})^2}$$

Probabilistic Models

1. Hypothesize 2 Components
   - Deterministic
   - Random Error
2. Example: Systolic blood pressure of newborns Is 6 Times the Age in days + Random Error
   - *SBP* = 6 x age(d) + ε
   - Random Error May Be Due to Factors Other Than age in days (e.g. Birthweight)

Types of
Probabilistic Models



Simple Regression

- Simple regression analysis is a statistical tool that gives us the ability to estimate the mathematical relationship between a dependent variable (usually called y) and an independent variable (usually called x).
- The dependent variable is the variable for which we want to make a prediction.
- While various non-linear forms may be used, simple linear regression models are the most common.

Introduction

- The primary goal of quantitative analysis is to use current information about a phenomenon to predict its future behavior.
- Current information is usually in the form of a set of data.
- In a simple case, when the data form a set of pairs of numbers, we may interpret them as representing the observed values of an independent (or predictor or explanatory) variable X and a dependent (or response or outcome) variable Y.

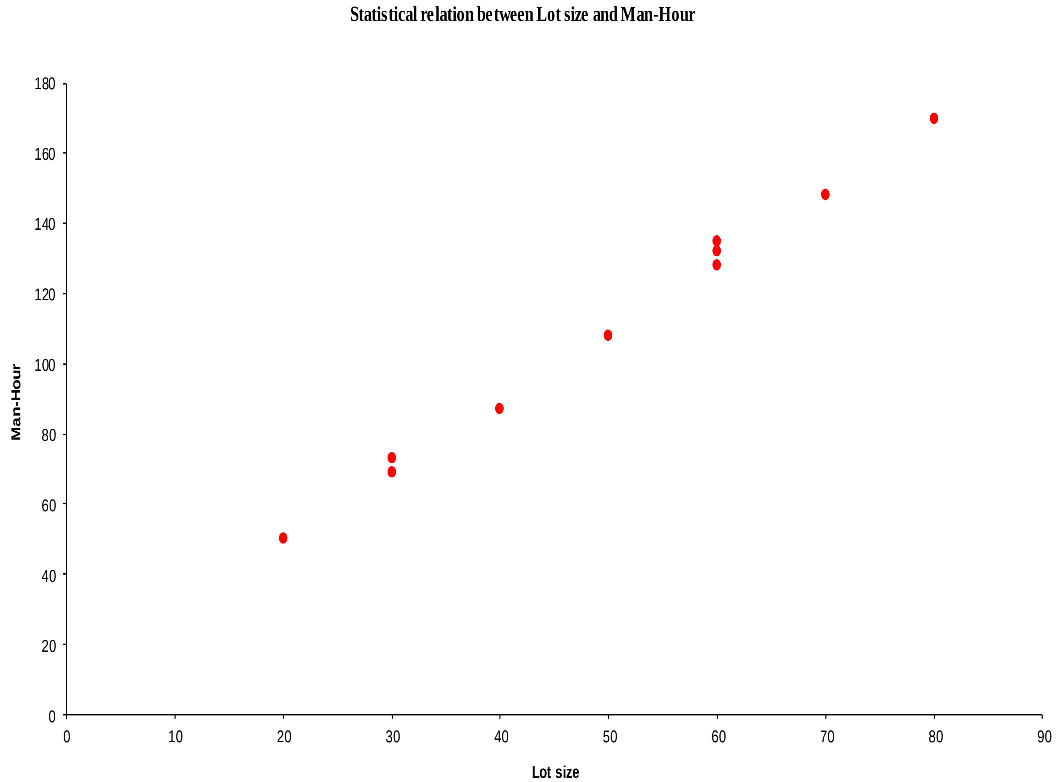
lot size	Man-hours
30	73
20	50
60	128
80	170
40	87
50	108
60	135
30	69
70	148
60	132

Introduction

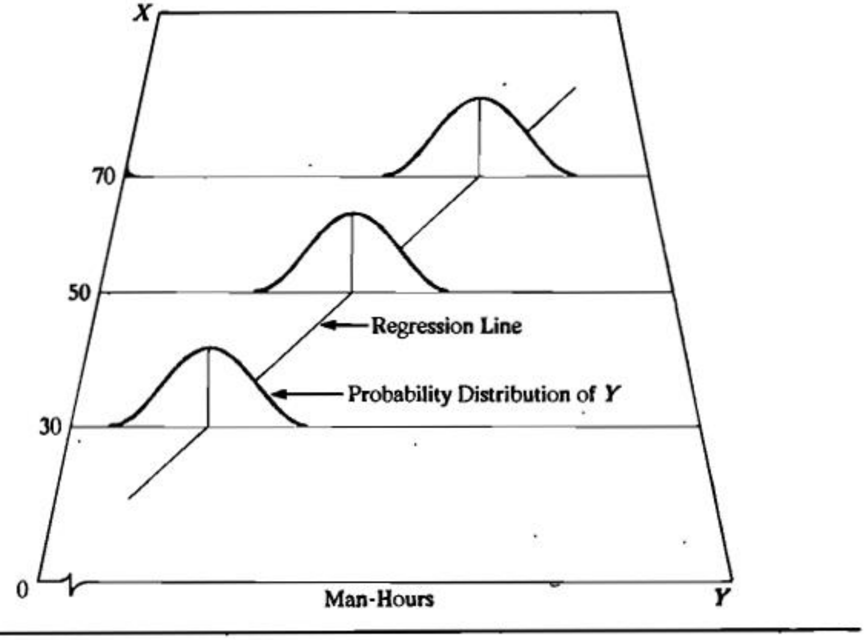
- The goal of the analyst who studies the data is to find a functional relation

 between the response variable y and the predictor variable x.

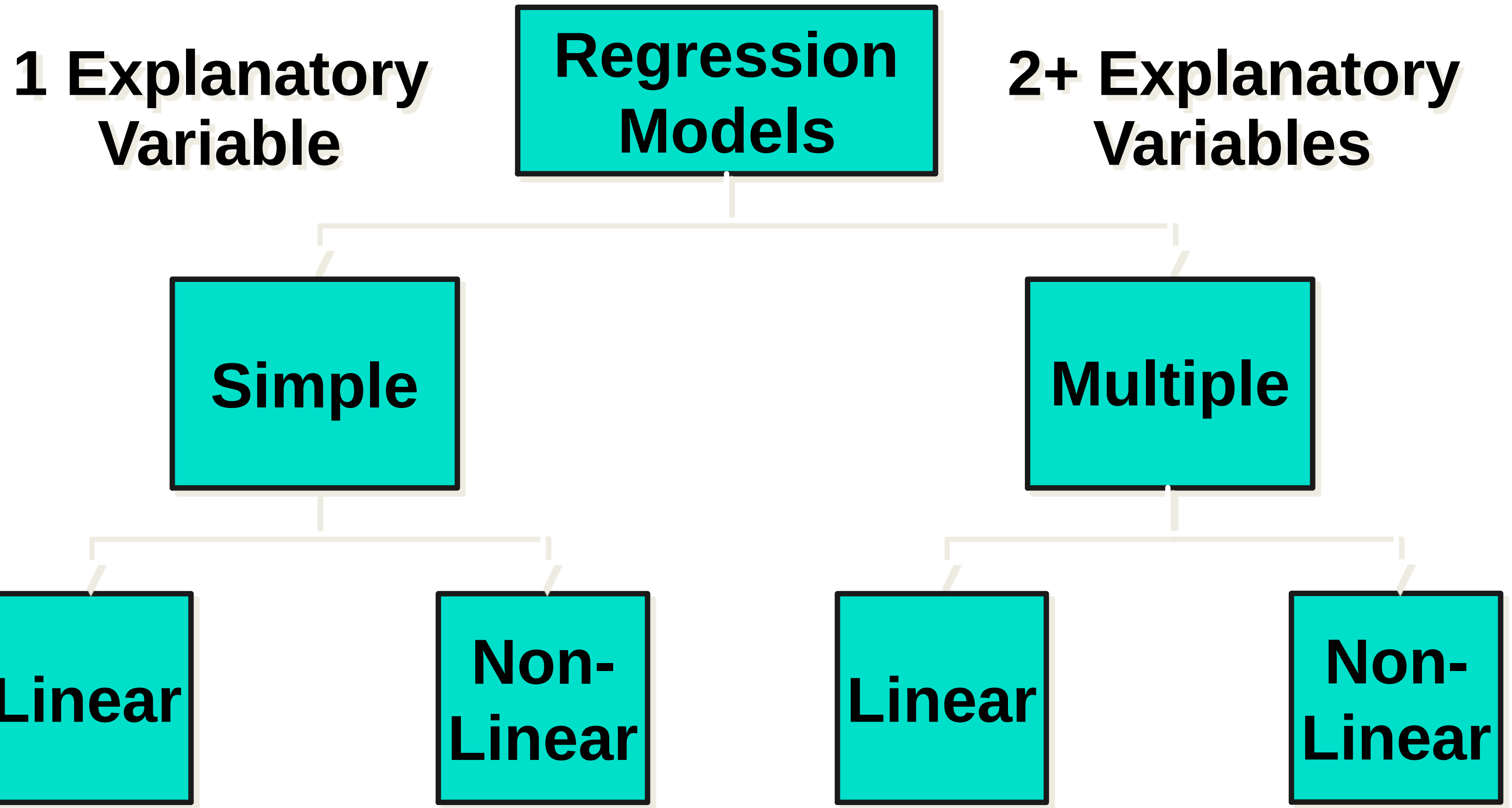
$$y = f(x)$$



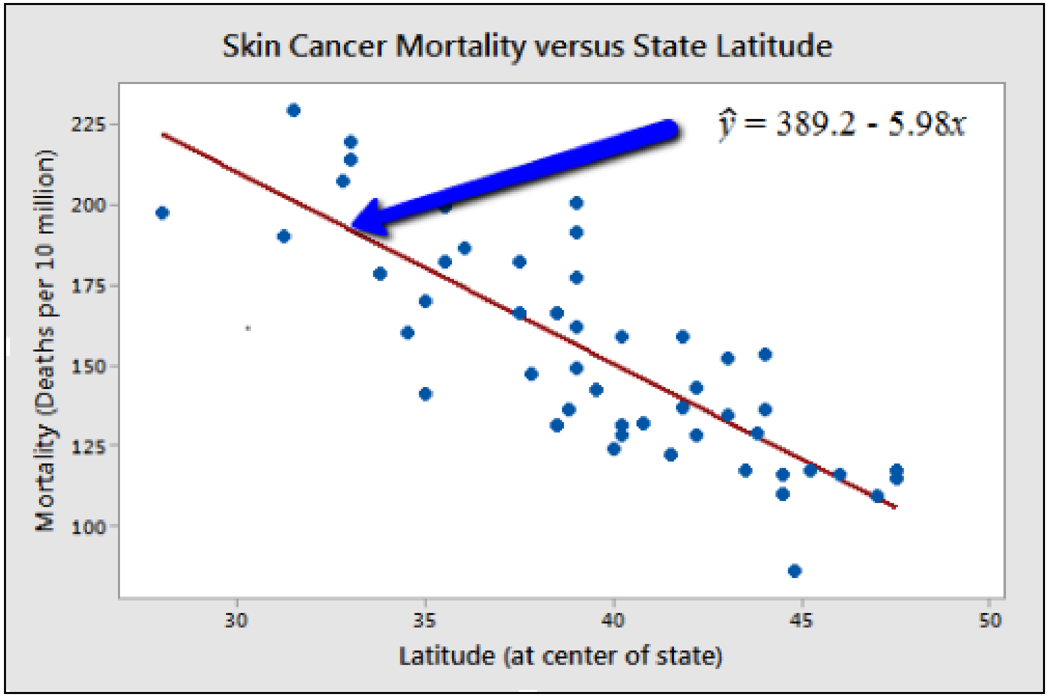
Pictorial Presentation of Linear Regression Model



Types of Regression Models



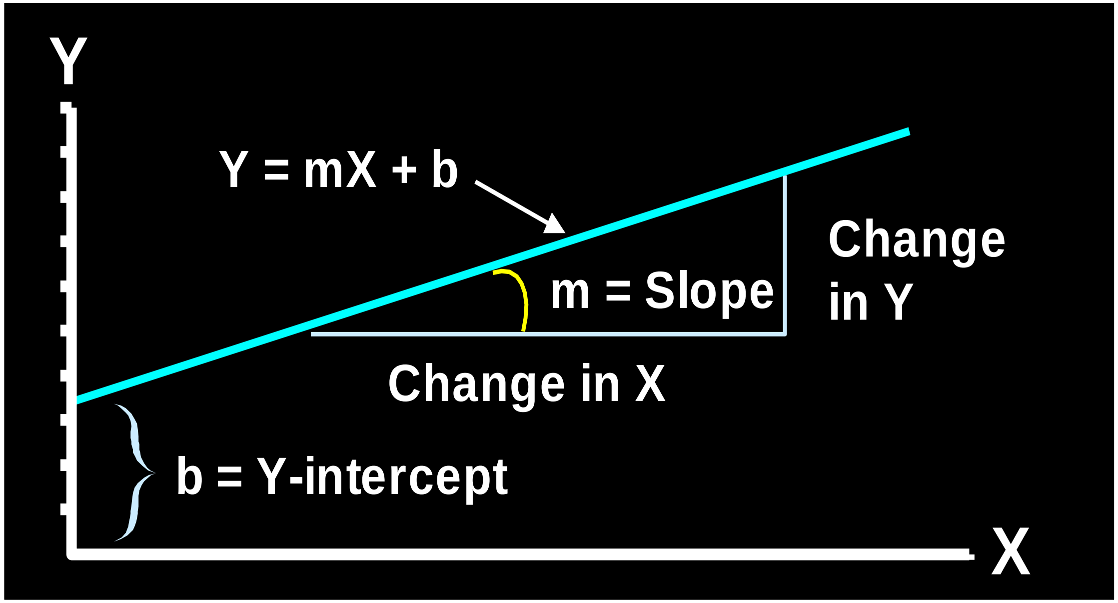
Linear Regression Model


Skin Cancer Mortality versus State Latitude
$\hat{y} = 389.2 - 5.98x$
Mortality (Deaths per 10 million)
225
200
175
150
125
100
30
35
40
45
50
Latitude (at center of state)

Assumptions

- Linear regression assumes that…
  - 1. The relationship between X and Y is linear
  - 2. Y is distributed normally at each value of X
  - 3. The variance of Y at every value of X is the same (homogeneity of variances)
  - 4. The observations are independent

Linear Equations



Linear Regression Model

- 1. Relationship Between Variables Is a Linear Function

Population Y-Intercept

Population Slope

Random Error

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Dependent (Response) Variable (e.g., CD+ c.)

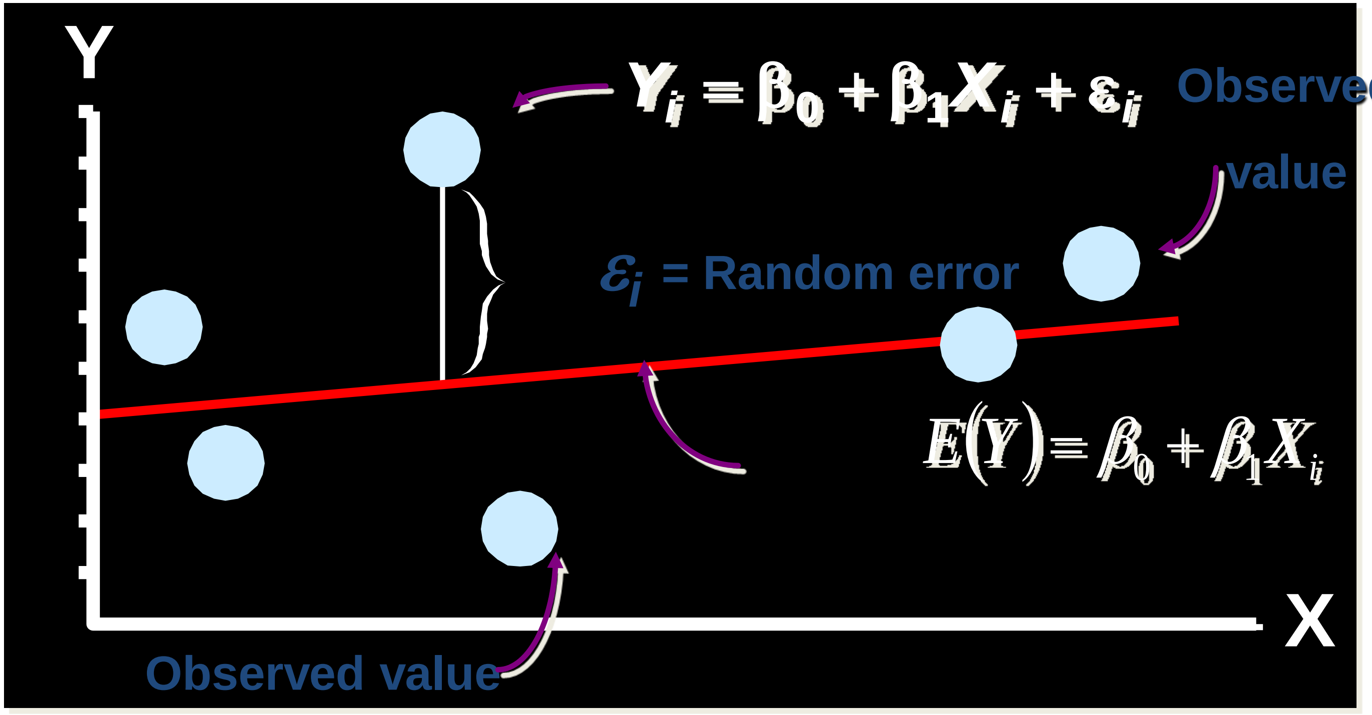
Independent (Explanatory) Variable (e.g., Years s. serocon.)

Meaning of Regression Coefficients

- General regression model
  1. β_0, and β_1 are parameters
  2. X is a known constant
  3. Deviations ε are independent N(o, σ^2)

- The values of the regression parameters β_0, and β_1 are not known. We estimate them from data.
- β_1 indicates the change in the mean response per unit increase in X.

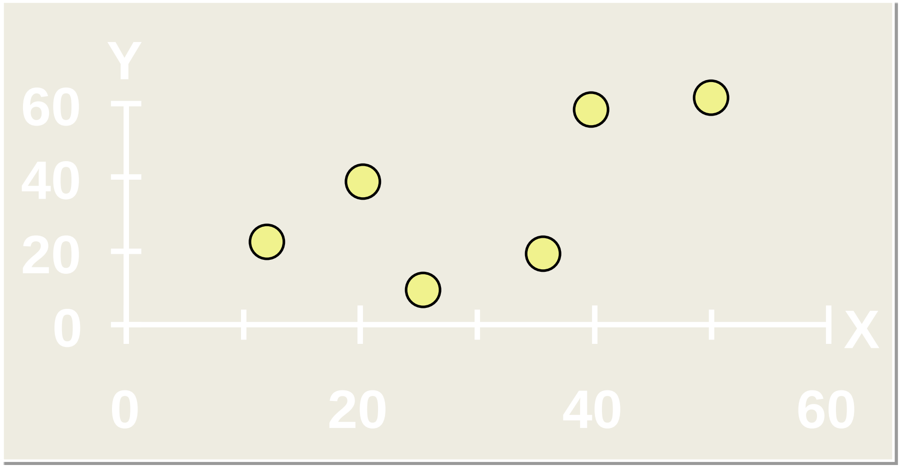
Population Linear Regression Model



Estimating Parameters: Least Squares Method

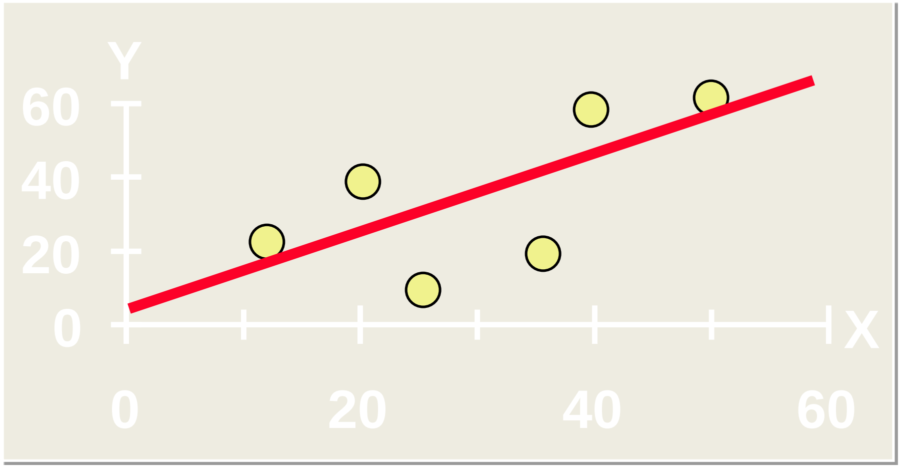
Scatter plot

- 1. Plot of All (X_i, Y_i) Pairs
- 2. Suggests How Well Model Will Fit



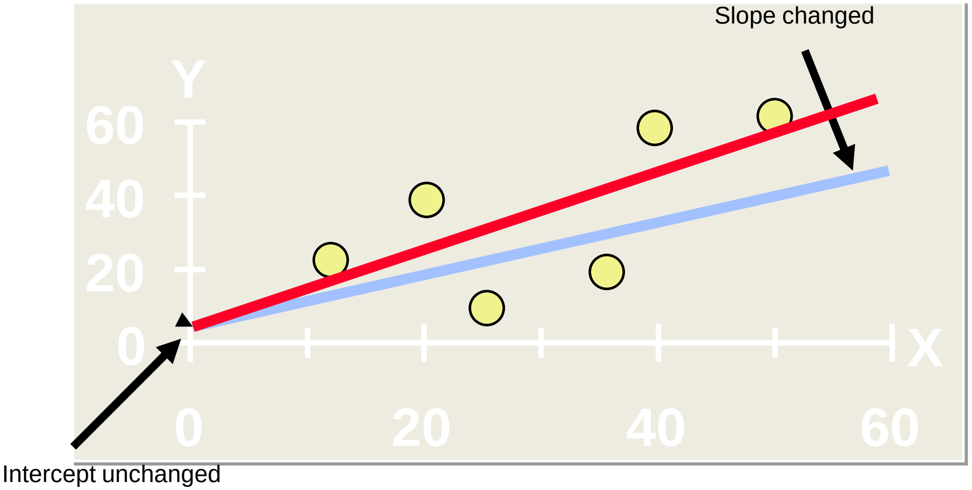
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



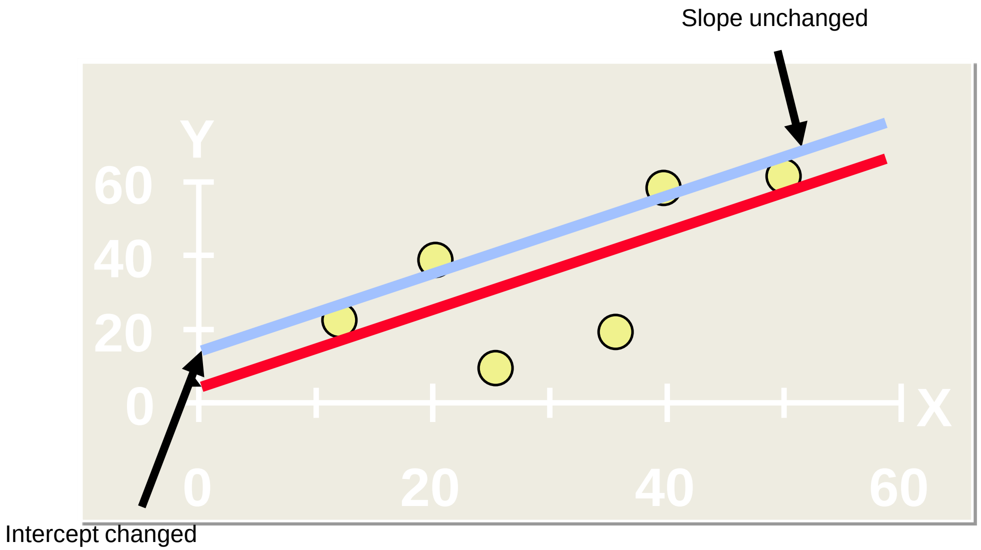
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



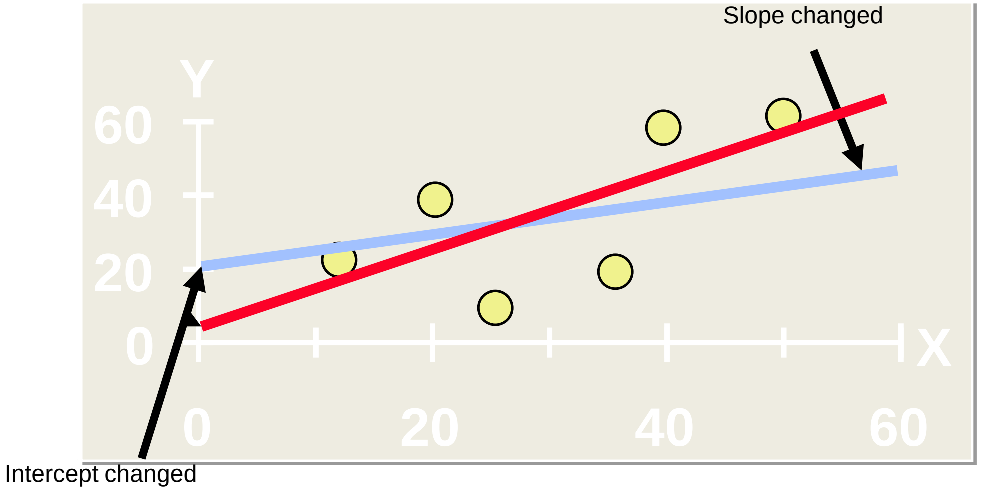
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



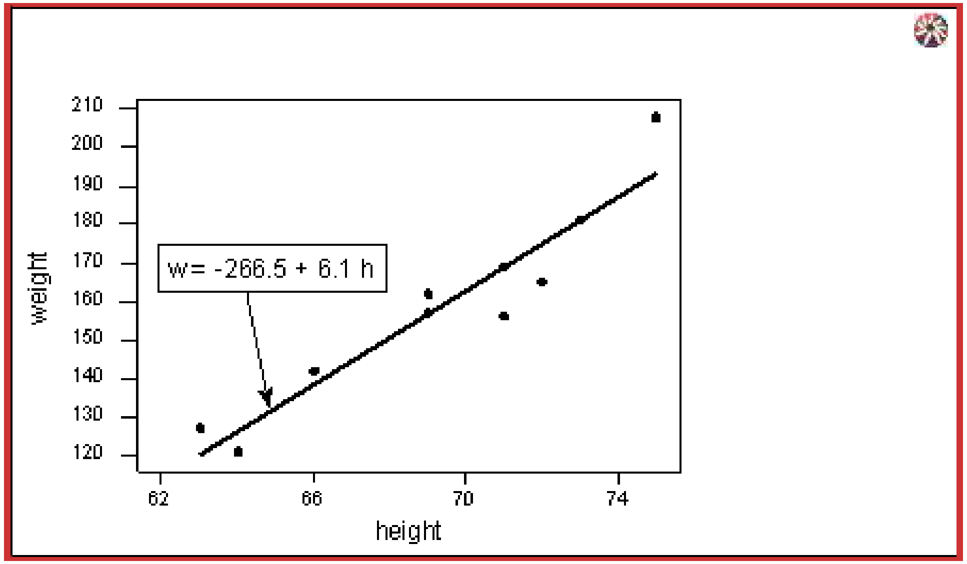
Thinking Challenge

How would you draw a line through the points? How do you determine which line 'fits best'?



What is the best fitting line

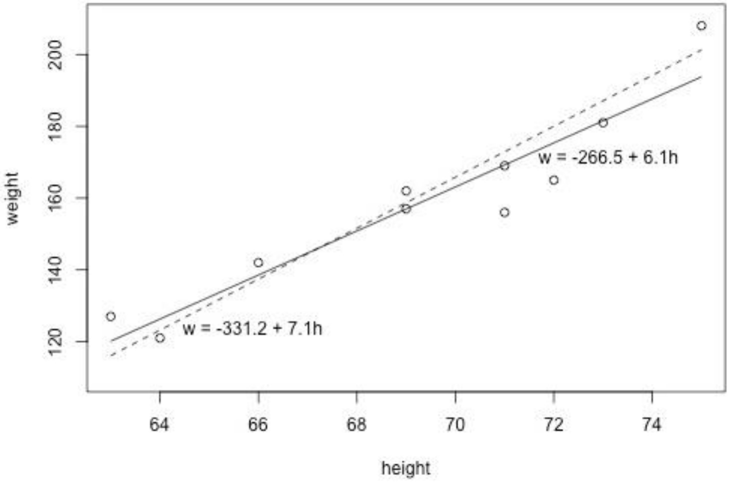
i	x_i	y_i	$\hat{y}_i$
1	63	127	120.1
2	64	121	126.3
3	66	142	138.5
4	69	157	157.0
5	69	162	157.0
6	71	156	169.2
7	71	169	169.2
8	72	165	175.4
9	73	181	181.5
10	75	208	193.8



$$\hat{y}_i = b_0 + b_1 x_i$$

- y_i denotes the observed response for experimental unit i
- x_i denotes the predictor value for experimental unit i
- $\hat{y}_i$ is the predicted response (or fitted value) for experimental unit i

Prediction Error



w = **-331.2 + 7.1** h (the dashed line)					
i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	63	127	116.1	10.9	118.81
2	64	121	123.2	-2.2	4.84
3	66	142	137.4	4.6	21.16
4	69	157	158.7	-1.7	2.89
5	69	162	158.7	3.3	10.89
6	71	156	172.9	-16.9	285.61
7	71	169	172.9	-3.9	15.21
8	72	165	180.0	-15.0	225.00
9	73	181	187.1	-6.1	37.21
10	75	208	201.3	6.7	44.89
					766.5

w = **-266.53 + 6.1376** h (the solid line)					
i	x_i	y_i	$\hat{y}_i$	$(y_i - \hat{y}_i)$	$(y_i - \hat{y}_i)^2$
1	63	127	120.139	6.8612	47.076
2	64	121	126.276	-5.2764	27.840
3	66	142	138.552	3.4484	11.891
4	69	157	156.964	0.0356	0.001
5	69	162	156.964	5.0356	25.357
6	71	156	169.240	-13.2396	175.287
7	71	169	169.240	-0.2396	0.057
8	72	165	175.377	-10.3772	107.686
9	73	181	181.515	-0.5148	0.265
10	75	208	193.790	14.2100	201.924
					597.4

$$e_i = y_i - \hat{y}_i$$

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Least Squares

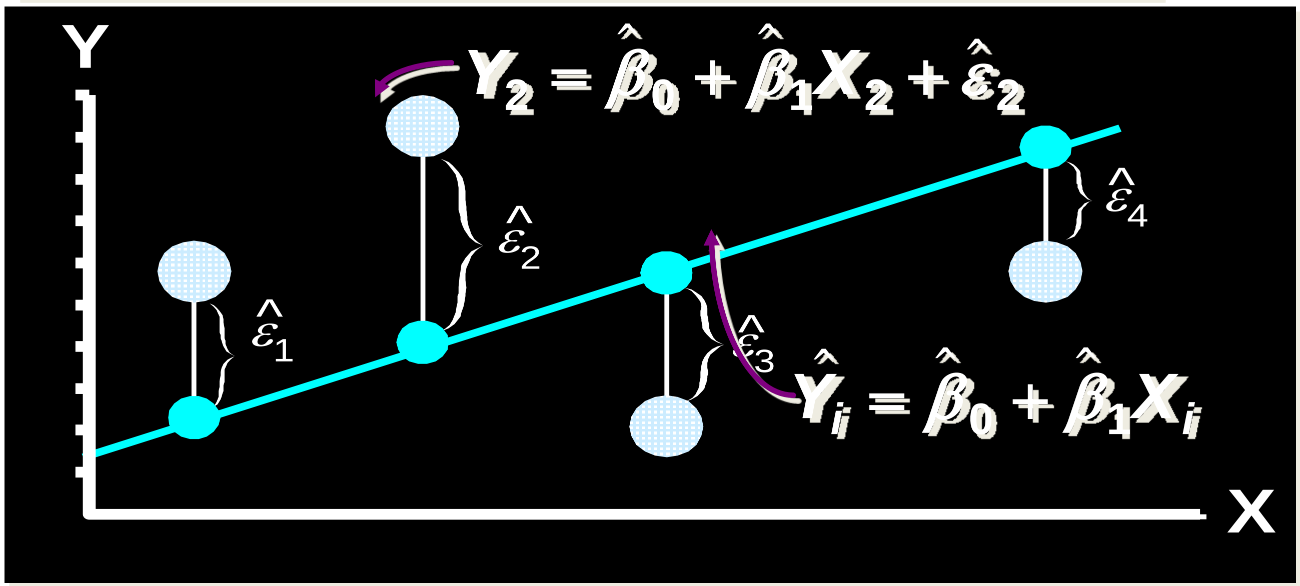
- 1. ‘Best Fit’ Means Difference Between Actual Y Values & Predicted Y Values Are a Minimum. *But* Positive Differences Off-Set Negative. So square errors!

$$\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

- 2. LS Minimizes the Sum of the Squared Differences (errors) (SSE)

Least Squares Graphically

LS minimizes $\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \hat{\varepsilon}_1^2 + \hat{\varepsilon}_2^2 + \hat{\varepsilon}_3^2 + \hat{\varepsilon}_4^2$



Coefficient Equations

- Prediction equation

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- Sample slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2}$$

- Sample Y - intercept

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Derivation of Parameters (1)

- Least Squares (L-S):
  Minimize squared error

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_i \right)^2$$

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_0} = \frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i \right)^2}{\partial \beta_0}$$

$$= -2\left(n\overline{y} - n\beta_0 - n\beta_1 \overline{x} \right)$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Derivation of Parameters (1)

- Least Squares (L-S):
  Minimize squared error

$$0 = \frac{\partial \sum \varepsilon_i^2}{\partial \beta_1} = \frac{\partial \sum \left(y_i - \beta_0 - \beta_1 x_i \right)^2}{\partial \beta_1}$$
$$= -2 \sum x_i \left(y_i - \beta_0 - \beta_1 x_i \right)$$
$$= -2 \sum x_i \left(y_i - \bar{y} + \beta_1 \bar{x} - \beta_1 x_i \right)$$

$$\beta_1 \sum x_i \left(x_i - \bar{x} \right) = \sum x_i \left(y_i - \bar{y} \right)$$
$$\beta_1 \sum \left(x_i - \bar{x} \right)\left(x_i - \bar{x} \right) = \sum \left(x_i - \bar{x} \right)\left(y_i - \bar{y} \right)$$
$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

Computation Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
X_1	Y_1	X_1^2	Y_1^2	X_1Y_1
X_2	Y_2	X_2^2	Y_2^2	X_2Y_2
:	:	:	:	:
X_n	Y_n	X_n^2	Y_n^2	X_nY_n
ΣX_i	ΣY_i	ΣX_i^2	ΣY_i^2	ΣX_iY_i

Interpretation of Coefficients

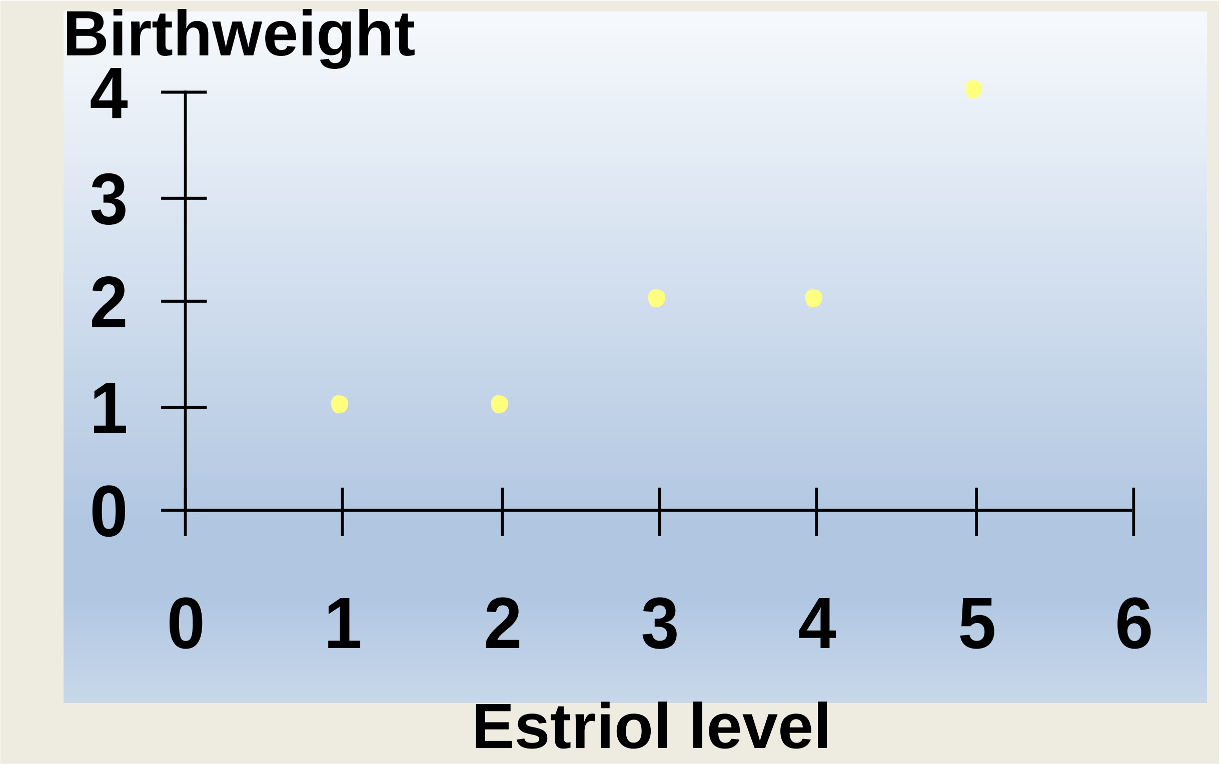
- 1. Slope (β_1)
  - Estimated *Y* Changes by β_1 for Each 1 Unit Increase in *X*
    - If β_1 = 2, then *Y* Is Expected to Increase by 2 for Each 1 Unit Increase in *X*
- 2. Y-Intercept (β_0)
  - Average Value of *Y* When *X* = 0
    - If β_0 = 4, then Average *Y* Is Expected to Be 4 When *X* Is 0

Parameter Estimation Example

- Obstetrics: What is the **relationship** between Mother's Estriol level & Birthweight using the following data?

Estriol (mg/24h)	**Birthweight** (g/1000)
1	1
2	1
3	2
4	2
5	4

Scatterplot
Birthweight vs. Estriol level



Parameter Estimation Solution Table

X_i	Y_i	X_i^2	Y_i^2	X_iY_i
1	1	1	1	1
2	1	4	1	2
3	2	9	4	6
4	2	16	4	8
5	4	25	16	20
15	10	55	26	37

Parameter Estimation Solution

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}} = \frac{37 - \frac{(15)(10)}{5}}{55 - \frac{(15)^2}{5}} = 0.70$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = 2 - (0.70)(3) = -0.10$$

How to estimate parameters

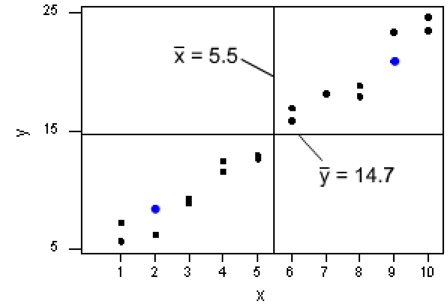
We minimize the equation for the sum of the squared prediction errors:

$$Q = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

(that is, take the derivative with respect to b_0 and b_1, set to 0, and solve for b_0 and b_1) and get the "**least squares estimates**" for b_0 and b_1:

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

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



the least squares line passes through the point $(\bar{x}, \bar{y})$, since when $x = \bar{x}$, then $y = b_0 + b_1 \bar{x} = \bar{y} - b_1 \bar{x} + b_1 \bar{x} = \bar{y}$.

Estimating the intercept and slope: least squares estimation

** Least Squares Estimation

A little calculus….

What are we trying to estimate? β, the slope, from

What's the constraint? We are trying to minimize the squared distance (hence the "least squares") between the observations themselves and the predicted values , or (<u>also called the "residuals", or left-over unexplained variability</u>)

Differencei = *yi* – (*βx* + *α*)   Differencei^2 = (*yi* – (*βx* + *α*)) ^{2}

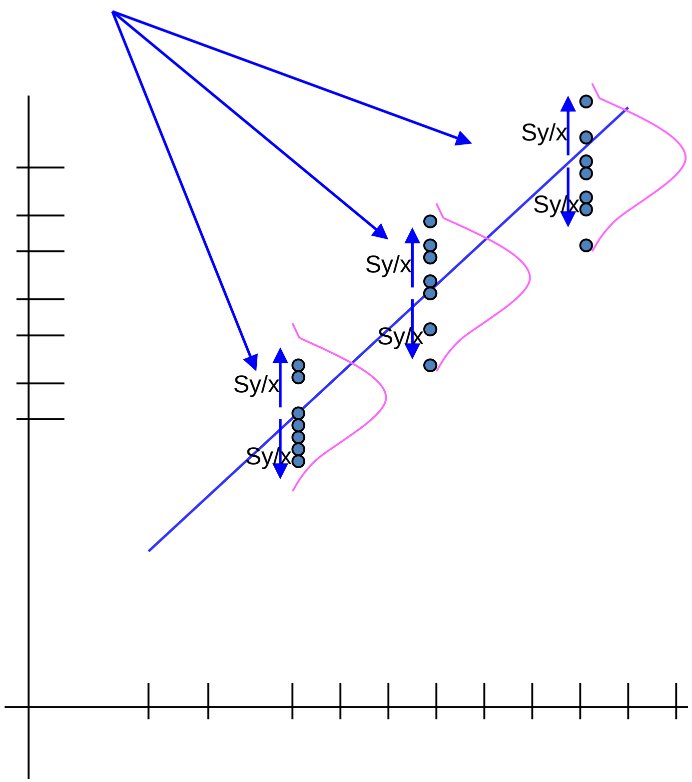
Find the *β* that gives the minimum sum of the squared differences. How do you maximize a function? Take the derivative; set it equal to zero; and solve. Typical max/min problem from calculus….

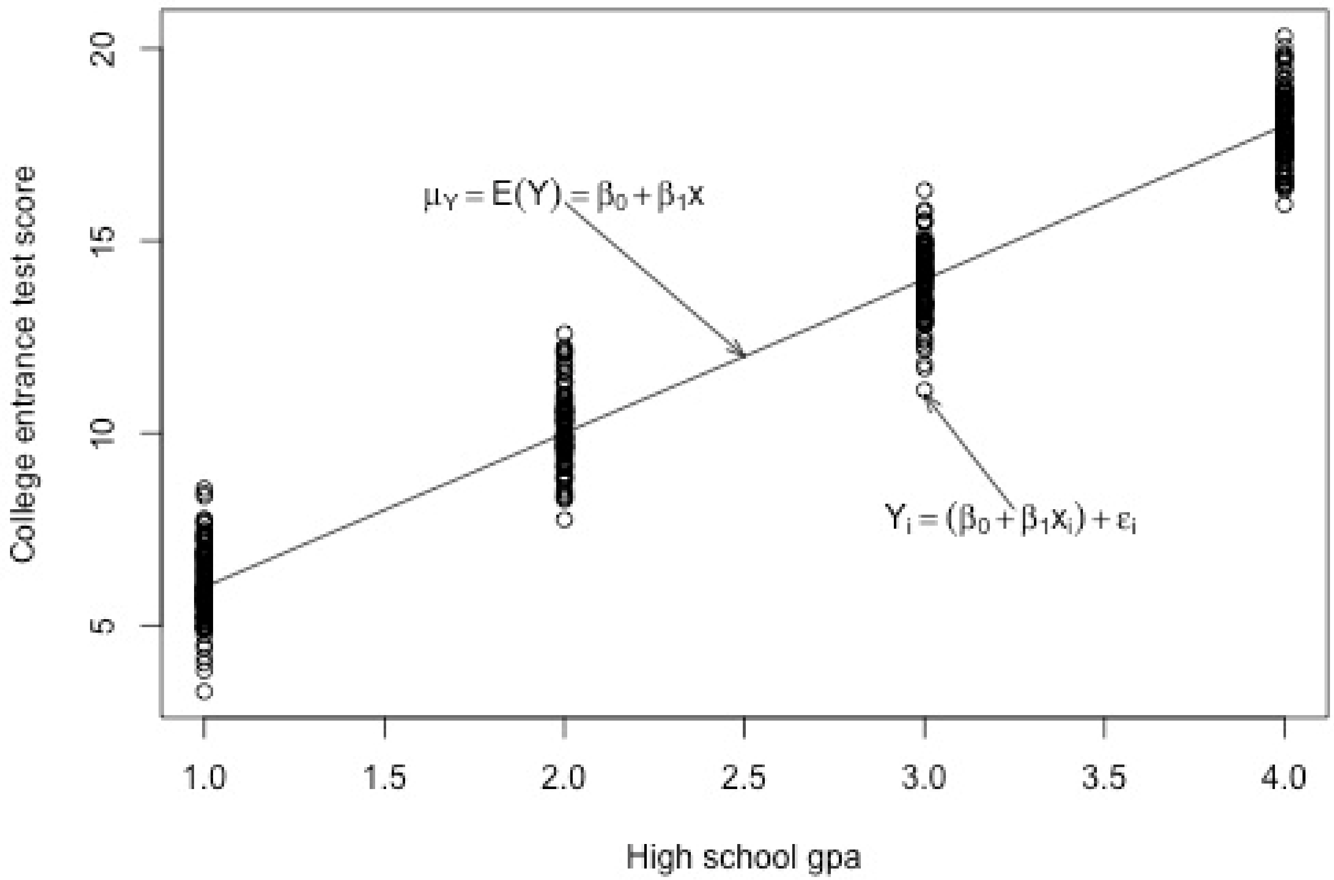
$$\frac{d}{d\beta}\sum_{i=1}^{n}(y_i-(\beta x_i+\alpha))^2 = 2(\sum_{i=1}^{n}(y_i-\beta x_i-\alpha)(-x_i))$$

$$2(\sum_{i=1}^{n}(-y_i x_i+\beta x_i{}^2+\alpha x_i)) = 0...$$

F ..

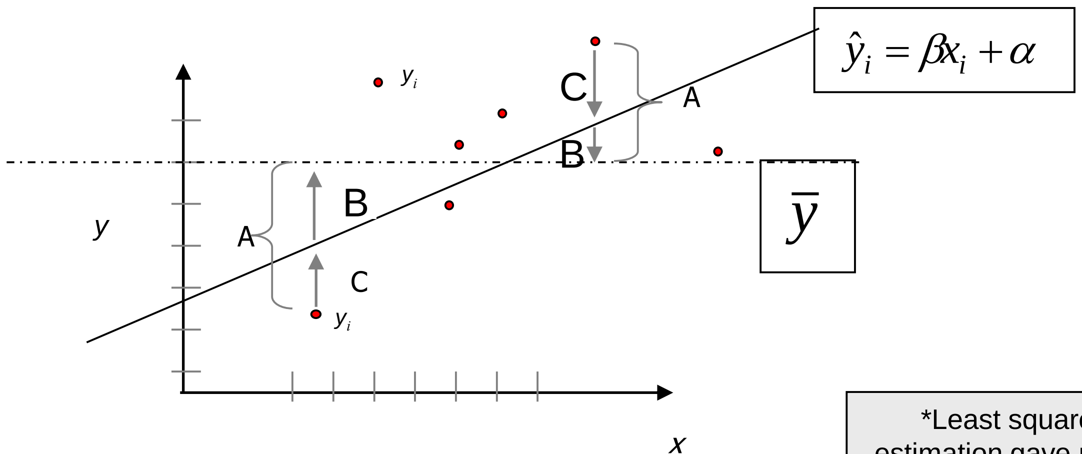
The standard error of Y given X is the average variability around the regression line at any given value of X. It is assumed to be equal at all values of X.





College entrance test score
High school gpa
$\mu_Y = E(Y) = \beta_0 + \beta_1 x$
$Y_i = (\beta_0 + \beta_1 x_i) + \varepsilon_i$
5
10
15
20
1.0
1.5
2.0
2.5
3.0
3.5
4.0

Regression Picture



*Least squares estimation gave us the line (β) that minimized C^2

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

R^2=SSreg/SStotal

A^2	B^2	C^2
SS_{total}	SS_{reg}	$SS_{residual}$
Total squared distance of observations from naïve mean of y	Distance from regression line to naïve mean of y	Variance around the regression line
Total variation	Variability due to x (regression)	Additional variability not explained by x—what least squares method aims to minimize

Regression Line

- If the scatter plot of our sample data suggests a linear relationship between two variables i.e.

  $$y = \beta_0 + \beta_1 x$$

  we can summarize the relationship by drawing a straight line on the plot.
- Least squares method give us the "best" estimated line for our set of sample data.

Regression Line

- We will write an estimated regression line based on sample data as

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

- The method of least squares chooses the values for b_0, and b_1 to minimize the sum of squared errors

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y - b_0 - b_1 x\right)^2$$

Regression Line

- Using calculus, we obtain estimating formulas:

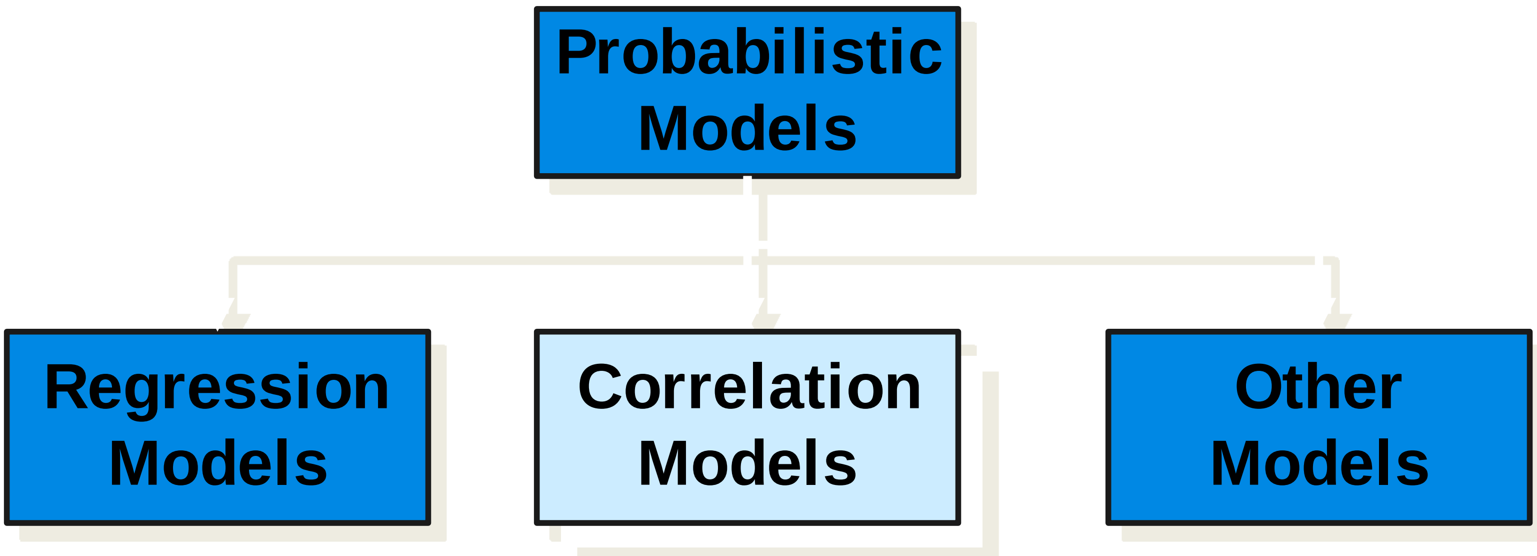
$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

or

$$b_1 = r\frac{S_y}{S_x}$$

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

Types of Probabilistic Models



Correlation vs. regression

- Both variables are treated the same in correlation; in regression there is a predictor and a response
- In regression the x variable is assumed non-random or measured without error
- Correlation is used in looking for relationships, regression for prediction

Correlation Models

- 1. Answer '**How Strong** Is the Linear Relationship Between 2 Variables?'
- 2. Coefficient of Correlation Used
  - Population Correlation Coefficient Denoted ρ (Rho)
  - Values Range from -1 to +1
  - Measures Degree of Association
- 3. Used Mainly for Understanding

Covariance

$$\operatorname{cov}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{n-1}$$

Interpreting Covariance

cov(X,Y) > 0 $\rightarrow$ X and Y are positively correlated

cov(X,Y) < 0 $\rightarrow$ X and Y are inversely correlated

cov(X,Y) = 0 $\rightarrow$ X and Y are independent

Correlation coefficient

- Pearson's Correlation Coefficient is standardized covariance (unitless):

$$r = \frac{covariance(x, y)}{\sqrt{\text{var } x}\sqrt{\text{var } y}}$$

Correlation

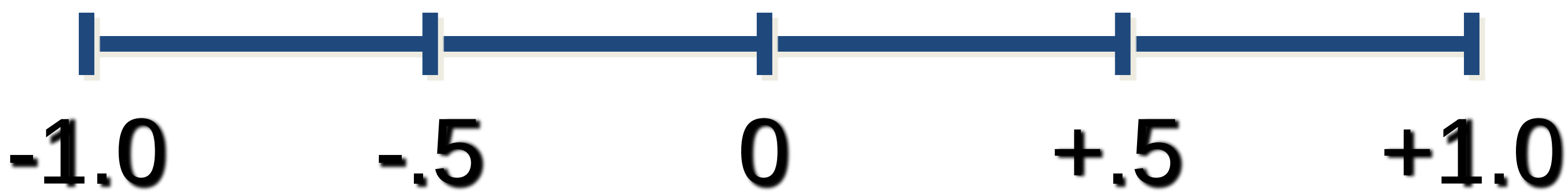
- Measures the relative strength of the *linear* relationship between two variables
- Unit-less
- Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

Sample Coefficient of Correlation

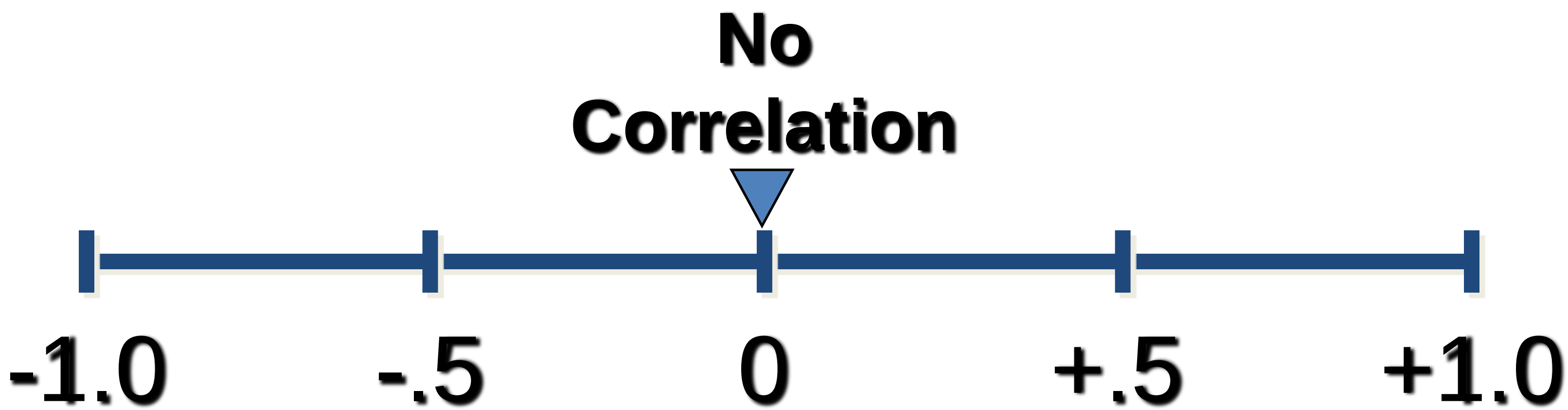
- 1. Pearson Product Moment Coefficient of Correlation between x and y:

$$r = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} \cdot \sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}} = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

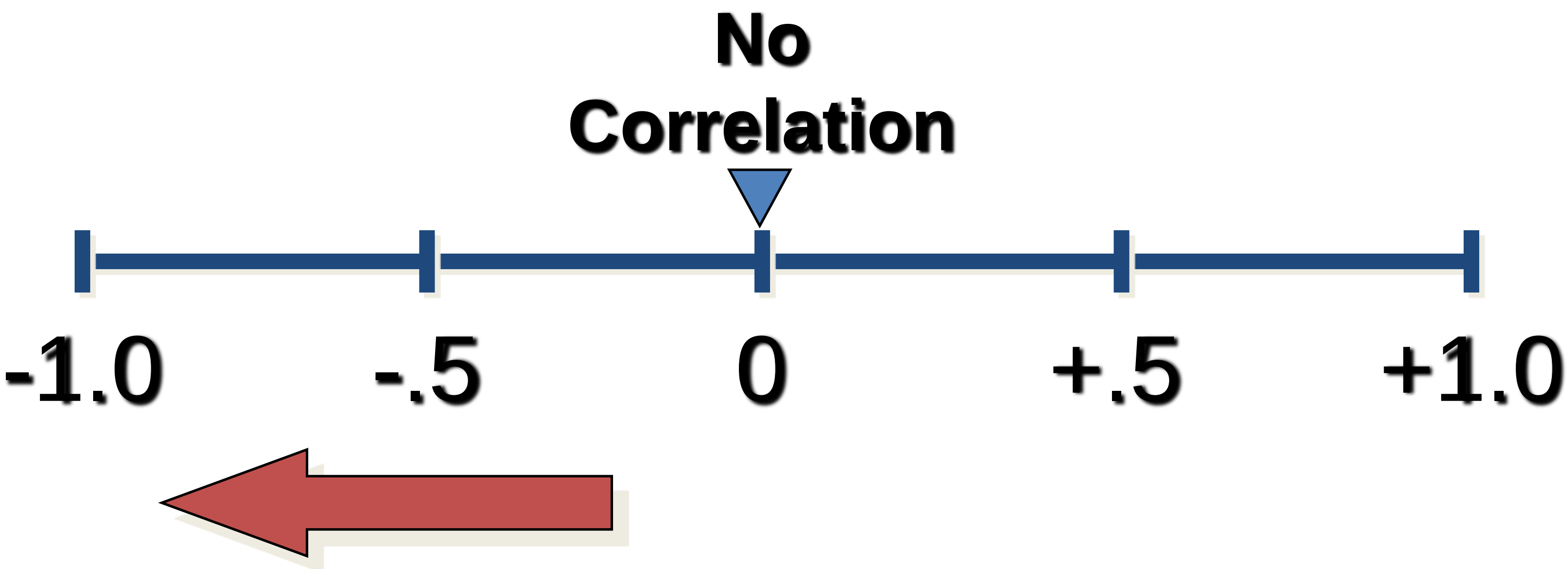
Coefficient of Correlation Values



Coefficient of Correlation Values

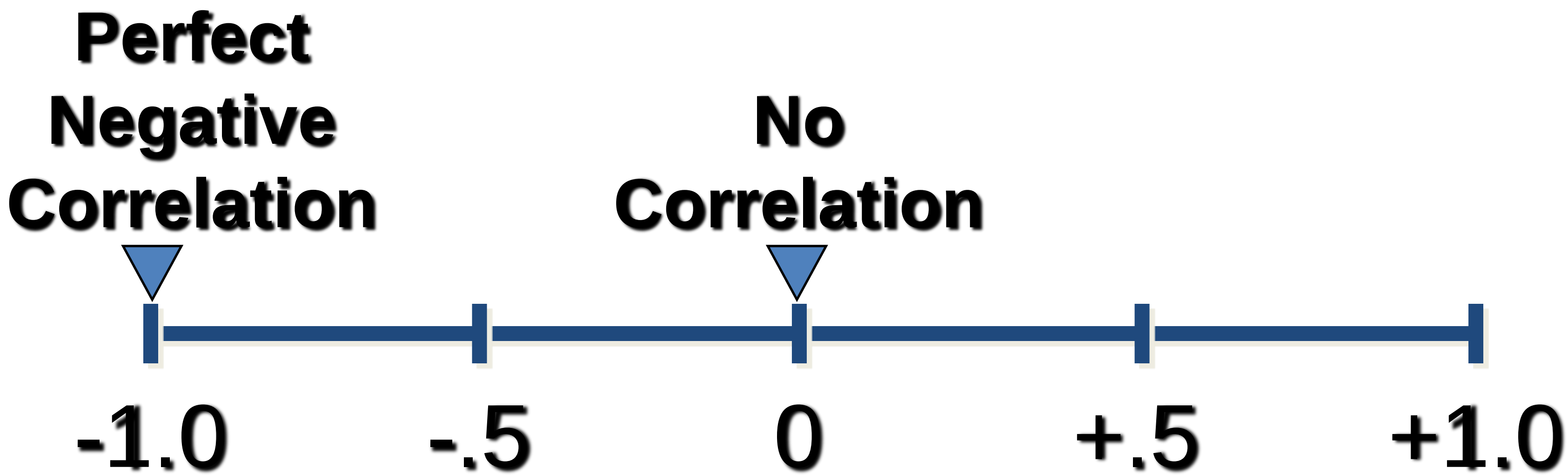


Coefficient of Correlation Values

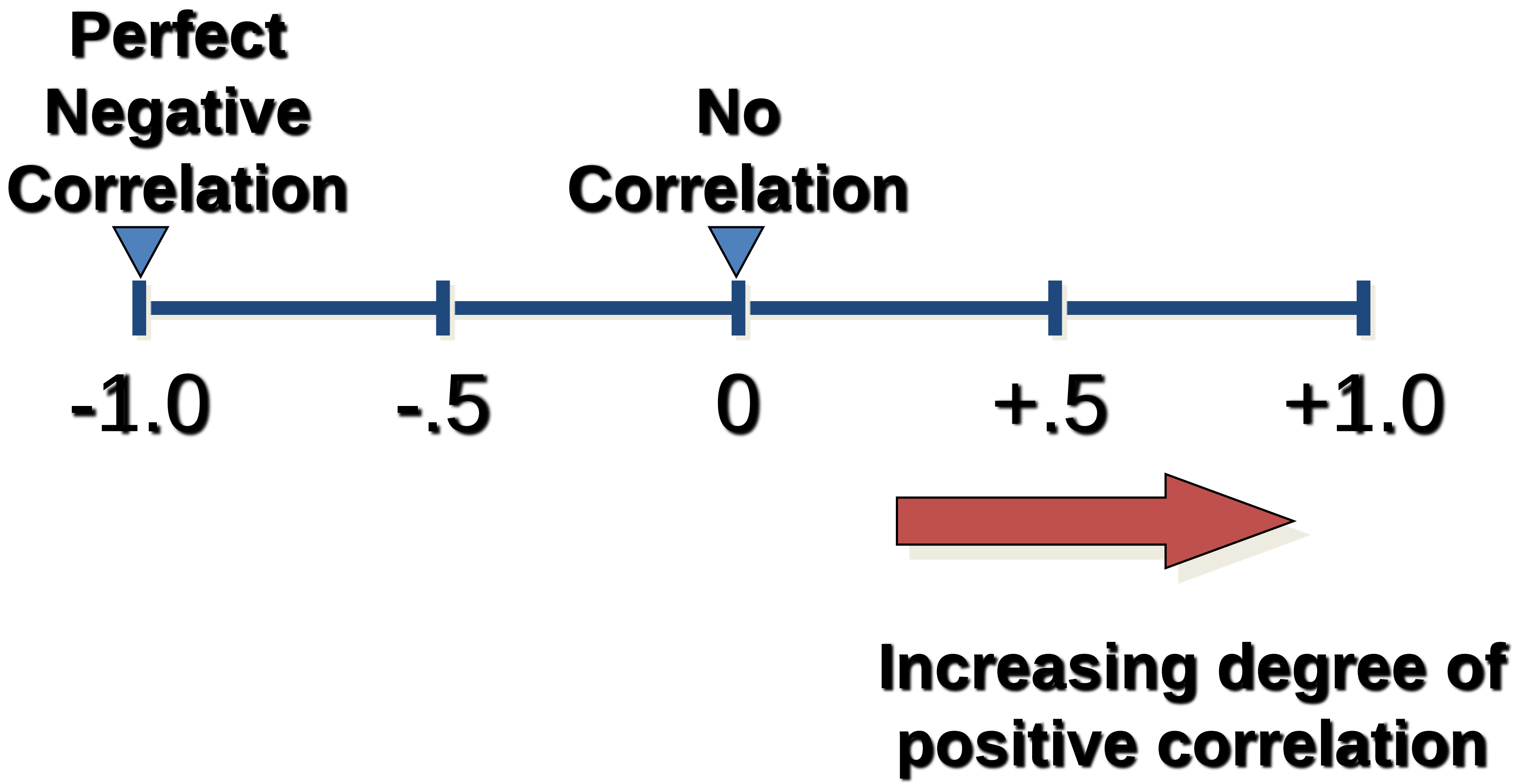


Increasing degree of negative correlation

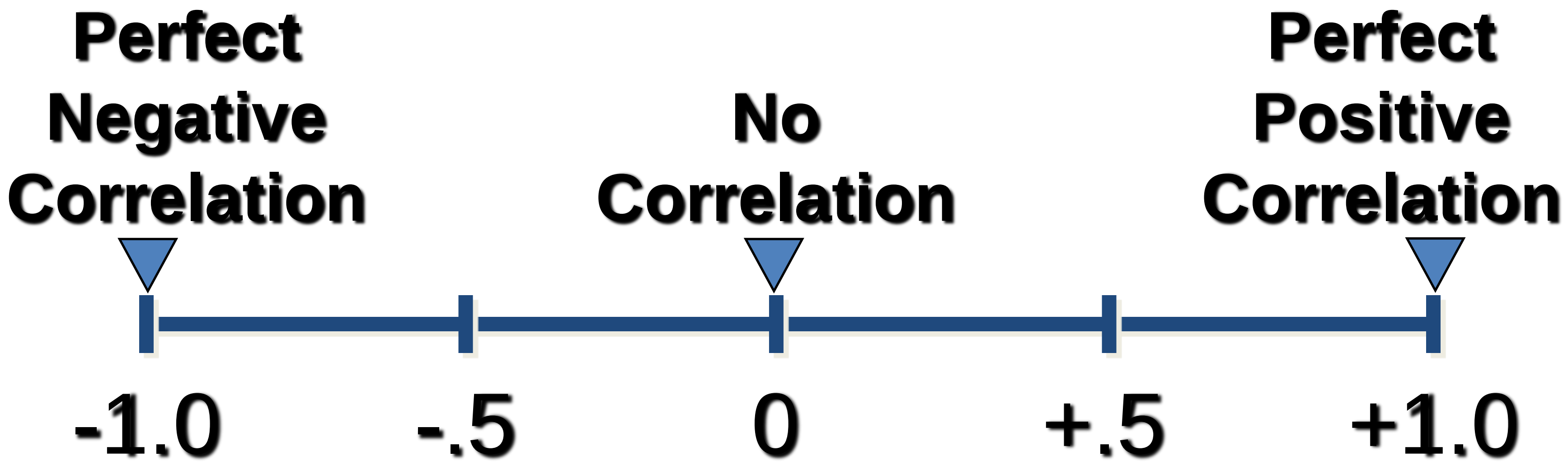
Coefficient of Correlation Values



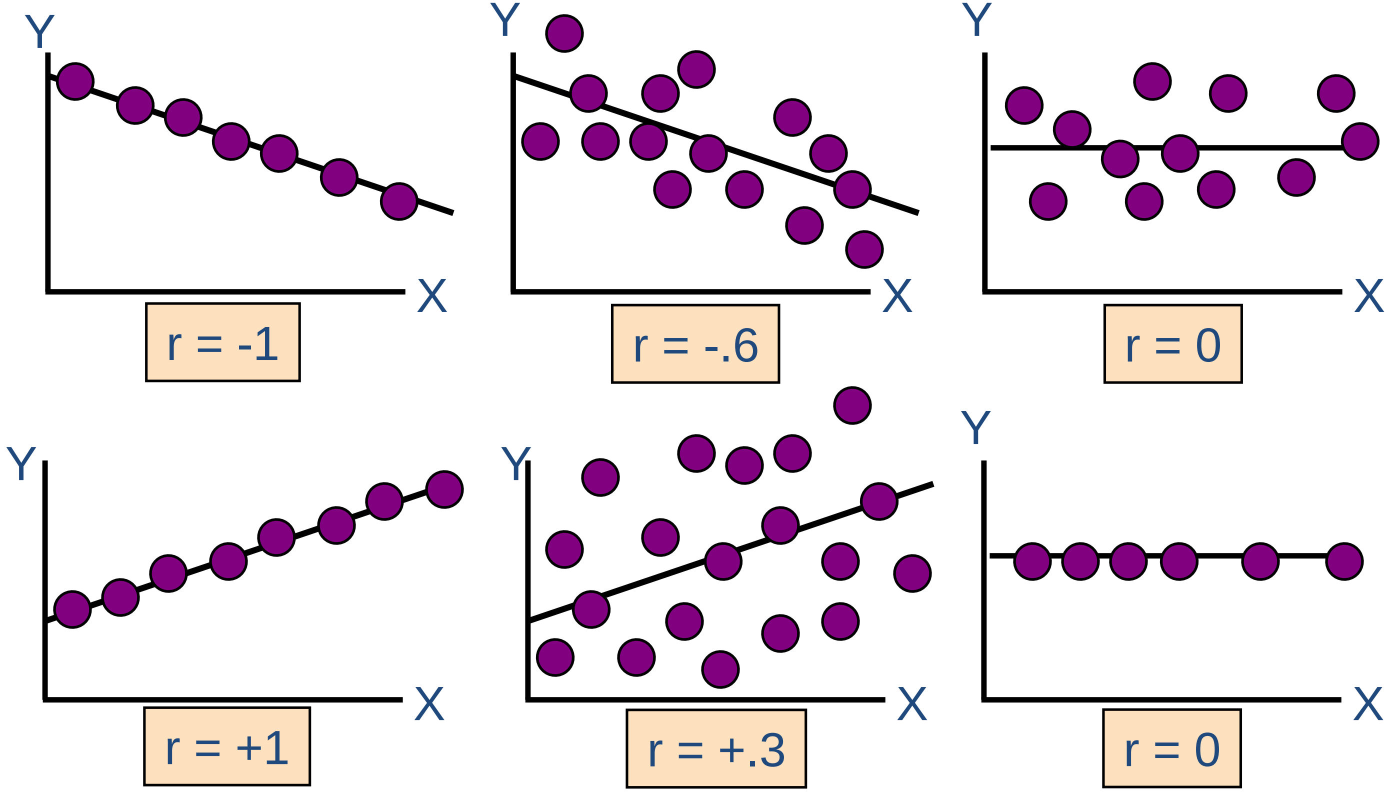
Coefficient of Correlation Values



Coefficient of Correlation Values

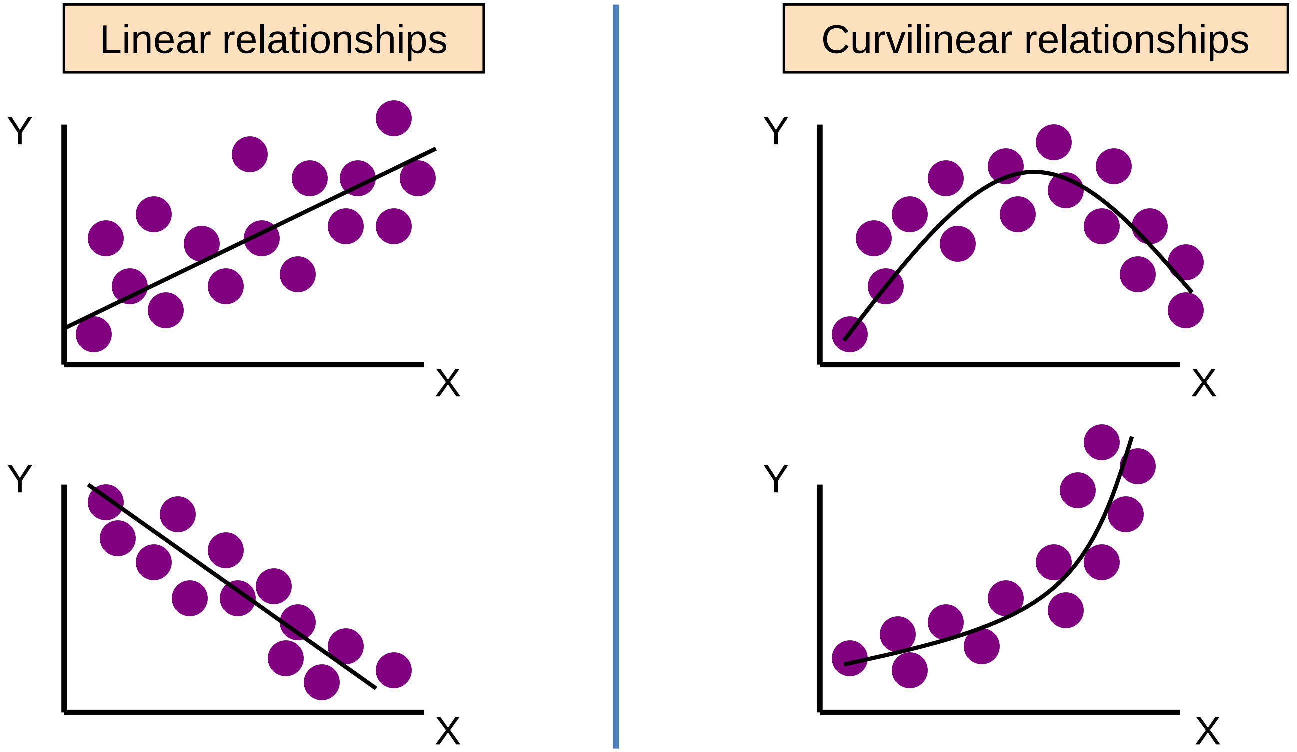


Scatter Plots of Data with Various Correlation Coefficients



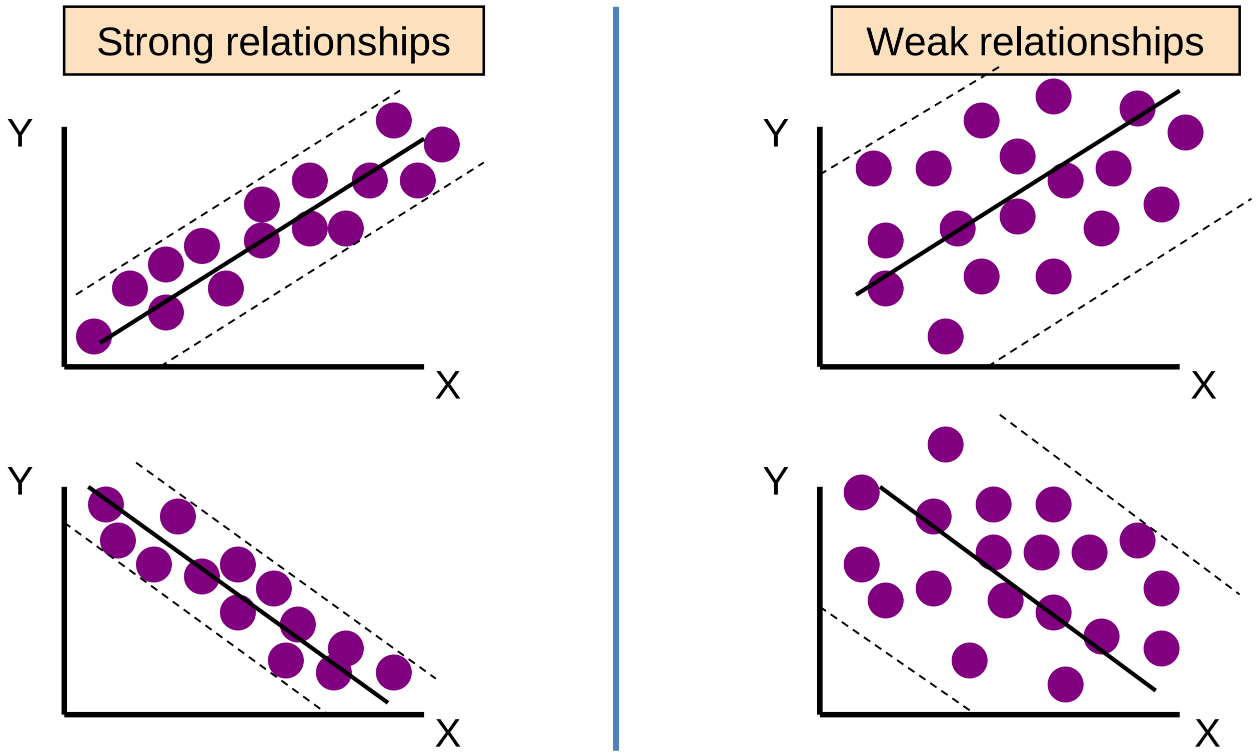
- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation



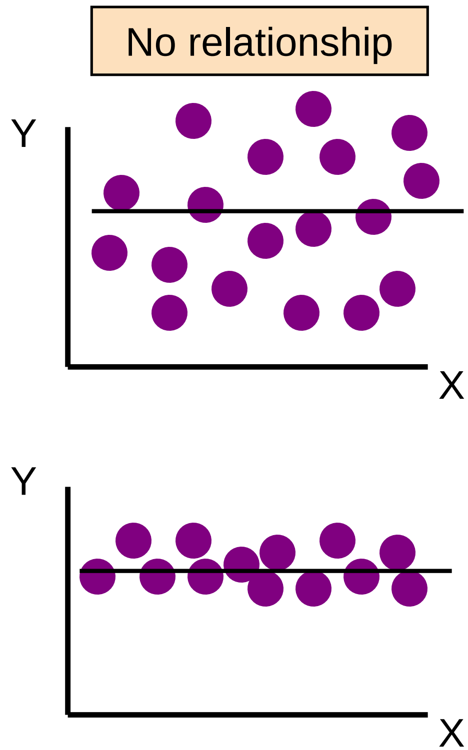
▪Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation



- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation



- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Calculating by hand…

$$\hat{r} = \frac{\mathrm{cov}ariance(x, y)}{\sqrt{\mathrm{var}\ x}\sqrt{\mathrm{var}\ y}} = \frac{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}}$$

Simpler calculation formula…

$$\hat{r} = \frac{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\cancel{n-1}}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\cancel{n-1}}}\sqrt{\dfrac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\cancel{n-1}}}} =$$

$$\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Numerator of covariance

$$\hat{r} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}}$$

Numerators of variance

Least Square estimation

Slope (beta coefficient) $\hat{\beta} = \dfrac{Cov(x, y)}{Var(x)}$

Intercept= $\text{Calculate}: \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$

Regression line always goes through the point: $(\bar{x}, \bar{y})$

Relationship with correlation

$$\hat{r} = \hat{\beta}\frac{SD_x}{SD_y}$$

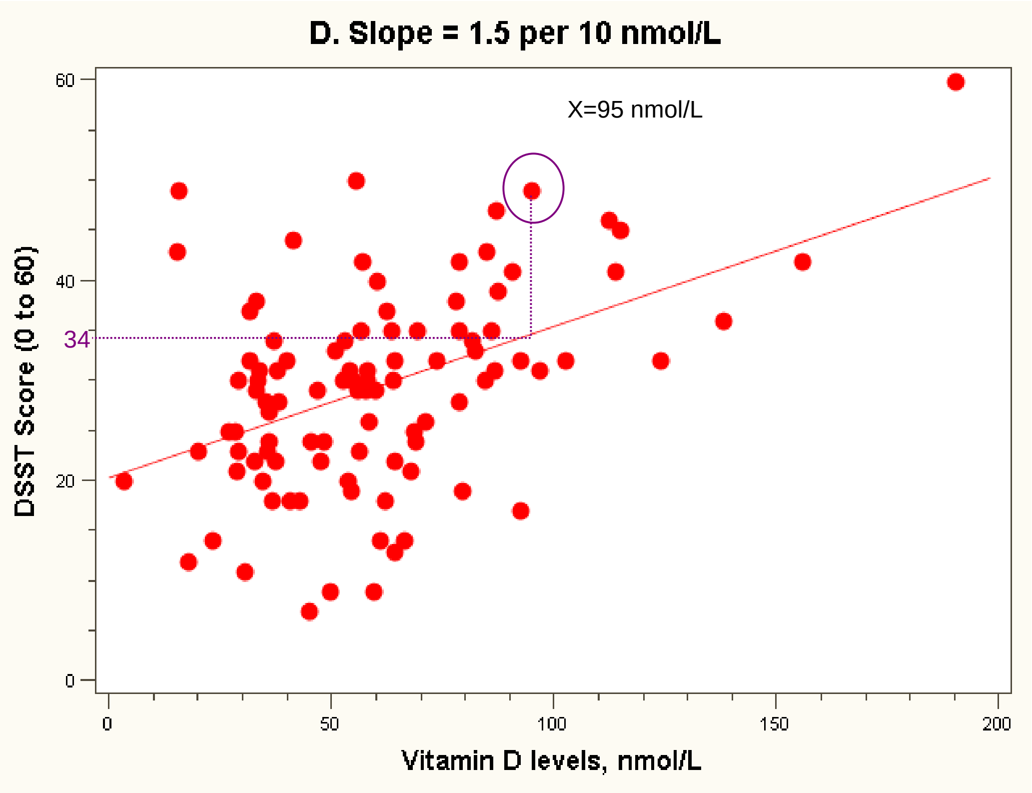
In correlation, the two variables are treated as equals. In regression, one variable is considered independent (=predictor) variable (X) and the other the dependent (=outcome) variable Y.

Residual Analysis: check assumptions

$$e_i = Y_i - \hat{Y}_i$$

- The residual for observation i, e_i, is the difference between its observed and predicted value
- Residuals are highly useful for studying whether a given regression model is appropriate for the data at hand.
- Check the assumptions of regression by examining the residuals
  - Examine for linearity assumption
  - Examine for constant variance for all levels of X (homoscedasticity)
  - Evaluate normal distribution assumption
  - Evaluate independence assumption
- Graphical Analysis of Residuals
  - Can plot residuals vs. X

Residual = observed - predicted

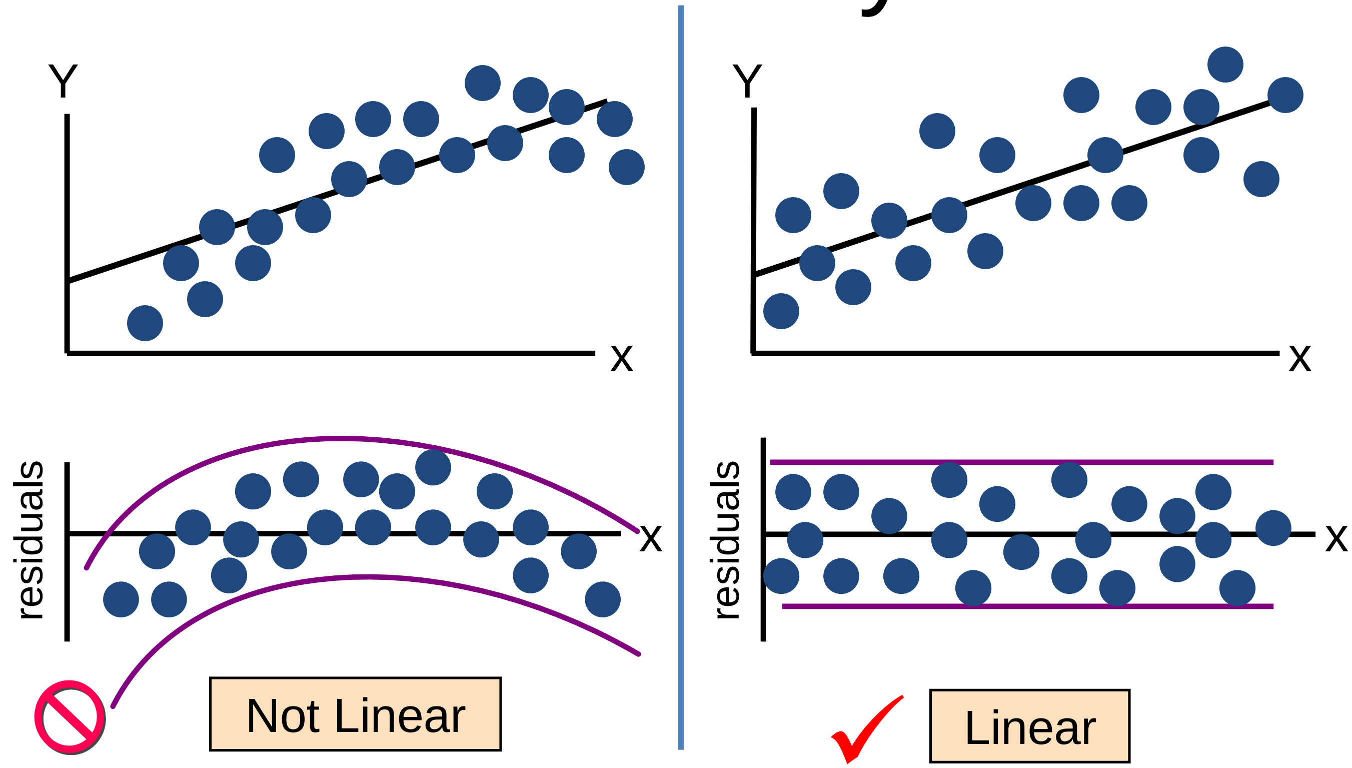


$$y_i = 48$$

$$\hat{y}_i = 34$$

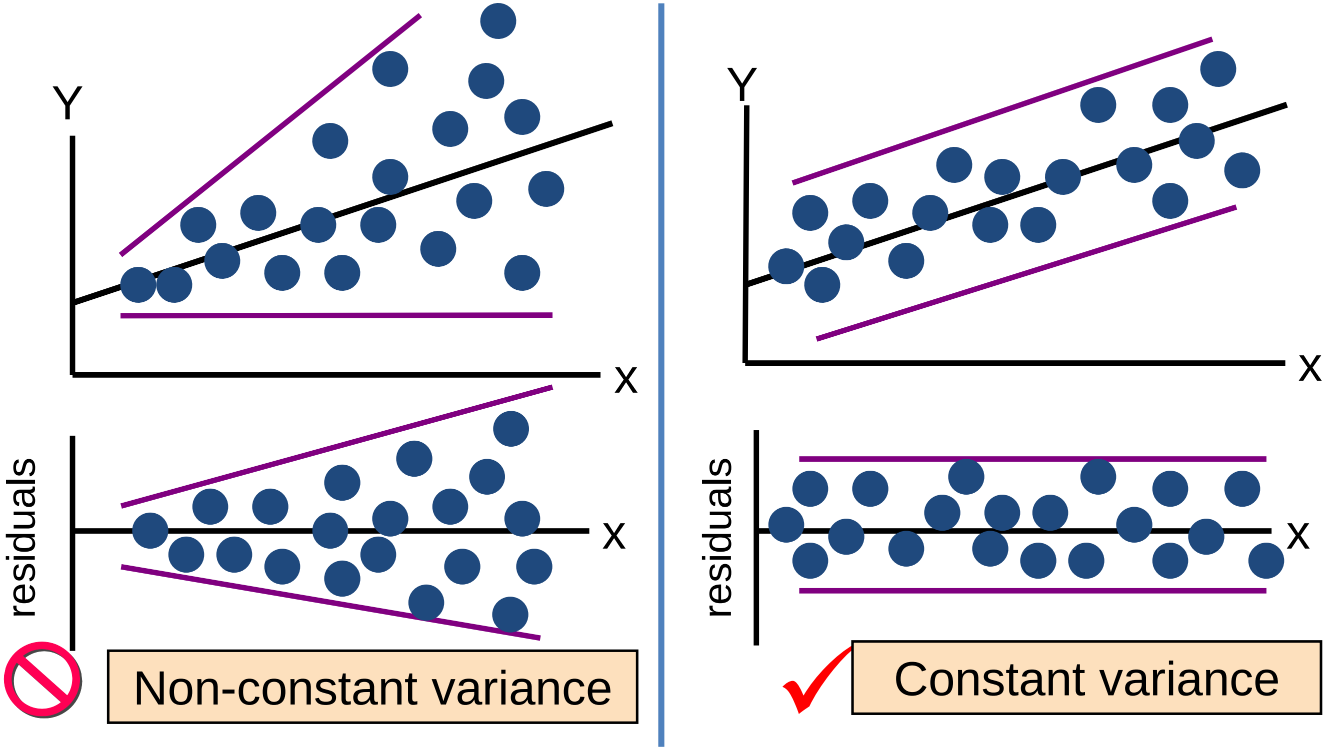
$$y_i - \hat{y}_i = 14$$

Residual Analysis for Linearity



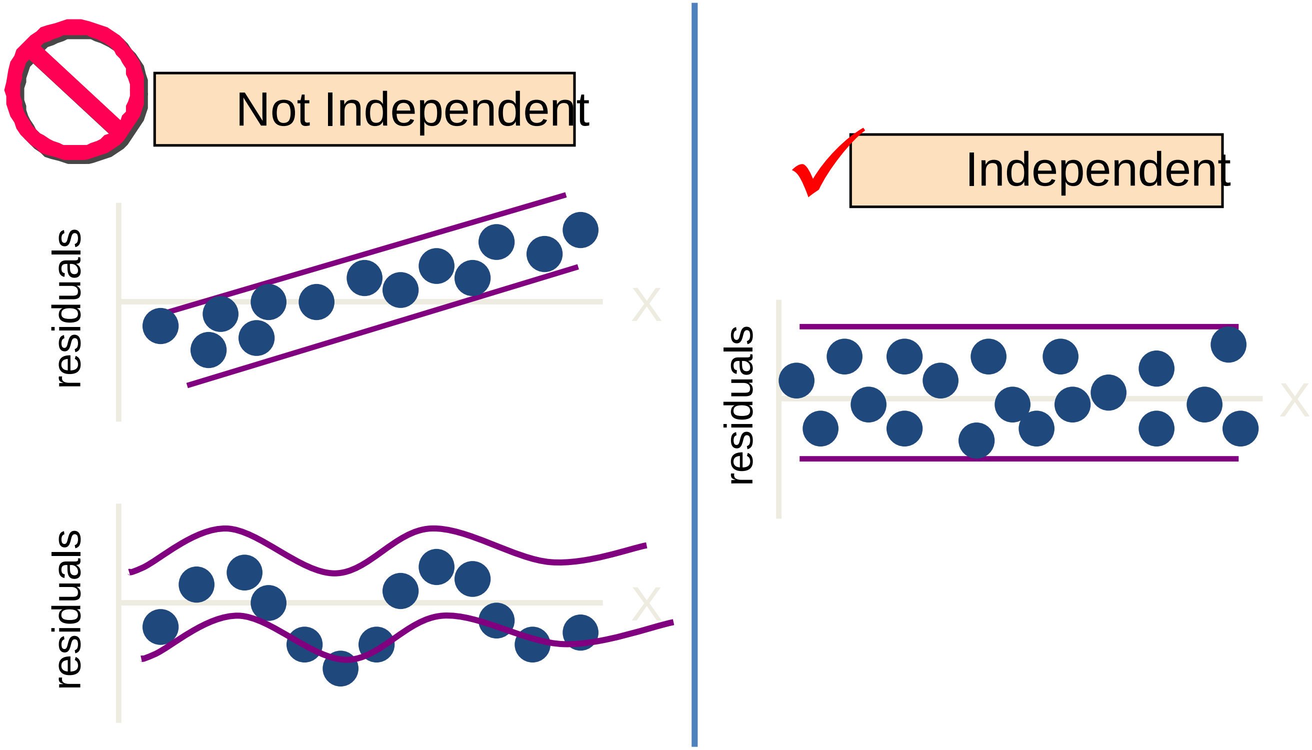
▪Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Residual Analysis for Homoscedasticity



▪Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Residual Analysis for Independence



▪Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Example: weekly advertising expenditure

y	x	y-hat	Residual (e)
1250	41	1270.8	-20.8
1380	54	1411.2	-31.2
1425	63	1508.4	-83.4
1425	54	1411.2	13.8
1450	48	1346.4	103.6
1300	46	1324.8	-24.8
1400	62	1497.6	-97.6
1510	61	1486.8	23.2
1575	64	1519.2	55.8
1650	71	1594.8	55.2

Estimation of the variance of the error terms, σ^2

- The variance σ^2 of the error terms ε_i in the regression model needs to be estimated for a variety of purposes.
  - It gives an indication of the variability of the probability distributions of y.
  - It is needed for making inference concerning regression function and the prediction of y.

Regression Standard Error

- To estimate σ we work with the variance and take the square root to obtain the standard deviation.
- For simple linear regression the estimate of σ^2 is the average squared residual.
- To estimate σ , use $s_{y.x}^{2} = \frac{1}{n-2}\sum e_i^2 = \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2$
- s estimates the standard deviation σ of the error term ε in the statistical model for simple linear regression.

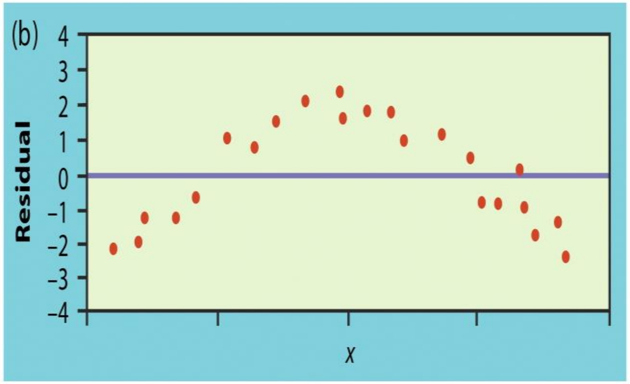
$$s_{y.x} = \sqrt{s_{y.x}^{2}}$$

Regression Standard Error

y	x	y-hat	Residual (e)	square(e)
1250	41	1270.8	-20.8	432.64
1380	54	1411.2	-31.2	973.44
1425	63	1508.4	-83.4	6955.56
1425	54	1411.2	13.8	190.44
1450	48	1346.4	103.6	10732.96
1300	46	1324.8	-24.8	615.04
1400	62	1497.6	-97.6	9525.76
1510	61	1486.8	23.2	538.24
1575	64	1519.2	55.8	3113.64
1650	71	1594.8	55.2	3047.04
y-hat = 828+10.8X			total	36124.76
			$S_{y.x}$	67.19818

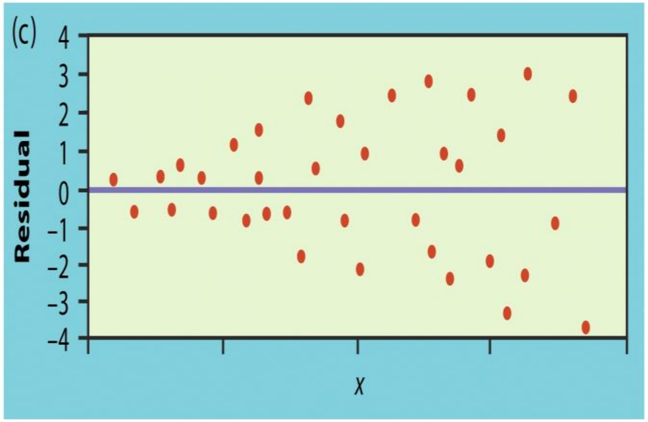
Residual plots

- The points in this residual plot have a curve pattern, so a straight line fits poorly



Residual plots

- The points in this plot show more spread for larger values of the explanatory variable *x*, so prediction will be less accurate when *x* is large.



Variable transformations

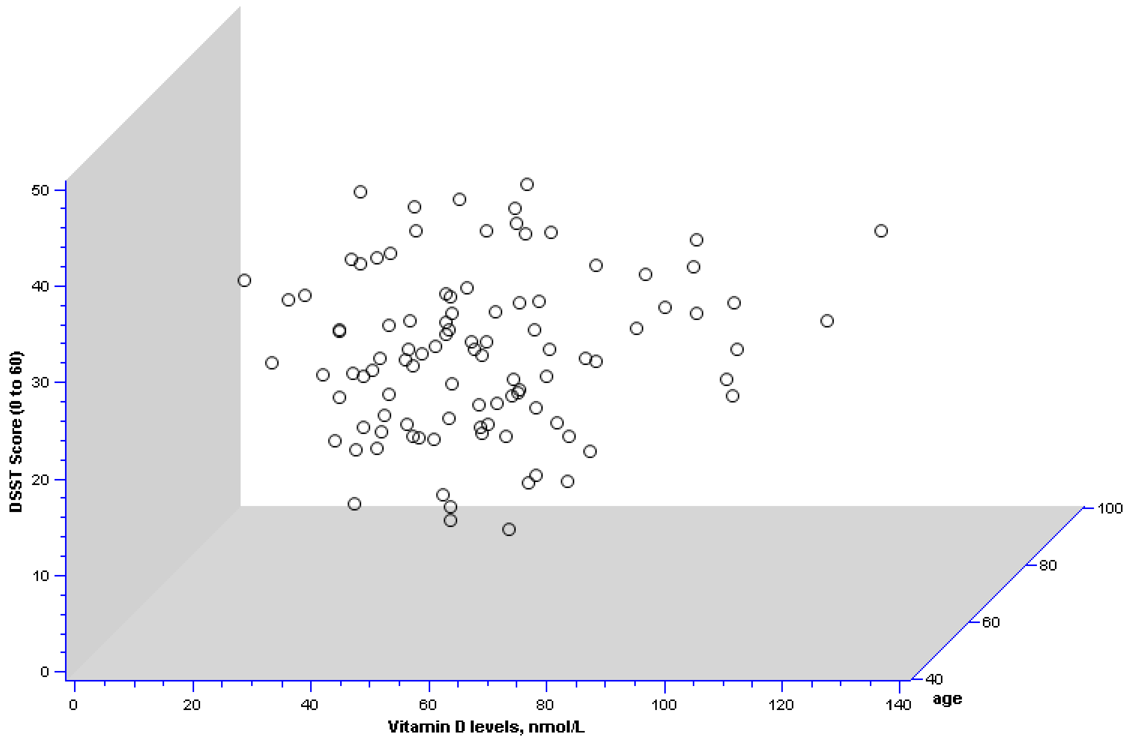
- If the residual plot suggests that the variance is not constant, a transformation can be used to stabilize the variance.
- If the residual plot suggests a non linear relationship between x and y, a transformation may reduce it to one that is approximately linear.
- Common linearizing transformations are:

$$\frac{1}{x}, \quad \log(x)$$

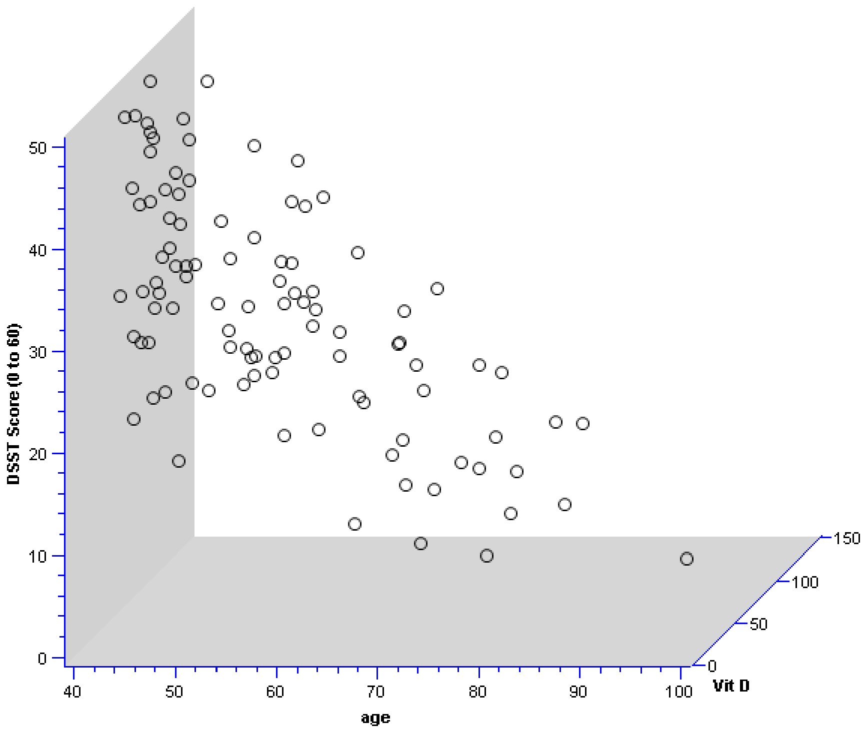
- Variance stabilizing transformations are:

$$\frac{1}{y}, \quad \log(y), \quad \sqrt{y}, \quad y^2$$

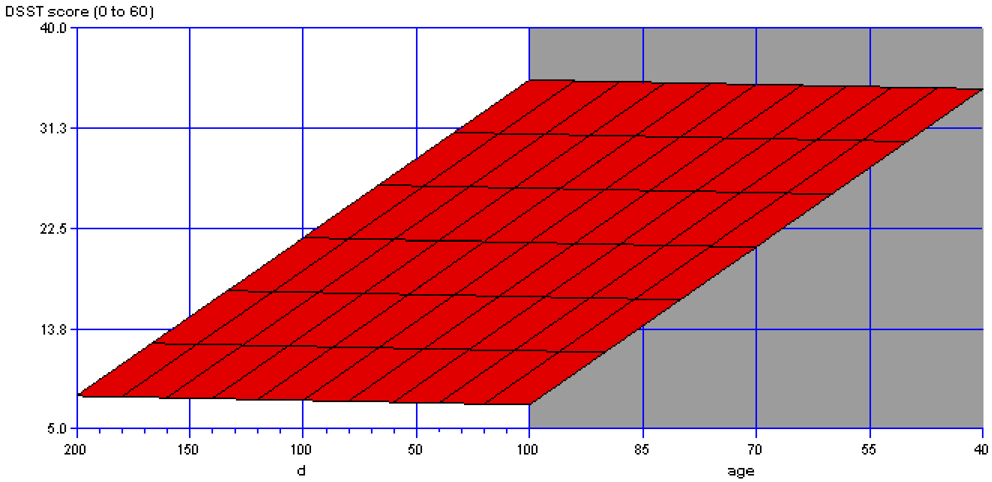
2 predictors: age and vit D…



Different 3D view…



Fit a plane rather than a line…



On the plane, the slope for vitamin D is the same at every age; thus, the slope for vitamin D represents the effect of vitamin D when age is held constant.