Data Analysis Course

Correlation and Regression(Version-1)

Venkat Reddy

Data Analysis Course

- Data analysis design document
- Introduction to statistical data analysis
- Descriptive statistics
- Data exploration, validation & sanitization
- Probability distributions examples and applications
- **Simple correlation and regression analysis**
- Multiple liner regression analysis
- Logistic regression analysis
- Testing of hypothesis
- Clustering and decision trees
- Time series analysis and forecasting
- Credit Risk Model building-1
- Credit Risk Model building-2

Note

- This presentation is just class notes. The course notes for Data Analysis Training is by written by me, as an aid for myself.
- The best way to treat this is as a high-level summary; the actual session went more in depth and contained other information.
- Most of this material was written as informal notes, not intended for publication
- Please send questions/comments/corrections to venkat@trenwiseanalytics.com or 21.venkat@gmail.com
- Please check my website for latest version of this document

-Venkat Reddy

Contents

What is need of correlation?

- What happens to Sweater sales with increase in temperature?
  - What is the strength of association between them?
- Ice-cream sales v.s temperature ?
  - What is the strength of association between them?
- Which one of these two is stronger? How to quantify the association?

What is Correlation

- It is a measure of association(linear association only)
- Formula for correlation coefficient

  r is the ratio of variance together and product of separate variances

  $r = cov(XY)/sd(x)*sd(y)$

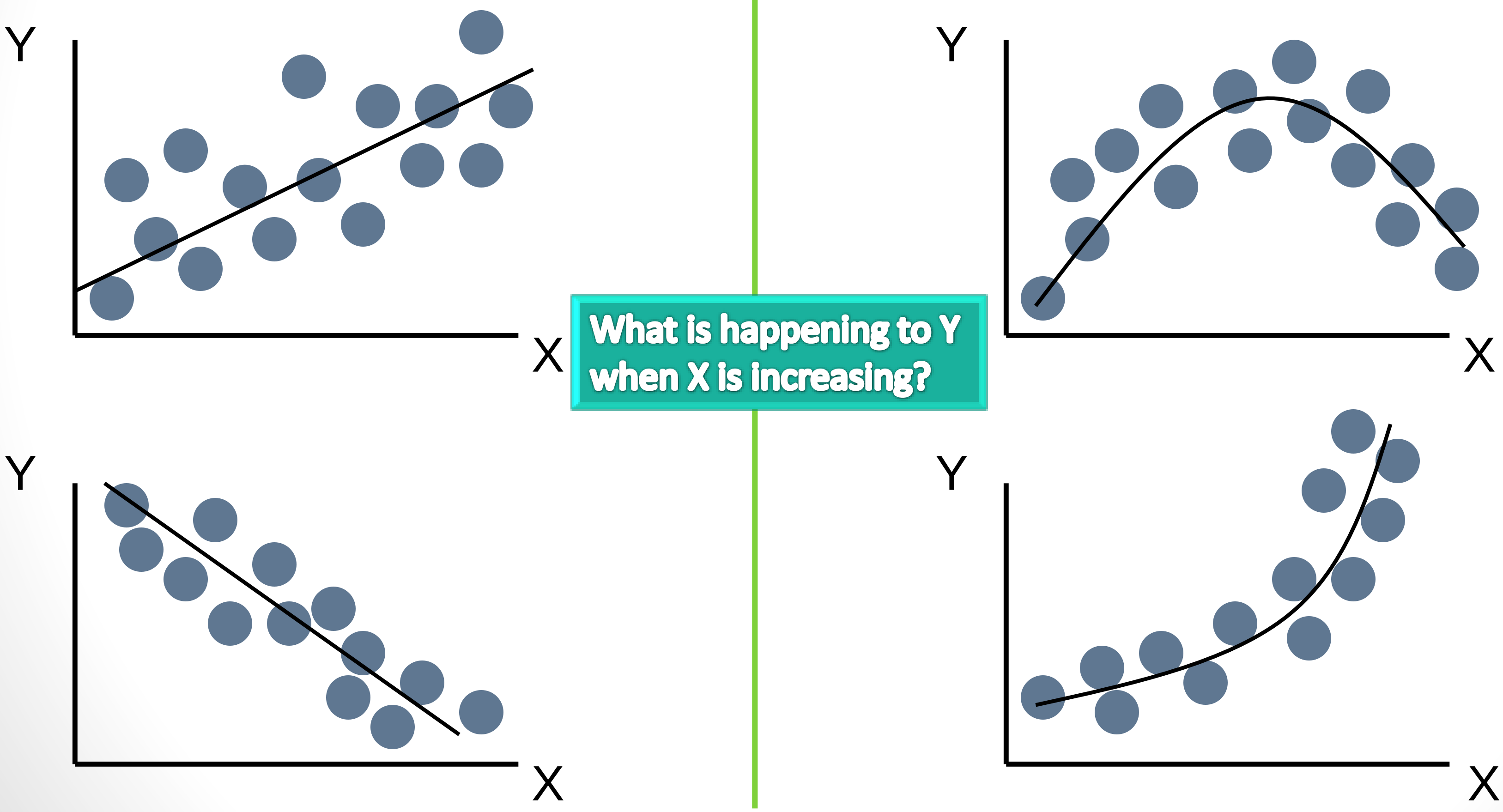
  $r = [n(\sum xy) - (\sum x)(\sum y)] / \{[n(\sum x^2) - (\sum x)^2][n(\sum y^2) - (\sum y)^2]\}^{0.5}$

Where *n* is the number of data pairs, x is the independent variable and y the dependent variable.

Type of relationship

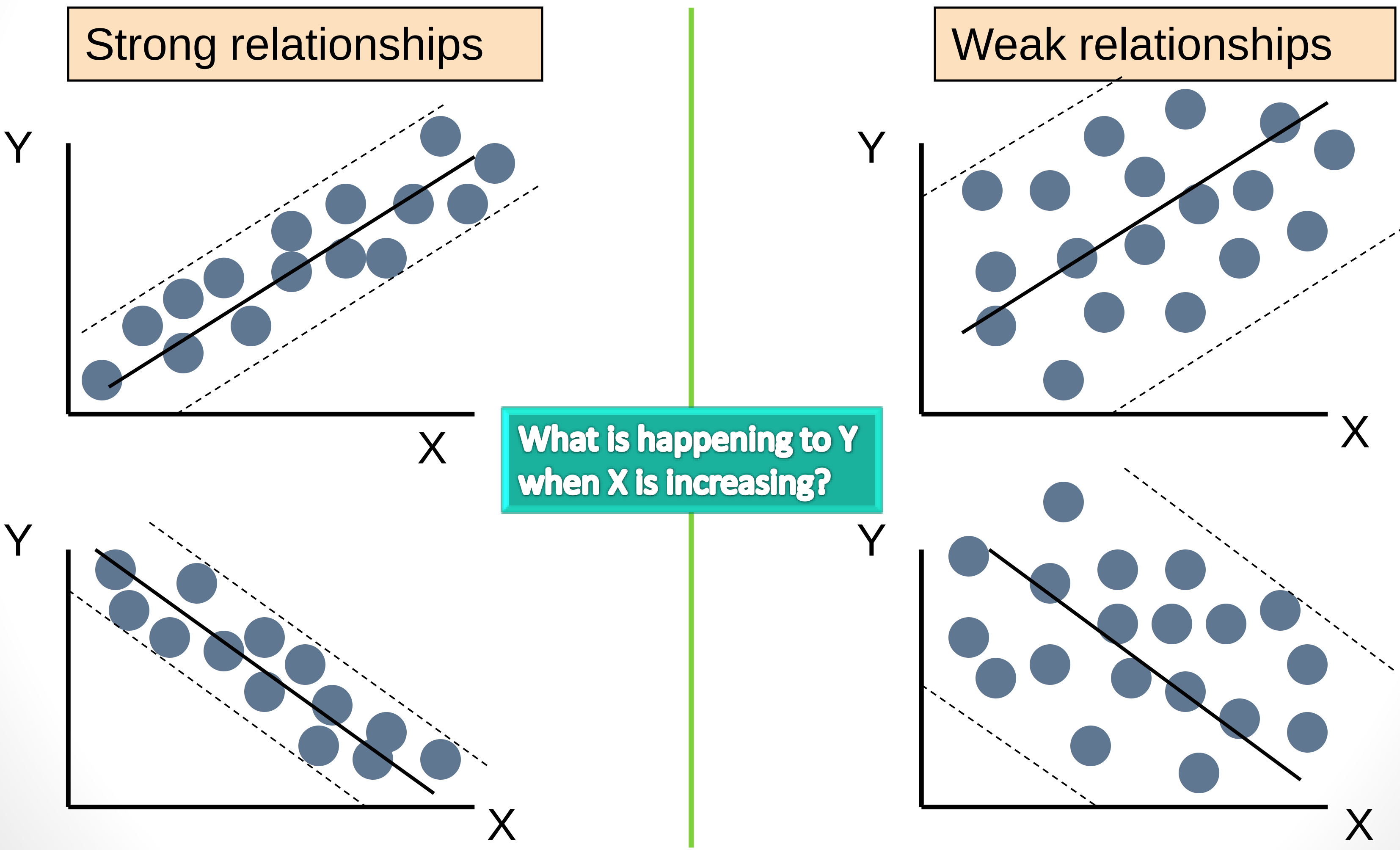
Linear relationships

Curvilinear relationships

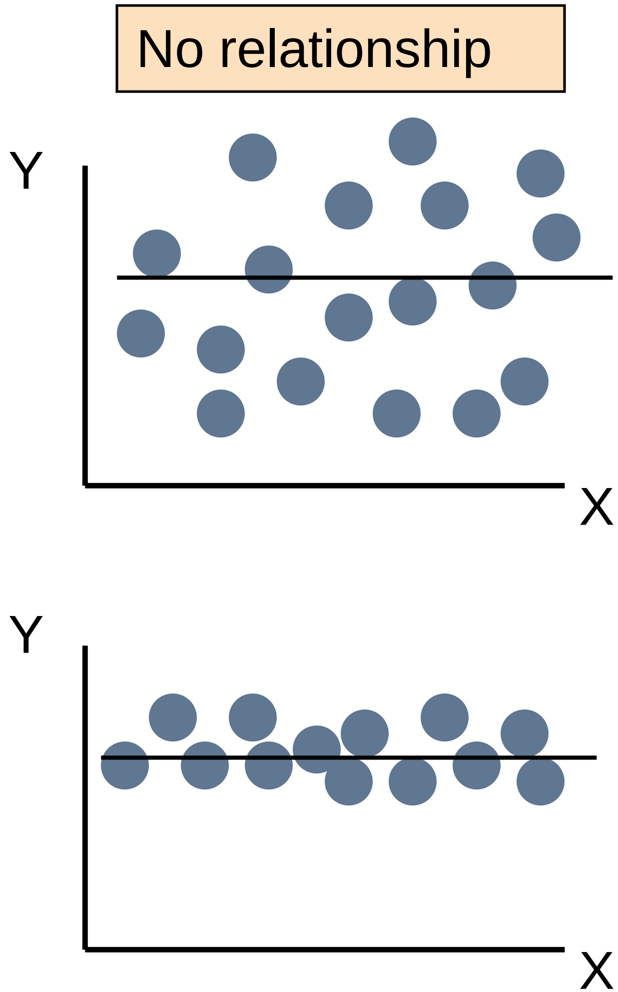


What is happening to Y when X is increasing?

Type of relationship



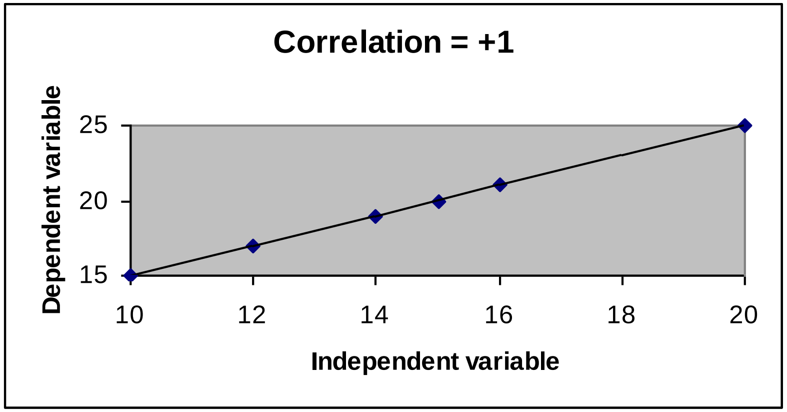
Type of relationship



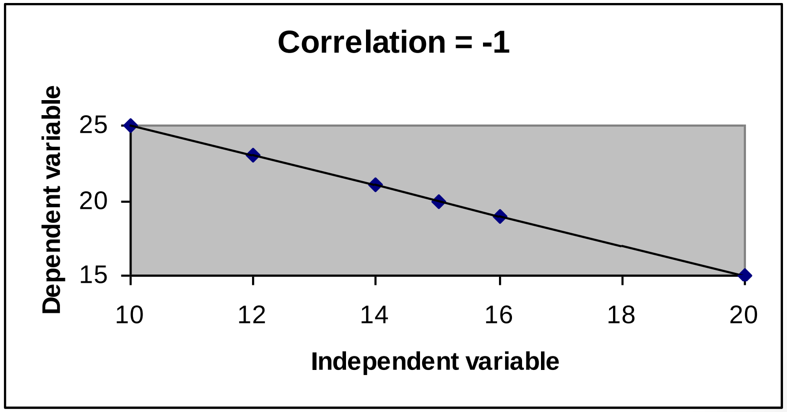
What is happening to Y when X is increasing?

Range of Correlation

- In case of exact positive linear relationship the value of r is ___.
- In case of a strong positive linear relationship, the value of *r* will be close to ___.

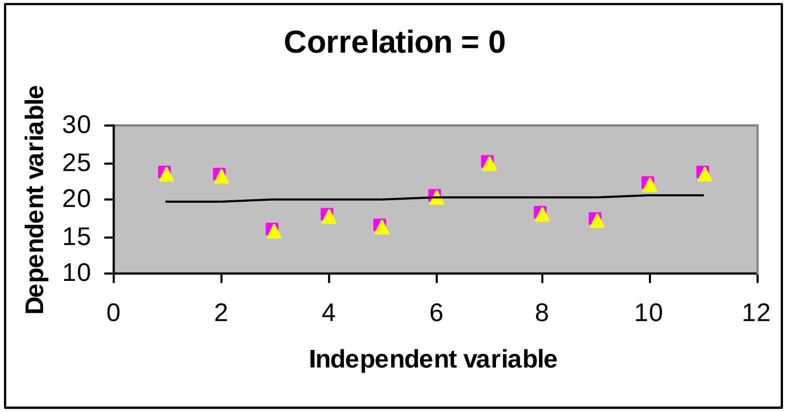


- In case of exact negative linear relationship the value of *r* is ___.
- In case of a strong negative linear relationship, the value of *r* will be close to ___.

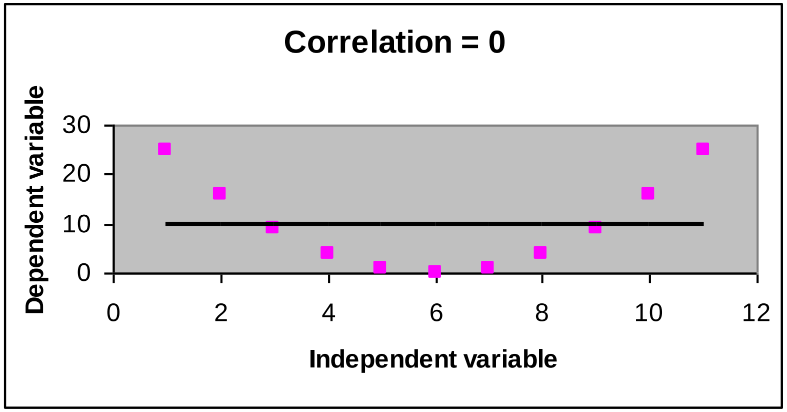


Range of Correlation

In case of a weak relationship the value of *r* will be close to ___.

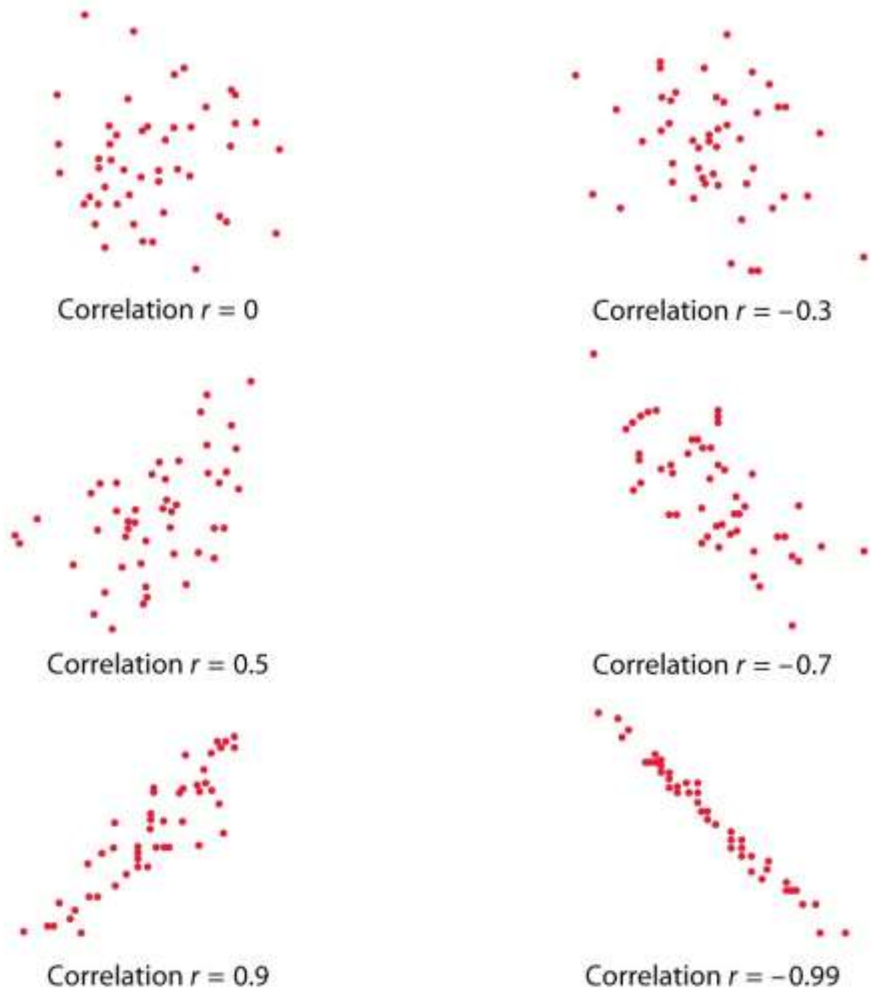


In case of nonlinear relationship the value of r will be close to ___.



Strength of Association

- Correlation 0 →No linear association
- Correlation 0 to 0.25 →Negligible positive association
- Correlation 0.25-0.5 → Weak positive association
- Correlation 0.5-0.75 →Moderate positive association
- Correlation >0.75 →Very Strong positive association
- What are the limits for negative correlation



Properties of Correlation

- $-1 \leq r \leq +1$
- r=0 represents no linear relationship between the two variables
- Correlation is unit free

Limitations:

- Though r measures how closely the two variables approximate a straight line, it does not validly measures the strength of nonlinear relationship
- When the sample size, n, is small we also have to be careful with the reliability of the correlation
- Outliers could have a marked effect on r

Lab

- Create these tables and find correlations

1

X	Y
-31	900
-25	625
-24	576
-19	361
-13	169
-6	36
-1	1
3	9
10	100
11	121
14	196
15	225
24	576
24	576
29	841

2

X	Y
1	2
2	4
3	3

X	Y
1	2
2	2.9
3	3

3

X	Y
10	14
17	25
22	23
21	31
24	29
34	60
25	19
31	35
45	45
33	38
60	50
46	56
47	45
48	70
50	750

Correlation - Limitations

1

X	Y
-31	900
-25	625
-24	576
-19	361
-13	169
-6	36
-1	1
3	9
10	100
11	121
14	196
15	225
24	576
24	576
29	841

r = -0.12

2

X	Y
1	2
2	4
3	3

r= 0.5

X	Y
1	2
2	2.9
3	3

r= 0.9

Correlation is a measure of linear relationship only

3

X	Y
10	14
17	25
22	23
21	31
24	29
34	60
25	19
31	35
45	45
33	38
60	50
46	56
47	45
48	70
50	750

r= 0.44

r= 0.86

- **Example1:** Y is both decreased and increased when X is increased. Correlation is -0.12, but this is not an appropriate measure of association
- **Example-2:** Correlation changed from 0.5 to 0.9 with a small change in the data. 'r' is not reliable when n is small
- **Example-3:** Correlation between X and Y is 0.44, correlation between X & Y is 0.86 if we exclude outlier

Correlation vs. Possible Relationships Between Variables

- **Direct cause and effect,** that is x cause y or water causes plant to grow.

- **Both cause and effect**, that y cause x or coffee consumption causes nervousness as well nervous people have more coffee.

- **Relationship caused by third variable;** Death due to drowning and soft drink consumption during summer. Both variables are related to heat and humidity (third variable). –This is dangerous (Why?)

- **Coincidental relationship;** Increase in the number of people exercising and increase in the number of people committing crimes. –This is even more dangerous (Why?)

- Correlation measures association and **not causation.**

Lab -Correlation

- Download stock price data from [here]
- Find the correlation between IBM open & Intel open price
- Can we apply correlation on this data? Draw a scatter plot between the stocks –What is the type of relationship
- When Intel stock open price increased
  - What happened to Microsoft open price?
  - What happened to IBM open price?
- Is there any correlation between closing price of three stocks?
- Correlation between the stocks with respect to change in price?

Correlation Practice

- Download Age vs Blood Pressure data

Age vs BP

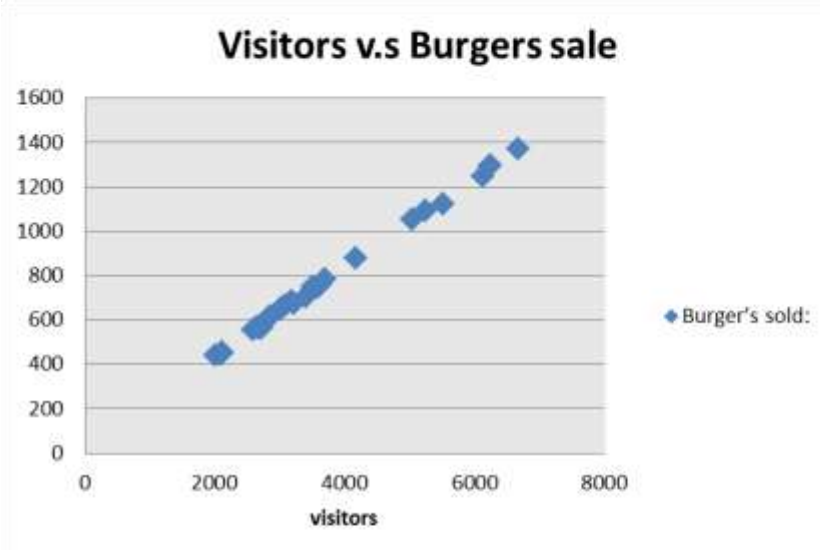
- Is there any association between age and blood pressure?
- Quantify the association between age and blood pressure.
- How strong is the association between age and blood pressure?
- What is your observation? As age increases does blood pressure increase or decrease?
- Are there any outliers? Is your measure of association reliable? Is it high/low due to outliers?
- What is the final verdict, in the given sample did you see a strong/moderate/no association between Age & BP

Regression

Why Regression?

Last 30 days data for a KFC shop in a Mall. Number of visitors v.s burgers sold (download it from here)

day	Mall visitors	Burger's sold:
1	2728	566
2	2098	444
3	2111	454
4	2009	440
5	3635	760
6	4171	881
7	5244	1091
8	3695	783
9	3088	666
10	2674	564
11	3591	750
12	3013	650
13	5045	1054
14	6118	1245
15	2851	616
16	2698	564
17	3015	652
18	3409	704
19	3179	683
20	5510	1125
21	6232	1294
22	3211	676
23	2582	557
24	2710	568



Number of visitors are expected to be 6500 tomorrow. How many burgers will be sold?

Data Analysis Course
Venkat Reddy

Regression

- Regression analysis is used to predict the value of one variable (the ***dependent variable***) on the basis of other variables (the ***independent variables***).
- 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.
- Dependent variable: denoted **Y**
- Independent variables: denoted $\mathbf{X_1, X_2, \ldots, X_k}$

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

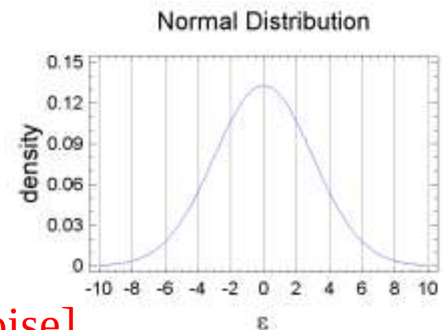
- Above model is referred to as simple linear regression. We would be interested in estimating β_0 and β_1 from the data we collect.

Simple Linear Regression Analysis

- If you know something about X, this knowledge helps you predict something about Y.

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

- Variables:
  - X = Independent Variable (we provide this)
  - Y = Dependent Variable (we observe this)
- Parameters:
  - β_0 = Y-Intercept
  - β_1 = Slope
  - ε ~ Normal Random Variable ($\mu_\varepsilon = 0$, $\sigma_\varepsilon = ???$) [Noise]



Assumptions of linear regression- When Can I fit the linear regression line

Linear regression assumes that…

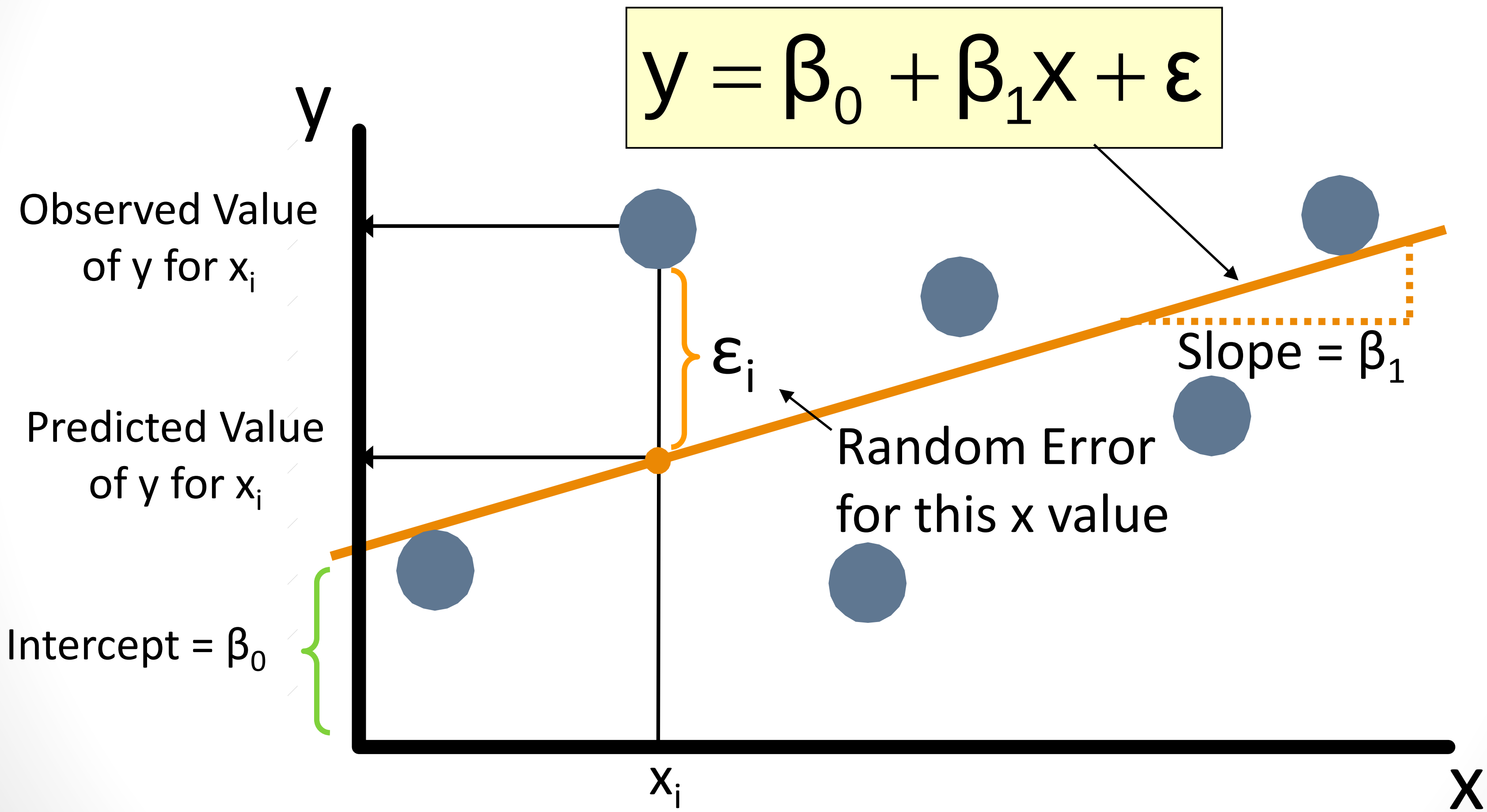
1. **The relationship between X and Y is linear:** Imagine a quadratic(parabolic) relation ship between X & Y. Does it make sense to fit a straight line through this data.

2. **Y is distributed normally at each value of X:** Imagine Y=0 whenever x is a multiple of 5. does it make sense to fit a straight line through this data. At each X, Y is normally distributed, which means at each X, Y value is around its overall mean value

3. **The variance of Y at every value of X is the same (homogeneity of variances):** Imagine data in the form of a cone, as we move away from origin the variance in Y is increasing drastically. Does it make sense to fit a straight line through this data?

4. **The observations are independent:** There is already one trend in the data If the observations are dependent. One trend line is not sufficient to model in this case

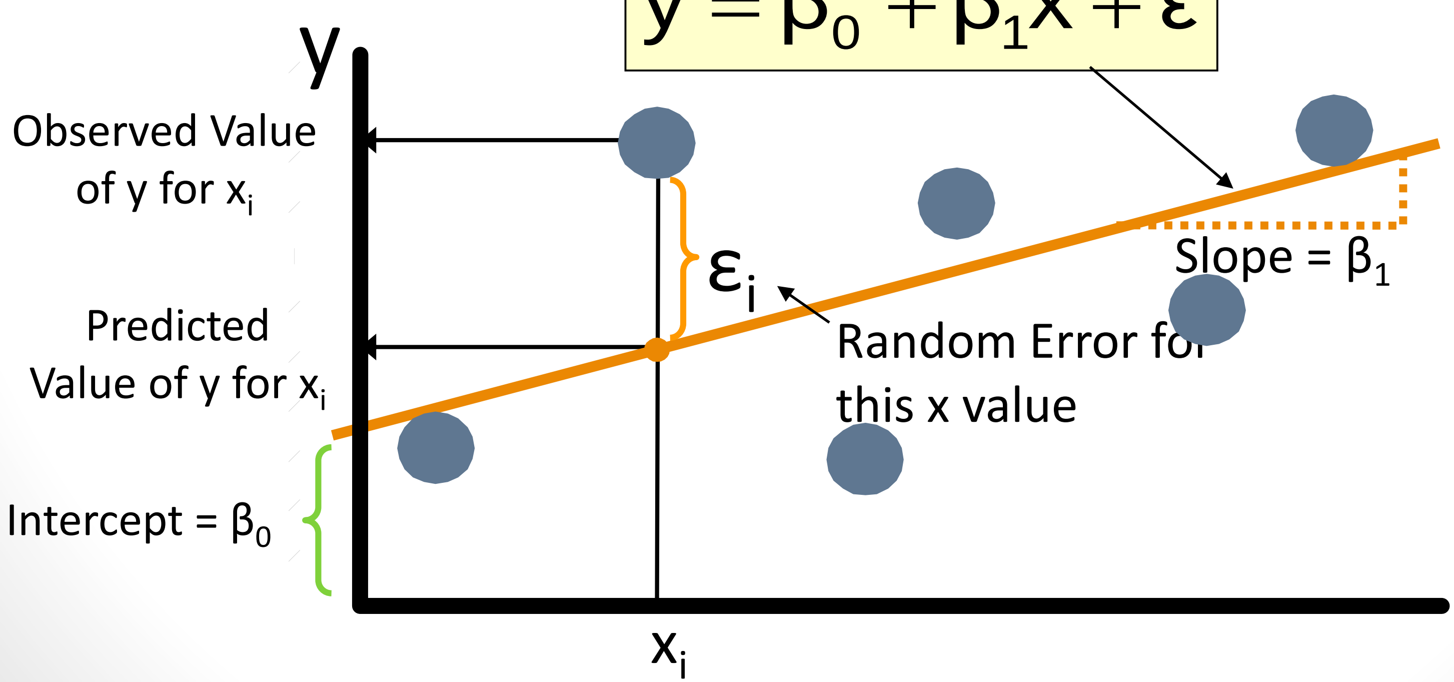
Regression line



Meaning of Beta

- Beta1 denotes the slope. What is slope? A slope of 2 means that every 1-unit change in X yields a 2-unit change in Y
- Beta1 is the estimated change in the average value of y as a result of a one-unit change in x
- Beta0 is the estimated average value of y when the value of x is zero

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



Least squares Estimation

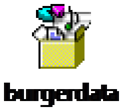
- X: x1, x2, x3, x4, x5, x6, x7,……..
- Y:y1, y2, y3, y4, y5, y6, y7…….
- Imagine a line through all the points
- Deviation from each point (residual or error)
- Square of the deviation
- Minimizing sum of squares of deviation

$$\sum e^2 = \sum (y - \hat{y})^2$$
$$= \sum (y - (b_0 + b_1 x))^2$$

b_0 and b_1 are obtained by finding the values of b_0 and b_1 that minimize the sum of the squared residuals

Lab: Burger Example

- Download burger data



- What is the mathematical relation between number of visitors and number of burgers sold?
- If what is the increase in the burger sales for every 100 visitors?
- If 5000 visitors are expected tomorrow, how much stock should I keep?
- If I want to sell 1000 burgers, how many visitors should I expect?
- Is it a good plan to increase the burger production above 3000?
- How reliable is this mathematical equation?

How good is my regression line?

- Take a regression line; Estimate y by substituting xi from data; Is it exactly same as yi?
- Remember no line is perfect
- There is always some error in the estimation
- Unless there is comprehensive dependency between predictor and response, there is always some part of response(Y) that can't be explained by predictor (x)
- So, total variance in Y is divided into two parts,
  - Variance that can be explained by x, using regression
  - Variance that can't be explained by x

Explained and Unexplained Variation

- Total variation is made up of two parts:

$$SST = SSE + SSR$$

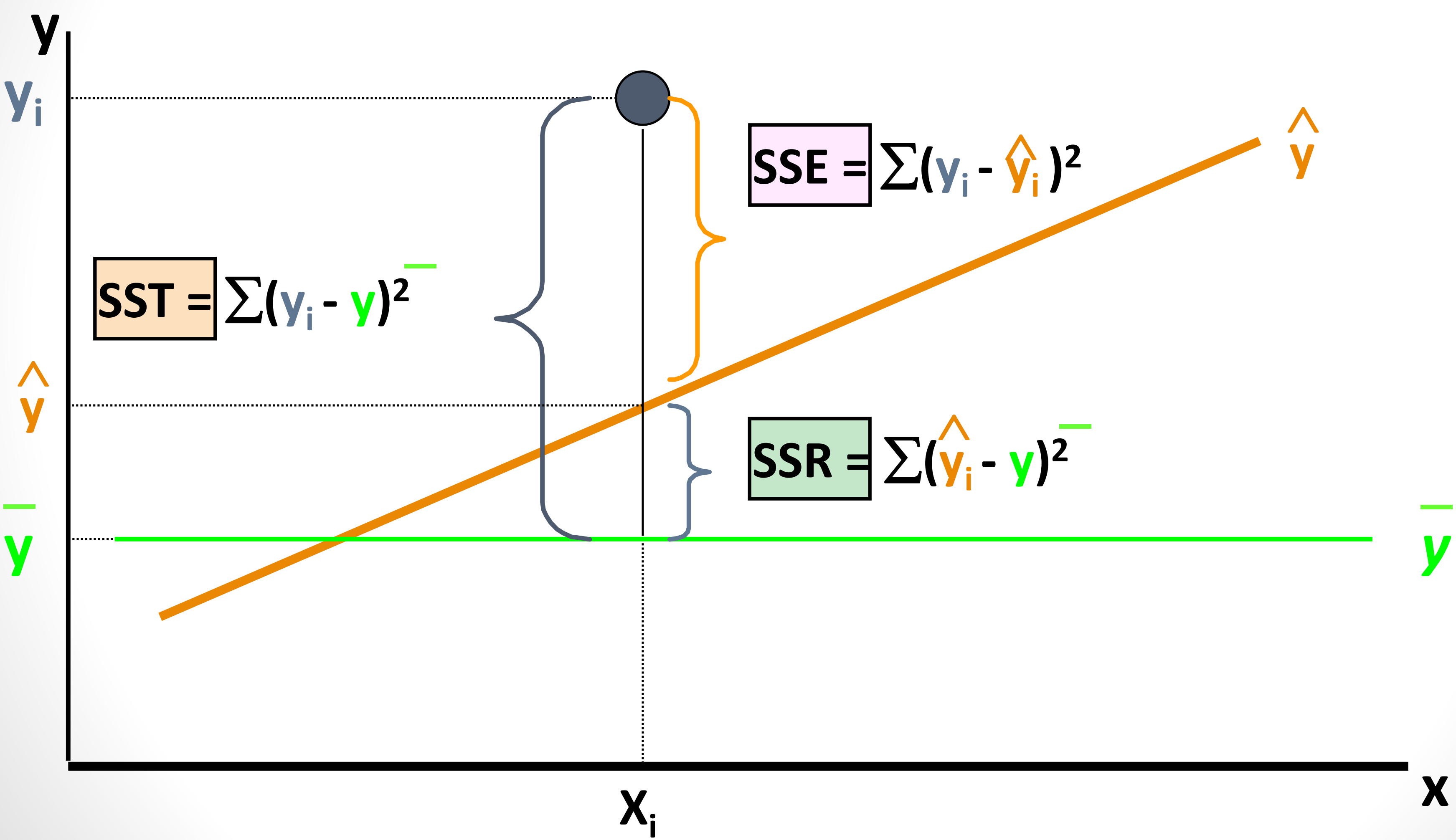
Total sum of Squares	Sum of Squares Error	Sum of Squares Regression
$SST = \sum (y - \bar{y})^2$	$SSE = \sum (y - \hat{y})^2$	$SSR = \sum (\hat{y} - \bar{y})^2$

SST :Measures the variation of the y_i values around their mean y

SSE : Variation attributable to factors other than the relationship between x and y

SSR: Explained variation attributable to the relationship between x and y

Explained and Unexplained Variation



Coefficient of determination

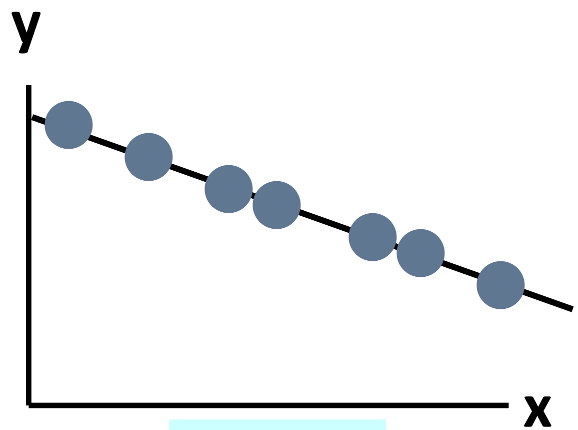
- A good fit will have
  - SSE (Minimum or Maximum?)
  - SSR (Minimum or Maximum?)
  - SSR/SSE(Minimum or Maximum?)
- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable

- The coefficient of determination is also called R-squared and is denoted as R^2

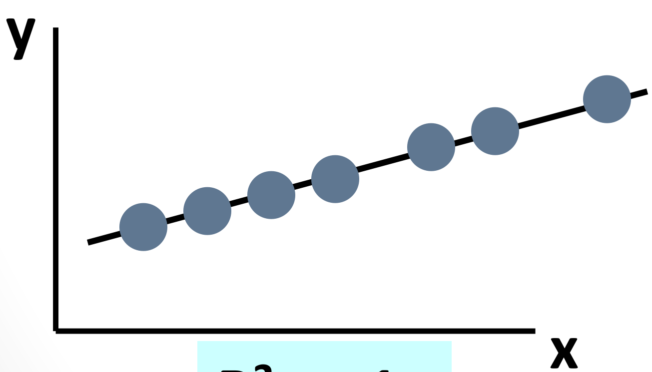
$$\boxed{R^2 = \frac{SSR}{SST}} \quad \text{where} \quad \boxed{0 \le R^2 \le 1}$$

In the single independent variable case, the coefficient of determination is equal to square of simple correlation coefficient

Type of relationship



$R^2 = 1$



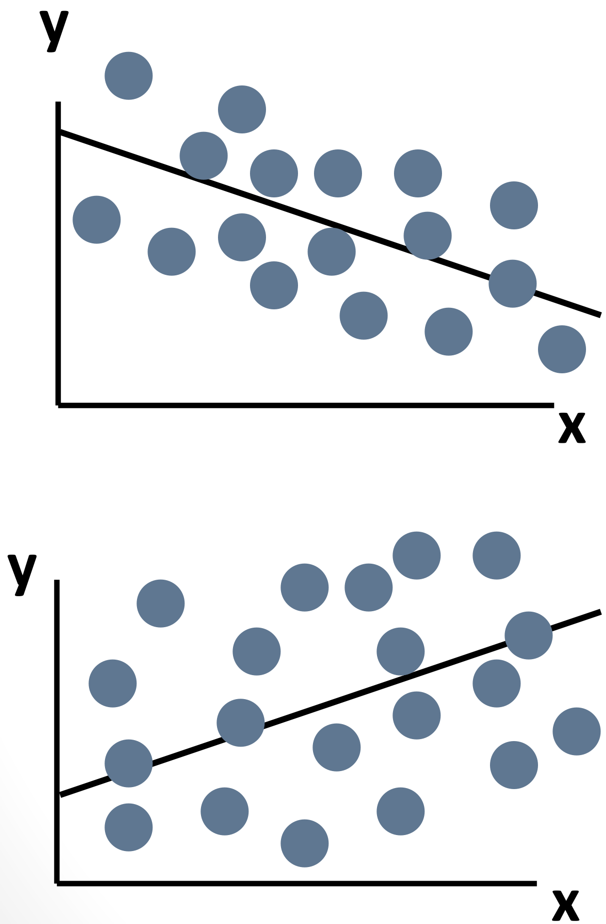
$R^2 = +1$

$R^2 = 1$

Perfect linear relationship between x and y:

100% of the variation in y is explained by variation in x

Type of relationship

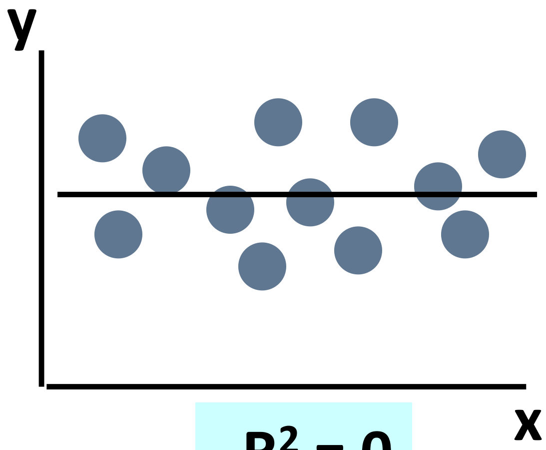


$0 < R^2 < 1$

Weaker linear relationship between x and y:

Some but not all of the variation in y is explained by variation in x

Type of relationship



$R^2 = 0$

$R^2 = 0$

No linear relationship between x and y:

The value of Y does not depend on x. (None of the variation in y is explained by variation in x)

Lab

- How good is the regression line for burger example?
- How good is the regression line?
- What is total sum of squares? How much model explained?
- What percentage of variation in Y(burgers sold) is explained by X(visitors)?
- In the age vs blood pressure example, estimate bp for a given age
- How good is the regression line?
- What is total sum of squares? How much model explained?
- What percentage of variation in Y(blood pressure) is explained by X(age)?

Standard Error of Estimate

- The standard deviation of the variation of observations around the regression line is estimated by

$$s_{\varepsilon} = \sqrt{\frac{SSE}{n-k-1}}$$

Where

SSE = Sum of squares error

n = Sample size

k = number of independent variables in the model

The Standard Deviation of the Regression Slope

- The standard error of the regression slope coefficient (b_1) is estimated by

$$s_{b_1} = \frac{s_\varepsilon}{\sqrt{\sum (x - \bar{x})^2}} = \frac{s_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}}$$

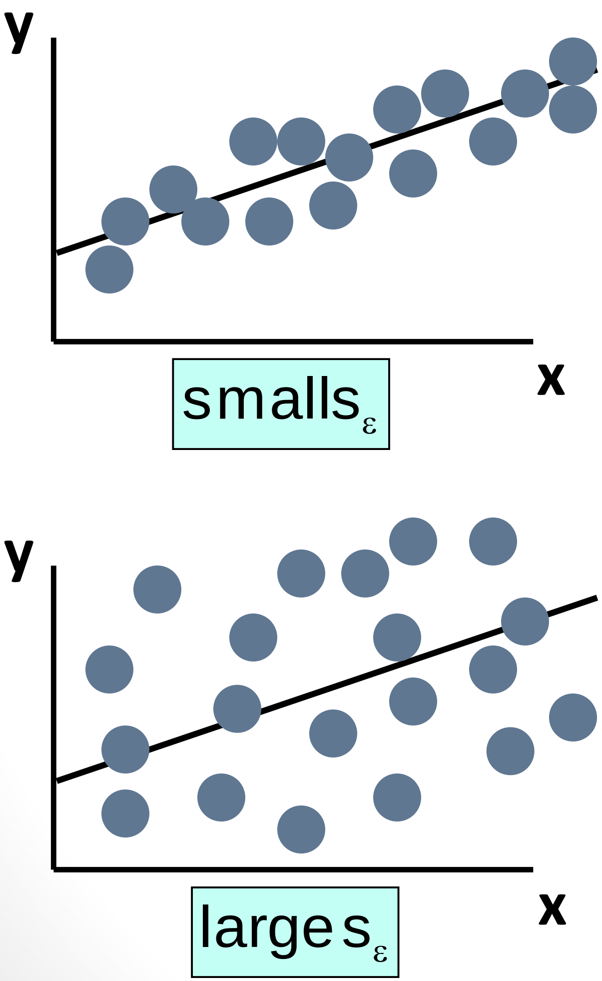
where:

s_{b_1} = Estimate of the standard error of the least squares slope

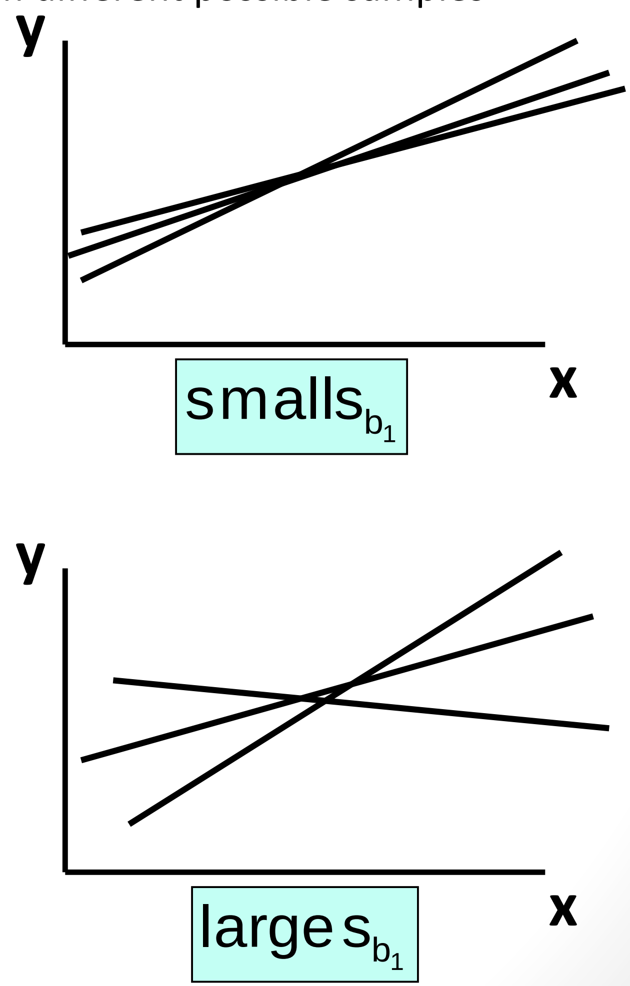
$s_\varepsilon = \sqrt{\frac{SSE}{n-2}}$ = Sample standard error of the estimate

Comparing Standard Errors

Variation of observed y values from the regression line



Variation in the slope of regression lines from different possible samples



Significance testing…

<u>Slope</u>

Distribution of slope ~ $T_{n-2}(\beta, s.e.(\hat{\beta}))$

H0:	$\beta 1 = 0$	(no linear relationship)
H1:	$\beta 1 \neq 0$	(linear relationship does exist)

$$T_{n-2} = \frac{\hat{\beta} - 0}{s.e.(\hat{\beta})}$$

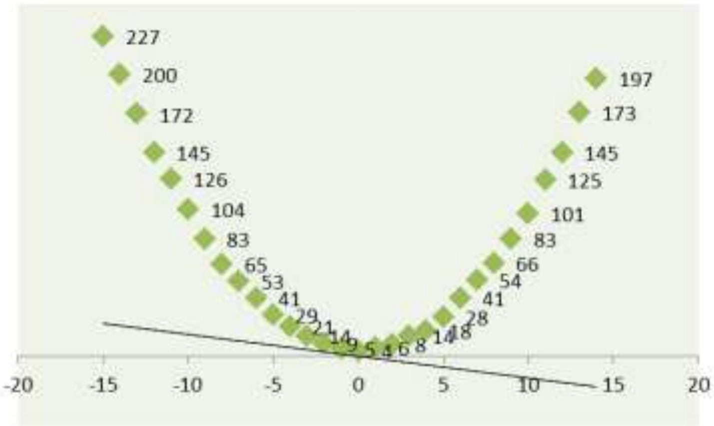
Reject or accept the null hypothesis based on above test statistic value.

Lab

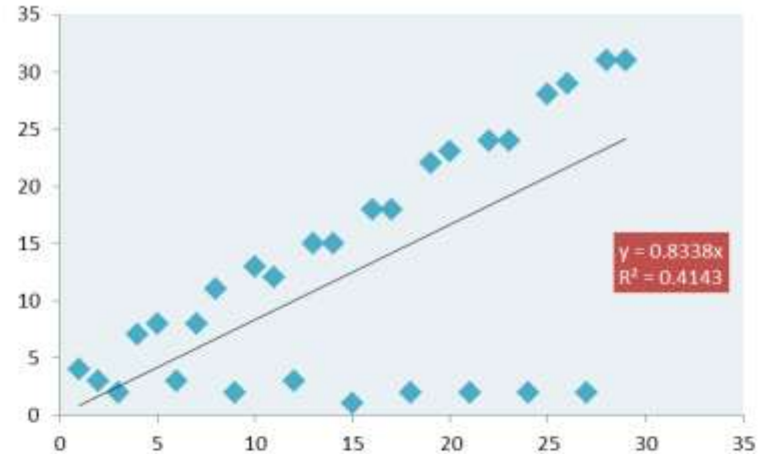
- In burger & BP examples
  - What is the standard error of estimates
  - What is the standard error estimate of beta
  - Compare two standard errors
- Which one of these two model is reliable?
- Are the coefficients significant?

When can I NOT fit a linear regression line

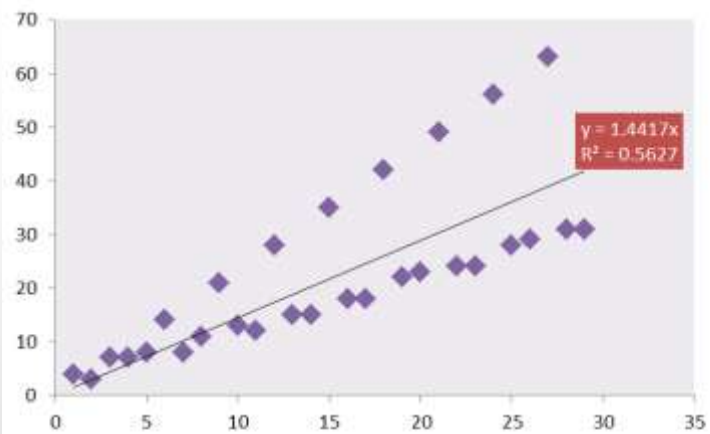
The relationship between X and Y is **NOT** linear



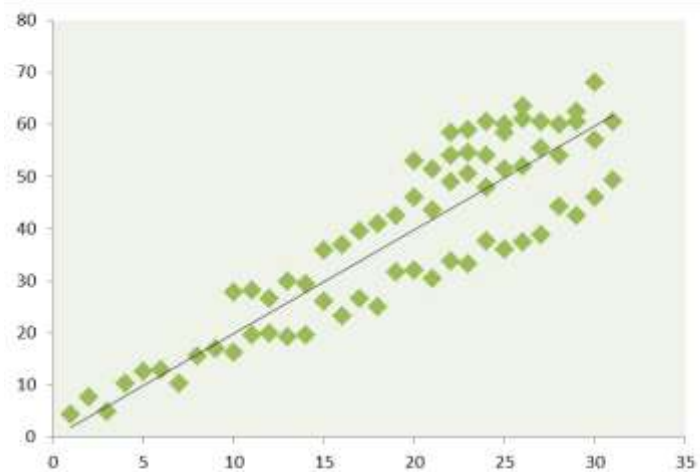
Y is **NOT** distributed normally at each value of X



The observations are **NOT** independent



Variance of Y at every value of X is **NOT** same



Venkat Reddy Konasani

Manager at Trendwise Analytics

venkat@TrendwiseAnalytics.com

21.venkat@gmail.com

+91 9886 768879