CORRELATION RESEARCH

By

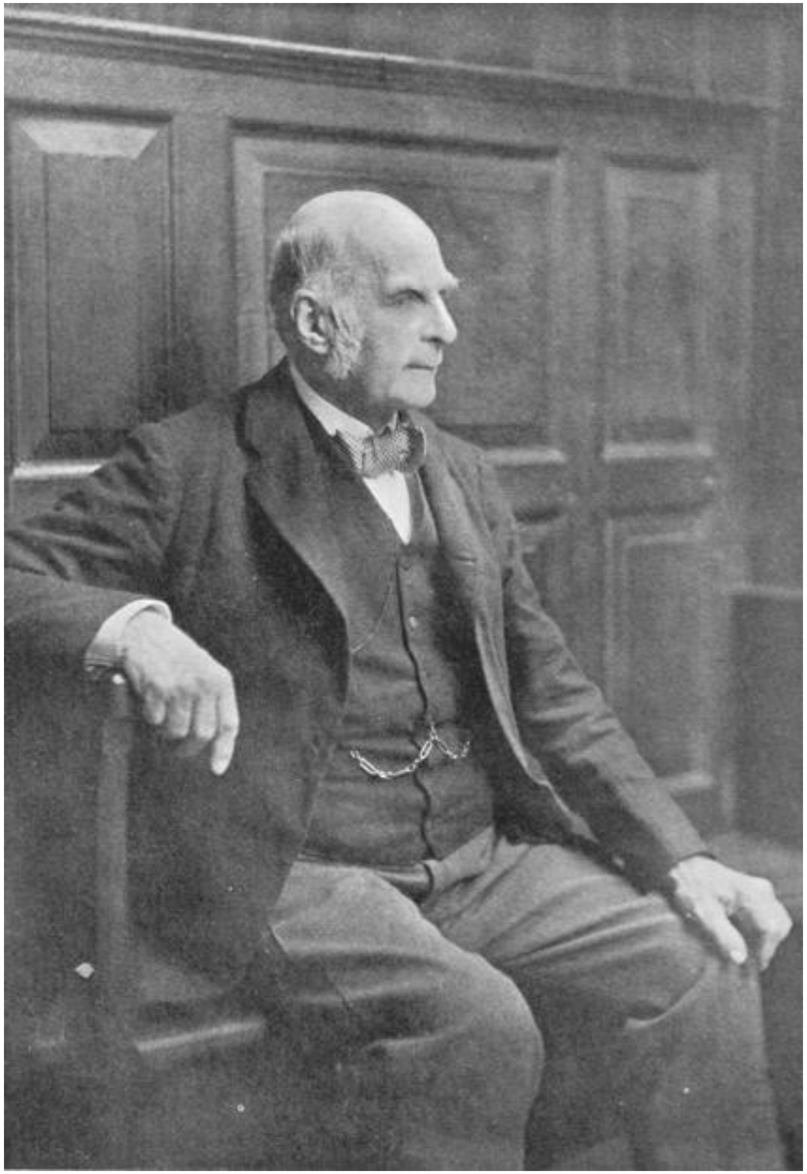
Prof. Rajeev Pandey
Department of Statistics
University of Lucknow
Lucknow

pandey_rajeev@lkouniv.ac.in **prof.rajeevlu@gmail.com**

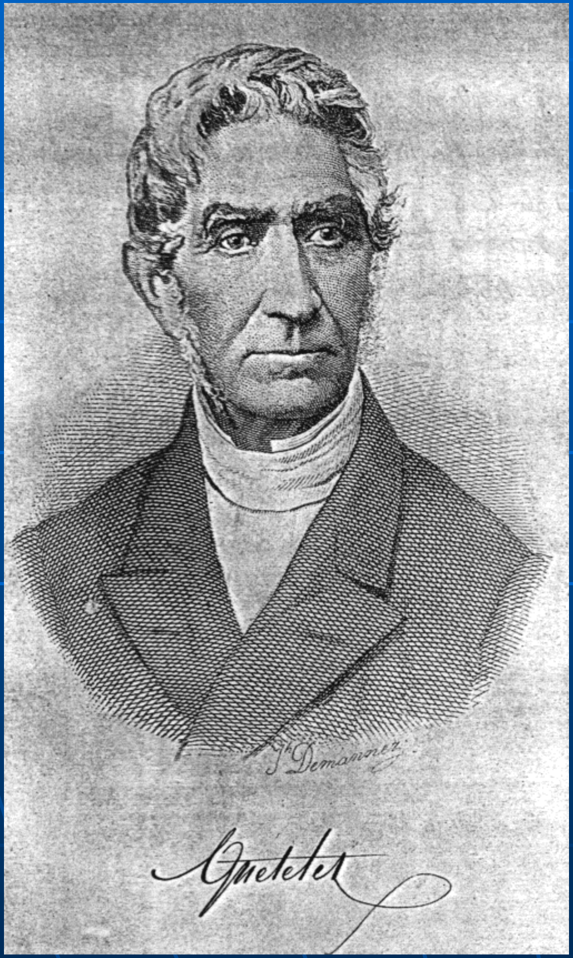
Sir Francis Galton

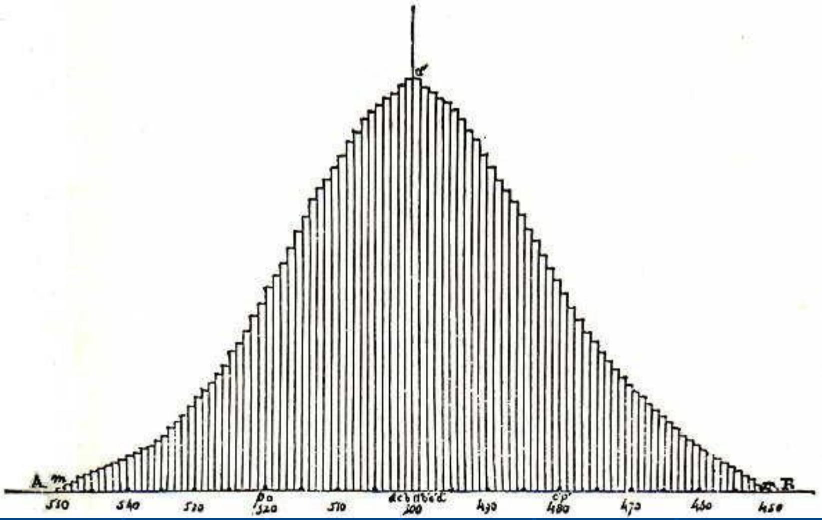
1822-1911

- **Obsessed with measurement**
- **Tried to measure everything from the weather to female beauty**
- **Invented correlation and regression**



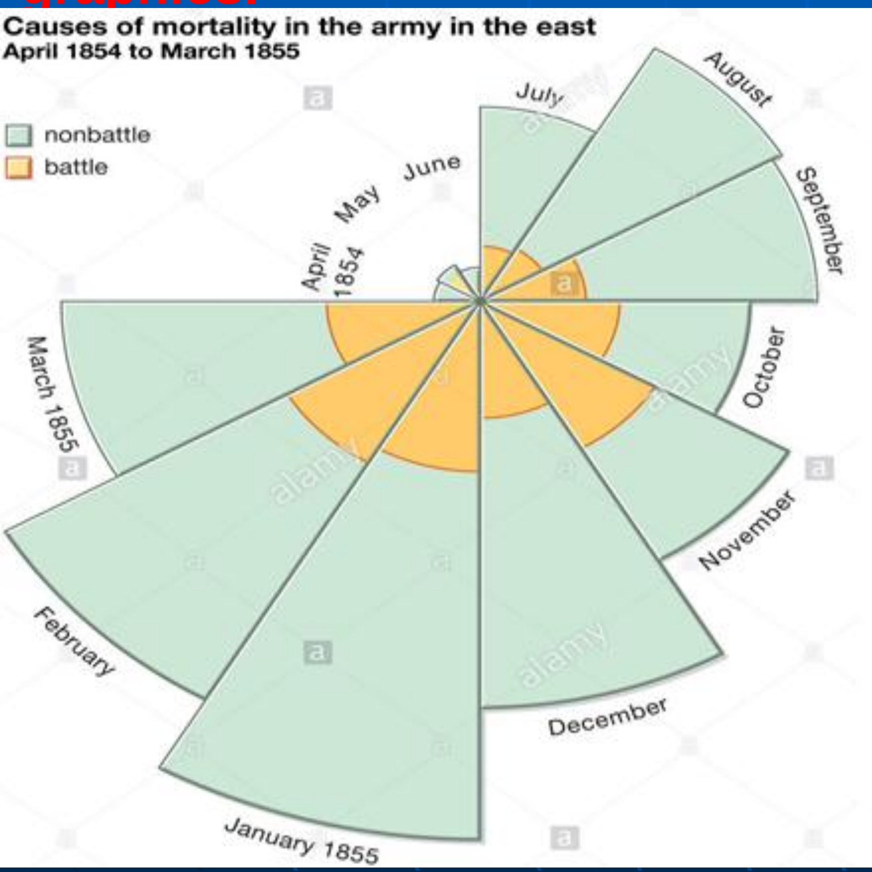
Ambitions for statistics





Adolphe Quetelet (1796-1874)
A Social Scientist
wanted statistics to be an experimental science of legislation.

Florence Nightingale exhibited a gift for mathematics from an early age and excelled in the subject under the tutelage of her father. Later, Nightingale became the first pioneer in the visual presentation of information and statistical graphics.



Florence Nightingale
(12 May 1820 – 13 August 1910)

Florence Nightingale

Differences between univariate and bivariate data.

Univariate Data	Bivariate Data
• involving a **single variable**	• involving **two variables**
• does not deal with causes or relationships	• deals with causes or relationships
• the major purpose of univariate analysis is to describe	• the major purpose of bivariate analysis is to explain
• central tendency - mean, mode, median • dispersion - range, variance, max, min, quartiles, standard deviation. • frequency distributions • bar graph, histogram, pie chart, line graph, box-and-whisker plot	• analysis of two variables simultaneously • correlations • comparisons, relationships, causes, explanations • tables where one variable is contingent on the values of the other variable. • independent and dependent variables
Sample question: How many of the students in the freshman class are female?	**Sample question:** Is there a relationship between the number of females in Computer Programming and their scores in Mathematics?

Correlation & Association

Multivariate Data Format

Unit	Variable		
	X_1	X_2 . . .	X_p
1	X_{11}	X_{12} . . .	X_{1p}
2	X_{21}	X_{22} . . .	X_{2p}
3	X_{31}	X_{32} . . .	X_{3p}
. . .	. . . .		
. . .	. . . .		
i	X_{i1}	X_{i2}	X_{ip}
. . .	....		
n	X_{n1}	X_{n2}	X_{np}

1. Dependence Methods

One or more variables (called ***criterion variables***) are predicted by a set of independent variables (called ***predictor variables***)

a. One criterion variable

(i) Correlation and Regression Analysis

Criterion Variable : Metric

Predictor Variables: Metric & Non-Metric

(ii) Logistic Regression

Criterion Variable : Non-Metric

Predictor Variables: Metric & Non-Metric

(iii) Discriminant Analysis

Criterion Variable : Non-Metric

Predictor Variables: Metric

b. Two or more criterion variables

(i) Canonical Analysis

Criterion Variable : Metric

Predictor Variables: Metric

(ii) Multivariate Analysis of Variance

Criterion Variable : Metric

Predictor Variables: Metric

How to determine similarity….?

Specify as many characteristics as possible and measure them on each unit. A single characteristic may not be sufficient.

Similarity is hard to define, but….

" we know it when see it"

Detecting similarity is a typical task in matching learning…..

2.Inter-dependence Methods

(i) Factor Analysis
(ii) Cluster Analysis
(iii) Multidimensional Scaling
(iv) Correspondence Analysis

Although we can anlyse each variable individually using methods available for univariate analysis but in multivariate we try to exploit information about inter-relationship among the variables to make several inferences which are not possible otherwise.

key concepts: Correlation in SPSS

- <u>Types of correlation</u>

Methods of studying correlation in SPSS

a) Scatter diagram

b) Karl pearson's coefficient of correlation

c) Spearman's Rank correlation coefficient

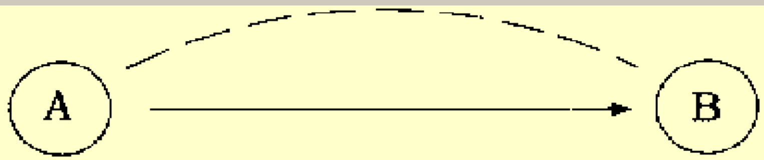
d) Kendall's Tau

Correlation

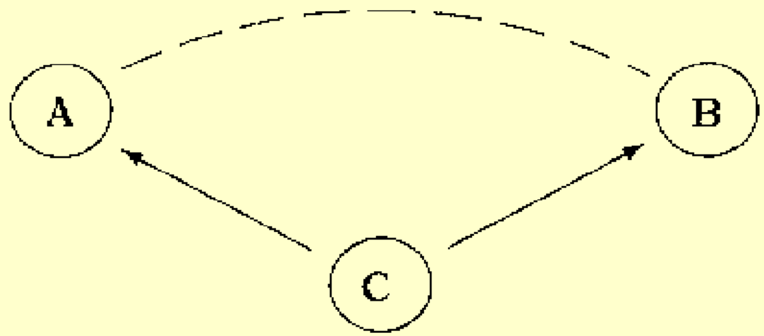
- Correlation is a statistical tool that helps to measure and analyze the degree of relationship between two variables.
- Correlation analysis deals with the association between two or more variables.

Correlation

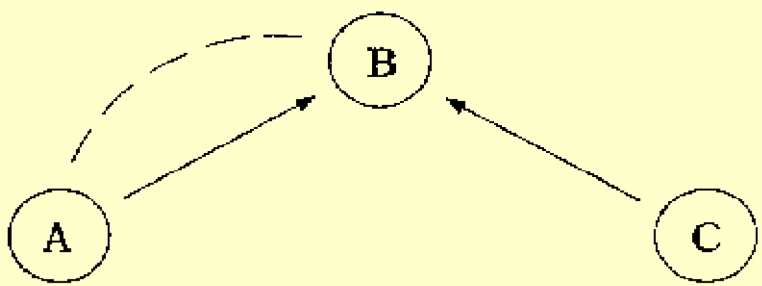
- **Correlation:** The degree of relationship between the variables under consideration is measure through the correlation analysis.
- The measure of correlation called the correlation coefficient .
- The degree of relationship is expressed by coefficient which range from correlation $(-1 \le r \le +1)$
- The direction of change is indicated by a sign.
- The correlation analysis enable us to have an idea about the degree & direction of the relationship between the two variables under study.



CAUSATION—Changes in A cause changes in B.



COMMON RESPONSE—Changes in both A and B are caused by changes in a third variable, C.



CONFOUNDING—Changes in B are caused both by changes in A and by changes in third variable C.

Correlation & Causation

- Causation means cause & effect relation.
- Correlation denotes the interdependency among the variables for correlating two phenomenon, it is essential that the two phenomenon should have cause-effect relationship,& if such relationship does not exist then the two phenomenon can not be correlated.
- If two variables vary in such a way that movement in one are accompanied by movement in other, these variables are called cause and effect relationship.
- Causation always implies correlation but correlation does not necessarily implies causation.

Spurious Relationship

- The final type of relationship could be spurious. The relationship between the jail population (X) and the crime rate (Y) could be associated with a third variable.
- The size of the jail population (X^1) could be related to the unemployment rate (X^2), which may be strongly associated with the crime rate (Y).

Types of Correlation
Type Ⅰ

- Correlation
  - Positive Correlation
  - Negative Correlation

Types of Correlation Type Ⅰ

- **Positive Correlation:** The correlation is said to be positive correlation if the values of two variables changing with same direction.

  Ex. Arrest Rate & Performance, clearance rate & Performance

- **Negative Correlation:** The correlation is said to be negative correlation when the values of variables change with opposite direction.

  Ex. Area Crime Rate & Performance.

Direction of the Correlation

- **Positive relationship** – Variables change in the same direction.
  - As X is increasing, Y is increasing
  - As X is decreasing, Y is decreasing
- E.g., As CLR increases, so does Performance.

Indicated by sign; (+) or (-).

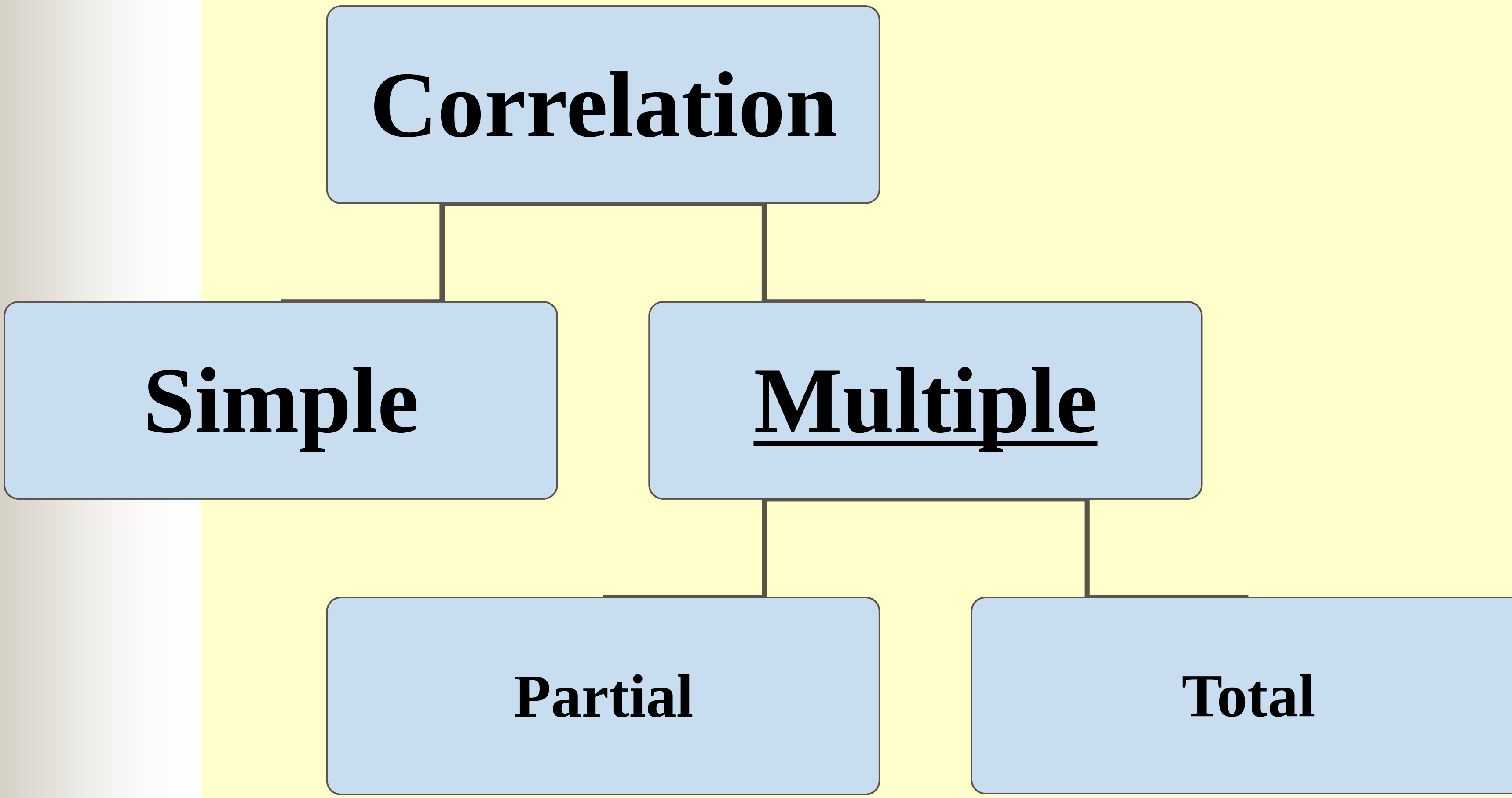
- **Negative relationship** – Variables change in opposite directions.
  - As X is increasing, Y is decreasing
  - As X is decreasing, Y is increasing
- E.g., As ACR increases, Performance decrease

More examples

- **Positive relationships**
- water consumption and temperature.
- study time and grades.

- **Negative relationships**:
- alcohol consumption and driving ability.
- Price & quantity demanded

Types of Correlation
Type II



Types of Correlation Type II

- **Simple correlation:** Under simple correlation problem there are only two variables are studied.
- **Multiple Correlation:** Under Multiple Correlation three or more than three variables are studied. Ex. $Q_d = f$ (ACR, CLR, Performance)
- **Partial correlation:** analysis recognizes more than two variables but considers only two variables keeping the other constant.
- **Total correlation:** is based on all the relevant variables, which is normally not feasible.

Types of Correlation
Type III

Correlation

- LINEAR
- NON LINEAR

Types of Correlation Type III

- **Linear correlation:** Correlation is said to be linear when the amount of change in one variable tends to bear a constant ratio to the amount of change in the other. The graph of the variables having a linear relationship will form a straight line.

  Ex X = 1, 2, 3, 4, 5, 6, 7, 8,

  Y = 5, 7, 9, 11, 13, 15, 17, 19,

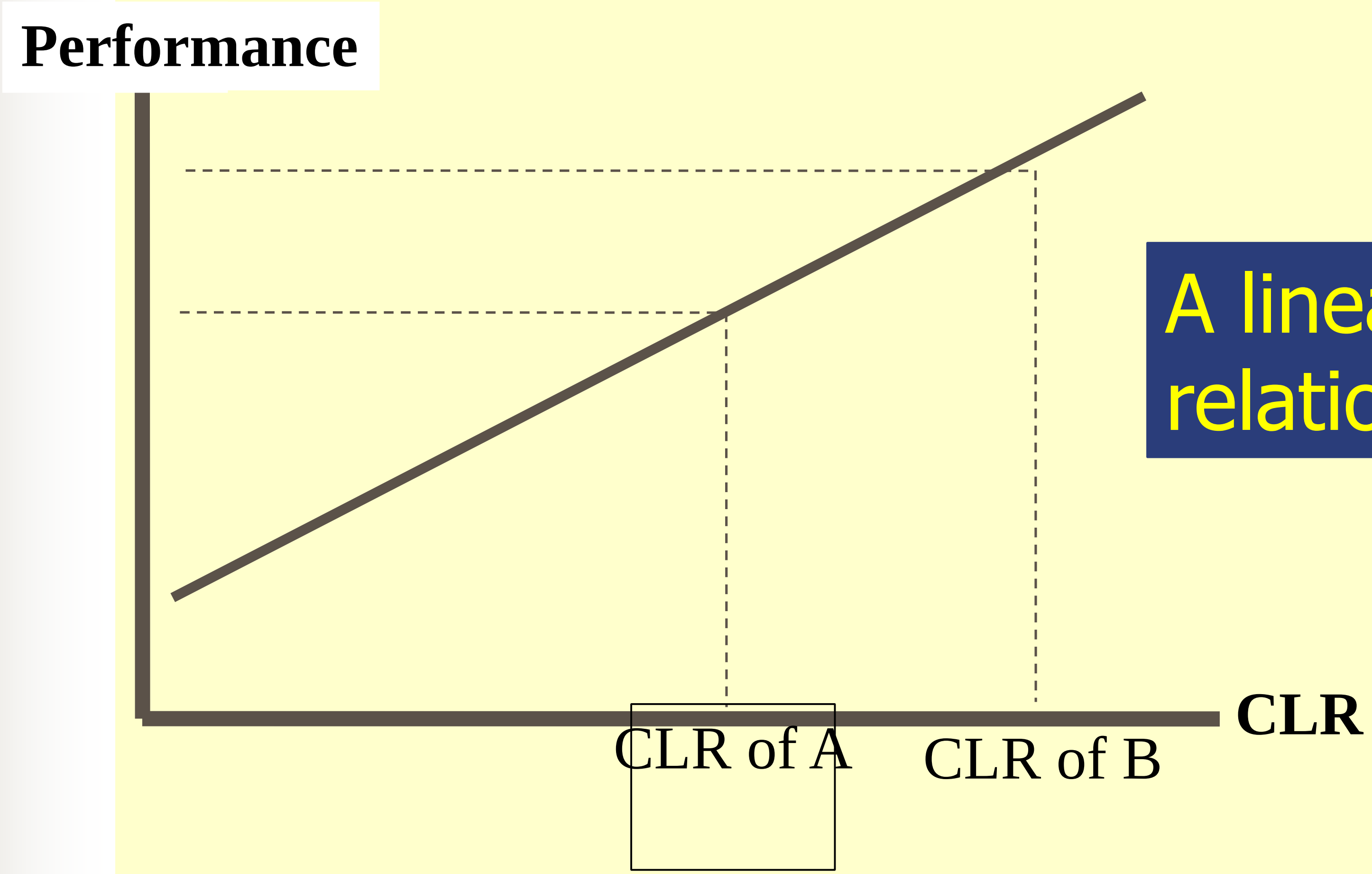
  $Y = 3 + 2x$

- **Non Linear correlation:** The correlation would be non linear if the amount of change in one variable does not bear a constant ratio to the amount of change in the other variable.

Scatter Diagram Method

- Scatter Diagram is a graph of observed plotted points where each points represents the values of X & Y as a coordinate. It portrays the relationship between these two variables graphically.

A perfect positive correlation

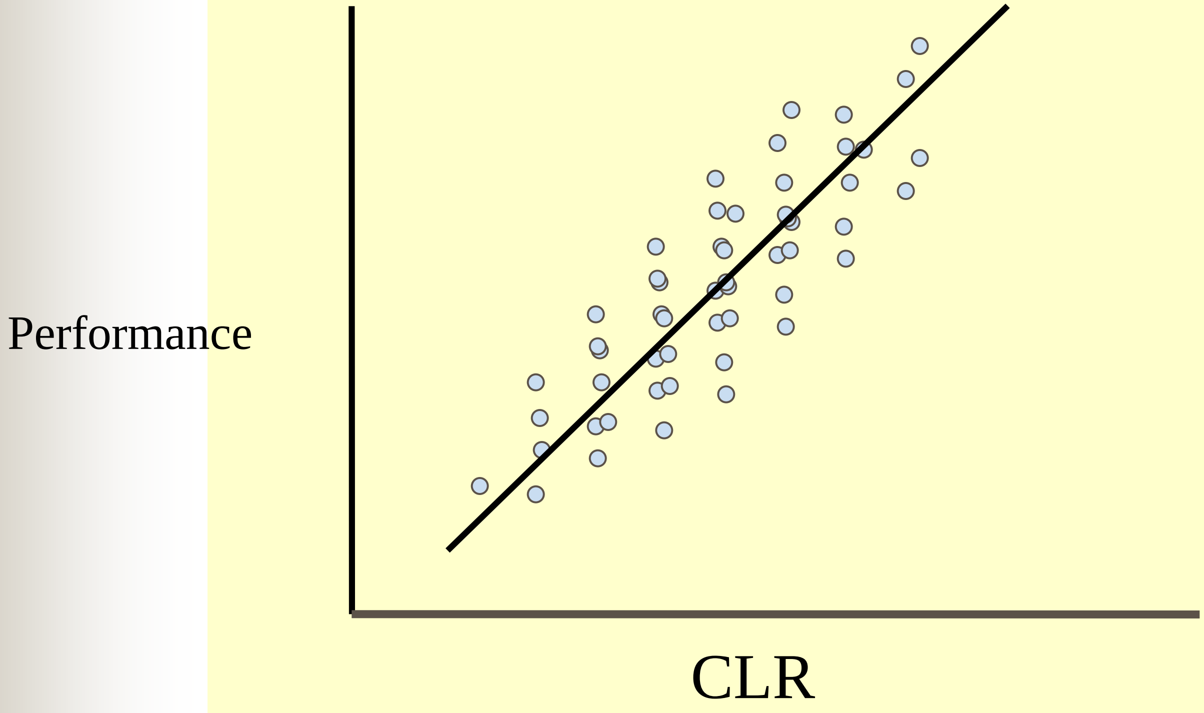


A linear relationship

High Degree of positive correlation

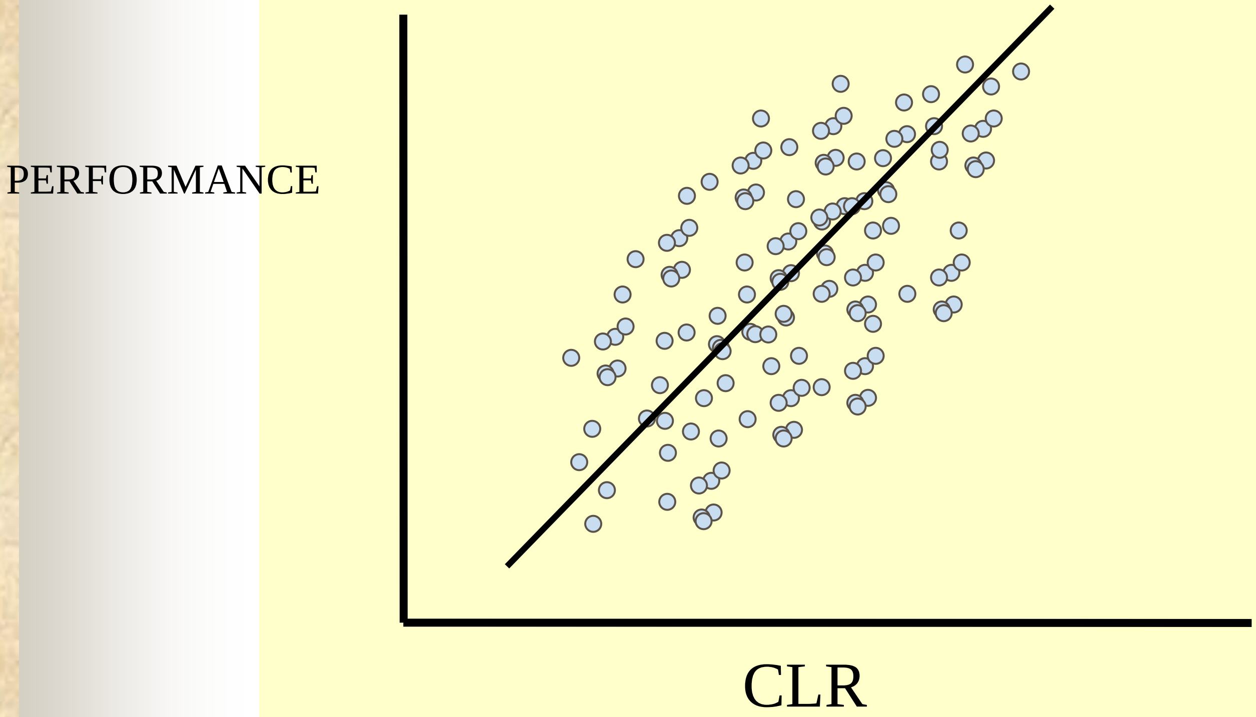
- Positive relationship

r = +.80



Degree of correlation

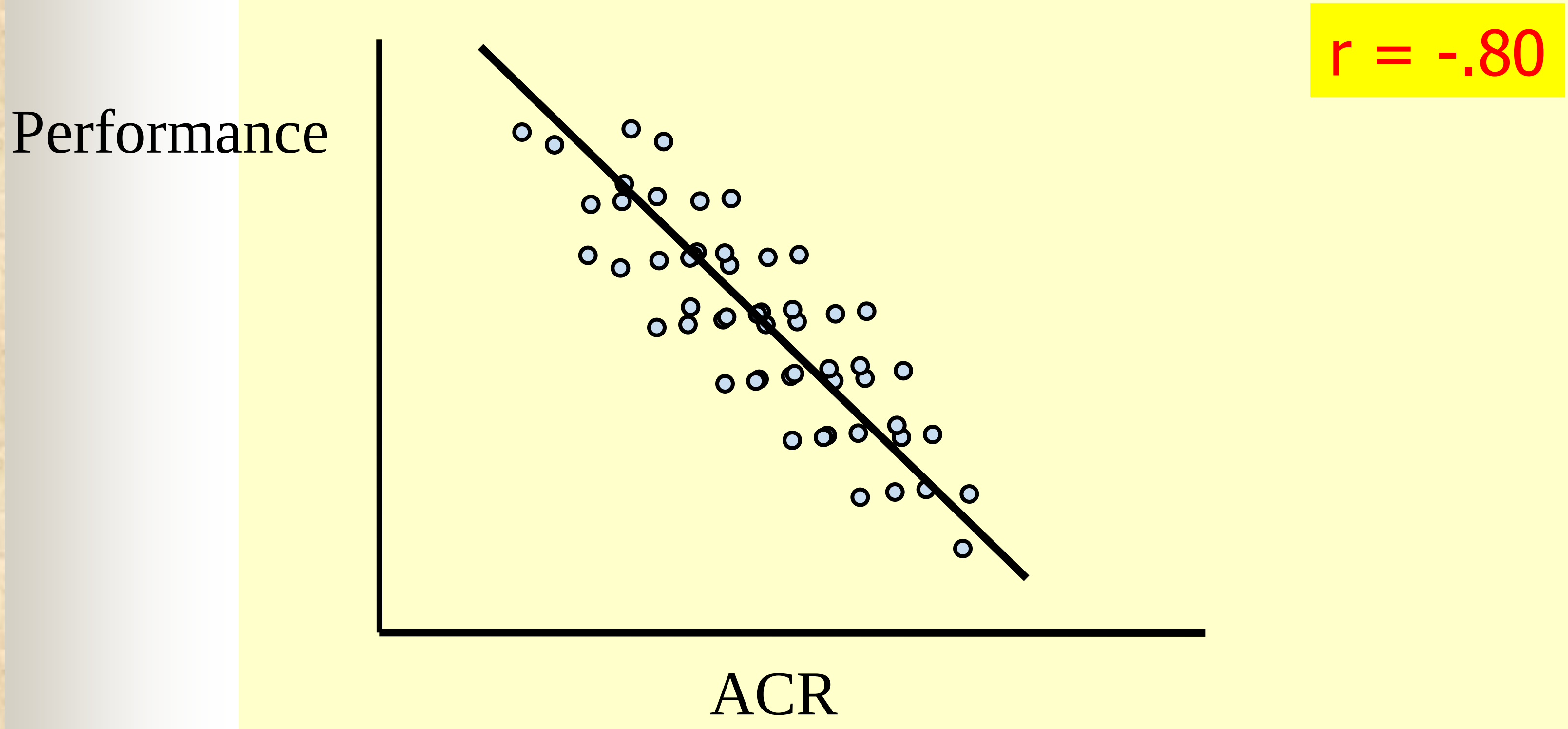
- **Moderate Positive Correlation**



r = + 0.4

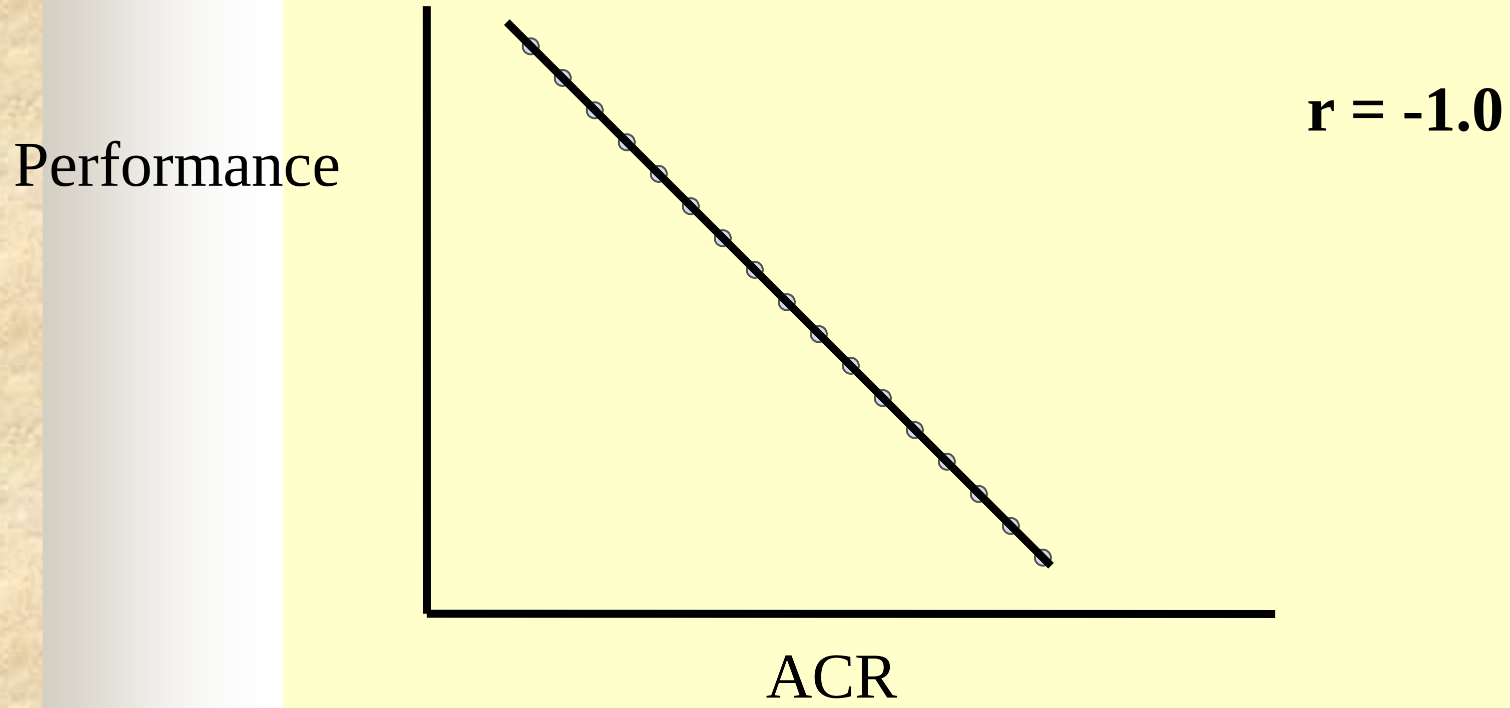
Degree of correlation

- **Moderate Negative Correlation**



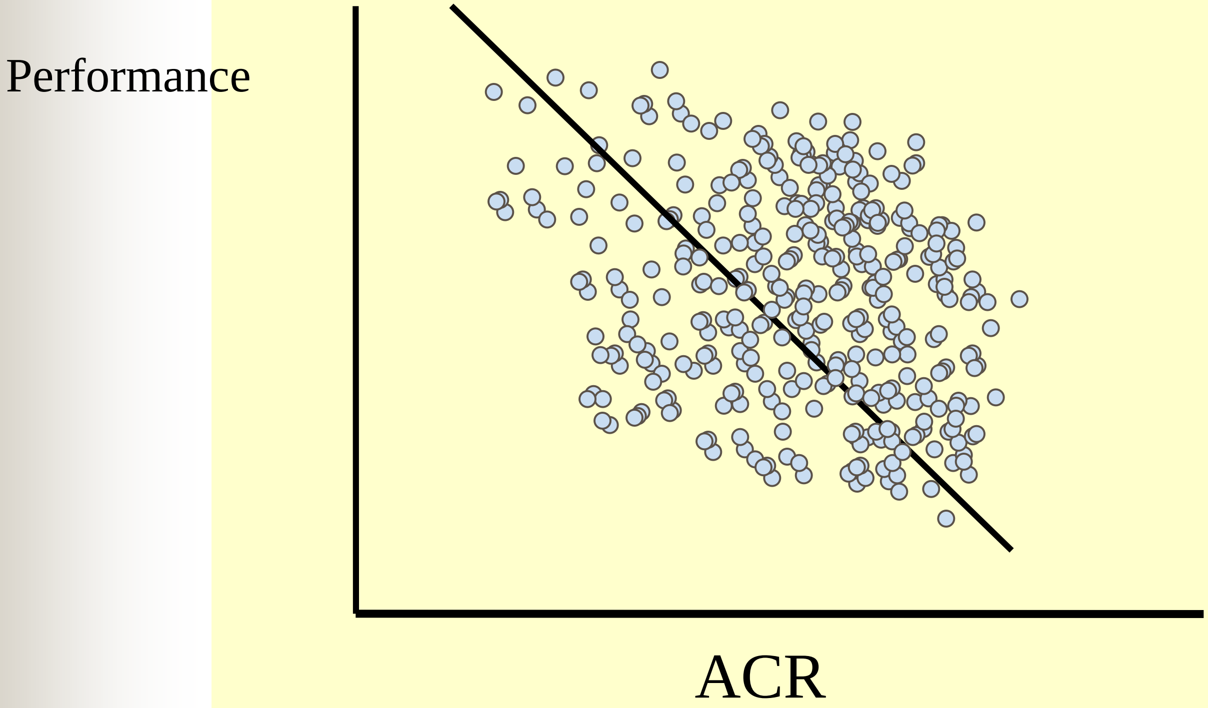
Degree of correlation

- **Perfect Negative Correlation**



Degree of correlation

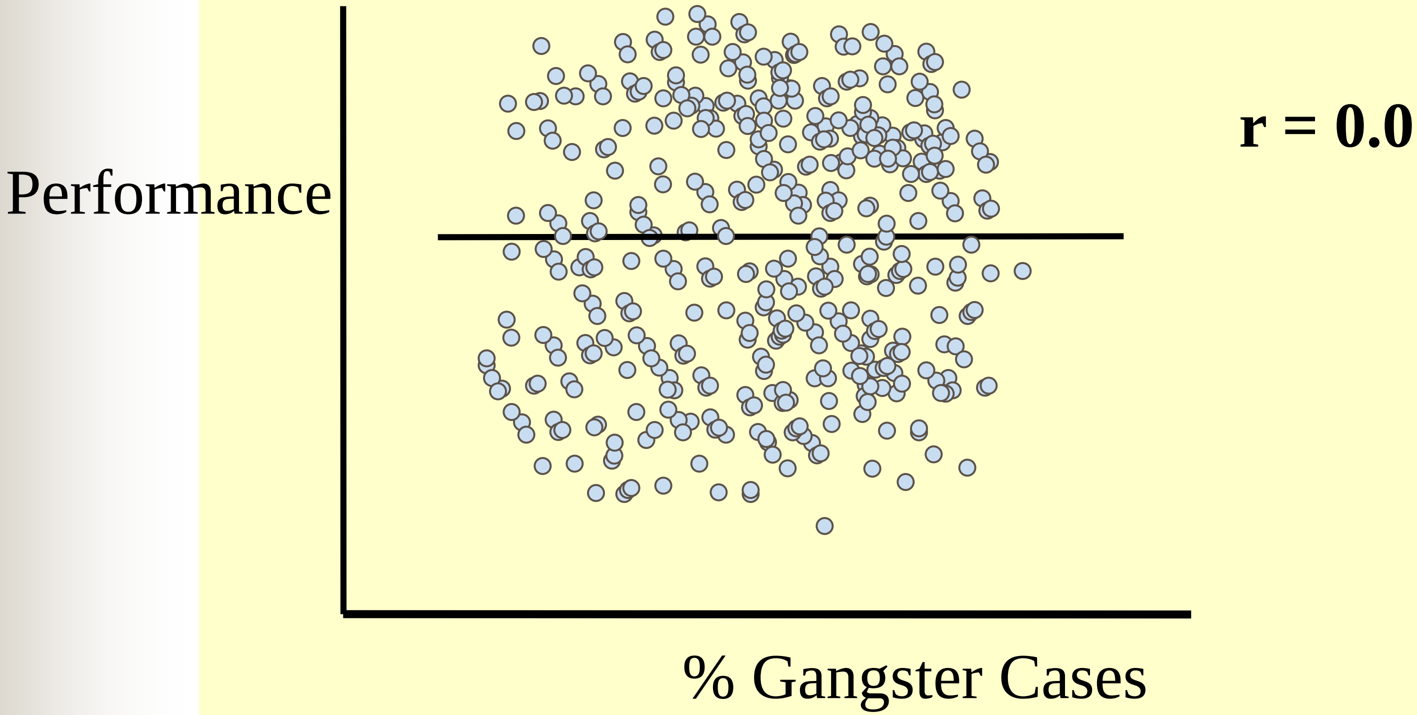
- **Weak negative Correlation**



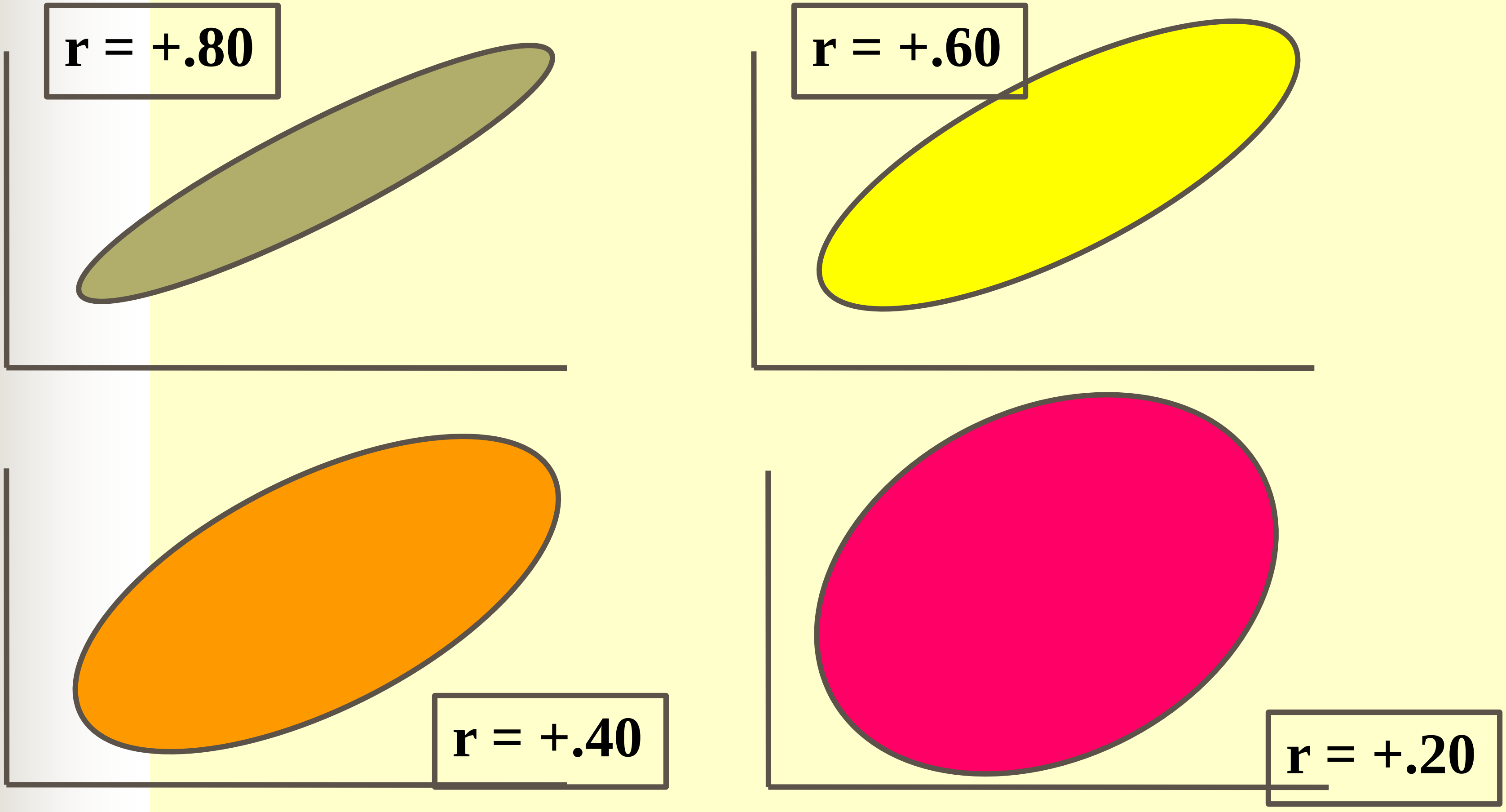
r = - 0.2

Degree of correlation

- **No Correlation (horizontal line)**



Degree of correlation (r)



Advantages of Scatter Diagram

- Simple & Non Mathematical method
- Not influenced by the size of extreme item
- First step in investing the relationship between two variables

Disadvantage of scatter diagram

Can not adopt the an exact degree of correlation

Karl Pearson's Coefficient of Correlation

- Pearson’s ‘r’ is the most common correlation coefficient.
- Karl Pearson’s Coefficient of Correlation denoted by ‘r’. The coefficient of correlation ‘r’ measure the degree of linear relationship between two variables say x & y.

Karl Pearson's Coefficient of Correlation

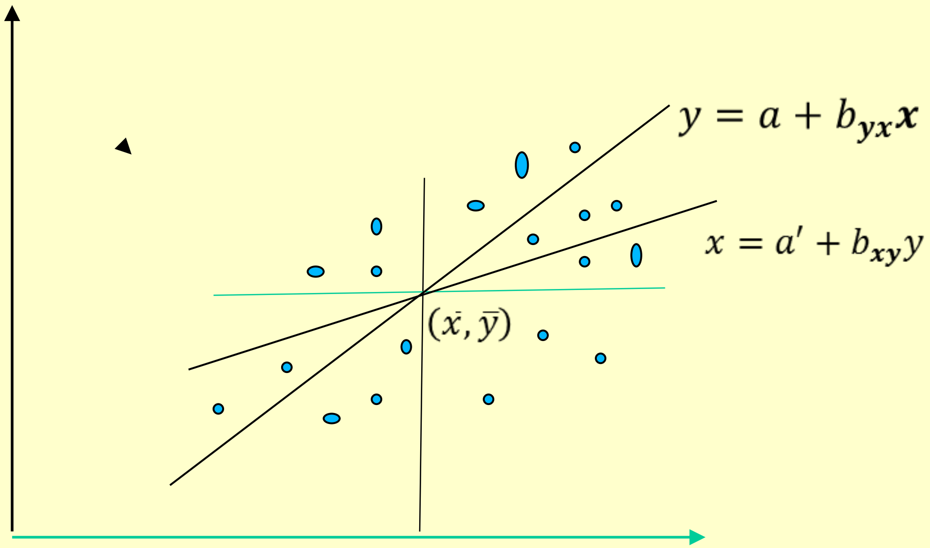
- Karl Pearson's Coefficient of Correlation denoted by "r"

$$-1 \leq r \leq +1$$

- Degree of Correlation is expressed by a value of Coefficient
- Direction of change is Indicated by sign

(- ve) or (+ ve)

Product Moment Correlation coefficient (Pearson's Correlation Coefficient)



- If we assume that Y depends on x and relationship is linear

$$y = a + b_{yx}\boldsymbol{x} \quad \textit{Regression of Y on x}$$

on the other hand if we assume that X depends on Y

$$x = a' + b_{xy}y \quad \textit{Regession of X on Y}$$

$$r_{xy} = \frac{covariance\ (X,Y)}{\sqrt{Variance\ (X).Variance(Y)}} = \frac{\frac{1}{n}\Sigma_1^n(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\Sigma_1^n(x_i - \bar{x})^2}.\sqrt{\frac{1}{n}\Sigma_1^n(y_i - \bar{y})^2}}$$

Interpretation of Correlation Coefficient (r)

- The value of correlation coefficient 'r' ranges from -1 to +1
- If r = +1, then the correlation between the two variables is said to be perfect and positive
- If r = -1, then the correlation between the two variables is said to be perfect and negative
- If r = 0, then there exists no correlation between the variables

Properties of Correlation coefficient

- The correlation coefficient lies between -1 & +1 symbolically $(-1 \leq r \leq 1)$
- The correlation coefficient is independent of the change of origin & scale.
- The coefficient of correlation is the geometric mean of two regression coefficient.

$$r = \sqrt{byx * bxy}$$

The one regression coefficient is (+ve) other regression coefficient is also (+ve) correlation coefficient is (+ve)

Assumptions of Pearson's Correlation Coefficient

- There is linear relationship between two variables, i.e. when the two variables are plotted on a scatter diagram a straight line will be formed by the points.
- Cause and effect relation exists between different forces operating on the item of the two variable series.

Advantages of Pearson's Coefficient

- It summarizes in one value, the degree of correlation & direction of correlation also.

Limitation of Pearson's Coefficient

- Always assume linear relationship
- Interpreting the value of r is difficult.
- Value of Correlation Coefficient is affected by the extreme values.
- Time consuming methods

Coefficient of Determination

- The convenient way of interpreting the value of correlation coefficient is to use of square of coefficient of correlation which is called Coefficient of Determination.
- The Coefficient of Determination = R^2.
- Suppose: $r = 0.9$, $R^2 = 0.81$ this would mean that 81% of the variation in the dependent variable has been explained by the independent variable. $\mathbf{Y = a + b_{yx}\, x. \ldots . . (1)}$

Independent Variable	Dependent Variable y	Estimated Value from (1) Y
x_1	y_1	Y_1
x_2	y_2	Y_2
. . .	. . .	. . .
. . .	. . .	. . .
x_n	y_n	y_n

$$\mathrm{Cov}(y, Y) = R^2$$

Coefficient of Determination

- The maximum value of R^2 is 1 because it is possible to explain all of the variation in y but it is not possible to explain more than all of it.
- Coefficient of Determination = Explained variation / Total variation

Coefficient of Determination: An example

- Suppose: $r = 0.60$

  $r = 0.30$ It does not mean that the first correlation is twice as strong as the second the 'r' can be understood by computing the value of r^2.

When $r = 0.60$ $r^2 = 0.36$ -----(1)

$r = 0.30$ $r^2 = 0.09$ -----(2)

This implies that in the first case 36% of the total variation is explained whereas in second case 9% of the total variation is explained .

Spearman's Rank Coefficient of Correlation

- When statistical series in which the variables under study are not capable of quantitative measurement but can be arranged in serial order, in such situation pearson's correlation coefficient can not be used in such case Spearman Rank correlation can be used.

1. *When two persons or judges give their ranks in same characteristics (variable).*
2. *When one person gives ranks to two different characteristics (variables).*

zone	Murder _FIR(Y)	Dacoity _Fir(X)	Y*X	Y-Mean(Y)	X-Mean(X)	(Y-Mean)²	(X-Mean)²	(Y-Mean)(X-Mean)	X²	Y²
Agra-1	1698	415	704670	227.87	-12.62	51927.02	159.39	-2876.92	172225	2883204
Allaha Bad - 2	1055	220	232100	-415.12	-207.62	172328.8	43108.14	86190.32	48400	1113025
Kanpur-3	1317	318	418806	-153.12	-109.62	23447.27	12017.64	16786.32	101124	1734489
Gorakhpur - 4	1095	397	434715	-375.12	-30.62	140718.8	937.89	11488.20	157609	1199025
Bareilly - 5	1518	514	780252	47.87	86.37	2292.016	7460.64	4135.20	264196	2304324
Meerut - 6	1881	823	1548063	410.87	395.37	168818.3	156321.39	162449.70	677329	3538161
Lucknow -7	2176	540	1175040	705.87	112.37	498259.5	12628.14	79322.70	291600	4734976
Varanasi -8	1021	194	198074	-449.12	-233.625	201713.3	54580.64	104926.82	37636	1042441
TOTAL	11761	3421	5491720			1259505	287213.87	462422.37	1750119	18549645
MEAN	1470.12	427.62	686465							

The correlation coefficient =

$Covariance\ (X,Y) \sqrt{}Variance\ (X).Variance\ (Y)$

$= 0.76$

Regression of Y (Murder) on X (Dacoity)

$Y = a + byx\ x$

$Y = 781.648 + 1.61x$

Regression of X (Dacoity) on Y (Murder)

$X = a' + bxy\ y$

$X = -111.91 + 0.367y$

$$\mathbf{r} = \sqrt{b_{yx} \cdot b_{xy}} \quad = \sqrt{}(1.61x\ 0.367) = 0.76$$

Spearman's Rank Correlation Coefficient

- $r_s = \dfrac{\sum_{i=1}^{n}(u_i-\overline{u})(v_i-\overline{v})}{\sqrt{\{\sum_{i=1}^{n}(u_i-\overline{u})^2\}\{\sum_{i=1}^{n}(v_i-\overline{v})^2\}}}$ (1)

  - Remark:
    - $u_i = rank(x_i)$ $v_i = rank(y_i)$
    - $d_i = u_i - v_i$ are the difference in ranks
    - n=number of pairs of X's and Y's.

Pearson's and Spearman's Correlation Coefficients

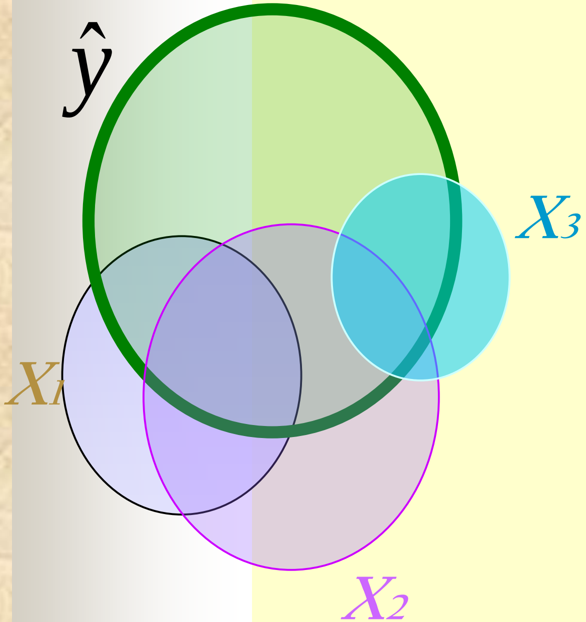
Zone	Murder_FIR (Y)	Dacoity_FIR (X)
1 - Agra	1698	415
2 - Allahabad	1055	220
3 - Kanpur	1317	318
4 - Gorakhpur	1095	397
5 - Bareilly	1518	514
6 - Merrut	1881	823
7 - Lucknow	2176	540
8 – Varanasi	1021	194

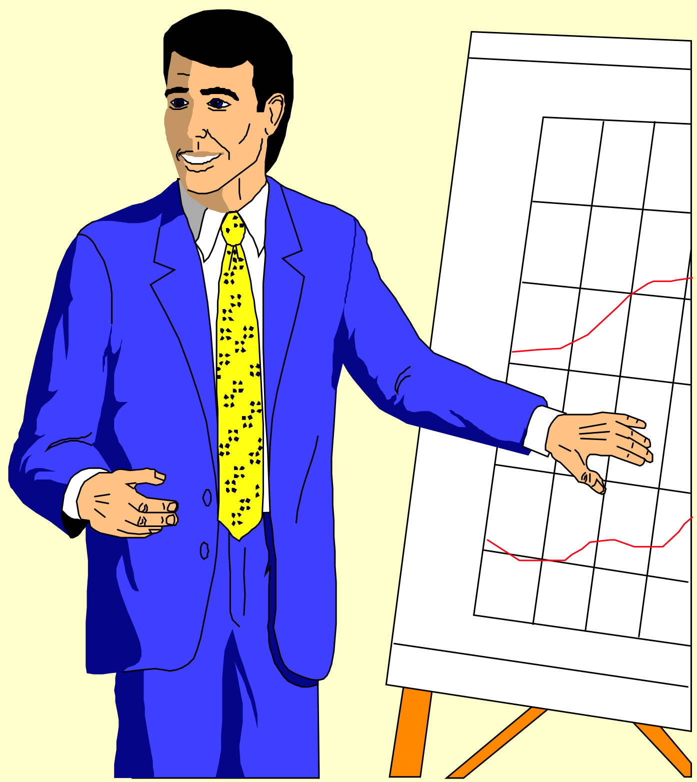
- Pearson's Correlation Coefficient = + 0.76
- Spearman's rho = + 0.39

N = 08 Zones

Regression Analysis

$$\hat{y} = a + b_1x_1 + b_2x_2 + b_3x_3 + \ldots + b_kx_k + \varepsilon$$





STATITICAL DATA ANALYSIS

COMMON TYPES OF ANALYSIS?

1. **Compare Groups**
   a. Compare Proportions (e.g., Chi Square Test—χ^2)
      - ✓ H_0: $P_1 = P_2 = P_3 = \ldots = P_k$

   b. Compare Means (e.g., Analysis of Variance)
      - ✓ H_0: $\mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$

2. **Examine Strength and Direction of Relationships**
   a. Bivariate (e.g., Pearson Correlation—r)
      - ✓ Between one variable and another: $Y = a + b_1 x_1$

   b. **Multivariate (e.g., Multiple Regression Analysis)**
      - ✓ **Between <u>one dep. var</u>. and <u>each of several</u> indep. variables, while holding all other indep. variables constant:**

$$Y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_k x_k$$

Simple and Multiple Regression Analysis

What does regression analysis do?

- Examines whether changes/differences in values of one variable (dependent variable Y) are linked to changes/differences in values of one or more other variables (independent variables X_1, X_2, etc.), **while controlling** for the changes in values of all other Xs.
  - E.g., Relationship between **ACR** and **Police Personals(No -X_1**) and gender X_2 for districts *who have the same levels of education, work experience, position level, seniority, etc.*
- The DV (Y) must be metric.
- The IVs (Xs) must be either metric or non metric var.
- Central Question Addressed:
  - Is Y(ACR) a function of X_1, X_2, etc.? How ?
  - Is there a relationship between Y and X_1, X_2 , etc., (in each case, after controlling for the effects of all other Xs)? In what way?
  - What is the relative impact of each X on Y, holding all other Xs constant (that is, all other Xs being equal)?

Simple and Multiple Regression Analysis

More specifically,

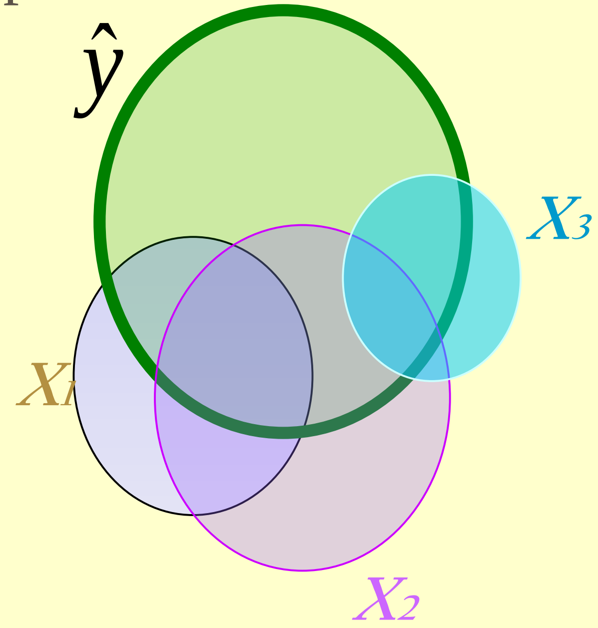
- Do values of Y tend to increase/decrease as values of X_1, X_2, etc. increase/decrease?

  If so,

  - By how much?

  And

  - How strong is the connection/relationship between Xs and Y?
    - what % of differences/variations in Y values (e.g., ACR) among study subjects can be explained by (or attributed to) differences in X values (e.g. years of service, years of their present posting, etc.)?



Simple and Multiple Regression Analysis

- NOTE: Once we can determine how values of Y change as a function of values of X_1, X_2, etc., we will also be able to **predict/estimate** the value of Y from specific values of X_1, X_2, etc.

$$Y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \ldots + b_k x_k + \varepsilon$$

^

- Therefore, regression analysis, in a sense, is about **ESTIMATING values of Y**, using information about values of Xs:
- Estimation, by definition, involves?
- The objective?
  - To minimize error in estimation.
  - Or, to compute estimates that are as close to the true/actual values as possible.

Simple and Multiple Regression Analysis

QUESTION: What is the simplest way to obtain an estimate for some population characteristic
(e.g., number of FIRs, Heinous FIRs etc. per districts)?

ANSWER:

1. **Select a representative sample from the population and**
2. **Compute the mean for that sample (e.g., compute the average number of FIRs for the sample District).**

Regression analysis can be viewed as a technique that often significantly improves the accuracy of estimation results relative to using the mean value.

So, suppose we were to estimate the number of FIRs for a particular district, based on information from a random sample of, say, n = 8 Zones in that district.

Simple and Multiple Regression Analysis

Estimating Number of FIRs*

i Zones	y_i Murder # FIR
1	1698
2	1055
3	1317
4	1095
5	1518
6	1881
7	2176
8	1021

$$\sum Y_i = 11761$$

$\hat{y} =$ **Estimate ?**

$$\hat{y} = \bar{y} = \frac{11761}{8} = 1470.125$$

QUESTION: Can we determine how much error in estimation we are committing by using $\bar{Y} = 1470.125$ as our estimate, for each of these ZNs?

Simple and Multiple Regression Analysis

Estimating Number of FIRs

i Zones	y_i **Murder # FIR**	$\hat{y} = \bar{y}$ **Estimate** for # of FIRs	Error in Estimation
1	**1698**	**1470.125**	**227.875**
2	**1055**	**1470.125**	**-415.125**
3	**1317**	**1470.125**	**-153.125**
4	**1095**	**1470.125**	**-375.125**
5	**1518**	**1470.125**	**47.875**
6	**1881**	**1470.125**	**410.875**
7	**2176**	**1470.125**	**705.875**
8	**1021**	**1470.125**	**-449.125**

$$\sum y_i = 11761 \qquad \hat{y} = \bar{y} = \frac{11761}{8} = 1470.125$$

Simple and Multiple Regression Analysis

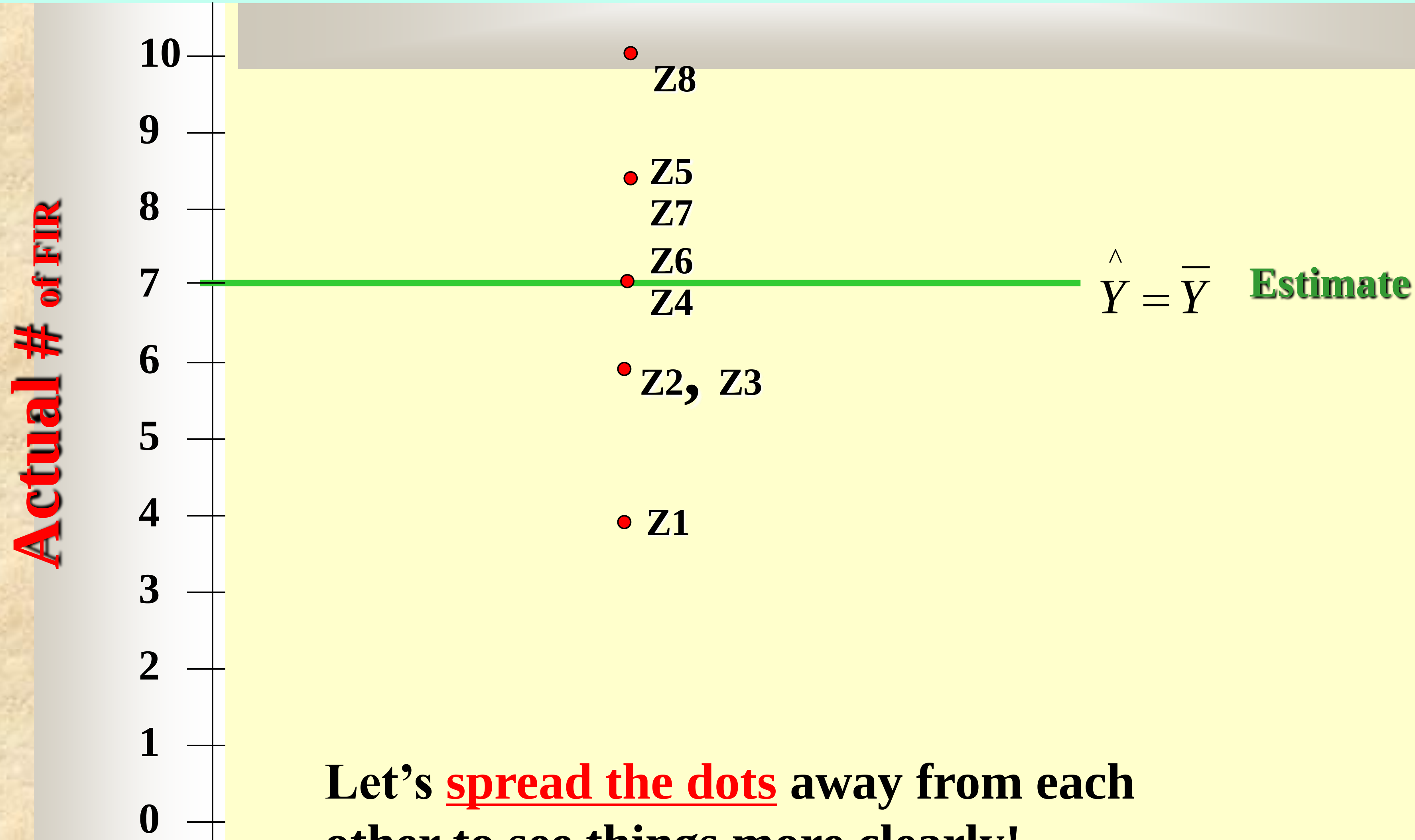
Estimating Number of FIRs

i Zones	y_i **Actual** # of FIRs	$\hat{y} = \bar{y}$ **Estimate** for # of FIRs	$y_i - \bar{y}$ Error in Estimation
1	**1698**	1470.125	227.875
2	**1055**	1470.125	-415.125
3	**1317**	1470.125	-153.125
4	**1095**	1470.125	-375.125
5	**1518**	1470.125	47.875
6	**1881**	1470.125	410.875
7	**2176**	1470.125	705.875
8	**1021**	1470.125	-449.125

$$\sum y_i = 11761 \quad \hat{y} = \bar{y} = \frac{11761}{8} = 1470.125$$

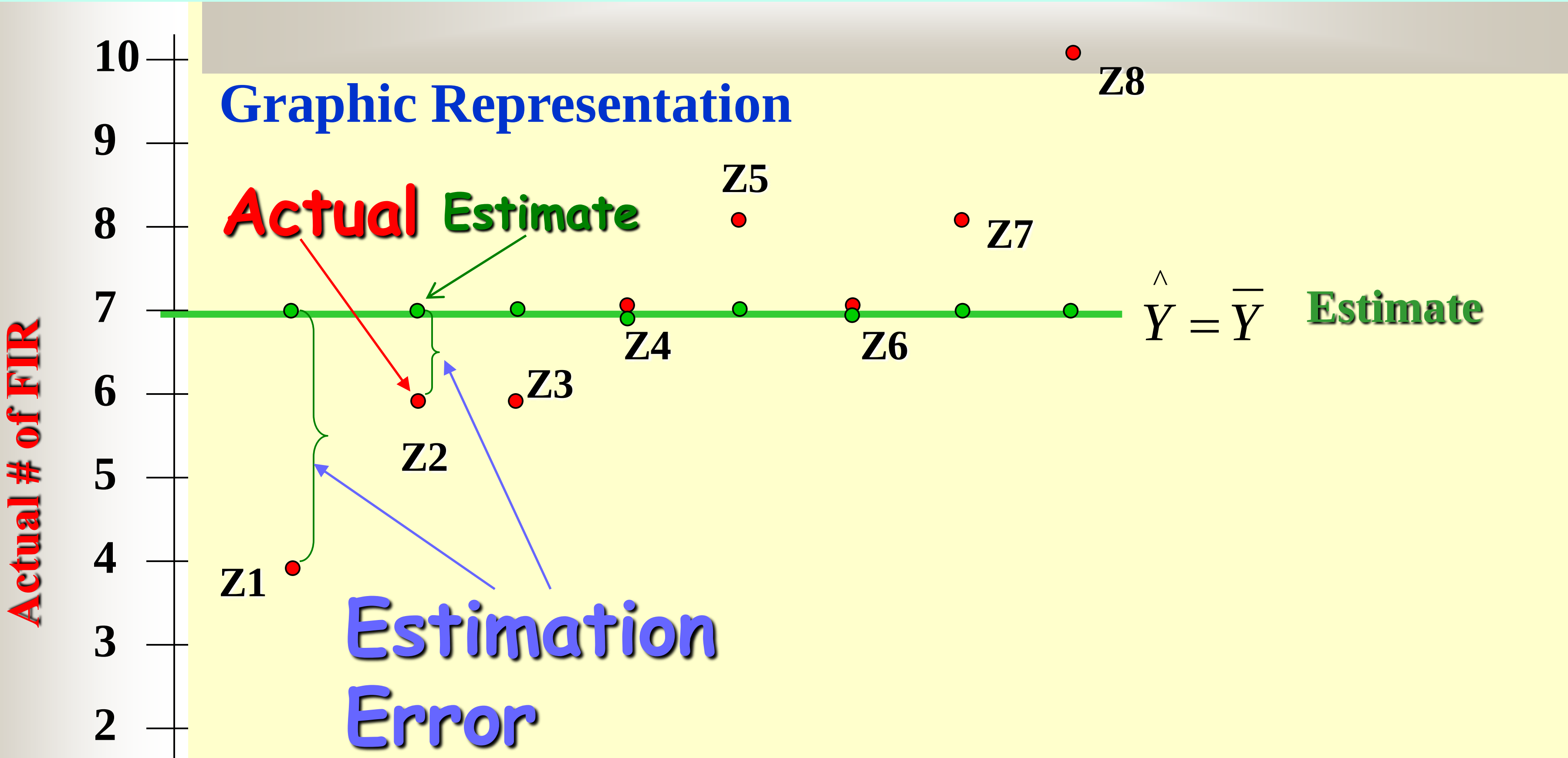
Lets now see all this graphically

Simple and Multiple Regression Analysis



Let's spread the dots away from each other to see things more clearly!

Simple and Multiple Regression Analysis



Can we determine the total estimation error for all 8 PSs?

Simple and Multiple Regression Analysis

i Zones	y_i **Actual** # of FIRs	$\hat{y}=\bar{y}$ **Estimate** for # of FIRs	$y_i-\bar{y}$ Error in Estimation
1	**1698**	1470.125	227.875
2	**1055**	1470.125	-415.125
3	**1317**	1470.125	-153.125
4	**1095**	1470.125	-375.125
5	**1518**	1470.125	47.875
6	**1881**	1470.125	410.875
7	**2176**	1470.125	705.875
8	**1021**	1470.125	-449.125

$\sum y_i = 11761$ $\quad \hat{y}=\bar{y}=\frac{11761}{8}=1470.125$ $\quad \sum(y_i-\bar{y})$ **= 0**

What would be the total estimation error for all 8 ZONEs combined?

Solution?

Simple and Multiple Regression Analysis

Estimating Number of FIRs

i Zones	Actual # of FIRs y_i	Estimate for # of FIRs $\hat{y} = \bar{y}$	Error in Estimation $y_i - \bar{y}$	Errors Squared $(y_i - \bar{y})^2$
1	1698	1470.125	227.875	**51927.02**
2	1055	1470.125	-415.125	**172328.8**
3	1317	1470.125	-153.125	**23447.27**
4	1095	1470.125	-375.125	**140718.8**
5	1518	1470.125	47.875	**2292.016**
6	1881	1470.125	410.875	**168818.3**
7	2176	1470.125	705.875	**498259.5**
8	1021	1470.125	-449.125	**201713.3**

$\sum y_i = 11761$ $\quad \hat{y} = \bar{y} = \frac{11761}{8} = 1470.125$ $\quad \sum (y_i - \bar{y}) = 0$ $\quad \sum (y_i - \bar{y})^2 = 1259505$

SST- Sum of Squares Total

Simple and Multiple Regression Analysis

1259505 **= SST = Index <u>for total (combined) amount of estimation error</u>**

for all Zones (observations) in the sample <u>when using the mean</u>

as the estimate.

- ✓ **SST is also the sum of squared deviations from the mean.**
  - ○ **<u>Remember</u> the formula for computing Variance?**

- **Objective in Estimation?**

  Minimize error, maximize precision.

- **Can we cut down the amount of estimation error (SST)? How?**

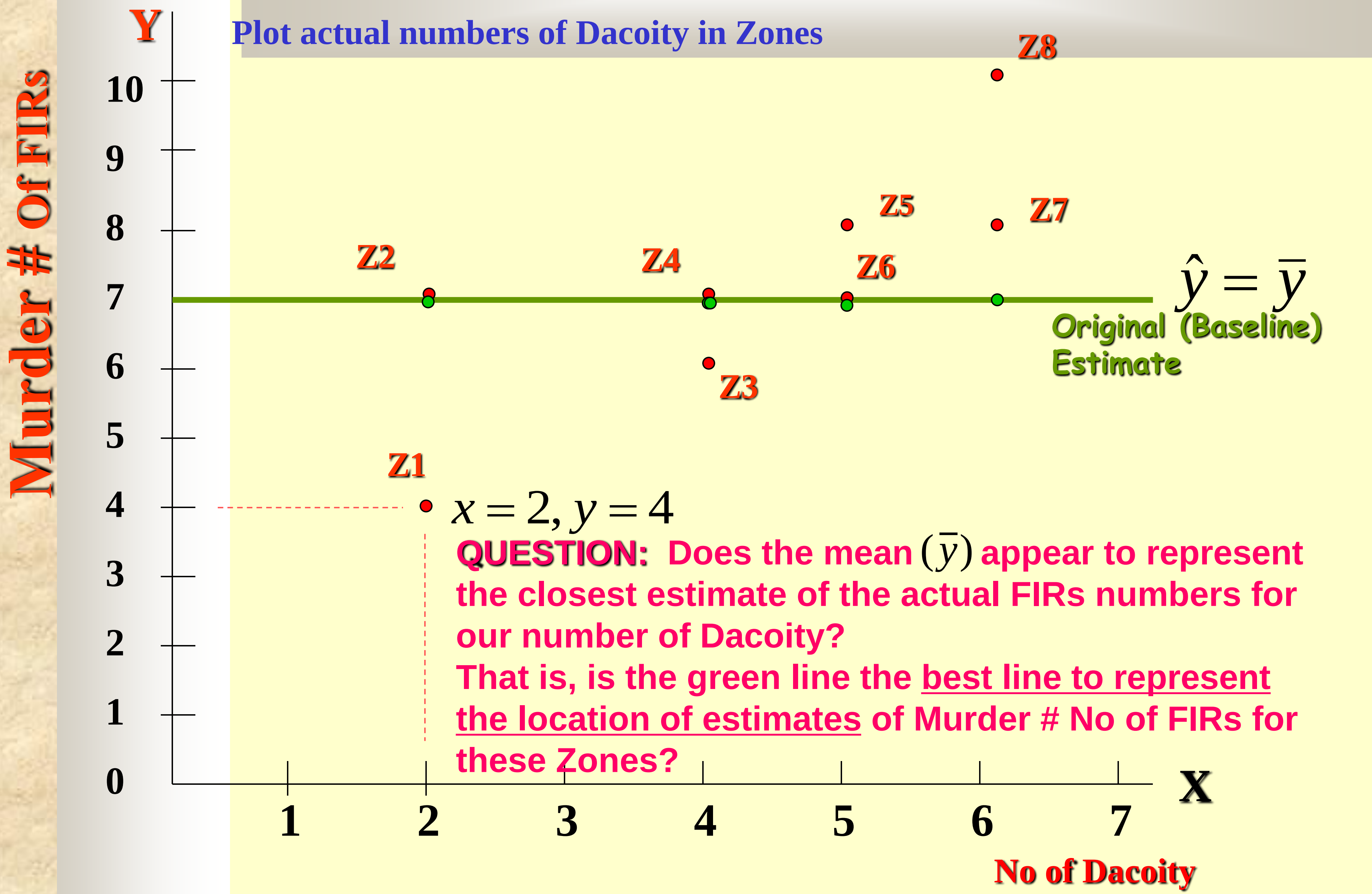
  Yes, we can, by using information about other variables suspected to be strong predictors (strongly related to) # of FIRs possessed by Zones (e.g., FIRs of Dacoity, Rape, Loot etc.)..

i Zones	y **Act ual** Murder# FIRs	x **No of** **FIR # Dacoity**
1	**1698**	**415**
2	**1055**	**220**
3	**1317**	**318**
4	**1095**	**397**
5	**1518**	**514**
6	**1881**	**823**
7	**2176**	**540**
8	**1021**	**194**

We now can attempt to estimate Murder # of FIRs from the information on no of Dacoity, rather than from its own mean.

Let's first see this graphically!

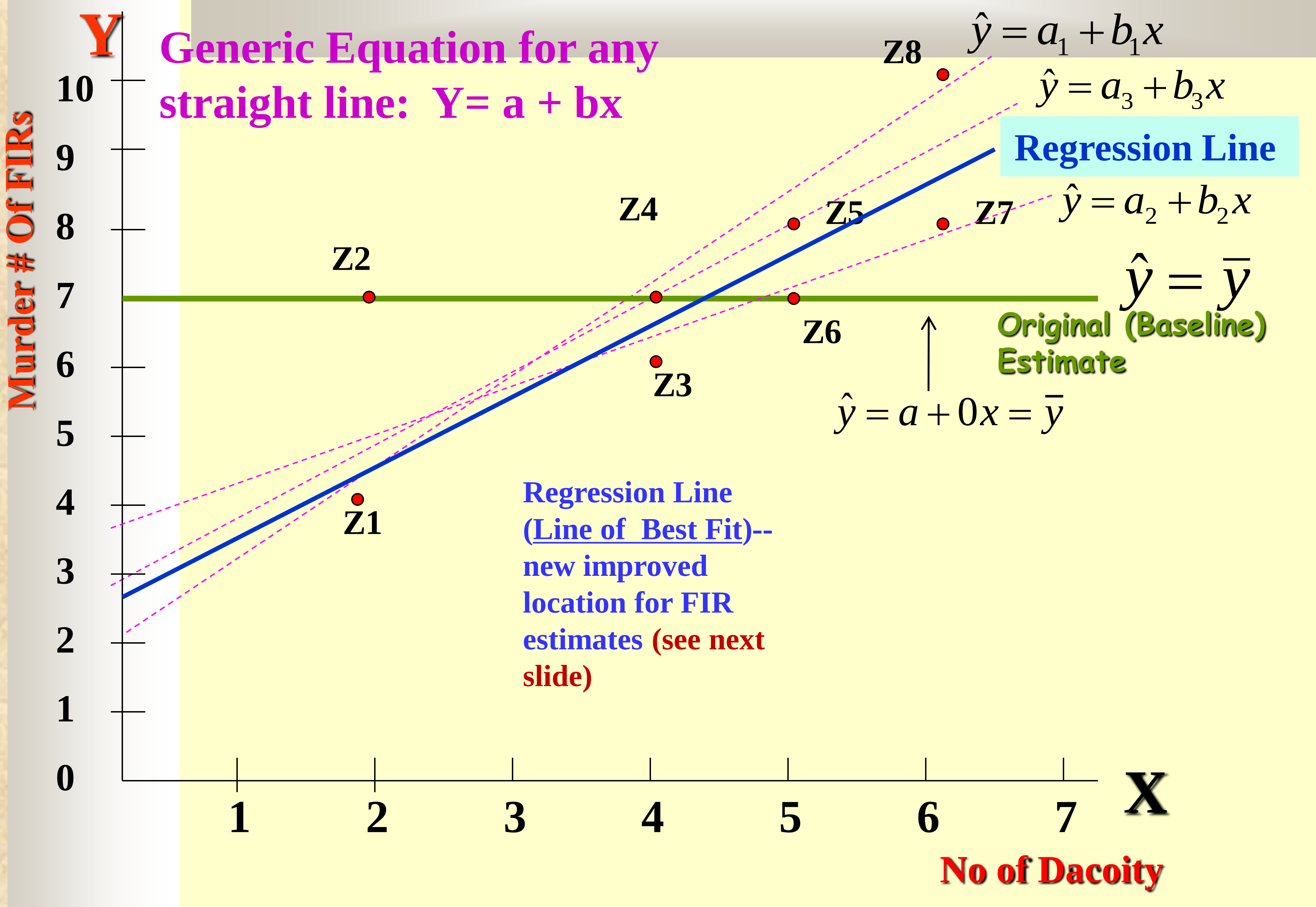
Simple and Multiple Regression Analysis



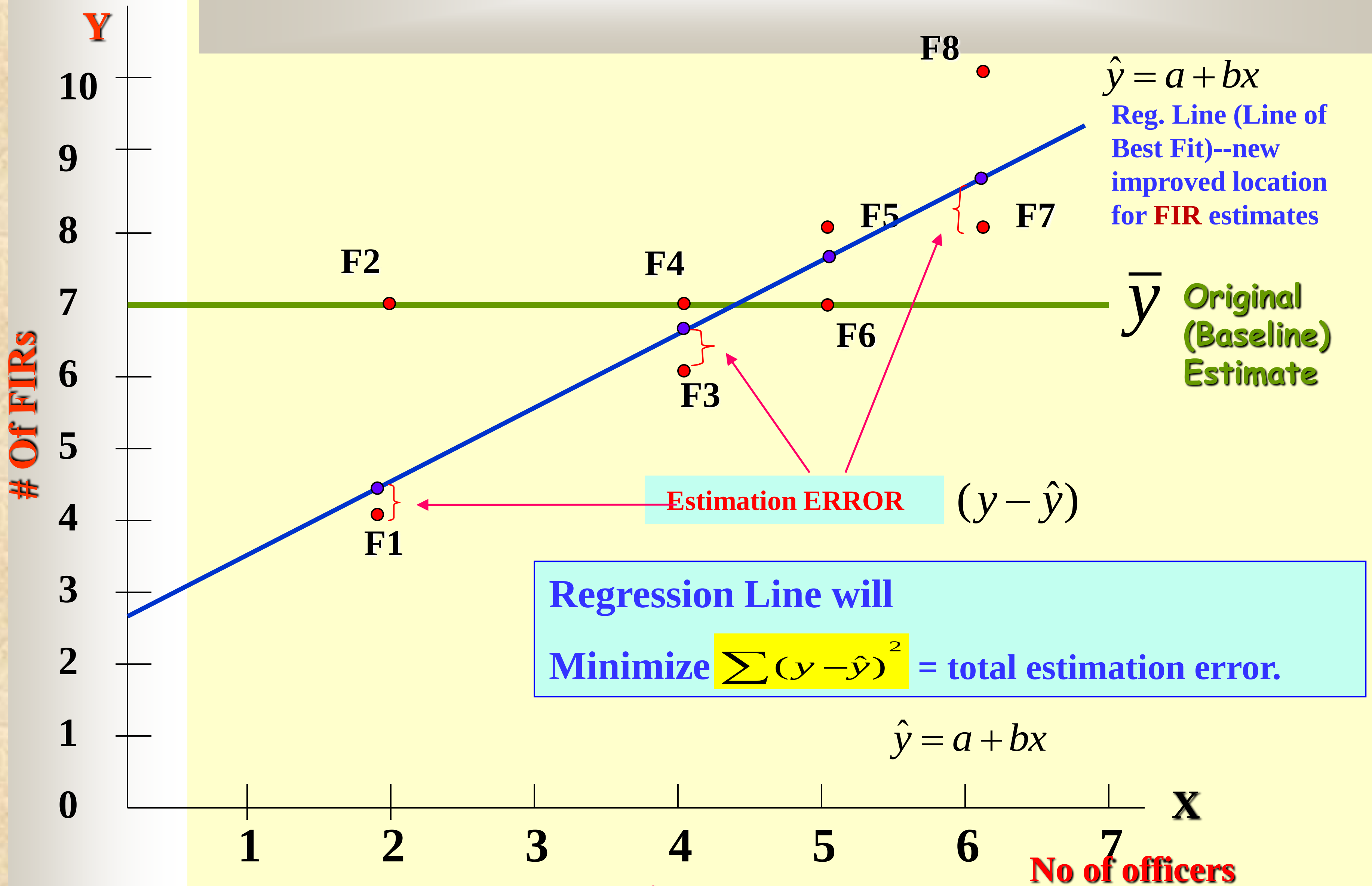
QUESTION: Does the mean $(\bar{y})$ appear to represent the closest estimate of the actual FIRs numbers for our number of Dacoity?

That is, is the green line the <u>best line to represent the location of estimates</u> of Murder # No of FIRs for these Zones?

Simple and Multiple Regression Analysis



Simple and Multiple Regression Analysis



Regression Line will

Minimize $\sum (y - \hat{y})^2$ **= total estimation error.**

But, how do we know the values *a* and *b* in (the reg. line)?

Actual Murders # of FIRs

EQUATION FOR REGRESSION LINE (LINE OF BEST FIT)—

Values of *a* and *b for the regression line*:

$$\hat{y} = a + bx \quad \begin{cases} b = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \\ a = \bar{y} - b\bar{x} \end{cases}$$

Let's use above formulas to compute the values of "*a*" and "*b*" for the regression line in our example.

We will need: $\bar{y}$, $\bar{x}$, $\Sigma(x-\bar{x})(y-\bar{y})$, **and** $\Sigma(x-\bar{x})^2$

Simple and Multiple Regression Analysis

We need: $\bar{y}$, $\bar{x}$, $\sum(x-\bar{x})(y-\bar{y})$, **and** $\sum(x-\bar{x})^2$

i Zones	y **Murder Actual #** FIRs	x **No Of Dacoity**	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	**1698**	**415**	**?**	**?**	**?**	**?**
2	**1055**	**220**	**?**	**?**	**?**	**?**
3	**1317**	**318**	**?**	**?**	**?**	**?**
4	**1095**	**397**	**?**	**?**	**?**	**?**
5	**1518**	**514**	**?**	**?**	**?**	**?**
6	**1881**	**823**	**?**	**?**	**?**	**?**
7	**2176**	**540**	**?**	**?**	**?**	**?**
8	**1021**	**194**	**?**	**?**	**?**	**?**

$\bar{Y} = \frac{11761}{8} = 1470.125$ $\bar{x} = \frac{3421}{8} = 427.625$ $\sum(x-\bar{x})(y-\bar{y}) = ?$ $\sum(x-\bar{x})^2 = ?$

Simple and Multiple Regression Analysis

We need: $\bar{y}$, $\bar{x}$, $\sum(x-\bar{x})(y-\bar{y})$, **and** $\sum(x-\bar{x})^2$

i Zones	y **Murder Actual #** FIRs	x **No Of Dacoity**	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
1	**1698**	**415**	-12.625	227.875	-2876.921875	159.390625
2	**1055**	**220**	-207.625	-415.125	86190.32813	43108.14063
3	**1317**	**318**	-109.625	-153.125	16786.32813	12017.64063
4	**1095**	**397**	-30.625	-375.125	11488.20313	937.890625
5	**1518**	**514**	86.375	47.875	4135.203125	7460.640625
6	**1881**	**823**	395.375	410.875	162449.7031	156321.3906
7	**2176**	**540**	112.375	705.875	79322.70313	12628.14063
8	**1021**	**194**	-233.625	-449.125	104926.8281	54580.64063

$\bar{Y} = \frac{11761}{8} = 1470.125$ $\bar{x} = \frac{3421}{8} = 427.625$ $\sum(x-\bar{x})(y-\bar{y}) = 462422.375$

$\sum(x-\bar{x})^2 = 287213.875$

Simple and Multiple Regression Analysis

REGRESSION LINE (LINE OF BEST FIT):

$$\hat{y} = a + b\underline{x} \begin{cases} b = \dfrac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \dfrac{462422.375}{287213.875} = 1.61 \\ a = \bar{y} - b\bar{x} = 1470.125 - 1.61(427.625) = 781.640 \end{cases}$$

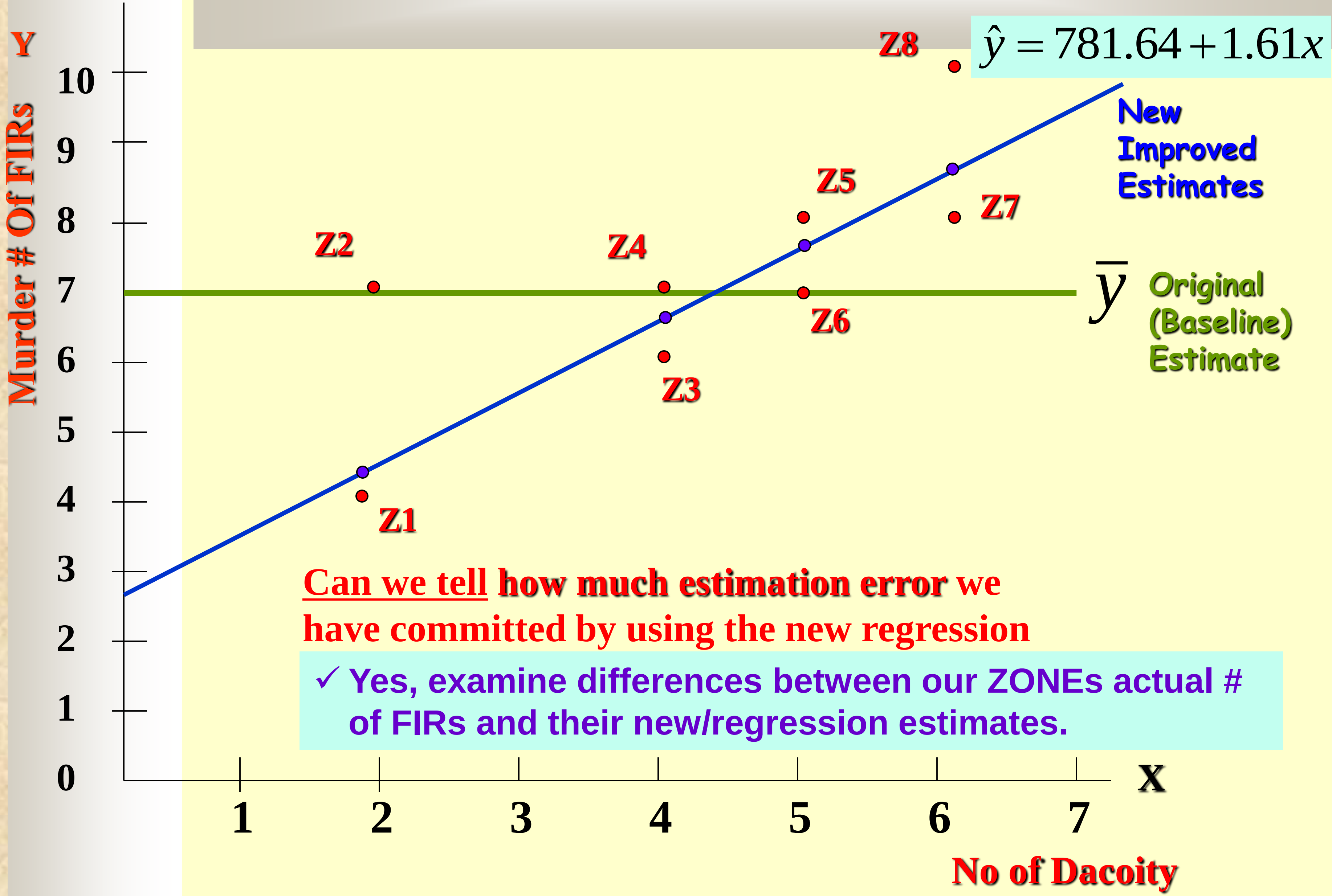
a =781.64 b = 1.61

$$\hat{y} = 781.64 + 1.61x$$

? ?

Y-Intercept **Regression Coefficient**

Simple and Multiple Regression Analysis



Can we tell **how much estimation error** we have committed by using the new regression

✓ Yes, examine differences between our ZONEs actual # of FIRs and their new/regression estimates.

Simple and Multiple Regression Analysis

$\hat{y} = 781.648 + 1.61x$

$\hat{y}$

i Zones	y **Murder Actual #** FIRs	x **No of Dacoity**	$\hat{y}$ Regression Estimate	$y-\hat{y}$ Error (Residual)	$(y-\hat{y})^2$ Errors Squared
1	**1698**	**415**	**?**	**?**	**?**
2	**1055**	**220**	**?**	**?**	**?**
3	**1317**	**318**	**?**	**?**	**?**
4	**1095**	**397**	**?**	**?**	**?**
5	**1518**	**514**	**?**	**?**	**?**
6	**1881**	**823**	**?**	**?**	**?**
7	**2176**	**540**	**?**	**?**	**?**
8	**1021**	**194**	**?**	**?**	**?**

$\Sigma(y-\hat{y})^2$

Simple and Multiple Regression Analysis

$\hat{y} = 781.648 + 1.61x$

$\hat{y} = 781.648 + 1.61(415) = 1449.798$

i Zones	y **Murder Actual #** FIRs	x No of Dacoity	$\hat{y}$ Regression Estimate	$y - \hat{y}$ Error (Residual)	$(y - \hat{y})^2$ Errors Squared
1	**1698**	**415**	1449.798	248.202	61604.23
2	**1055**	**220**	1135.848	-80.848	6536.39
3	**1317**	**318**	1293.628	23.372	546.25
4	**1095**	**397**	1420.818	-325.818	106157.37
5	**1518**	**514**	1609.188	-91.188	8315.25
6	**1881**	**823**	2106.678	-225.678	50930.56
7	**2176**	**540**	1651.048	524.952	275574.60
8	**1021**	**194**	1093.988	-72.988	5327.25

$514991.9 = \sum (y - \hat{y})^2$

SSE = Sum of Squares Error (SS Residual)

Simple and Multiple Regression Analysis

Total Baseline Error using the mean (SS Total) 1259505

New or Remaining Error (SS Error or SS Residual) 514991

QUESTION: How much of the original estimation error have we explained away (eliminated) by using the regression model (instead of the mean)?

1259505-514991= **744514** (SS **Regression** or SS Explained)

QUESTION: What % of estimation error have we explained (eliminated by using the regression model?

R^2 = 744514 / 1259505 = 0.591 or **60%** What is this called?

% of differences in # of FIRs among ZONEs that is explained by differences in their No of dacoity.

What does the remaining 40% represent?

Percent of variation (differences) in number of FIRs owned by Zones that can be accounted for by: (a) all other potential predictors not included in the model, beyond No of dacoity, and (b) unexplainable random/chance variations.

Simple and Multiple Regression Analysis

R^2 = SS Regression / SS Total = 0.591 = 60%

R^2 is a measure of our success regarding accuracy of our estimation effort.

- ✓ R^2 = <u>% of estimation error</u> that we have been able to <u>explain away</u> by using the regression model, instead of using the mean.
- ✓ R^2 indicates <u>how much better we can predict Y</u> from information about Xs, rather than from using its own mean.
- ✓ R^2 = % of differences (variations) in Y values that is explained by (attributable to) differences in X values.

Note: When dealing with <u>only two variables</u> (<u>a single X</u> and Y):

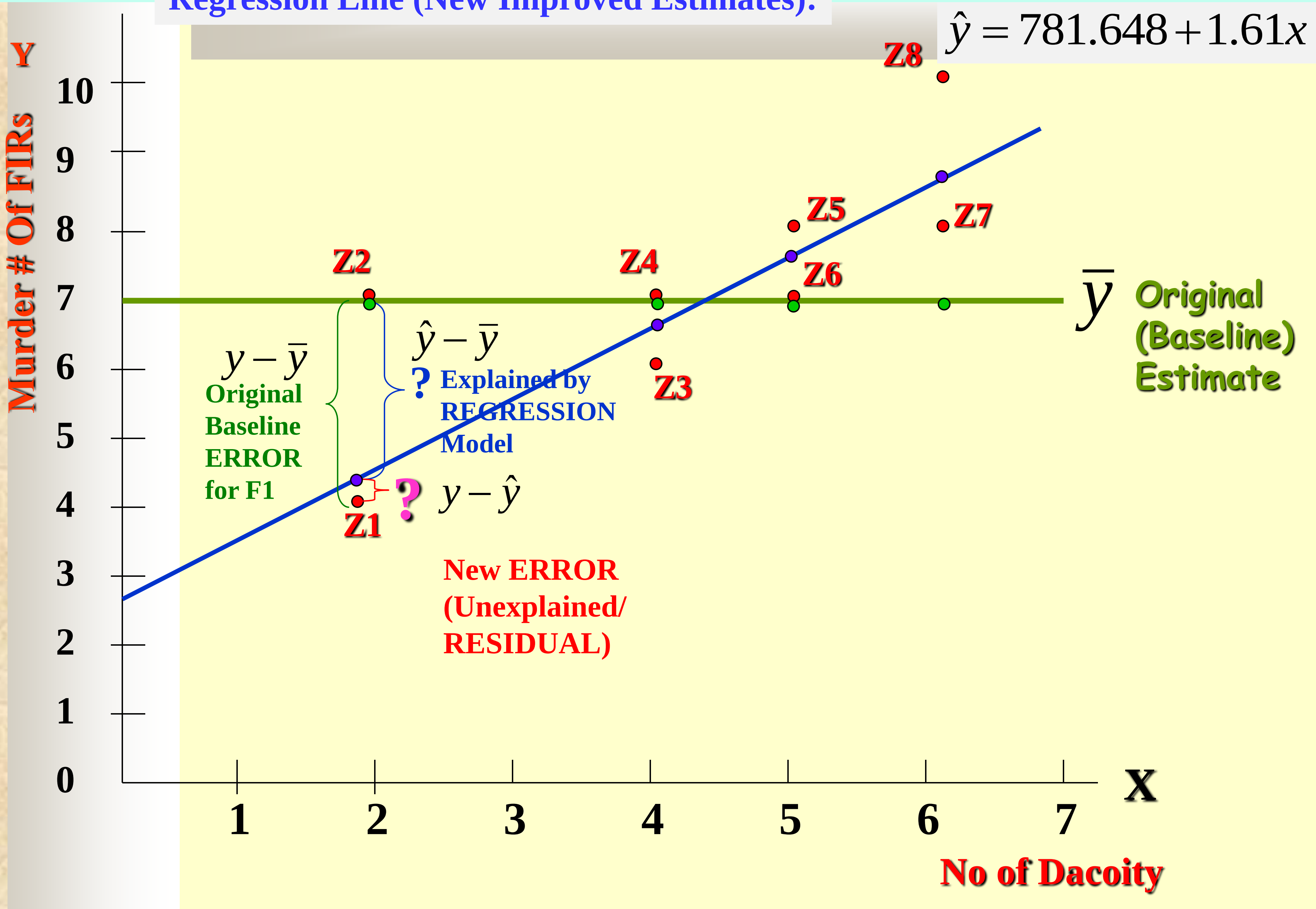
$$r = \sqrt{R^2} = \sqrt{0.591} = 0.769$$

Pearson Correlation of Y with X_1 (NOT controlling for any other var.)

Let's now examine all this graphically!

Simple and Multiple Regression Analysis

Regression Line (New Improved Estimates):



Simple and Multiple Regression Analysis

SSE = The amount of estimation error for the 8 ZONEs when using simple regression (i.e., a regression model that includes only information about No of Dacoity).

Can we reduce the amount of estimation error (SSE) to an even lower level and, thus, improving the estimation process? How?

Yes, by adding information on a second variables suspected to be strongly related to Murder # of FIRs (e.g., No of Rape Cases-X_2).

•

I ZONEs	y_i **Murder** **Actual** # FIRs	x_1 **No of Dacoity**	x_2 **No of Rape Cases**
1	**1698**	**415**	**2984**
2	**1055**	**220**	**2064**
3	**1317**	**318**	**2144**
4	**1095**	**397**	**4074**
5	**1518**	**514**	**4653**
6	**1881**	**823**	**4374**
7	**2176**	**540**	**4383**
8	**1021**	**194**	**2340**

We now can attempt to estimate Murder # of FIRs from our information on No of Dacoity and No of Rape cases!

Our regression model will now be a linear plane, rather than a straight line!

Generic Equation for a linear plane: $\hat{y} = a + b_1 x_1 + b_2 x_2$

Let's examine the regression plane for our example graphically.

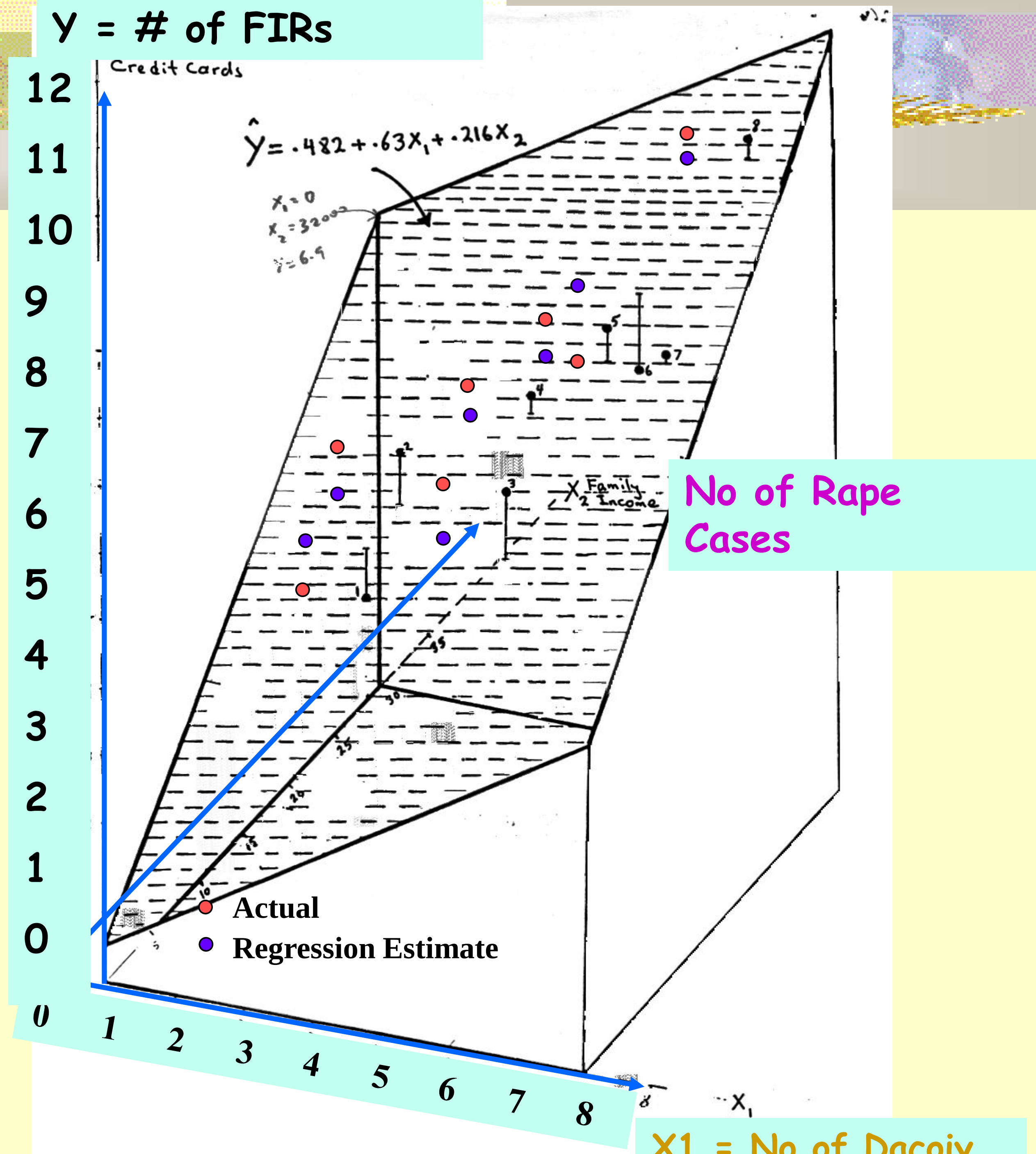
$$\hat{y} = a + b_1 x_1 + b_2 x_2$$

Formulas are available for computing values of a, b_1 and b_2

MULTIPLE REGRESSION MODEL FOR OUR EXAMPLE:

$$\hat{y} = 774.367 + .76 x_1 + 0.013 x_2$$

Let's now see how much error in estimation we are committing by using this multiple regression model.



Graphic Representation of the Multiple Regression Model

Simple and Multiple Regression Analysis

$$\hat{y} = 774.367 + .76x_1 + .013x_2$$

$\hat{y}$

i PSs	y **Actual #** FIR	x_1 **No of Dacoity**	x_2 **No of Rape Cases**	$\hat{Y}$ Regression Estimate	$y - \hat{y}$ Error (Residual)	$(y - \hat{y})^2$ Errors Squared
1	**1698**	**415**	**2984**	**?**	**?**	**?**
2	**1055**	**220**	**2064**	**?**	**?**	**?**
3	**1317**	**318**	**2144**	**?**	**?**	**?**
4	**1095**	**397**	**4074**	**?**	**?**	**?**
5	**1518**	**514**	**4653**	**?**	**?**	**?**
6	**1881**	**823**	**4374**	**?**	**?**	**?**
7	**2176**	**540**	**4383**	**?**	**?**	**?**
8	**1021**	**194**	**2340**	**?**	**?**	**?**

$$\sum (y - \hat{y})^2$$

Simple and Multiple Regression Analysis

$\hat{y} = 774.367 + .76x_1 + .013x_2$ $\hat{y} = 774.367 + .76x_1 + .013x_2$

i Zones	y_i **Murder Actual** # FIRs	x_1 **No of Dacoity**	x_2 **No of Rape Cases**	$\hat{Y}$ Regression Estimate	$y - \hat{y}$ Error (Residual)	$(y - \hat{y})^2$ Errors Squared
1	**1698**	**415**	**2984**	1089.78	608.22	369931.57
2	**1055**	**220**	**2064**	1209.89	-154.89	23990.91
3	**1317**	**318**	**2144**	1045.22	271.78	73864.37
4	**1095**	**397**	**4074**	1129.05	-34.05	1159.40
5	**1518**	**514**	**4653**	1225.50	292.50	85556.25
6	**1881**	**823**	**4374**	1456.71	424.29	184289.90
7	**2176**	**540**	**4383**	1241.75	934.25	872823.06
8	**1021**	**194**	**2340**	952.23	68.77	4729.31

SSE = Sum of Squares Error (Residual) $\longrightarrow$ $1616344.77 = \sum(y - \hat{y})^2$

Unique (additional) contribution of X_2 (No of Rape cases) beyond X_1 = ?

Simple and Multiple Regression Analysis

The MULTIPLE REGRESSION MODEL FOR OUR EXAMPLE:

$$\hat{y} = 774.367 + 0.76x_1 + 0.013x_2$$

?

?

Y-Intercept, "*a*"

(NOTE: Only when all Xs can meaningfully take on value of zero, the intercept will have a meaningful/direct/ practical interpretation. Otherwise, it is simply an aid in increasing accuracy of estimation.

b_1 and b_2 = Regression Coefficients

0.76: Among ZONEs , an increase in number of Dacoity by one would, on average, result in .76 more Murde FIRs

0.013: Among ZONEs , number of Rape cases increase by 1, results in an average increase of 0.013 Rape FIRs.

"*b*"s represent effect of each X on Y when all other Xs are controlled for/held constant/taken into account

- **i.e., after impacts of all other variables are accounted for (*remember the high blood pressure-hearing problem connection*?)**

Simple and Multiple Regression Analysis

The MULTIPLE REGRESSION MODEL FOR OUR EXAMPLE:

$$\hat{y} = 774.367 + 0.76x_1 + 0.013x_2$$

what is our new R^2?

R^2 = 0.64 or 64%

Percent of differences in ZONEs' number of Murder FIRs that is explained by differences in No of Dacoity and number of Rape cases FIRs

The Remaining 36%?

Percent of variation in number of FIRs that can be accounted for by (a) <u>all other relevant factors</u> not included in the model, beyond No of Dacoity and Rape cases FIRs and (b) <u>unexplainable random/chance</u> variations.

Simple and Multiple Regression Analysis

I ZONEs	y_i **Murder Actual** # FIRs	x_1 **No of Dacoity**	x_2 **No of Rape Cases**	x_3 **No of Loot Cases**
1	**1698**	**415**	**2984**	**3379**
2	**1055**	**220**	**2064**	**2512**
3	**1317**	**318**	**2144**	**2371**
4	**1095**	**397**	**4074**	**2878**
5	**1518**	**514**	**4653**	**3167**
6	**1881**	**823**	**4374**	**5397**
7	**2176**	**540**	**4383**	**5121**
8	**1021**	**194**	**2340**	**3052**

We now can attempt to estimate Murder # of FIRs from our information on No of Dacoity , No of Rape cases and no of Loots!

Our regression model will now be a 4D figure rather than a straight line!

Generic Equation for a linear plane: $$\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

Simple and Multiple Regression Analysis

The MULTIPLE REGRESSION MODEL FOR OUR EXAMPLE:

$$\hat{y} = 392.195 + 0.17x_1 + 0.023x_2 + 0.694x_3$$

what is our new R^2?

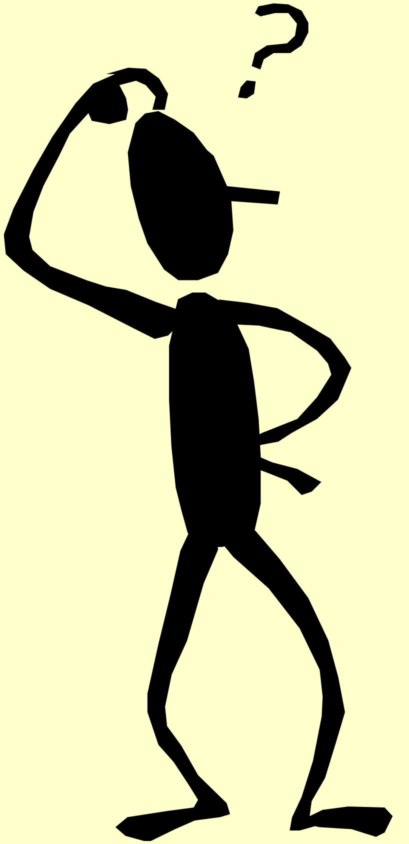
R^2 = 0.749 or 75%

Percent of differences in ZONEs' number of Murder FIRs that is explained by differences in No of Dacoity, number of Rape cases FIRs and no of FIRs in Loot……..

The Remaining 25%?

Percent of variation in number of FIRs that can be accounted for by (a) <u>all other relevant factors</u> not included in the model, beyond No of Dacoity and Rape cases FIRs and (b) <u>unexplainable random/chance</u> variations.

QUESTIONS OR COMMENTS?



Thank You