Linear Regression Analysis

Dr. Linta Rose

rose.l@incois.gov.in

Recall: Covariance

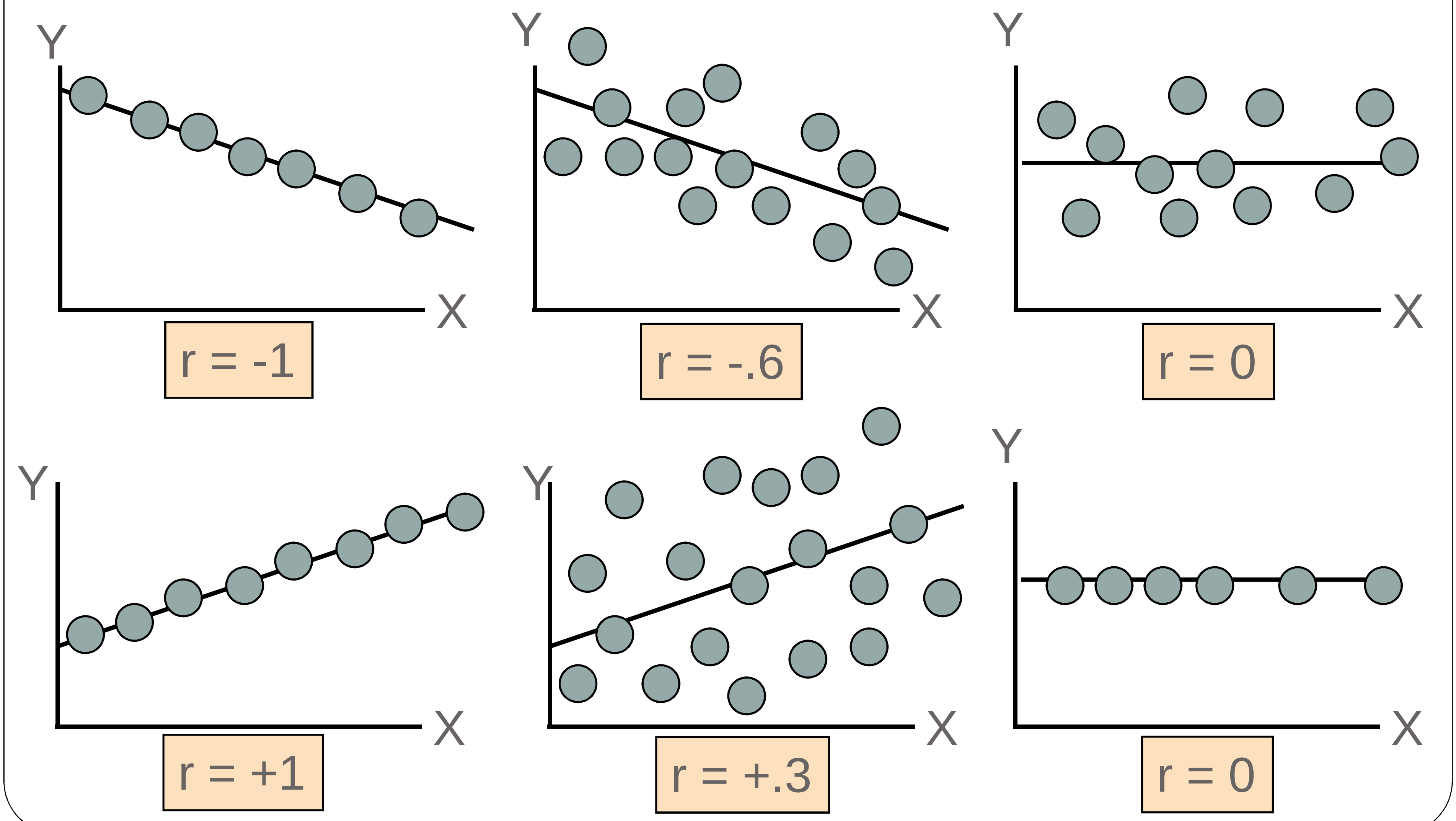
$$\operatorname{cov}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{X})(y_i-\bar{Y})}{n-1}$$

Correlation coefficient

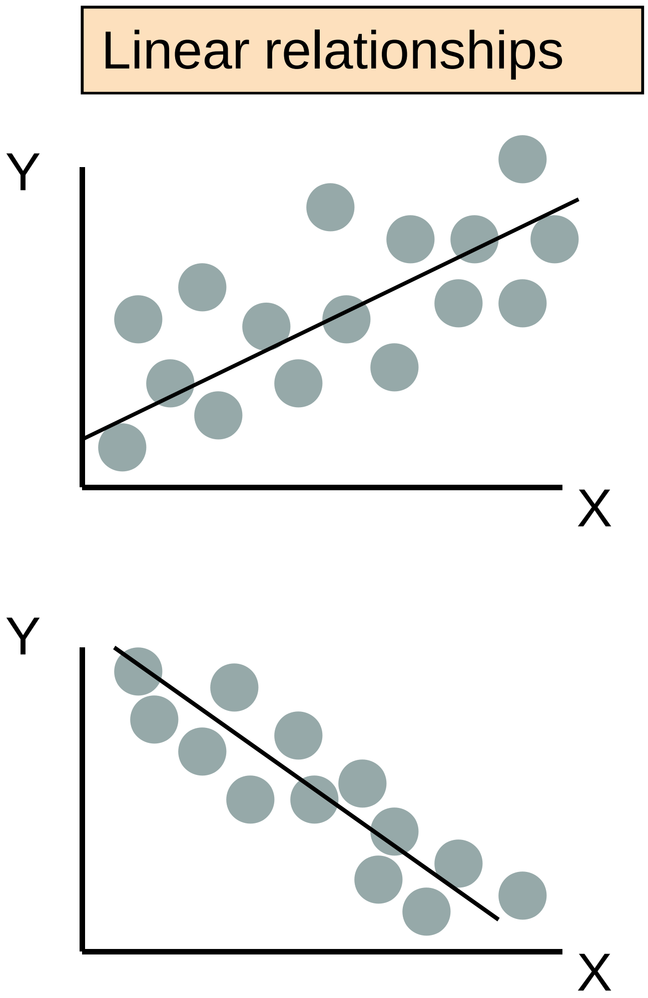
- Pearson's Correlation Coefficient is standardized covariance (unitless):

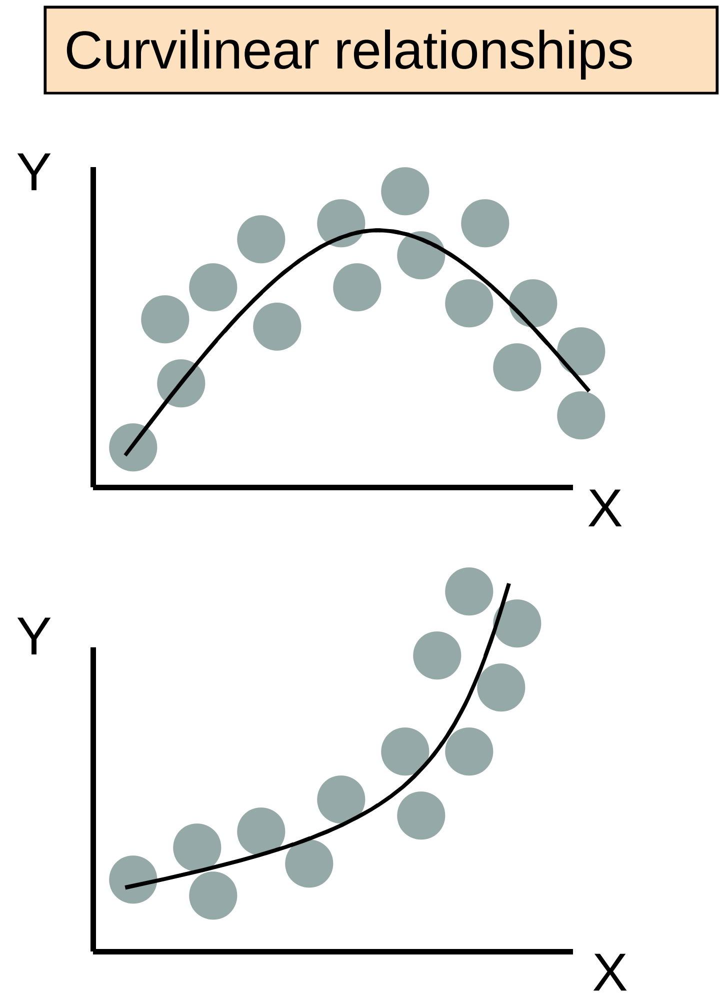
$$r=\frac{\operatorname{cov}ariance(x,y)}{\sqrt{\operatorname{var}x}\sqrt{\operatorname{var}y}}$$

Scatter Plots of Data with Various Correlation Coefficients

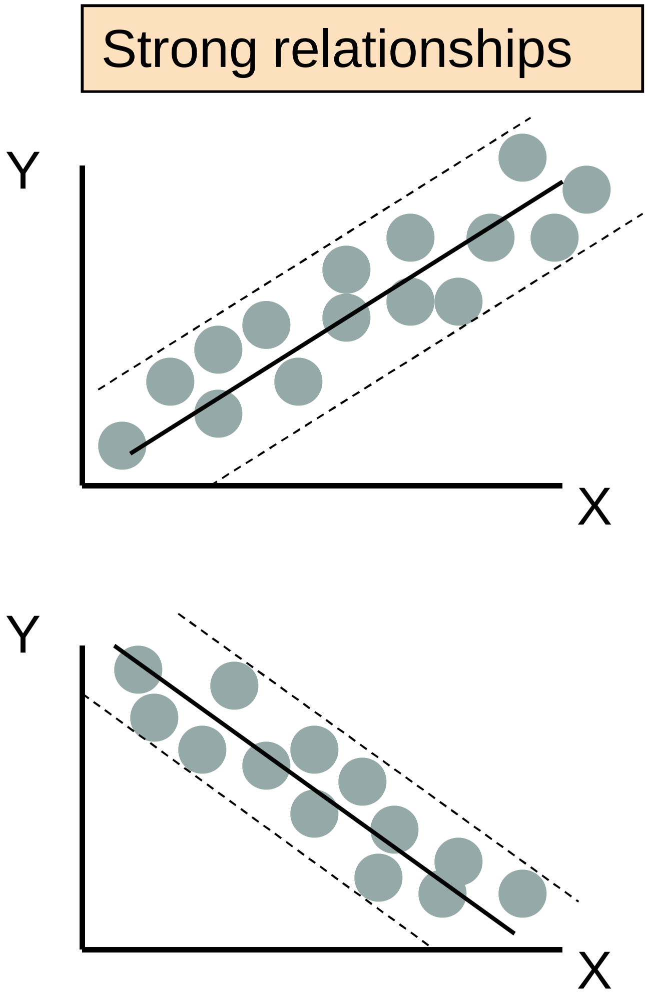


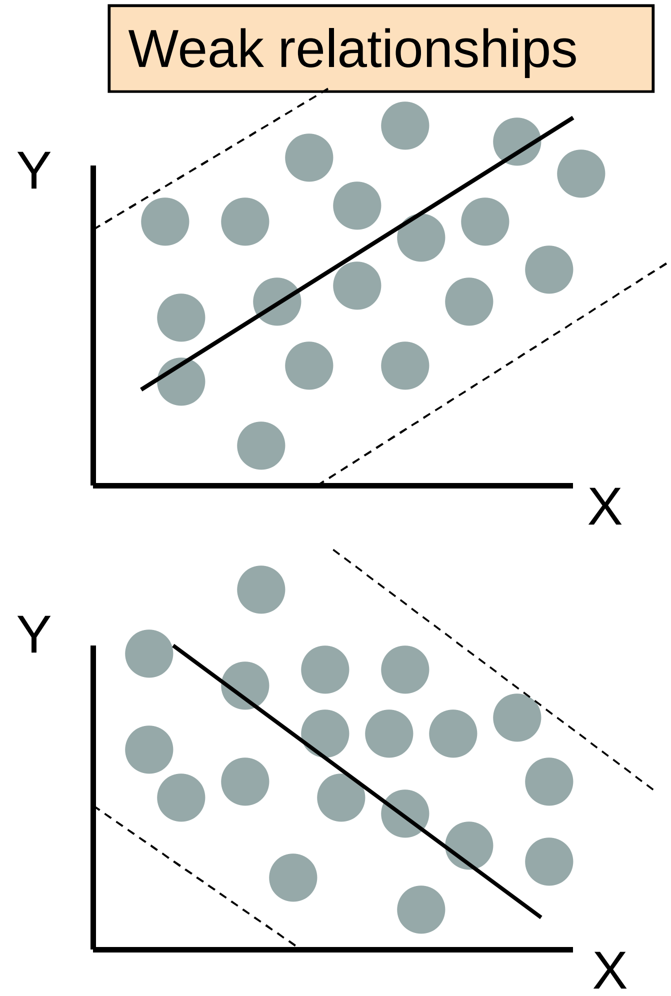
Linear Correlation



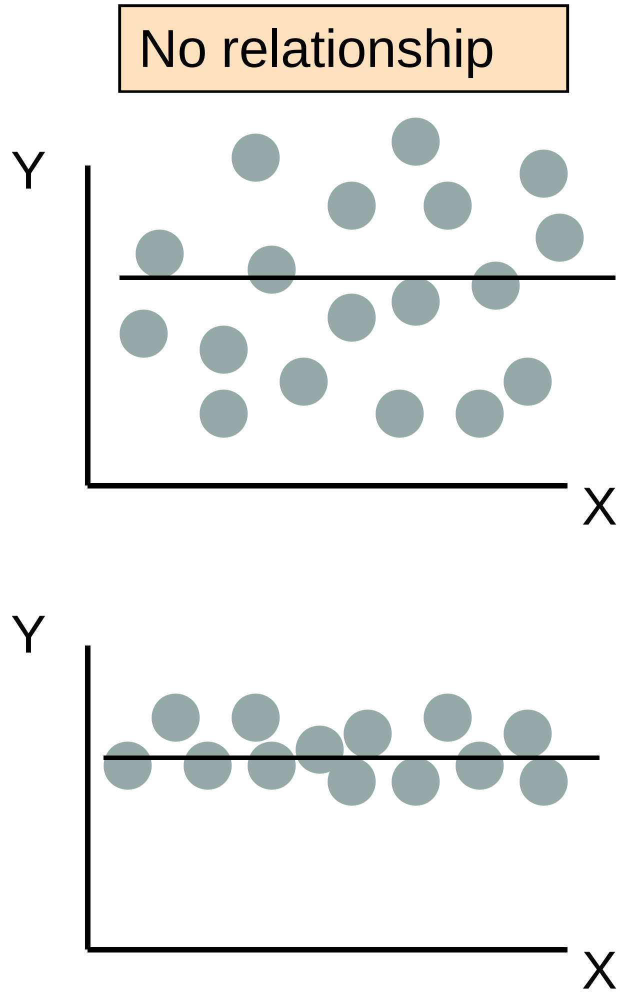


Linear Correlation





Linear Correlation



Linear regression

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*.

Prediction

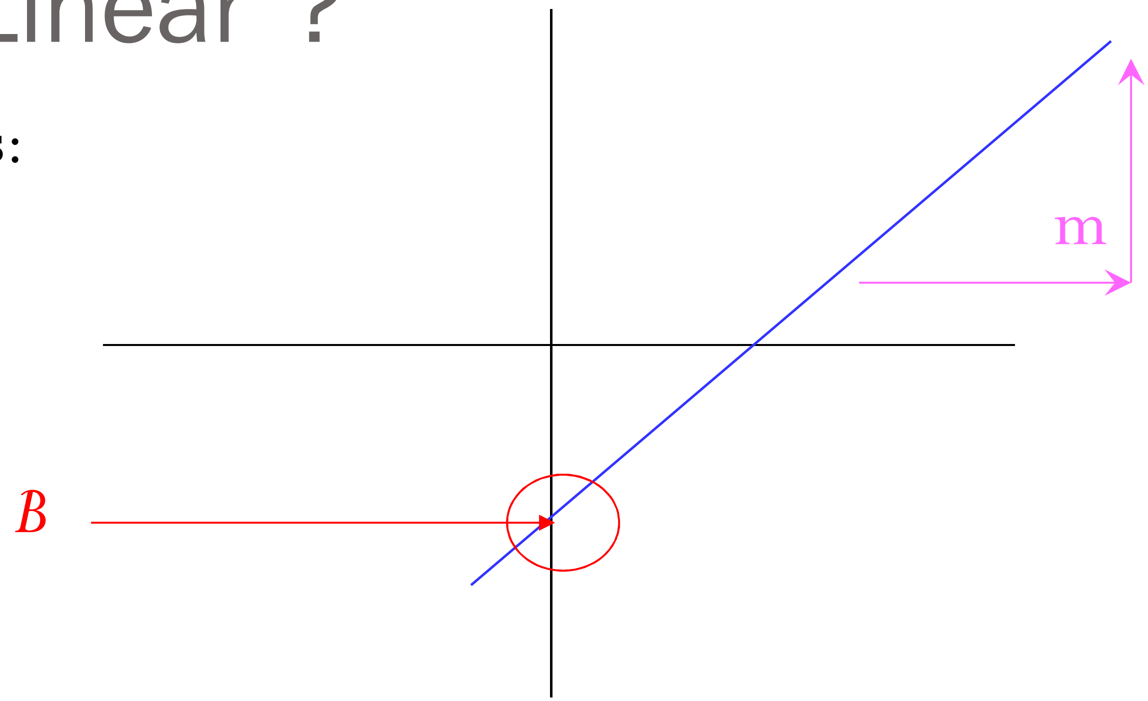
If you know something about X, this knowledge helps you predict something about Y.

Uses of Regression Analysis

- Regression analysis serves Three major purposes.
  1. Description
  2. Control
  3. Prediction
- The several purposes of regression analysis frequently overlap in practice

What is "Linear"?

- Remember this:
- $Y = mX + B$?



What's Slope?

A slope of 2 means that every 1-unit change in X yields a 2-unit change in Y.

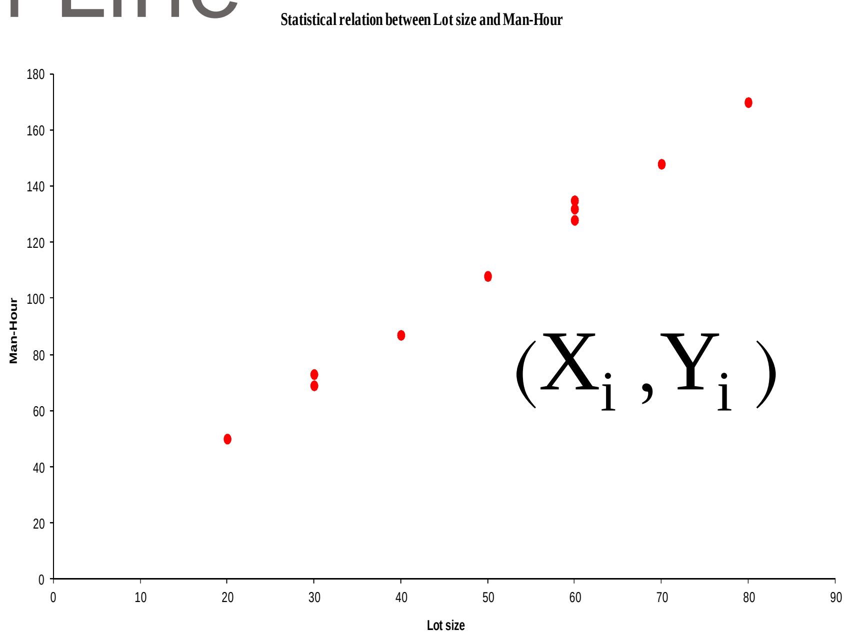
Predicted value for an individual…

$$\underbrace{\hat{y} = b_0 + b_1 x}_{\text{Fixed – exactly on the line}} + \boxed{\text{random error}_i}$$

Follows a normal distribution

- The values of the regression parameters b_0, and b_1 are not known. We estimate them from 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

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

Minimise the sum of square of errors

- Using Calculus

$$\frac{\partial(SSE)}{\partial b_0} = 0 \qquad \frac{\partial(SSE)}{\partial b_1} = 0$$

- Solve for b_0, and b_1 to get the position of the line

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

or

$$b_1 = r\frac{S_y}{S_x} \qquad b_0 = \overline{y} - b_1\overline{x}$$

The Fit Parameters

Define sums of squares:

$$S_x = \Sigma(x_i - \bar{x})^2$$

$$S_y = \Sigma(y_i - \bar{y})^2$$

The quality of fit is parameterized by r^2 the correlation coefficient

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

Estimation of Mean Response

- Fitted regression line can be used to estimate the mean value of y for a given value of x.
- Example
  - The weekly advertising expenditure (x) and weekly sales (y) are presented in the following table.

y	x
1250	41
1380	54
1425	63
1425	54
1450	48
1300	46
1400	62
1510	61
1575	64
1650	71

Point Estimation of Mean Response

- From previous table we have:

$$n = 10 \qquad \sum x = 564 \qquad \sum x^2 = 32604$$
$$\sum y = 14365 \qquad \sum xy = 818755$$

- The least squares estimates of the regression coefficients are:

$$b_1 = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - (\sum x)^2} = \frac{10(818755) - (564)(14365)}{10(32604) - (564)^2} = 10.8$$

$$b_0 = 1436.5 - 10.8(56.4) = 828$$

Point Estimation of Mean Response

- The estimated regression function is:

$$\hat{y} = 828 + 10.8x$$

$$\text{Sales} = 828 + 10.8\ \text{Expenditur e}$$

- This means that if the weekly advertising expenditure is increased by $1 we would expect the weekly sales to increase by $10.8.

Point Estimation of Mean Response

- Fitted values for the sample data are obtained by substituting the x value into the estimated regression function.
- For example if the advertising expenditure is $50, then the estimated Sales is:

$$Sales = 828 + 10.8(50) = 1368$$

- This is called the point estimate (forecast) of the mean response (sales).

Residual

- The difference between the observed value y_i and the corresponding fitted value $\hat{y}_i$

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

- Residuals are highly useful for studying whether a given regression model is appropriate for the data at hand.

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

Regression Standard Error

- Approximately 95% of the observations should fall within plus/minus 2*standard error of the regression from the regression line, which is also a quick approximation of a 95% prediction interval.

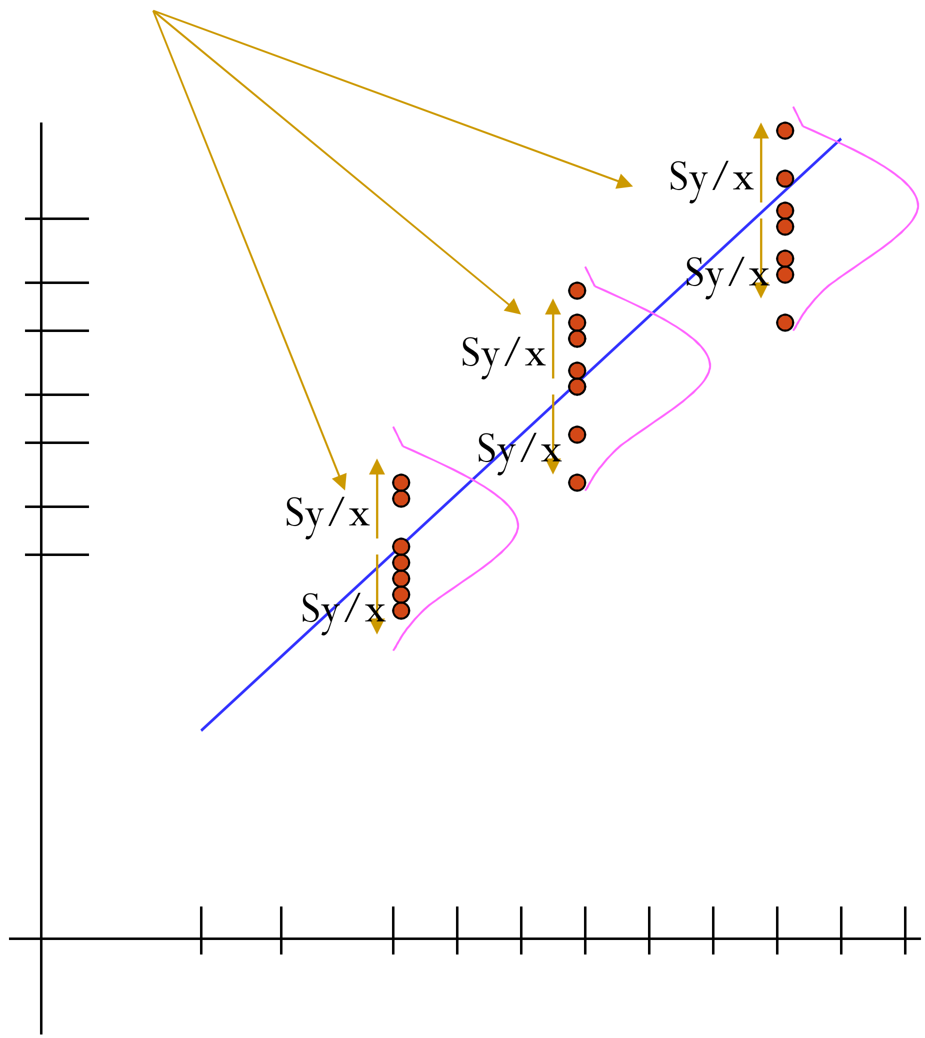
- For simple linear regression standard error is the square root of the average squared residual.

$$s_{y.x}{}^2 = \frac{1}{n-2}\sum e_i^2 = \frac{1}{n-2}\sum (y_i - \hat{y}_i)^2$$

- To estimate standard error, use $s_{y.x} = \sqrt{s_{y.x}{}^2}$

- s estimates the standard deviation of the error term ε in the statistical model for simple linear regression.

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.



Regression Standard Error

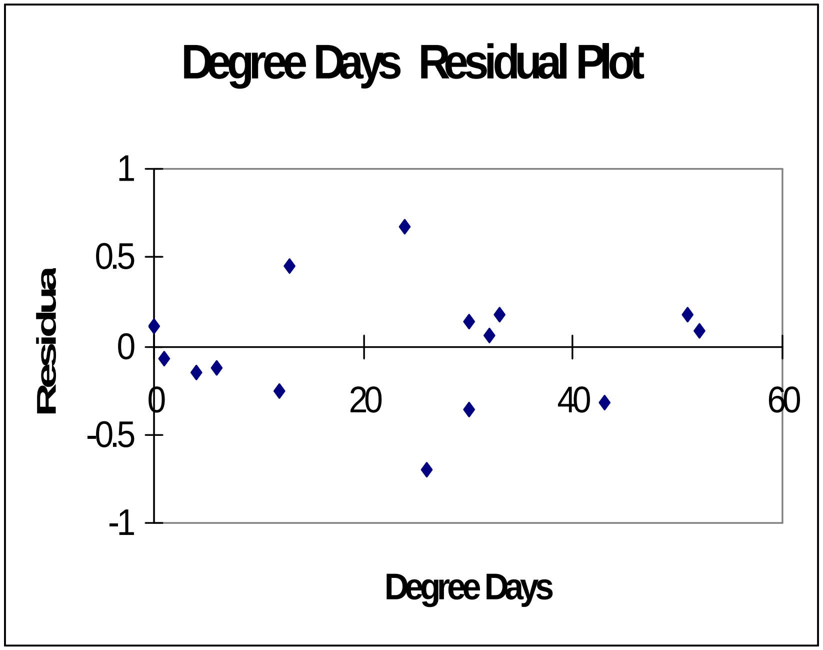
y	x	y-hat	Residual (e)	square(e)
1250	41	1270.8	-20.8	432.64
1380	54	1411.2	-31.2	973.44
1425	63	1508.4	-83.4	6955.56
1425	54	1411.2	13.8	190.44
1450	48	1346.4	103.6	10732.96
1300	46	1324.8	-24.8	615.04
1400	62	1497.6	-97.6	9525.76
1510	61	1486.8	23.2	538.24
1575	64	1519.2	55.8	3113.64
1650	71	1594.8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

Analysis of Residual

- To examine whether the regression model is appropriate for the data being analyzed, we can check the residual plots.
- Residual plots are:
  - Plot a histogram of the residuals
  - Plot residuals against the fitted values.
  - Plot residuals against the independent variable.
  - Plot residuals over time if the data are chronological.

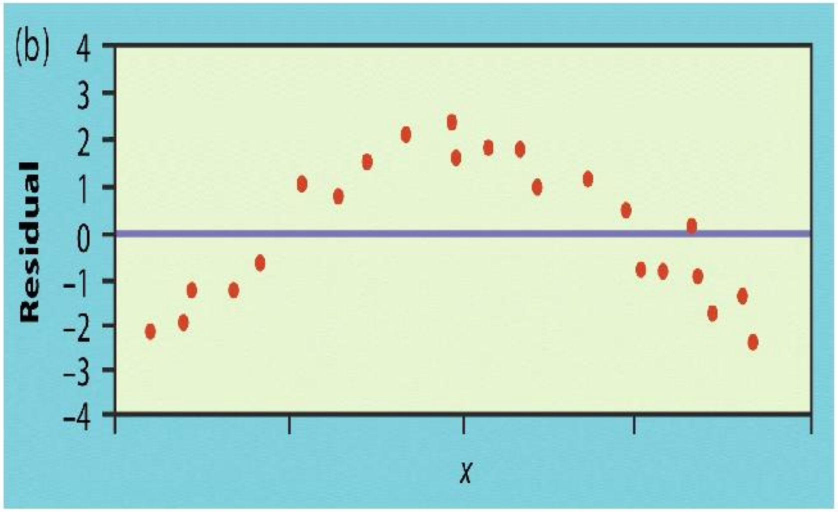
Residual plots

- The residuals should have no systematic pattern.
- The residual plot to right shows a scatter of the points with no individual observations or systematic change as *x* increases.



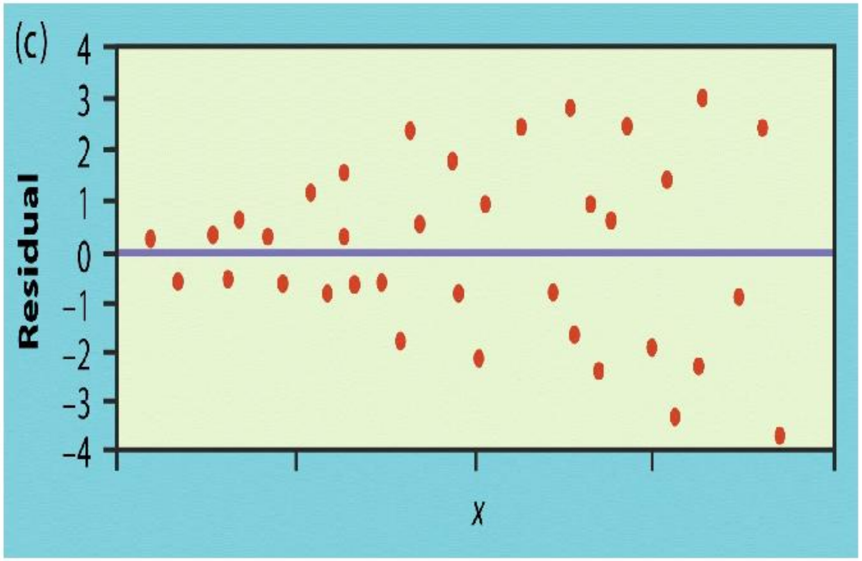
Residual plots

- 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.



ANOVA

- Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.
- An **ANOVA** test is a way to find out if survey or experiment results are significant.
- Compares the samples on the basis of their means