TOPIC 3. INTRODUCTION TO LINEAR REGRESSION

Linear Correlation

- Pearson's correlation coefficient: shows linear correlation between two continuous variables

  Covariance (sample):

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

x_i and y_i are observations of two continuous variables, $\bar{X}$ and $\bar{Y}$ are sample means of the two continuous variables, and n is the sample size.

Interpretation of Covariance

$\mathrm{cov}(X, Y) > 0$: X and Y are positively correlated

$\mathrm{cov}(X, Y) < 0$: X and Y are inversely correlated

$\mathrm{cov}(X, Y) = 0$: X and Y are uncorrelated (independent when the joint distribution of X and Y is Normal)

Correlation Coefficient (Population)

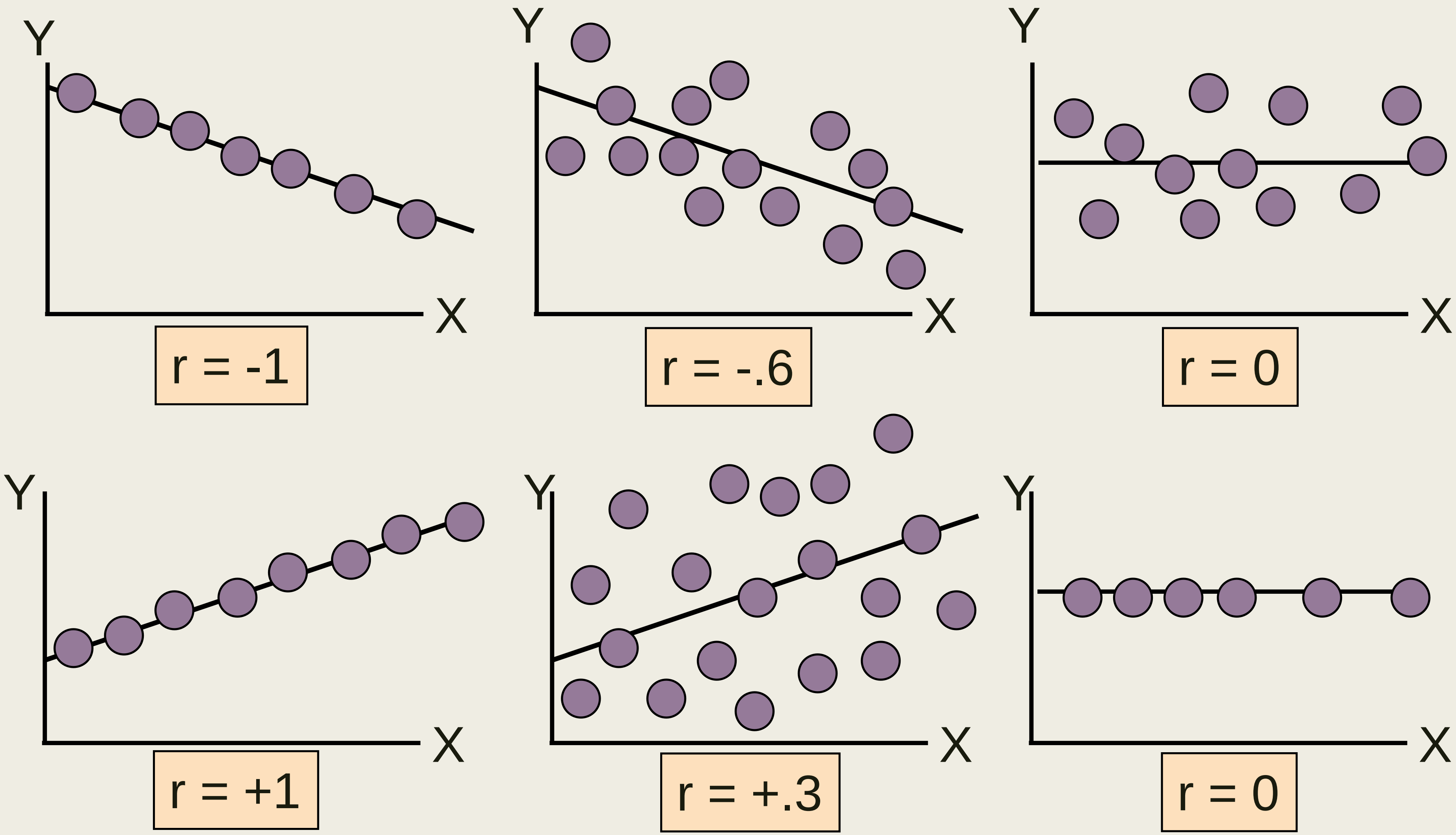
Pearson’s correlation coefficient (population) can be defined as:

$$r = \frac{cov(X, Y)}{\sqrt{var(X)} \cdot \sqrt{var(Y)}}$$

Interpretation of Correlation Coefficient

- Measures the relative strength of *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 linear relationship

Visualizing Correlation Coefficients

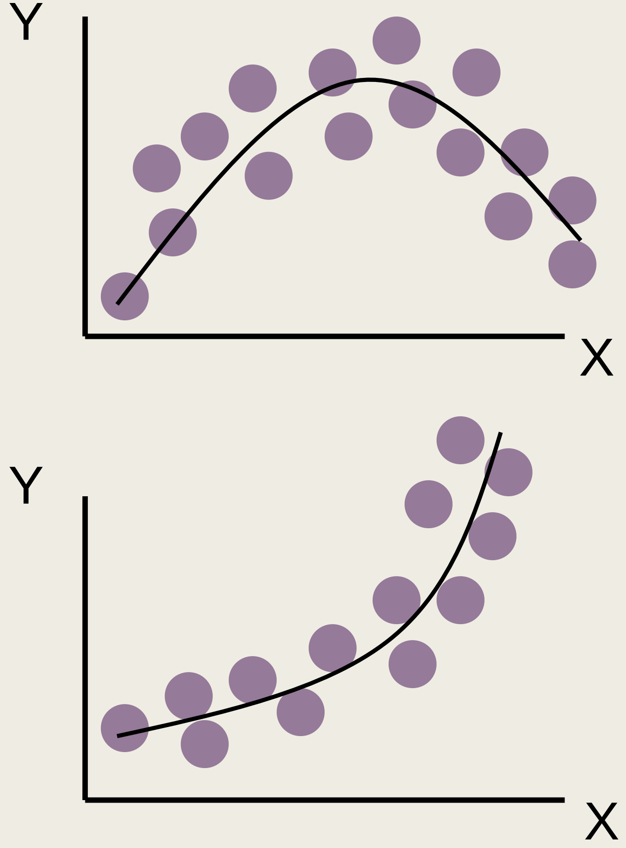


Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation – Part I

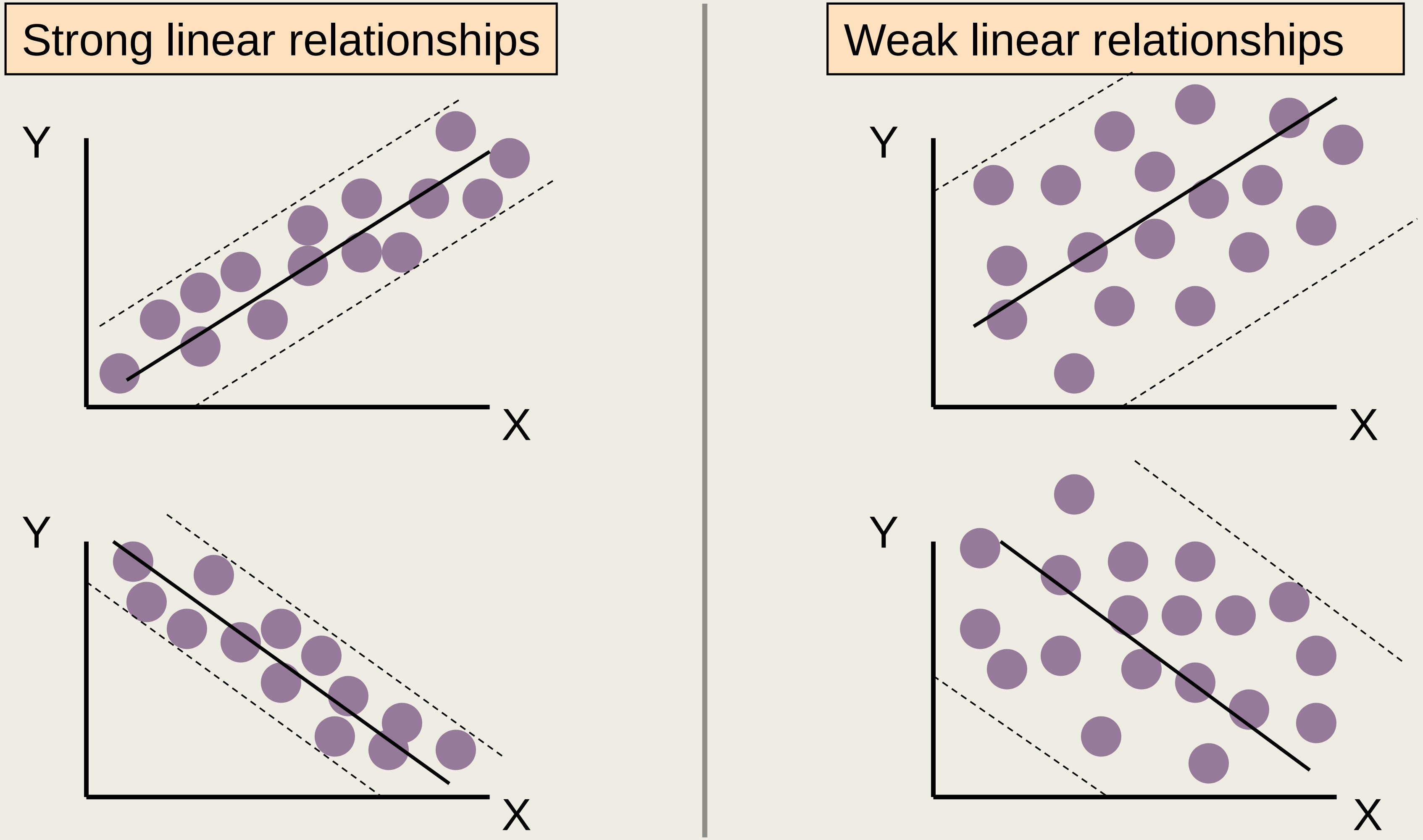
Linear relationships

Curvilinear relationships (?)



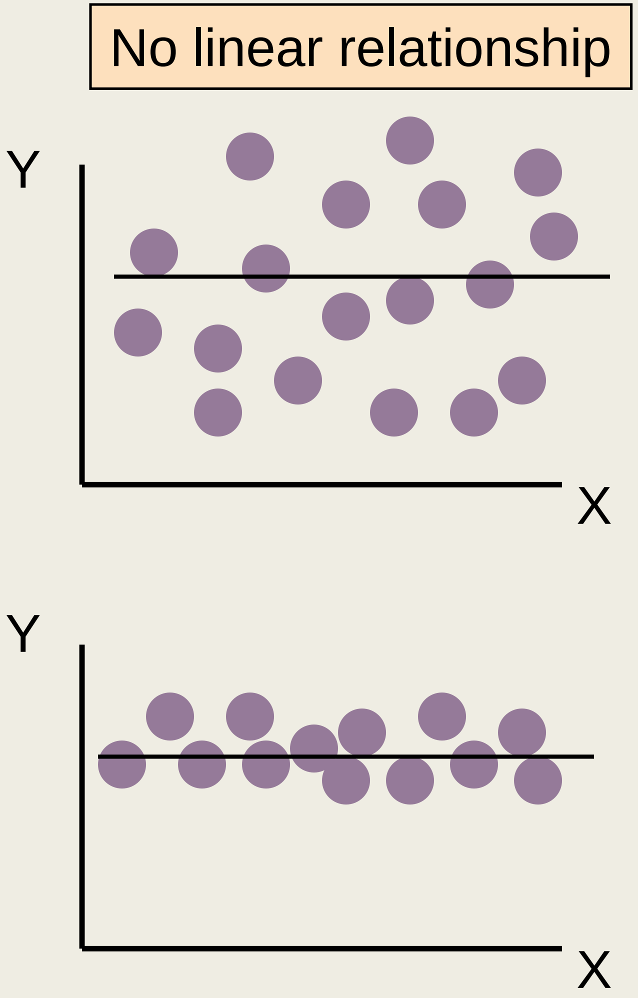
- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Linear Correlation – Part II



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

Linear Correlation – Part III



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

Calculation of Sample Correlation Coefficient

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

Continue...

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

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

$$\longrightarrow$$

Numerator of sample covariance

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

Numerators of sample variance

Distribution of Correlation Coefficient

If X and Y are uncorrelated variables and follow a bivariate normal distribution, a function of their sample correlation coefficient follows a t-distribution with $n-2$ degrees of freedom.

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

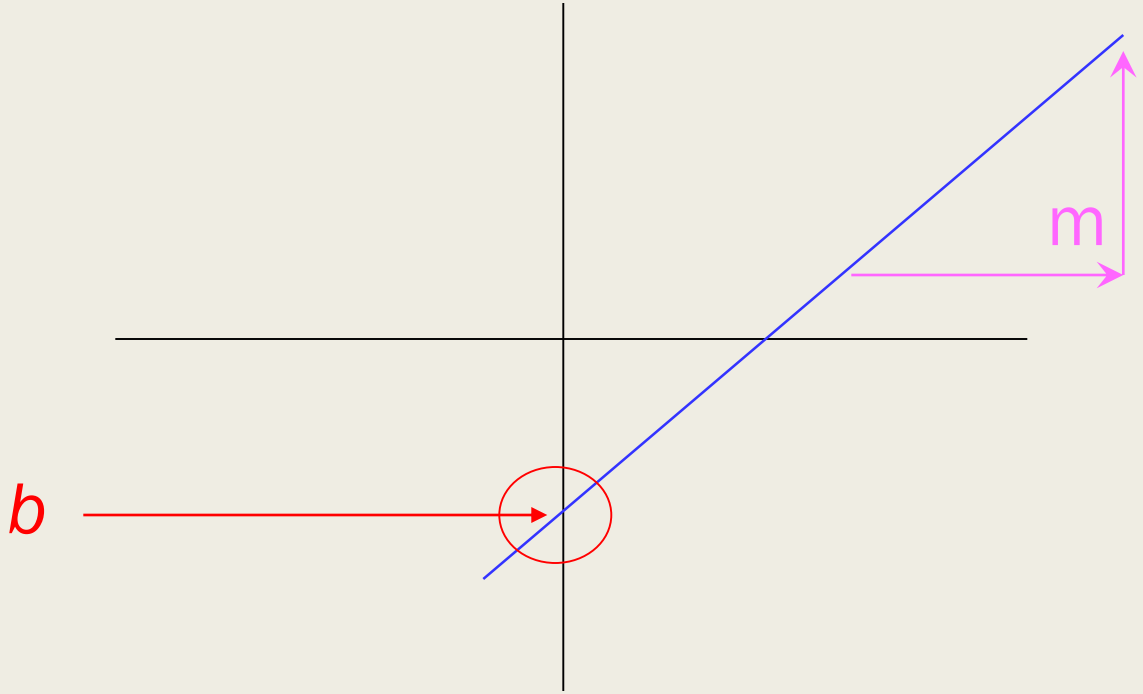
Simple Linear Regression

- Explore a linear model for estimating mean of *Y* or predicting *Y* using inputs *X*.

- In regression, one variable is considered independent (=predictor) variable (*X*) and the other the dependent (=response) variable *Y*.

Visualization of Simple Linear Regression

- Remember this:
- *$y=mx+b$?*



Slope and Intercept

- In an equation of a line, $y = mx+b$, m is the slope and b is the intercept.
- A slope of 3 means that any 1 unit change in x will result in a 3 units change in y.
- An intercept of 2 means that if x is equal to 0 then y is equal to 2.

Prediction by Linear Regression Model

- We target on obtaining the optimal value of m and b so that if we have any x, we can most accurately estimate y

- After estimating m and b, we will predict y which is unknown based on the obtained value of x using the following relation.

- $E(Y_i|x_i) = b + mx_i$

Model for Simple Linear Regression

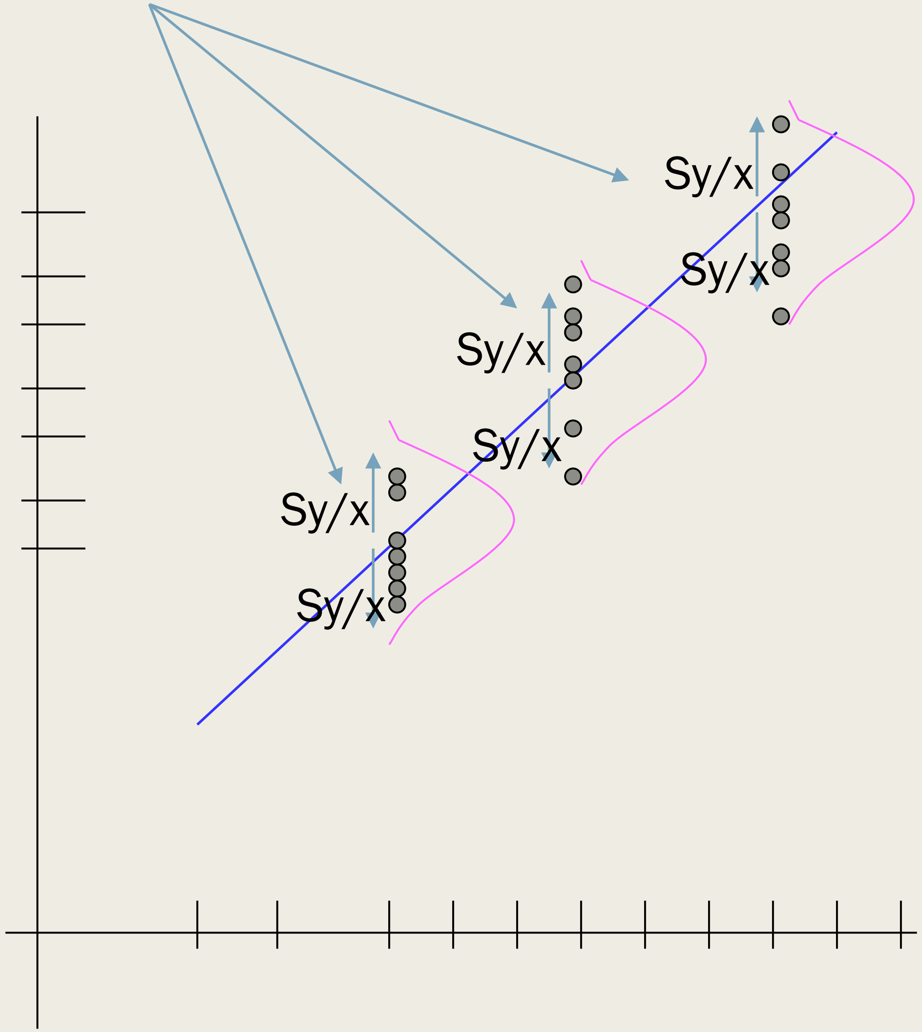
$$Y_i = \underbrace{b \quad + \quad mx_i}_{\text{Fixed – exactly on the line}} \quad + \quad \boxed{\text{random error}_\text{i}}$$

Follows a normal distribution with mean 0 and variance σ^2

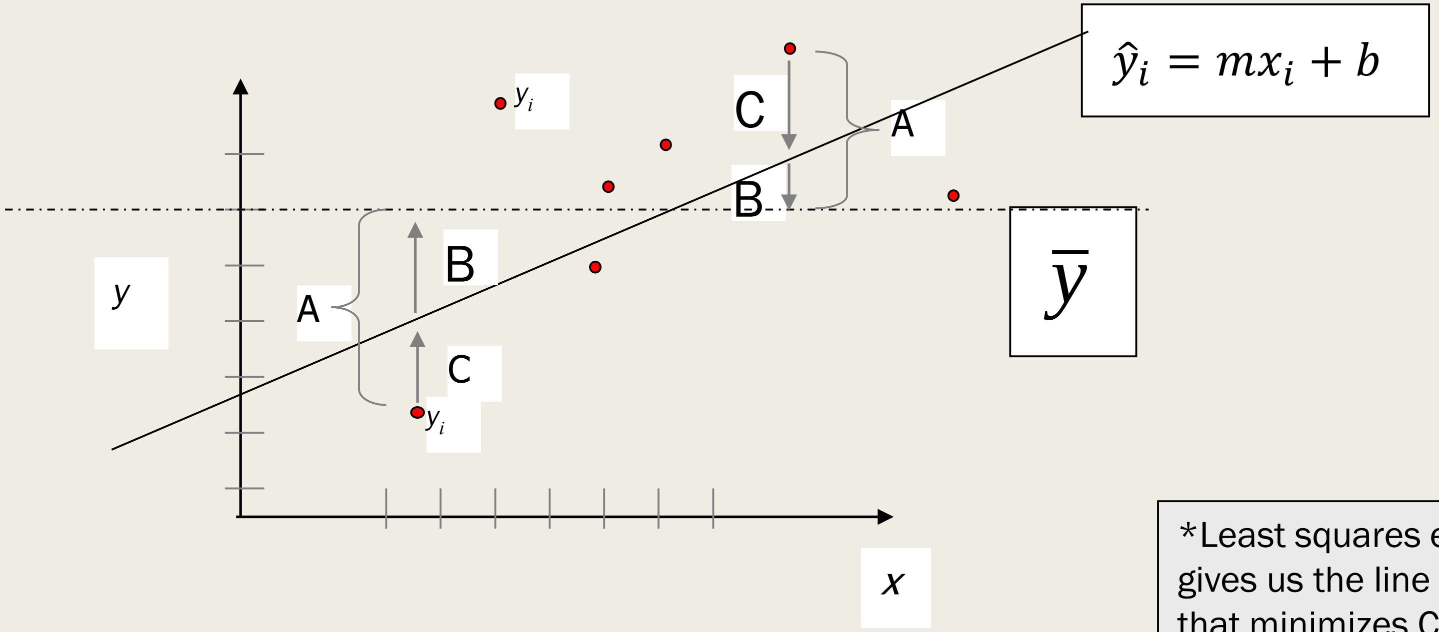
Assumptions in Linear Regression

- Linear Regression Model assumes
  - *The relationship between x and Y is linear*
  - *Y is distributed normally at each value of x*
  - *The variance of Y at every value of x is the same*
  - *The observations are independent*

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



Regression Model Interpretation



*Least squares estimation gives us the line (m and b) that minimizes 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
SS_{total}
Total squared distance of observations from naive (overall) mean of y
Total variability

B^2
SS_{reg}
Distance from regression line to naïve mean of y
Variability due to x (regression)

C^2
$SS_{residual}$
Variance around the regression line
Additional variability not explained by x

Estimating Slope and Intercept – Least Square Estimation

Least Squares Estimation

What are we trying to estimate? ***m*, the slope, and b, the intercept**

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>

$\text{Difference}_i = y_i - (mx + b)$ => $\text{Difference}_i^2 = (y_i - (mx + b))^2$

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

$$\frac{\partial}{\partial m}\sum_{i=1}^{n}(y_i - (mx_i + b))^2 = 2\sum_{i=1}^{n}(y_i - mx_i - b)(-x_i)$$

$$2(\sum_{i=1}^{n}(-y_i x_i + mx_i^2 + bx_i)) = 0$$

Results

Slope: $\hat{m} = \frac{sample\ Cov(x,y)}{sample\ Var(x)} = \frac{SS_{xy}}{SS_x}$

Intercept: $\hat{b} = \bar{y} - \hat{m}\bar{x}$

Regression lines always go through the point $(\bar{x}, \bar{y})$

Relationship with Correlation

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

$$\hat{r} = \widehat{m}\,\frac{SD_x}{SD_y} \ \text{ or } = \widehat{m}\,\frac{\sqrt{SS_x}}{\sqrt{SS_y}}$$

Where SD is sample standard deviation, i.e. $SD_x = \sqrt{\frac{SS_x}{n-1}}$, $SD_y = \sqrt{\frac{SS_y}{n-1}}$

Significance testing

Slope

Distribution of slope ~ $t_{n-2}(m, \text{s.e.}(\hat{m}))$, s.e. means standard error

$$\boxed{\begin{array}{l}\mathrm{H}_0: m = 0 \text{ (no linear relationship)} \\ \mathrm{H}_1: m \neq 0 \text{ (linear relationship exists)}\end{array}}$$

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

Standard Error of Slope

$$s.e.(\widehat{m}) = \sqrt{\frac{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}{SS_x}}$$

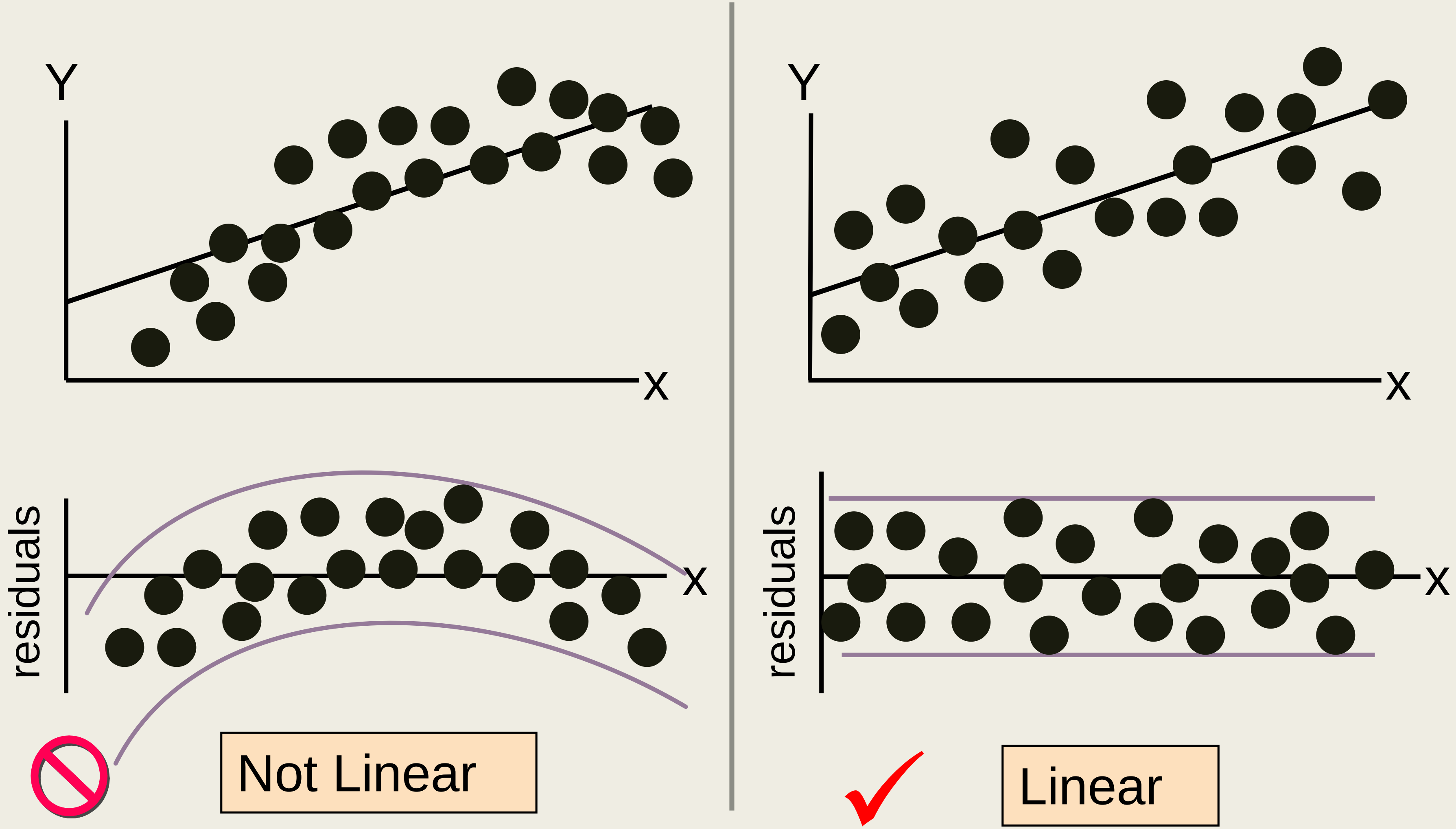
where $SS_x = \sum_{i=1}^{n}(x_i - \bar{x})^2$ and $\hat{y}_i = \hat{b} + \widehat{m}x_i$

Residual Analysis: Check Assumptions

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

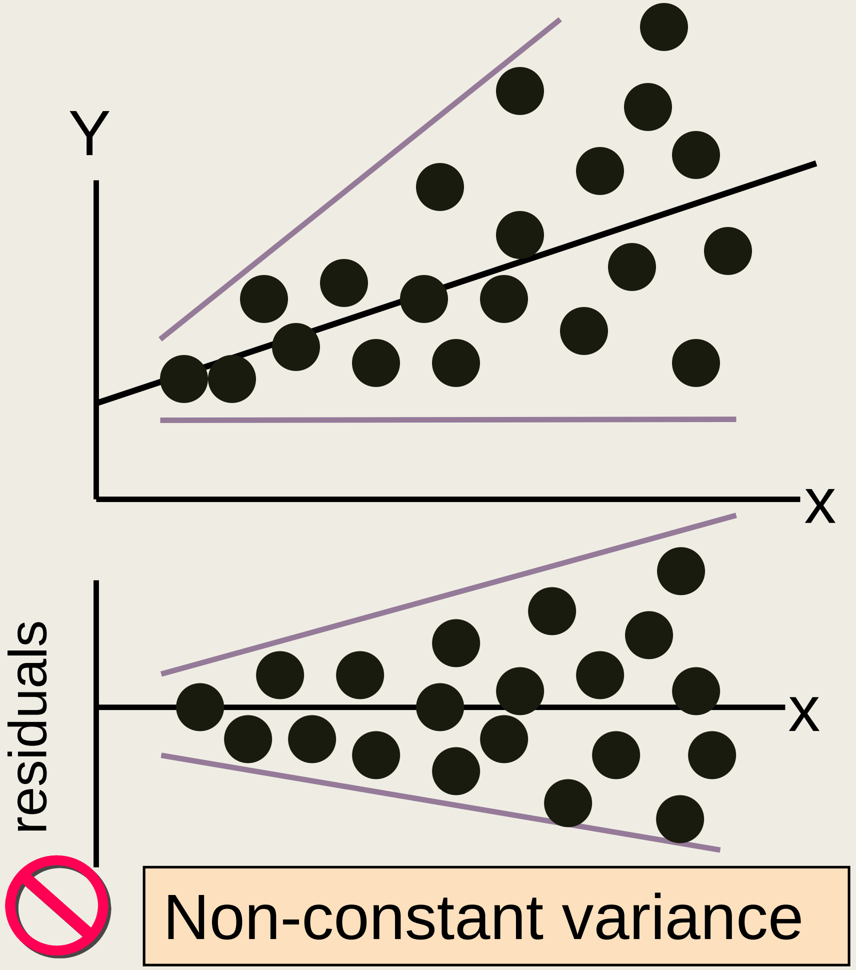
- The residual for observation *i*, e_i, is the difference between its observed and predicted value
- Check the assumptions of regression by examining the residuals
  - *Examine for linearity assumption*
  - *Examine for constant variance for all levels of X (Homoscedasticity)*
  - *Evaluate normal distribution assumption*
  - *Evaluate independence assumption*
- Graphical Analysis of Residuals

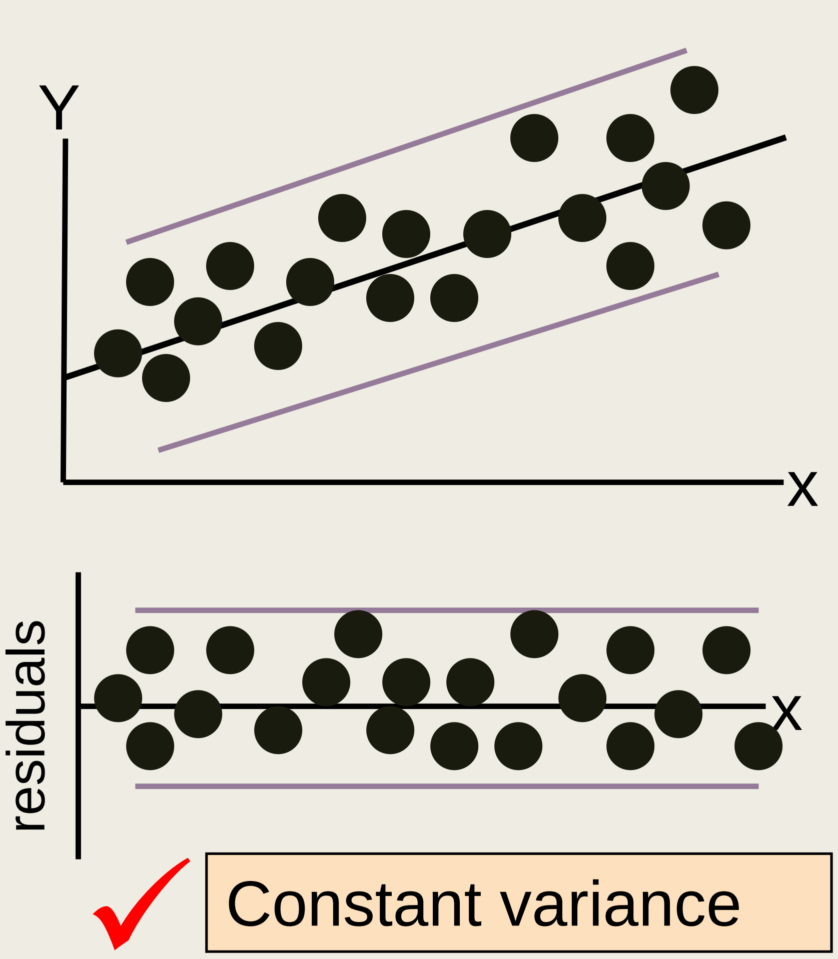
Residual Analysis for Linearity



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

Residual Analysis for Homoscedasticity

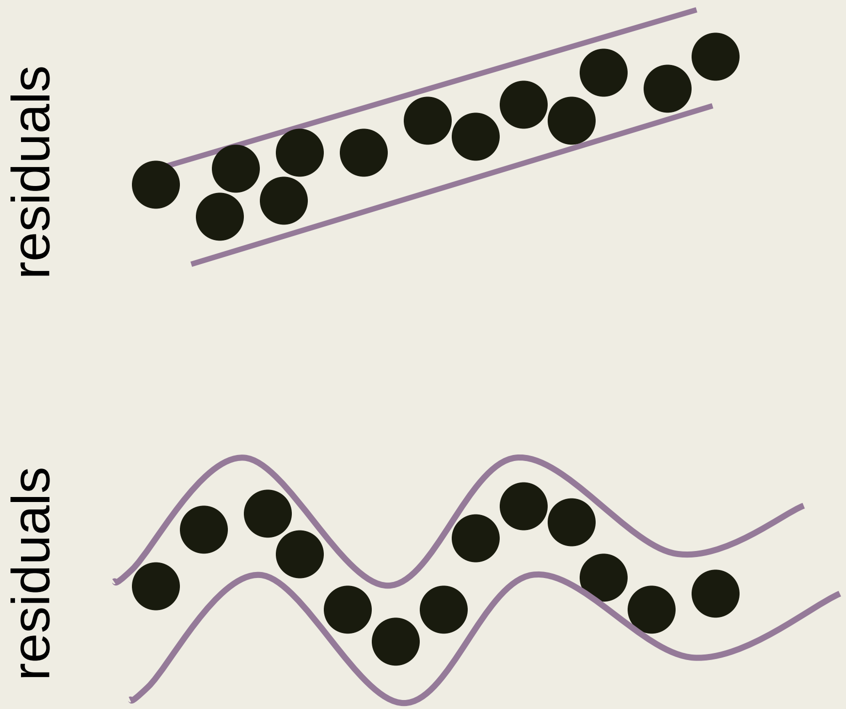




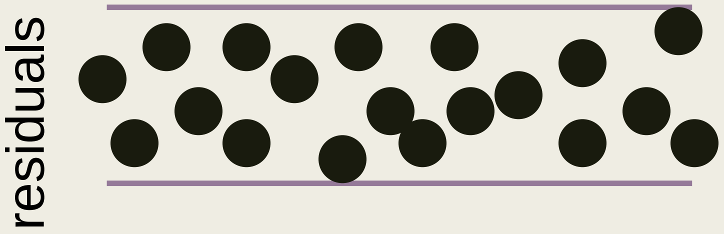
- Slide from: Statistics for Managers Using Microsoft® Excel 4th Edition, 2004 Prentice-Hall

Residual Analysis for Independence

Not Independent



Independent



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

Multiple Linear Regression

- Confounders in estimating target y
- Multiple predictors x
- Example: $y = m_1x_1 + m_2w + m_3z + b$

Each regression coefficient is the amount of change in the outcome variable that would be expected per one-unit change of the predictor, if all other variables in the model were held constant.

An initial taste, more will be taught in advanced courses.