Statistics 2
Chapter 7
Linear Regression and Corelation

Chapter 7
Linear Regression and Correlation

GOALS

When you have completed this chapter, you will be able to:

ONE

Draw a scatter diagram.

TWO

Understand and interpret the terms *dependent* variable and *independent* variable.

THREE

Calculate and interpret the coefficient of correlation, the coefficient of determination, and the standard error of estimate.

FOUR

Conduct a test of hypothesis to determine if there is a difference among block means.

Chapter 7 *continued*

Linear Regression and Correlation

GOALS

When you have completed this chapter, you will be able to:

FIVE

Calculate the least squares regression line and interpret the slope and intercept values.

SIX

Construct and interpret a confidence interval and and prediction interval for the dependent variable.

SEVEN

Set up and interpret an ANOVA table.

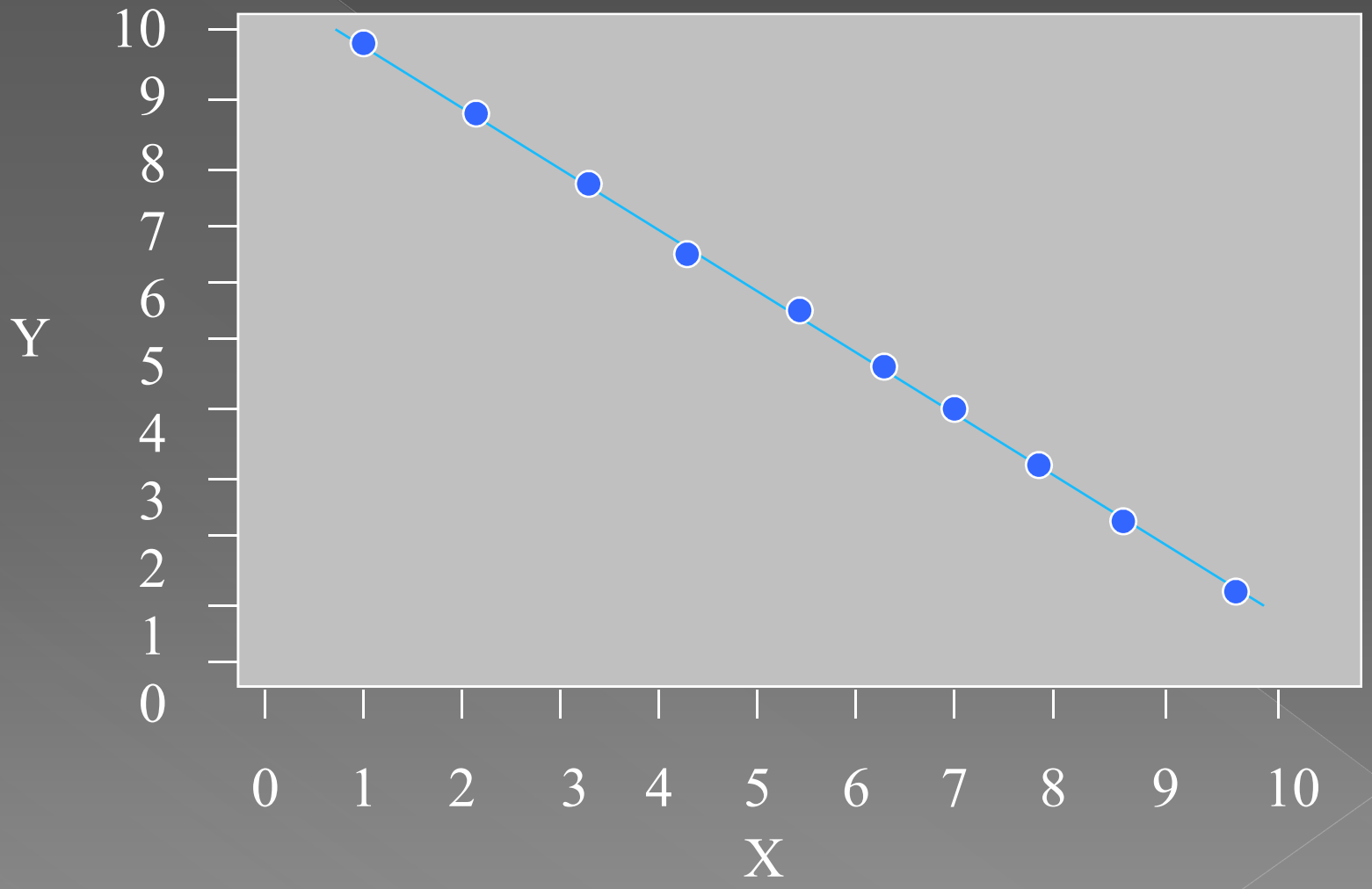
Correlation Analysis

- Correlation Analysis: A group of statistical techniques used to measure the strength of the relationship (correlation) between two variables.
- Scatter Diagram: A chart that portrays the relationship between the two variables of interest.
- Dependent Variable: The variable that is being predicted or estimated.
- Independent Variable: The variable that provides the basis for estimation. It is the predictor variable.

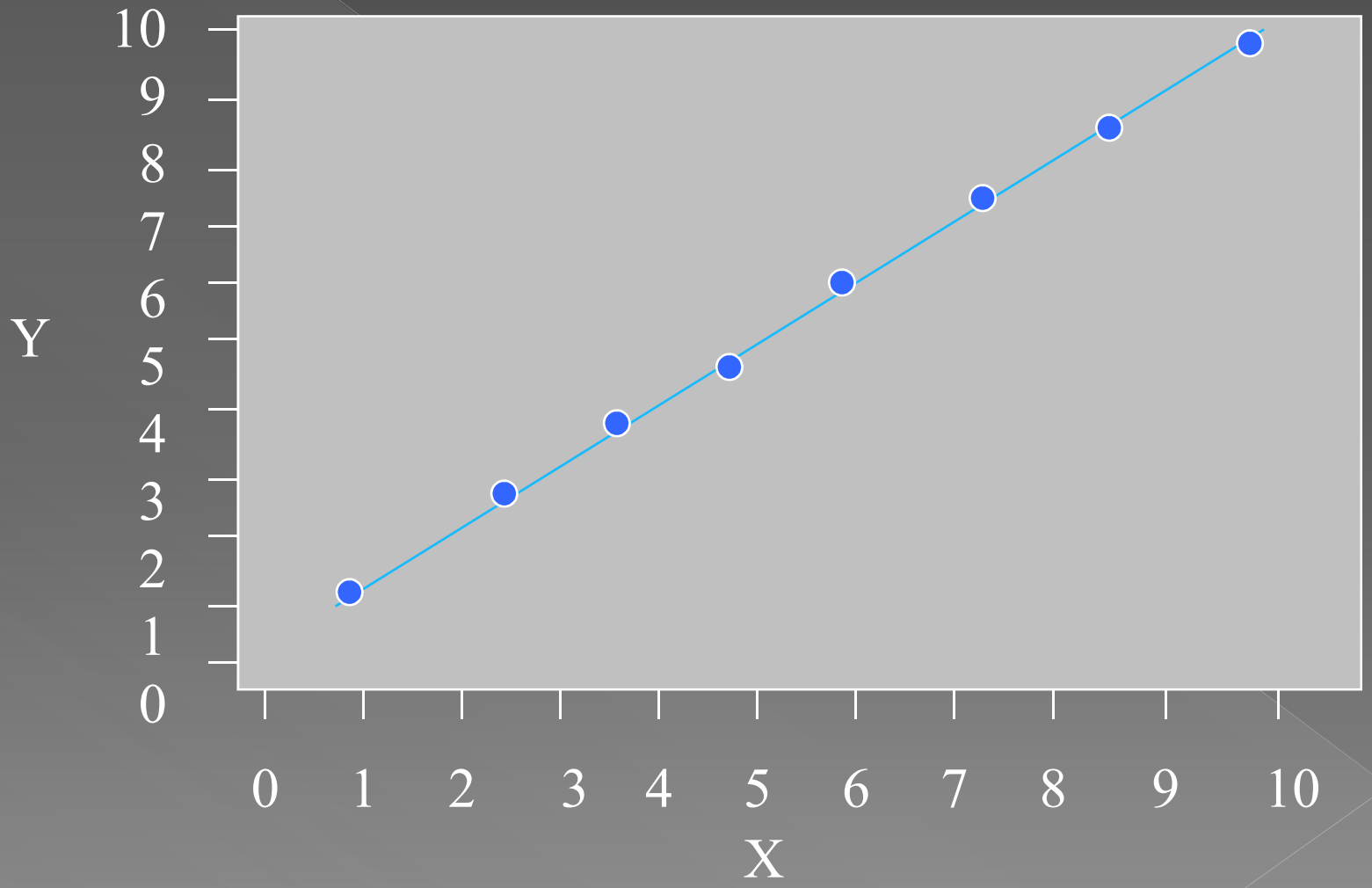
The Coefficient of Correlation, r

- The Coefficient of Correlation (r) is a measure of the strength of the relationship between two variables.
  - It requires interval or ratio-scaled data (variables).
  - It can range from -1.00 to 1.00.
  - Values of -1.00 or 1.00 indicate perfect and strong correlation.
  - Values close to 0.0 indicate weak correlation.
  - Negative values indicate an inverse relationship and positive values indicate a direct relationship.

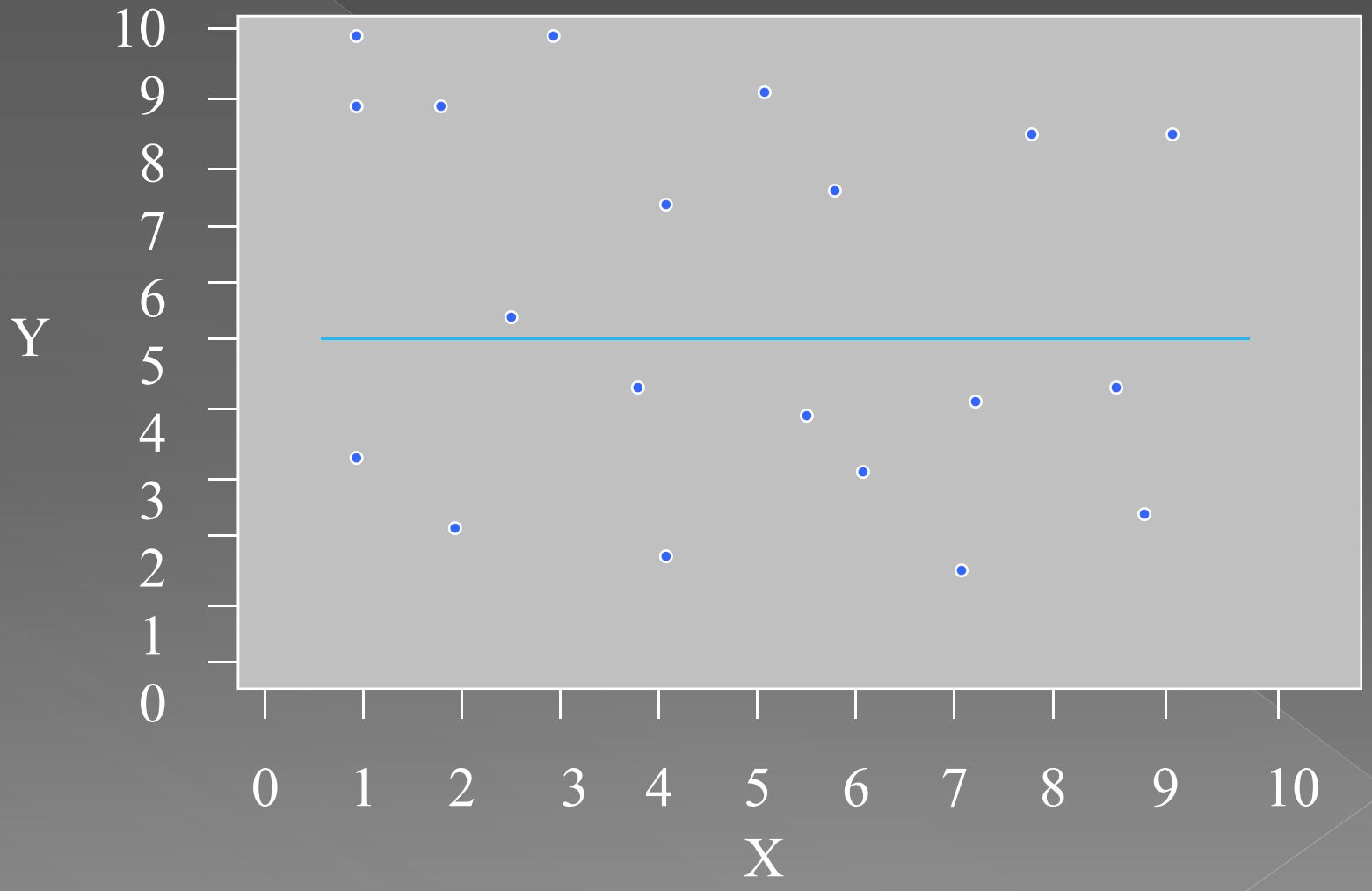
Perfect Negative Correlation



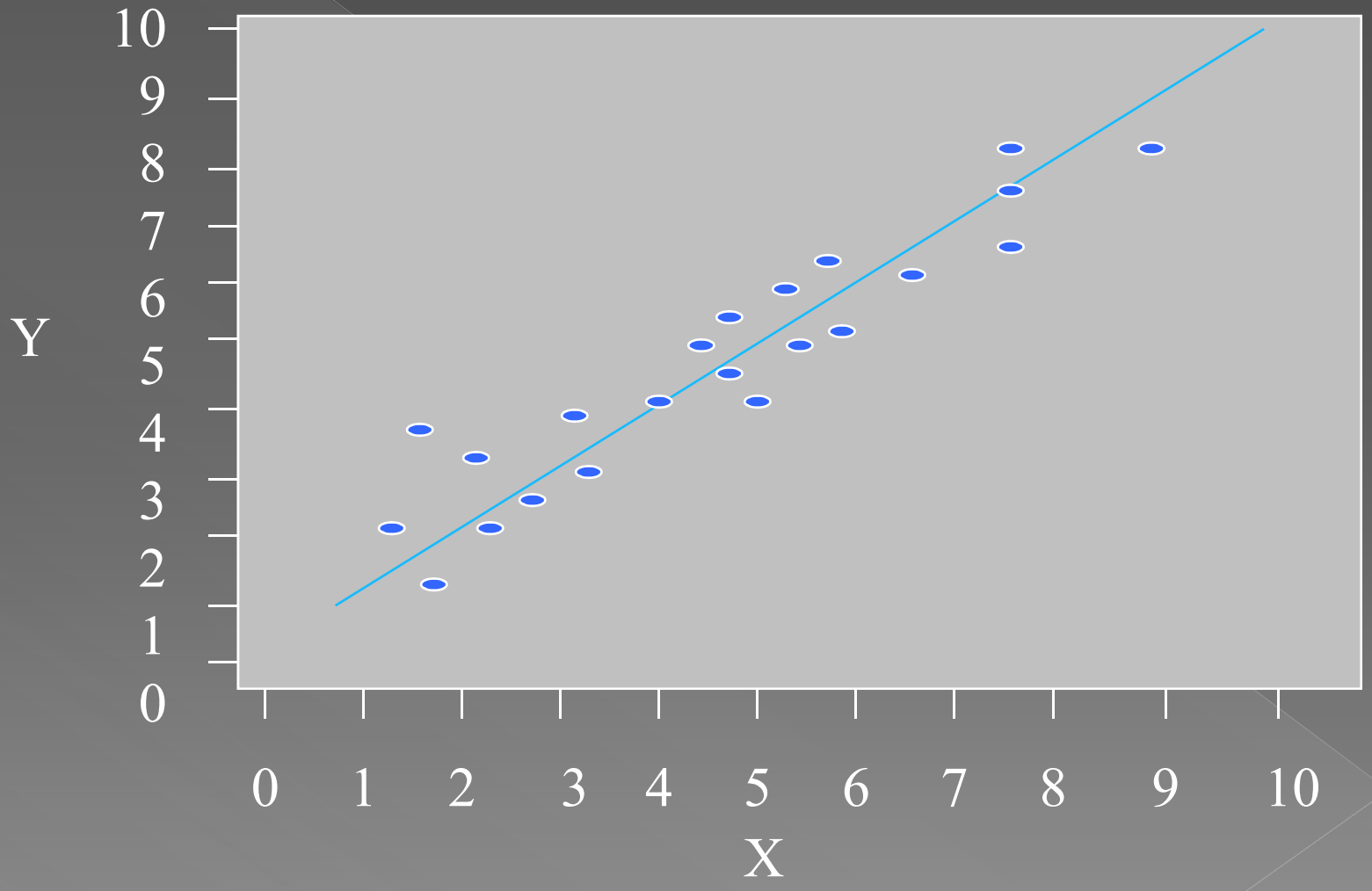
Perfect Positive Correlation



Zero Correlation



Strong Positive Correlation



Formula for r

$$r = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{\sqrt{\left[n(\Sigma X^2) - (\Sigma X)^2\right]\left[n(\Sigma Y^2) - (\Sigma Y)^2\right]}}$$

Coefficient of Determination

- The Coefficient of Determination, r^2 - the proportion of the total variation in the dependent variable Y that is explained or accounted for by the variation in the independent variable X.
  - The coefficient of determination is the square of the coefficient of correlation, and ranges from 0 to 1.

EXAMPLE 1

- Dan Ireland, the student body president at Toledo State University, is concerned about the cost of textbooks. To provide insight into the problem he selects a sample of eight textbooks currently on sale in the bookstore. He decides to study the relationship between the number of pages in the text and the cost. Compute the correlation coefficient.

EXAMPLE 1 *continued*

Book	Pages	Cost($)
1	500	28
2	700	25
3	800	33
4	600	24
5	400	23
6	500	27
7	600	21
8	800	31

EXAMPLE 1 *continued*

- r=.614 (verify)
- Test the hypothesis that there is no correlation in the population. Use a .02 significance level.
- Step 1: H_0 The correlation in the population is zero. H_1 The correlation in the population is not zero.
- Step 2: H_0 is rejected if t>3.143 or if t<-3.143, df=6, α =.02

EXAMPLE 1 *continued*

- The test statistic is t=1.9055, computed by

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

- with (n-2) degrees of freedom
- Step 4: H_0 is not rejected

Regression Analysis

- Purpose: to determine the regression equation; it is used to predict the value of the dependent variable (Y) based on the independent variable (X).
- Procedure: select a sample from the population and list the paired data for each observation; draw a scatter diagram to give a visual portrayal of the relationship; determine the regression equation.

Regression Analysis

- the regression equation: $Y' = a + bX$, where:
- Y' is the average predicted value of Y for any X.
- a is the Y-intercept, or the estimated Y value when $X=0$
- b is the slope of the line, or the average change in Y' for each change of one unit in X

$$b = \frac{n(\Sigma XY) - (\Sigma X)(\Sigma Y)}{n(\Sigma X^2) - (\Sigma X)^2}$$

- the least squares principle is used to obtain a & b:

$$a = \frac{\Sigma Y}{n} - b\frac{\Sigma X}{n}$$

EXAMPLE 2

- Develop a regression equation for the information given in EXAMPLE 1 that can be used to estimate the selling price based on the number of pages.
- By the least squares principle, b=.01714 and a=16.00175
- Y' =16.00175 + .01714X

The Standard Error of Estimate

- The standard error of estimate measures the scatter, or dispersion, of the observed values around the line of regression
- The formulas that are used to compute the standard error:

$$S_{Y \cdot X} = \sqrt{\frac{\Sigma (Y - Y')^2}{n - 2}}$$

$$= \sqrt{\frac{\Sigma Y^2 - a(\Sigma Y) - b(\Sigma XY)}{n - 2}}$$

Assumptions Underlying Linear Regression

- For each value of X, there is a group of Y values, and these Y values are *normally distributed*.
- The *means* of these normal distributions of Y values all lie on the straight line of regression.
- The *standard deviations* of these normal distributions are equal.
- The Y values are statistically independent. This means that in the selection of a sample, the Y values chosen for a particular X value do not

Confidence Interval

- The confidence interval for the mean value of Y for a given value of X is given by:

$$Y' \pm t \cdot (S_{Y \cdot X}) \sqrt{\frac{1}{n} + \frac{(X - \overline{X})^2}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}}$$

Prediction Interval

- The prediction interval for an individual value of Y for a given value of X is given by:

$$Y' \pm t \cdot (S_{Y \cdot X}) \sqrt{1 + \frac{1}{n} + \frac{(X - \overline{X})^2}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}}$$

EXAMPLE 3

- Use the information from EXAMPLE 1 to:
  - Compute the standard error of estimate:

$$S_{Y \cdot X} = 3.471$$

  - Develop a 95% confidence interval for all 650 page textbooks: [24.03, 30.25] Verify
  - Develop a 95% prediction interval for a 650 page text: [18.09, 36.19] Verify

More on the Coefficient of Determination

$$r^2 = \frac{\text{Total variation - unexplained variation}}{\text{Total variation}}$$

$$= \frac{\Sigma(Y - \bar{Y})^2 - \Sigma(Y - Y')^2}{\Sigma(Y - \bar{Y})^2}$$

$$\text{Regression} = SSR = \Sigma(Y' - \bar{Y})^2$$

$$\text{Error variation} = SSE = \Sigma(Y - Y')^2$$

$$\text{Total variation} = SS\ total = \Sigma(Y - \bar{Y})^2$$

More on the Coefficient of Determination

$$r^2 = \frac{SSR}{SS\ total} = 1 - \frac{SSE}{SS\ total}$$

$$S_{Y \cdot X} = \sqrt{\frac{SSE}{n-2}}$$