Simple Linear Regression Analysis

The Simple Linear Regression Model and the Least Squares Point Estimates

- The **dependent** (or response) variable is the variable we wish to understand or predict
- The **independent** (or predictor) variable is the variable we will use to understand or predict the dependent variable
- **Regression analysis** is a statistical technique that uses observed data to relate the dependent variable to one or more independent variables
- The objective is to build a regression model that can describe, predict and control the dependent variable based on the independent variable

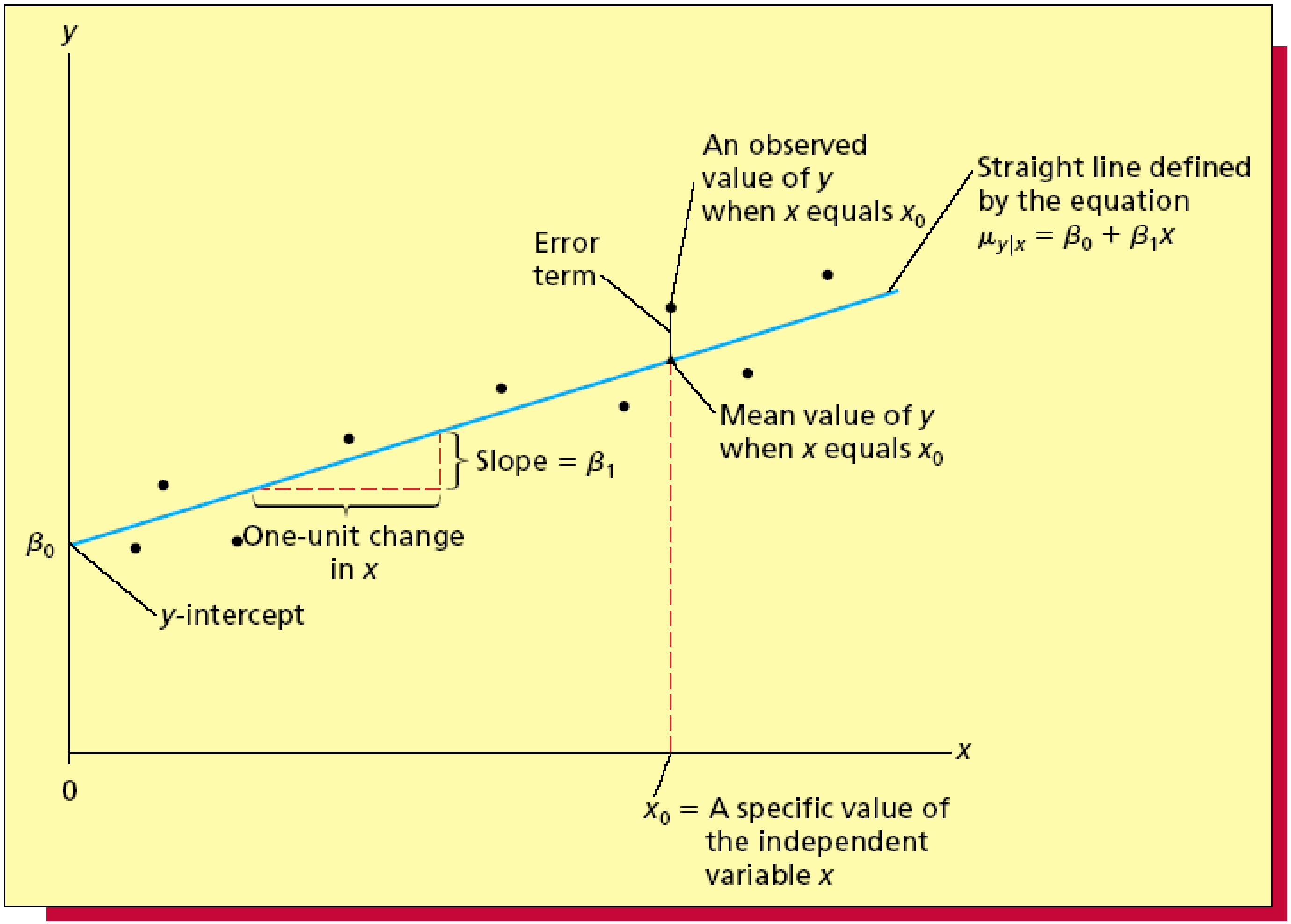
Form of The Simple Linear Regression Model

- $y = \beta_0 + \beta_1 x + \varepsilon$
- $\mu_y = \beta_0 + \beta_1 x + \varepsilon$ is the mean value of the dependent variable y when the value of the independent variable is x
- β_0 is the y-intercept; the mean of y when x is 0
- β_1 is the slope; the change in the mean of y per unit change in x
- ε is an error term that describes the effect on y of all factors other than x

Regression Terms

- β_0 and β_1 are called regression parameters
- β_0 is the y-intercept and β_1 is the slope
- We do not know the true values of these parameters
- So, we must use sample data to estimate them
- b_0 is the estimate of β_0 and b_1 is the estimate of β_1

The Simple Linear Regression Model Illustrated



The Least Squares Point Estimates

- Estimation/prediction equation
  $\hat{y} = b_0 + b_1 x$
- Least squares point estimate of the slope β_1

$$b_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

- Least squares point estimate of y-intercept β_0

$$b_0 = \bar{y} - b_1 \bar{x} \qquad \bar{y} = \frac{\sum y_i}{n} \qquad \bar{x} = \frac{\sum x_i}{n}$$

The Tasty Sub Shop Case #1

y_i	x_i	x_i^2	$x_i y_i$
527.1	20.8	$(20.8)^2 = 432.64$	$(20.8)(527.1) = 10963.68$
548.7	27.5	$(27.5)^2 = 756.25$	$(27.5)(548.7) = 15089.25$
767.2	32.3	$(32.3)^2 = 1{,}043.29$	$(32.3)(767.2) = 24780.56$
722.9	37.2	$(37.2)^2 = 1{,}383.84$	$(37.2)(722.9) = 26891.88$
826.3	39.6	$(39.6)^2 = 1{,}568.16$	$(39.6)(826.3) = 32721.48$
810.5	45.1	$(45.1)^2 = 2{,}034.01$	$(45.1)(810.5) = 36553.55$
1040.7	49.9	$(49.9)^2 = 2{,}490.01$	$(49.9)(1040.7) = 51930.93$
1033.6	55.4	$(55.4)^2 = 3{,}069.16$	$(55.4)(1033.6) = 57261.44$
1090.3	61.7	$(61.7)^2 = 3{,}806.89$	$(61.7)(1090.3) = 67271.51$
1235.8	64.6	$(64.6)^2 = 4{,}173.16$	$(64.6)(1235.8) = 79832.68$
$\sum y_i = 8603.1$	$\sum x_i = 434.1$	$\sum x_i^2 = 20{,}757.41$	$\sum x_i y_i = 403{,}296.96$

The Tasty Sub Shop Case #2

- From last slide,
  - $\Sigma y_i = 8,603.1$
  - $\Sigma x_i = 434.1$
  - $\Sigma x^2_i = 20,757.41$
  - $\Sigma x_i y_i = 403,296.96$
- Once we have these values, we no longer need the raw data
- Calculation of b_0 and b_1 uses these totals

The Tasty Sub Shop Case #3 (Slope b_1)

$$SS_{xy} = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}$$

$$= 403{,}296.96 - \frac{(434.1)(8{,}603.1)}{10} = 29{,}836.389$$

$$SS_{xx} = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$= 120{,}757.41 - \frac{(434.1)^2}{10} = 1{,}913.129$$

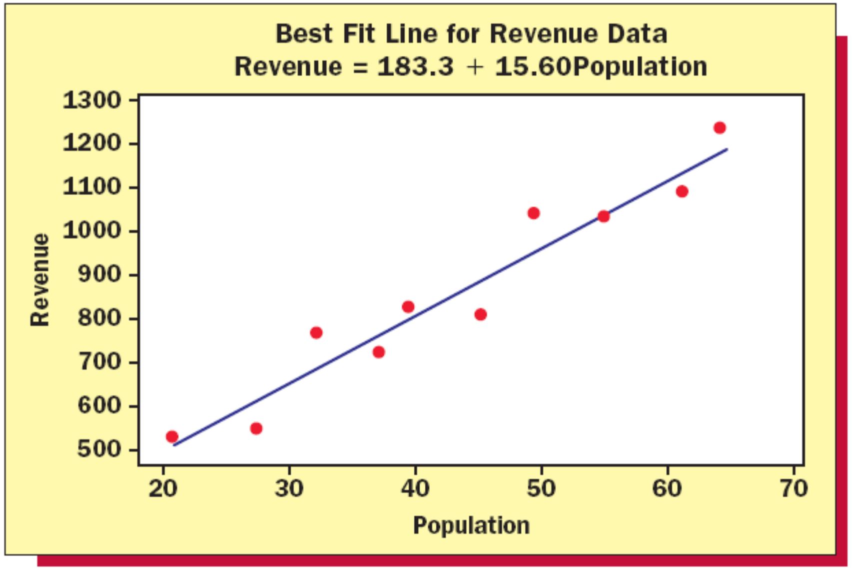
$$b_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{29{,}836.389}{1{,}913.129} = 15.596$$

The Tasty Sub Shop Case #4 (y-Intercept b_0)

$$\bar{y} = \frac{\sum y_i}{n} = \frac{8{,}603.1}{10} = 860.31$$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{434.1}{10} = 43.41$$

$$\begin{aligned} b_0 &= \bar{y} - b_1\bar{x} \\ &= 860.31 - (15.596)(43.41) \\ &= 183.31 \end{aligned}$$

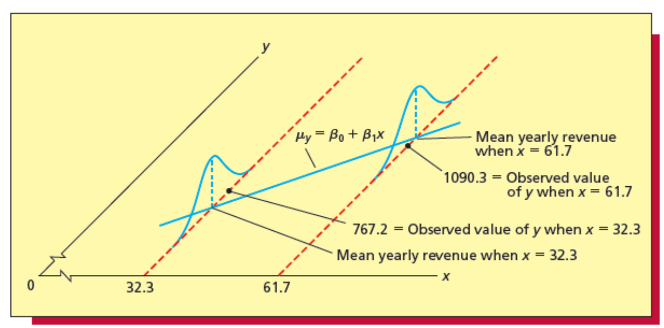


- Prediction (x = 20.8)
- $\hat{y} = b_0 + b_1x = 183.31 + (15.59)(20.8)$
- $\hat{y} = 507.69$
- Residual is 527.1 – 507.69 = 19.41

Figure 14.5

Model Assumptions and the Standard Error

1. **Mean of Zero**
   At any given value of x, the population of potential error term values has a mean equal to zero
2. **Constant Variance Assumption**
   At any given value of x, the population of potential error term values has a variance that does not depend on the value of x
3. **Normality Assumption**
   At any given value of x, the population of potential error term values has a normal distribution
4. **Independence Assumption**
   Any one value of the error term ε is statistically independent of any other value of ε



Sum of Squares

- Sum of squared errors

$$SSE = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

- Mean square error
  - Point estimate of the residual variance σ^2

$$s^2 = MSE = \frac{SSE}{n\text{-}2}$$

- Standard error
  - Point estimate of residual standard deviation σ

$$s = \sqrt{MSE} = \sqrt{\frac{SSE}{n\text{-}2}}$$

Testing the Significance of the Slope and y-Intercept

- A regression model is not likely to be useful unless there is a significant relationship between x and y
- To test significance, we use the null hypothesis:

  H0: $\beta_1 = 0$

- Versus the alternative hypothesis:

  Ha: $\beta_1 \neq 0$

Testing the Significance of the Slope #2

Alternative	Reject H_0 If	*p*-Value
H_a: $\beta_1 > 0$	$t > t_\alpha$	Area under t distribution right of t
H_a: $\beta_1 < 0$	$t < -t_\alpha$	Area under t distribution left of t
H_a: $\beta_1 \neq 0$	$\lvert t\rvert > t_{\alpha/2}$*	Twice area under t distribution right of $\lvert t\rvert$

* That is $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$

Testing the Significance of the Slope #3

- Test Statistics $t = \dfrac{b_1}{s_{b_1}}$ where $s_{b_1} = \dfrac{s}{\sqrt{SS_{xx}}}$

- 100(1-α)% Confidence Interval for β_1
  [$b_1 \pm t_{\alpha/2} S_{b1}$]

- t_α, $t_{\alpha/2}$ and p-values are based on n–2 degrees of freedom

Confidence and Prediction Intervals

- The point on the regression line corresponding to a particular value of x_0 of the independent variable x is $\hat{y} = b_0 + b_1x_0$
- It is unlikely that this value will equal the mean value of y when x equals x_0
- Therefore, we need to place bounds on how far the predicted value might be from the actual value
- We can do this by calculating a confidence interval mean for the value of y and a prediction interval for an individual value of y

Distance Value

- Both the confidence interval for the mean value of y and the prediction interval for an individual value of y employ a quantity called the distance value
- The distance value for a particular value x_0 of x is

$$\frac{1}{n}+\frac{(x_0-\bar{x})^2}{SS_{xx}}$$

- The distance value is a measure of the distance between the value x_0 of x and $\bar{x}$
- Notice that the further x_0 is from $\bar{x}$, the larger the distance value

A Confidence and Prediction Interval for a Mean Value of y

- Assume that the regression assumption holds
- The formula for a 100(1-α) confidence interval for the mean value of y is as follows:

$$[\hat{y} \pm t_{\alpha/2} s\sqrt{\text{Distance value}}\,]$$

- The formula for a 100(1-α) prediction interval for an individual value of y is as follows:

$$[\hat{y} \pm t_{\alpha/2} s\sqrt{1 + \text{Distance value}}\,]$$

- This is based on n-2 degrees of freedom

Which to Use?

- The prediction interval is useful if it is important to predict an individual value of the dependent variable
- A confidence interval is useful if it is important to estimate the mean value
- The prediction interval will always be wider than the confidence interval

14.5 Simple Coefficient of Determination and Correlation

- How useful is a particular regression model?
- One measure of usefulness is the simple coefficient of determination
- It is represented by the symbol r^2

Calculating The Simple Coefficient of Determination

1. Total variation is given by the formula $\Sigma(y_i-\bar{y})^2$
2. Explained variation is given by the formula $\Sigma(\hat{y}_i-\bar{y})^2$
3. Unexplained variation is given by the formula $\Sigma(y_i-\hat{y})^2$
4. Total variation is the sum of explained and unexplained variation
5. r^2 is the ratio of explained variation to total variation

The Simple Correlation Coefficient

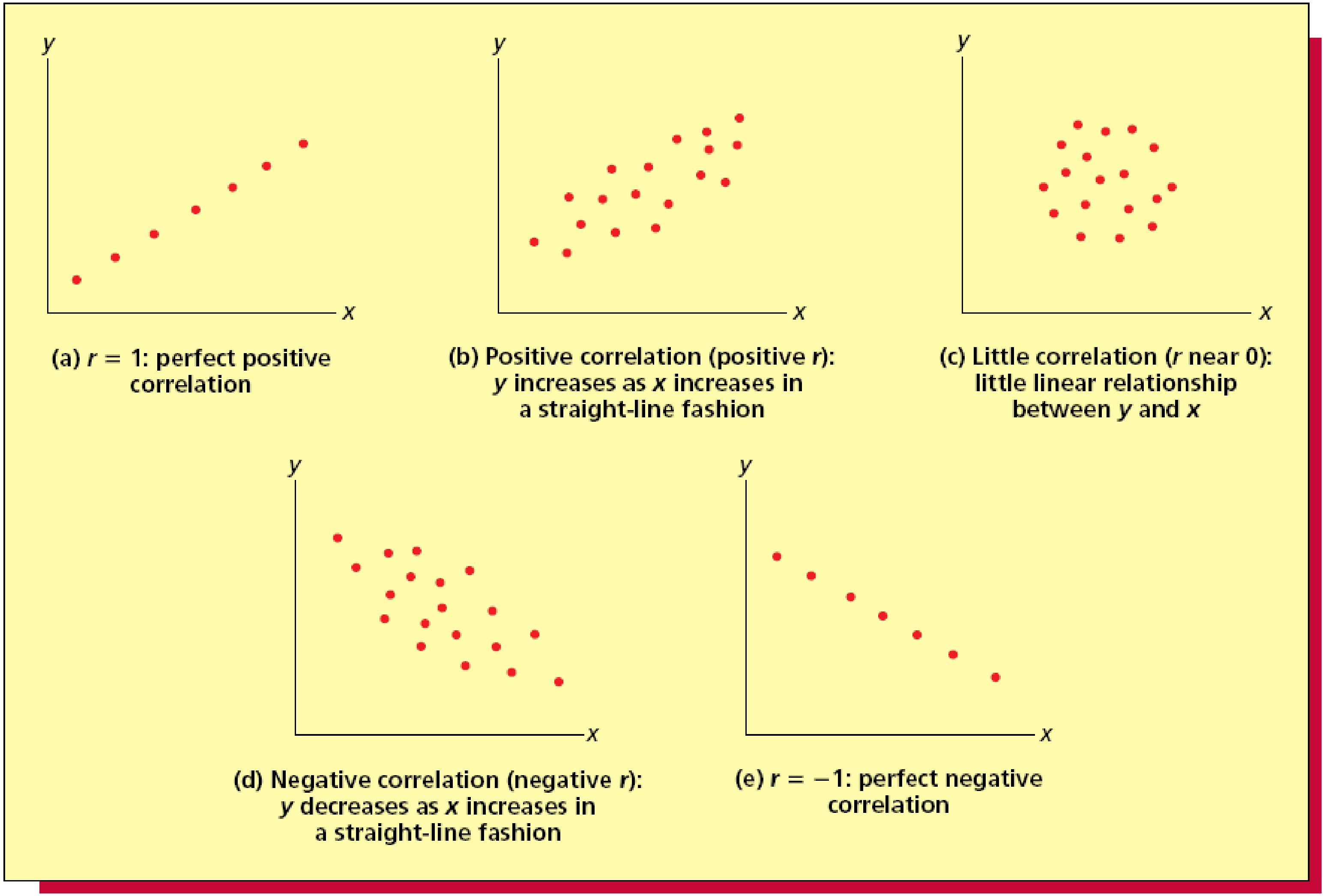
- The simple correlation coefficient measures the strength of the linear relationship between y and x and is denoted by r

$$r = +\sqrt{r^2} \text{ if } b_1 \text{ is positive, and}$$

$$r = -\sqrt{r^2} \text{ if } b_1 \text{ is negative}$$

- Where b_1 is the slope of the least squares line

Different Values of the Correlation Coefficient



Testing the Significance of the Population Correlation Coefficient

- The simple correlation coefficient (r) measures the linear relationship between the observed values of x and y from the sample
- The population correlation coefficient (ρ) measures the linear relationship between all possible combinations of observed values of x and y
- r is an estimate of ρ

Testing ρ

- We can test to see if the correlation is significant using the hypotheses

  $H_0: \rho = 0$
  $H_a: \rho \neq 0$

- The statistic is $t = \dfrac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$

- This test will give the same results as the test for significance on the slope coefficient b_1

An *F* Test for Model

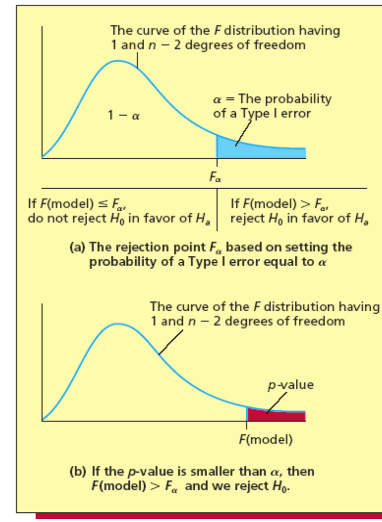
- For simple regression, this is another way to test the null hypothesis

  $H_0: \beta_1 = 0$

- This is the only test we will use for multiple regression
- The F test tests the significance of the overall regression relationship between x and y

Mechanics of the F Test

- To test H_0: $\beta_1 = 0$ versus Ha: $\beta_1 \neq 0$ at the α level of significance
- Test statistics based on F



(a) The rejection point F_α based on setting the probability of a Type I error equal to α

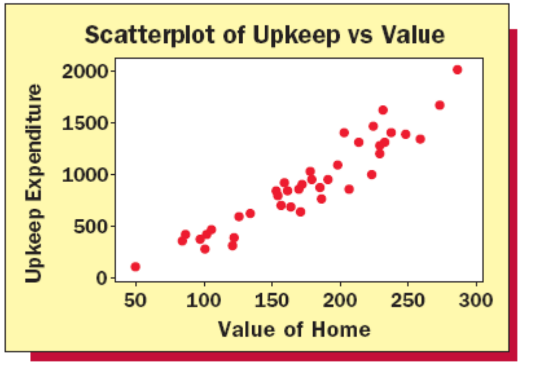
(b) If the p-value is smaller than α, then F(model) $> F_\alpha$ and we reject H_0.

$$F = \frac{\text{Explained variation}}{(\text{Unexplained variation})/(n-2)}$$

- Reject H_0 if F(model) > F_α or p-value < α
- F_α is based on 1 numerator and n-2 denominator degrees of freedom

The QHIC Case

- Quality Home Improvement Center (QHIC) operates five stores
- Wish to study relationship between home value and yearly expenditure on home upkeep
- Random sample of 40 homeowners
  - Intercept = –348.3921
  - Slope 7.2583



Residual Analysis

- Checks of regression assumptions are performed by analyzing the regression residuals
- Residuals (e) are defined as the difference between the observed value of y and the predicted value of y, $e = y - \hat{y}$
  - Note that e is the point estimate of ε
- If regression assumptions valid, the population of potential error terms will be normally distributed with mean zero and variance σ^2
- Different error terms will be statistically independent

Residual Analysis #2

- Residuals should as if they are randomly and independently selected from normal populations with mean zero and variance σ^2
- With any real data, assumptions will not hold exactly
- Mild departures do not affect our ability to make statistical inferences
- In checking assumptions, we are looking for pronounced departures from the assumptions
- So, only require residuals to approximately fit the description above

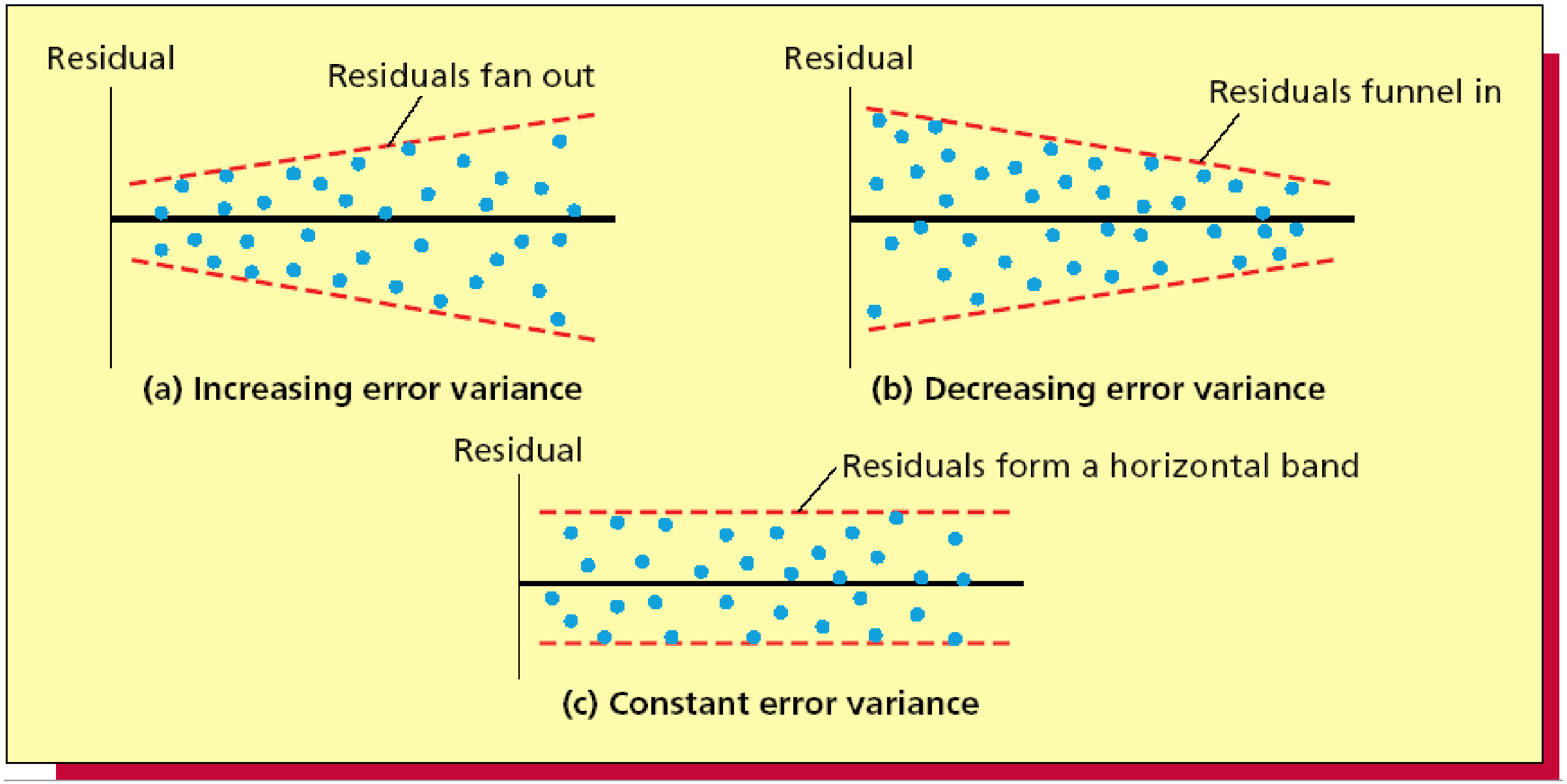
Residual Plots

1. Residuals versus independent variable
2. Residuals versus predicted y's
3. Residuals in time order (if the response is a time series)

Constant Variance Assumptions

- To check the validity of the constant variance assumption, examine residual plots against
  - The x values
  - The predicted y values
  - Time (when data is time series)
- A pattern that fans out says the variance is increasing rather than staying constant
- A pattern that funnels in says the variance is decreasing rather than staying constant
- A pattern that is evenly spread within a band says the assumption has been met

Constant Variance Visually



Assumption of Correct Functional Form

- If the relationship between x and y is something other than a linear one, the residual plot will often suggest a form more appropriate for the model
- For example, if there is a curved relationship between x and y, a plot of residuals will often show a curved relationship

Normality Assumption

- If the normality assumption holds, a histogram or stem-and-leaf display of residuals should look bell-shaped and symmetric
- Another way to check is a normal plot of residuals
  - Order residuals from smallest to largest
  - Plot $e_{(i)}$ on vertical axis against $z_{(i)}$
    - $Z_{(i)}$ is the point on the horizontal axis under the z curve so the area under this curve to the left is (3i-1)/(3n+1)
- If the normality assumption holds, the plot should have a straight-line appearance

Independence Assumption

- Independence assumption is most likely to be violated when the data are time-series data
    - If the data is not time series, then it can be reordered without affecting the data
    - Changing the order would change the interdependence of the data
- For time-series data, the time-ordered error terms can be autocorrelated
    - Positive autocorrelation is when a positive error term in time period i tends to be followed by another positive value in i+k
    - Negative autocorrelation is when a positive error term in time period i tends to be followed by a negative value in i+k
- Either one will cause a cyclical error term over time

Independence Assumption Visually