Lecture 23:
Least-Square Regression Line; Residual Plot

Chapter 3: Bivariate, Multivariate Data and Distributions

Review

1. Scatter plots
   - –Used to plot sample data points for bivariate data (x, y)
   - –Plot the *(x,y) pairs directly on a rectangular coordinate*
   - –Qualitative visual representation of the relationships between the two variables
   - –*no precise statement can be made*

2. Pearson's (linear) Correlation Coefficient, r,
   - – measures the direction (+ or -) and strength of linear relationship between x and y observations.
   - – Properties
   - -- Concerns and Cautions about r.

3. Association does not imply Causation…

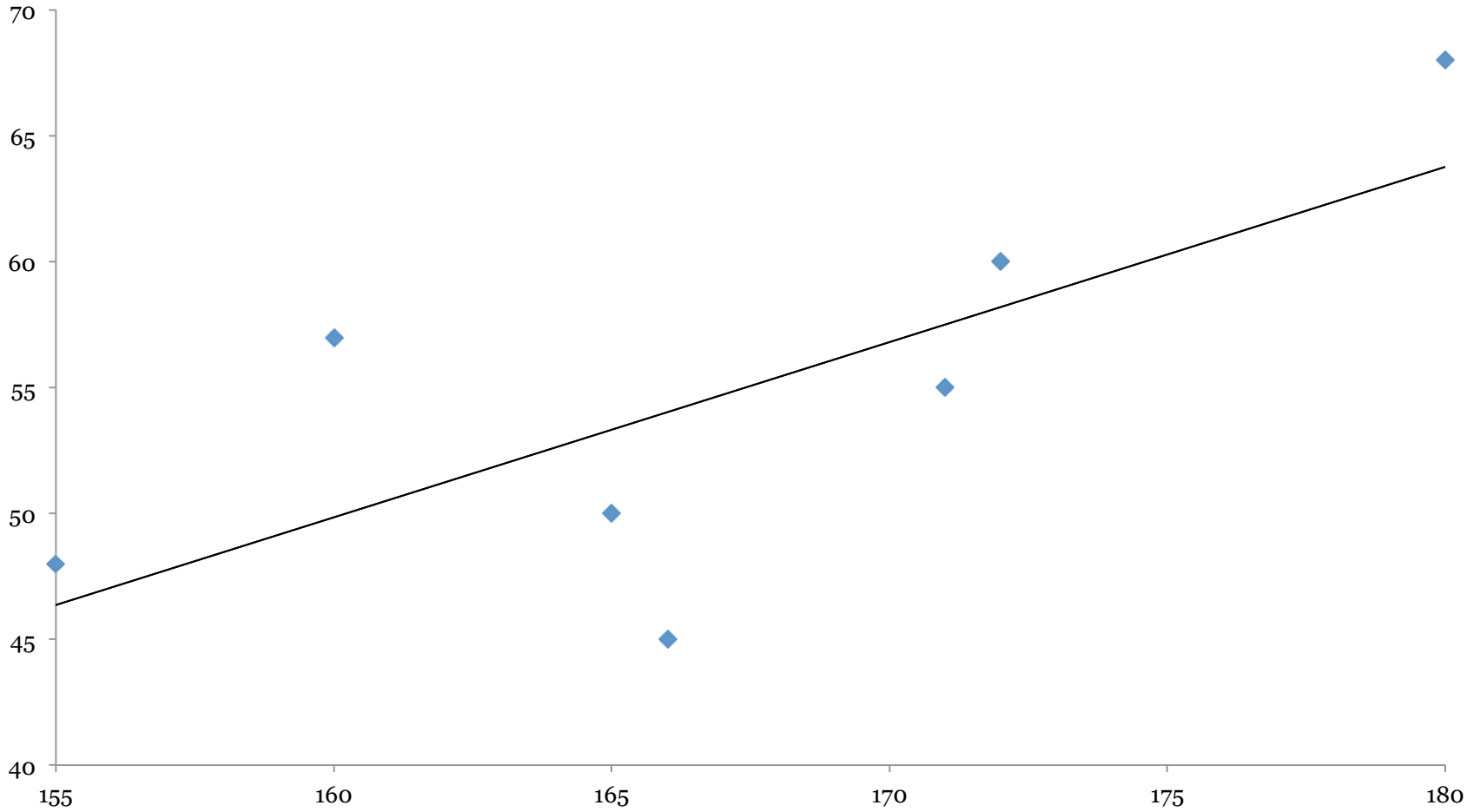
3.3
Fitting a Line (Regression line)

- If our data shows a linear relationship between *X* and *Y*, we want to find the line which best describes this linear relationship
  - Called a Regression Line

- Equation of straight line: $\hat{y} = a + b\,x$
  - a is the intercept (where it crosses the y-axis)
  - b is the slope (rate)

- Idea:
  - Find the line which best fits the data
  - Use this line to predict what happens to *Y* for given values of *X* (How does *Y* change as *X* changes?)

Example



Least Squares Regression Line

- Regression line is: $\hat{y} = -61.53 + 0.696x$
  - How do we know this is the right line?
  - What makes it best?
- The line above is the *Least Squares Regression Line*
  - It is the line which makes the vertical distances from the data points to the line as small as possible
  - Uses the concept of sums of squares
    - Small sums of squares is good → Least Squares!
    - See previous slide.

Finding the Least Squares Regression Line

- The Least Squares method relies on taking partial derivatives with respect to the slope and intercept which provides a solvable pair of equations called *normal equations* (see page 116)
- The solution gives:

$$b = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} = \frac{S_{xy}}{S_{xx}},$$

$$a = \overline{y} - b\overline{x}$$

Alternate calculations

- Understanding the regression line:

$$b = r\left(\frac{s_y}{s_x}\right) \qquad a = \bar{y} - b\bar{x}$$

- Meaning of slope
  - If I change *X* by 1 unit, how much does *Y* change?
  - "Rise over run" concept
  - Directly related to the correlation
- Meaning of intercept
  - Mostly of the time we don't care
    - It's simply a feature of the line
  - Sometimes has meaning
    - What should *Y* be if X=0
    - Want to use line for prediction

Assessing the fit

- How effectively does the least squares line summarize the relationship between *X* and *Y*? OR how well does the line fit the data?
- We assess the fit of the line by looking at how much variation in *Y* is explained by the regression line on *X*
  - Again this is the concept of sums of squares

Breaking up Sums of Squares

- Total Sums of Squares = *SST(O)* = $\sum (y_i - \bar{y})^2$
  - measures the total variation in *Y*
- Break into two pieces:
  - Regression Sums of Squares = *SSR* = $\sum (\hat{y}_i - \bar{y})^2$
    - Part we are explaining with our regression line
  - Error Sums of Squares = *SSE* = $\sum (y_i - \hat{y}_i)^2$
    - Part we can't explain, unexplained variation
    - Leftover, Residual, or Error
    - Also called SSResid
- If *SSE* is small, we can assume our fit is good
  - But how small is small?

Coefficient of Determination

- r^2 is given by: $r^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
- Notice when *SSR* is large (or SSE small), we have explained a large amount of the variation in *Y*
- Multiplying r^2 by 100 gives the percent of variation attributed to the linear regression between *Y* and *X*
  - Percent of variation explained!
- Also, just the square of the correlation!

Standard Deviation LS line

- Mean Squared Error about the LS line (with sample size = n):

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

- Standard Deviation about the LS line:

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

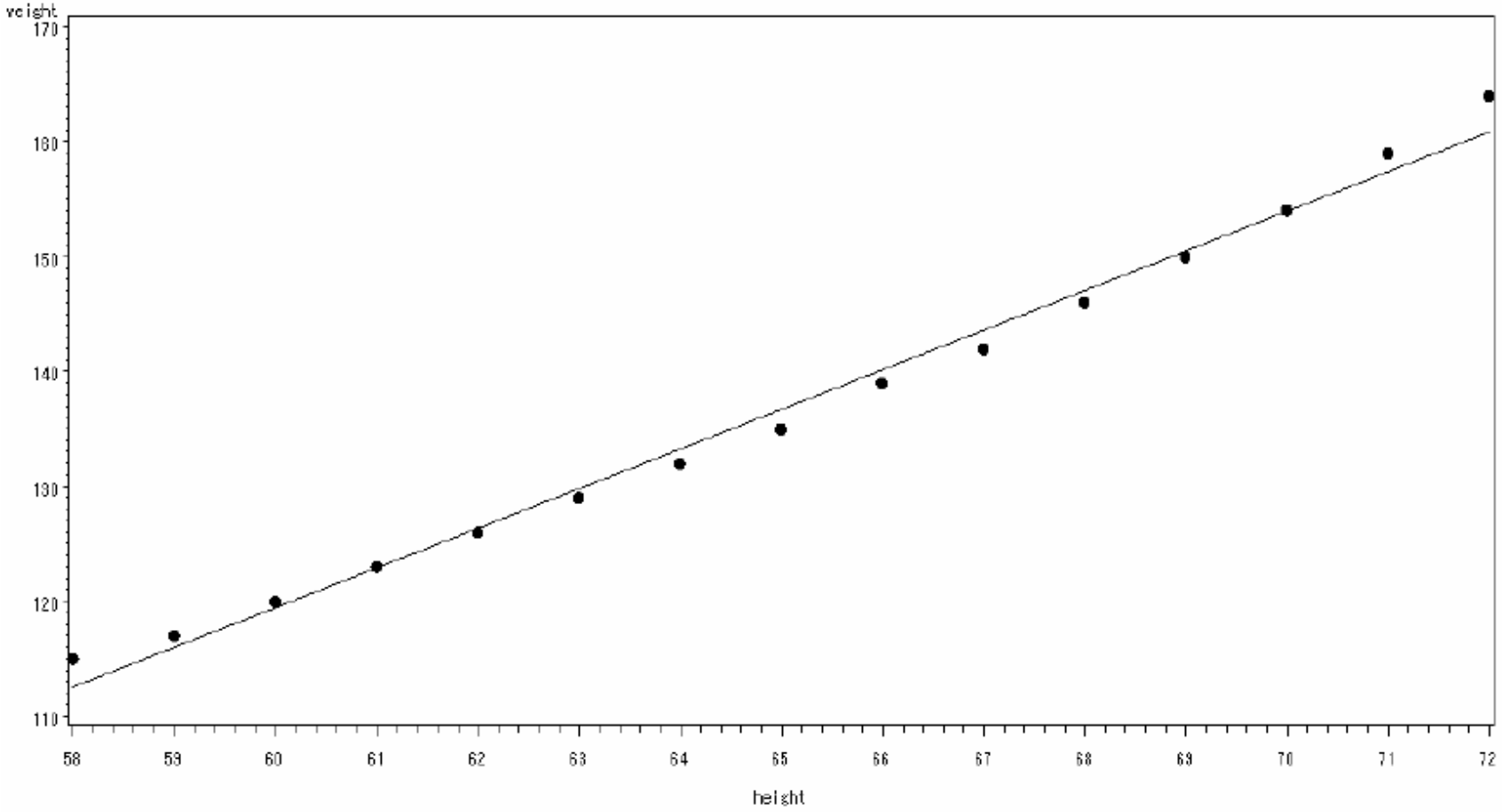
  n – 2 comes from the degrees of freedom.
- It is the typical amount by which an observation varies about the regression line
- Also called “root MSE” or the square root of the Mean Square Error

Example—Height and Weight

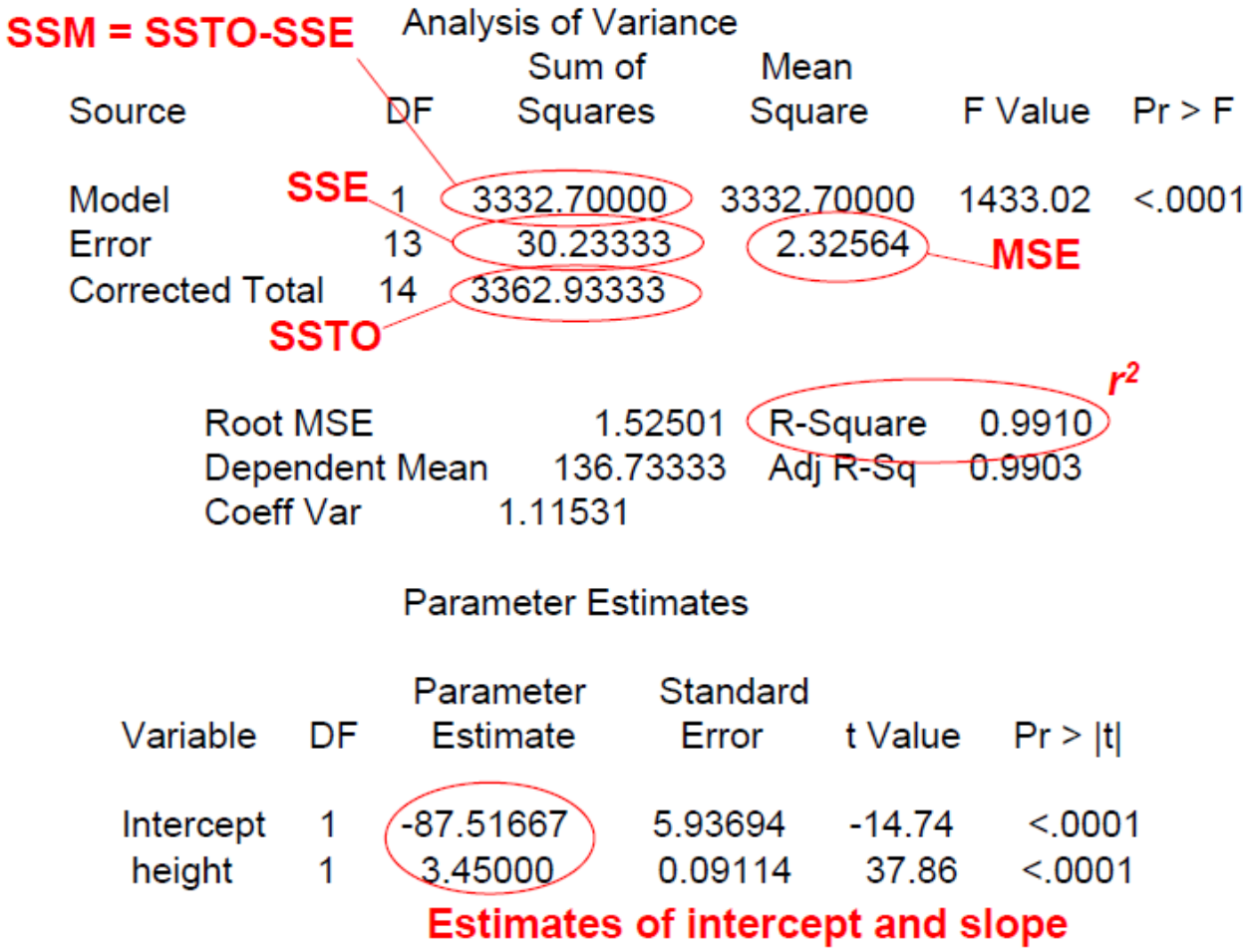
- The following data set gives the average heights and weights for American women aged 30-39 (source: *The World Almanac and Book of Facts, 1975)*. Total observations *15*.

(in) x_i	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
(lb) y_i	115	117	120	123	126	129	132	135	139	142	146	150	154	159	164

Example—Height and Weight



Example—Height and Weight



Example—Height and Weight

- What is the estimated regression line?

$$\hat{y} = a + bx = -87.52 + 3.45x$$

- Using the line, predict the weight of women 73in tall.

$$a = 3.45 \quad b = -87.52$$

$$3.45 \times 73 - 87.52 = 164.33 \text{ (lb)}$$

So our prediction would be 164.33 lb

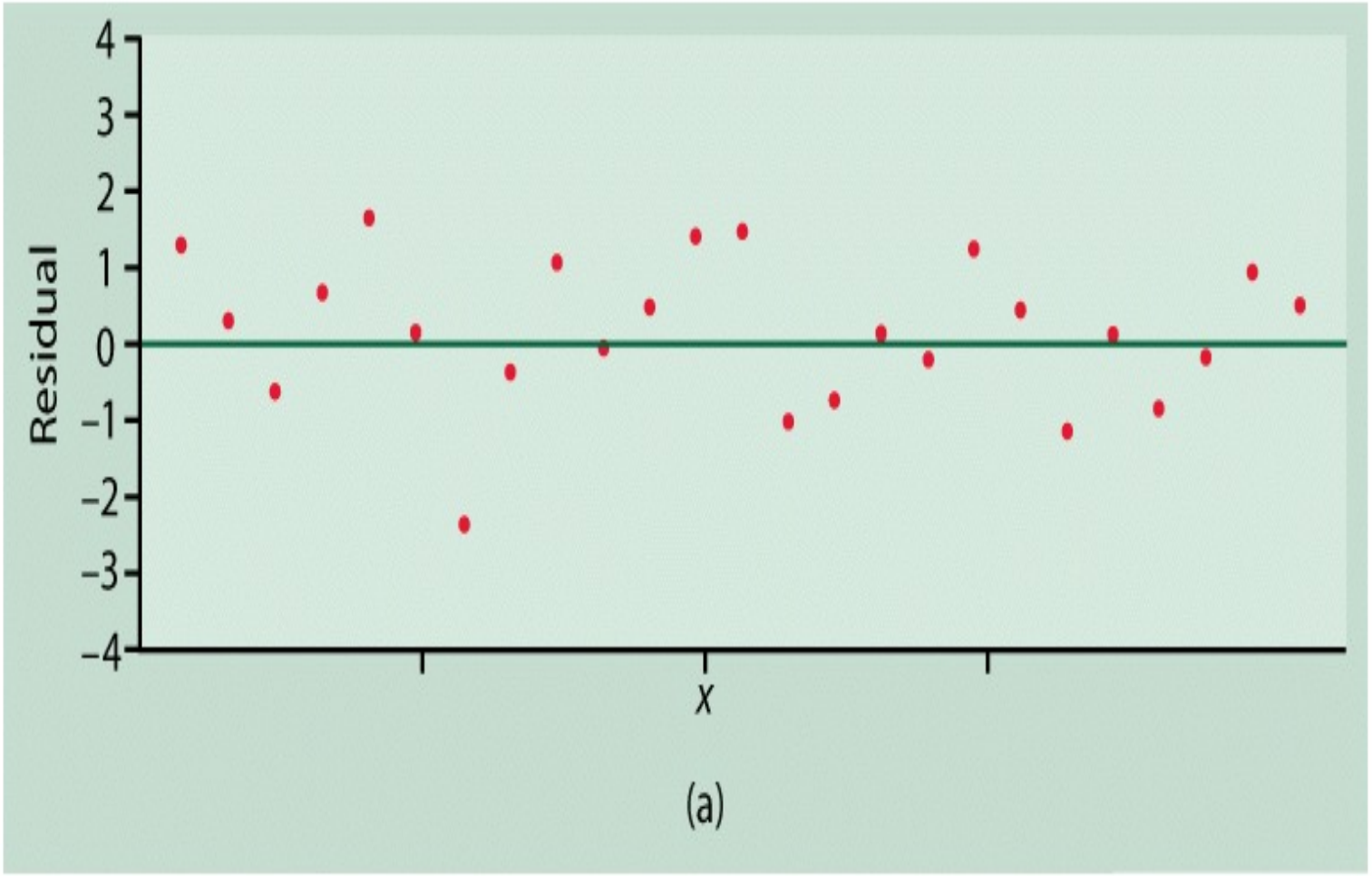
Another way to assess appropriateness: Residual Plots

- The residuals can be used to assess the appropriateness of a linear regression model. A residual of y_i is given as

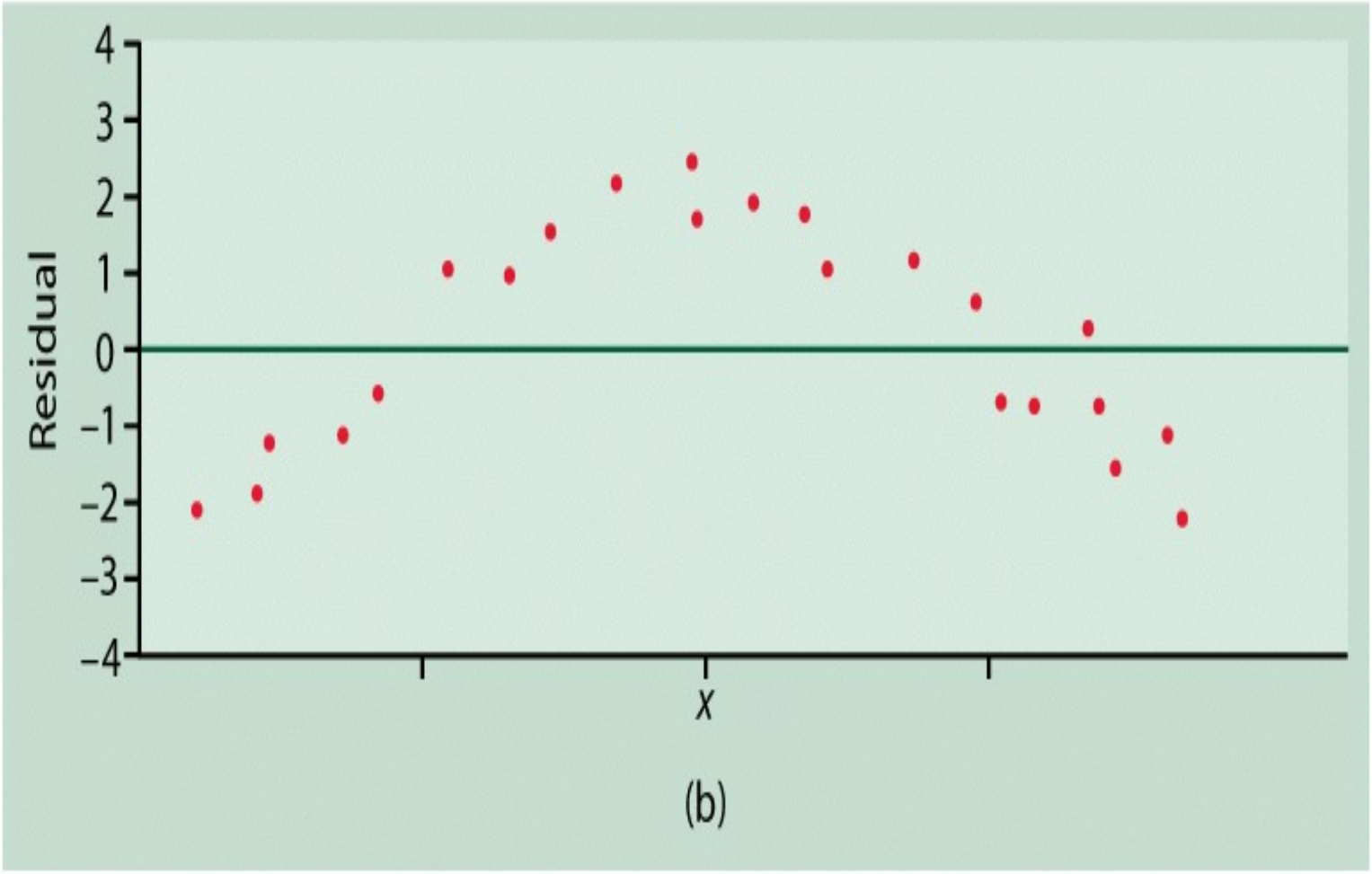
$$e_i = y_i - \hat{y}_i$$

- Specifically, *a **residual plot***, plotting the residuals against x, gives a good indication of whether the model is working
  - The residual plot should not have any pattern but should show a random scattering of points
  - If a pattern is observed, the linear regression model is probably not appropriate.

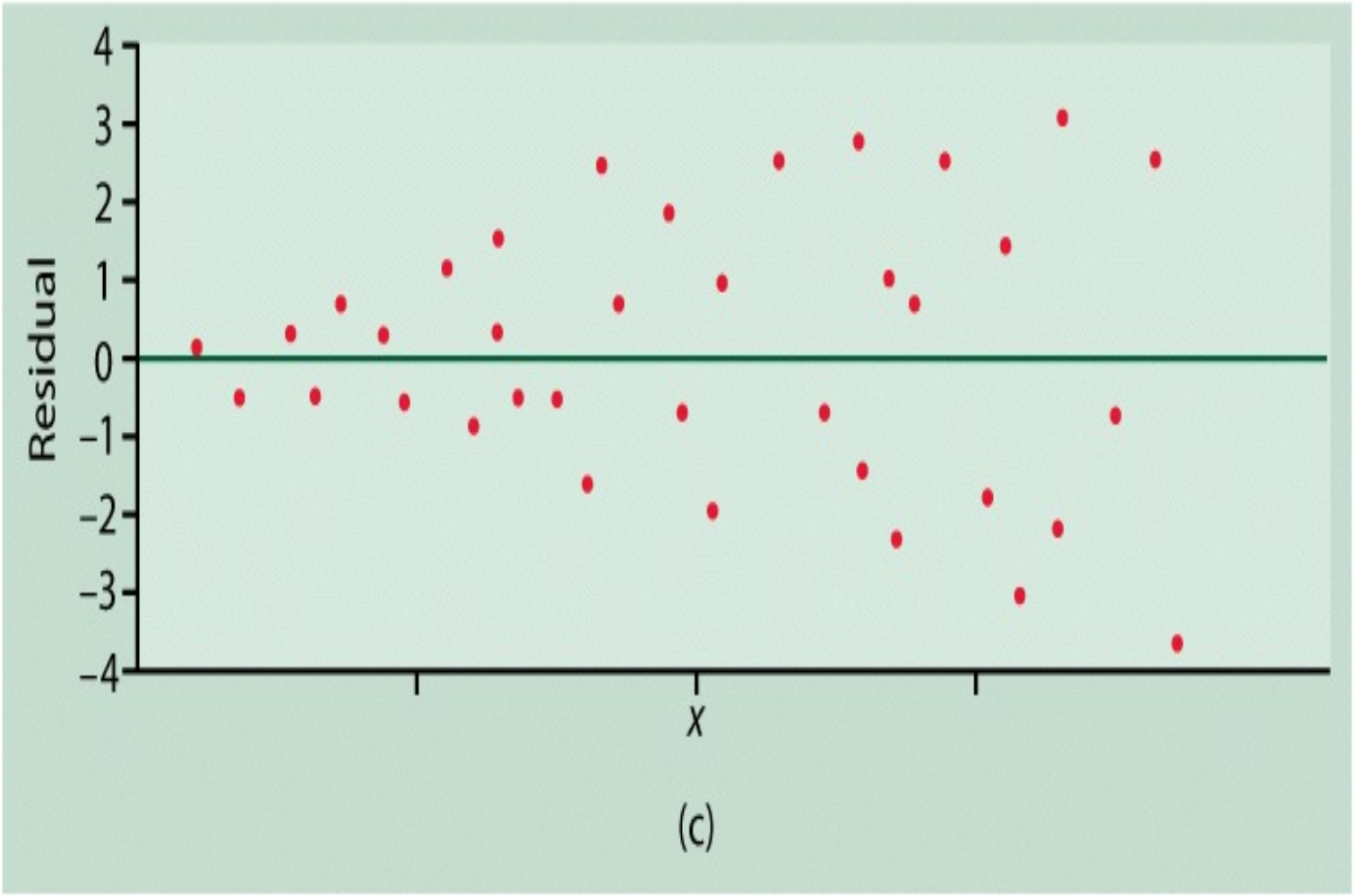
Examples—good



Examples—linearity violation



Examples—
constant variance violation



If Bad residual plots… try transformations!

- Can think of transformations as simple mathematical manipulations.
  - Suppose x doesn't predict y well but x is a very good predictor of log y.
  - Before we ever start, let new_y = log y! Then just fit a linear regression between x and new_y!!!
  - Is it still linear?
    - Yes! x and new_y should have a nice linear relationship
    - No! If I want to describe the "real" relationship between x and y I need to "undo" the transformation. Meaning explain it as a logarithm.
- Also see power transformations in Section 3.4 (self-reading, not covered in exams)

After Class …

- Review Section 3.1 through 3.4
- Read Section 3.5

- **Next Wed (5pm)**, Hw#10.
- **Next Wed,** Lab #6.