ST 352
Section 2.4
Residuals & Residual Plots

Return to the clear cut example:

Landslides are common events in tree-growing regions of the Pacific Northwest, so their effect on timber growth is of special concern to forest The article *"Effects of Landslide Erosion on Subsequent Douglas Fir Growth and Stocking Levels in the Western Cascades, Oregon"* (Soil Scie Society of American Journals (1984)) reported on the results of a study in which growth in a landslide area was compared with growth in a previously clear-cut area. Here we consider clear-cut growth areas only. Data on the age of the tree and its most recent 5-year height growth cm) are given below:

age (years)	5-yr growth (cm)	age (years)	5-yr growth (cm)
5	70	13	305
9	150	13	335
9	260	14	290
10	230	14	340
10	255	15	225
11	165	15	300
11	225	18	380
12	340	18	400

1. Residuals

   a. What was the 5-year growth of the 12-year old tree in this sample?

   b. Using the least-squares regression line on the previous handout ($\hat{y} = 4.3516 + 21.3217x$), what would you **predict** the 5-year growth of the 12-year old tree to be?

   c. The difference between the observed growth and the **predicted** growth (expected growth) for any tree is called the ***residual***. That is:

      residual = observed value – predicted value

      Calculate the residual for the 12-year old tree.

   d. Residuals can be calculated for each of the trees in the sample. Below is a table the age of the tree (years), the 5-year growth of the tree (cm), the **predicted** 5-year growth, and their residuals.

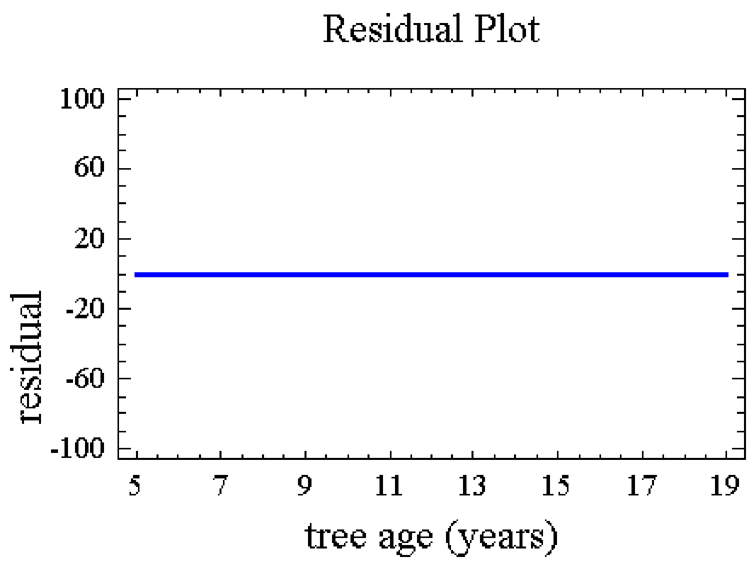
age	growth	predicted values	residuals	age	growth	values	residuals
5	70	110.96	-40.9601	13	305	281.534	23.4663
9	150	196.247	-46.2469	13	335	281.534	53.4663
9	260	196.247	63.7531	14	290	302.855	-12.8554
10	230	217.569	12.4314	14	340	302.855	37.1446
10	255	217.569	37.4314	15	225	324.177	-99.1771
11	165	238.89	-73.8903	15	300	324.177	-24.1771
11	225	238.89	-13.8903	18	380	388.142	-8.14214
12	340	260.212	79.788	18	400	388.142	11.8579

   e. What would you expect the sum of all 16 residuals to be? Therefore, what would you expect the average of the residuals to be? Verify this.

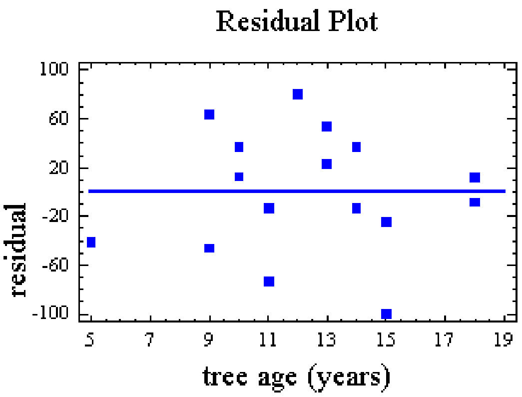
2. Residual plot

   a. A residual plot is a scatter plot of the residuals versus the explanatory variable, with the residuals on the y-axis and the explanatory variable (age) on the x-axis. Using the idea from #1 part (e), we would place 0 in the middle of the y-axis when making a scale for the residuals on the y-axis,.

Using the data above, draw the residual plot by hand.



b. Compare your residual plot to the residual plot from STATGRAHICS:



A "good" residual plot will look as if the points (residuals) are randomly scattered. That is, **<u>there will be no patterns to the residuals</u>**. Does that appear to be the case here?

c. The uses of the Residual Plot

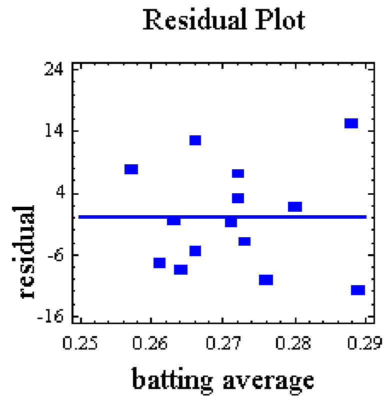
1. To check to see if a <u>linear</u> relationship exists between the explanatory and response variable. If the residuals form a non-linear pattern, the relationship between the explanatory and response variable will be non-linear.

2. To check to see if the variation in the residuals remains constant over the entire range of the explanatory variable – more on this in Chapter 10.

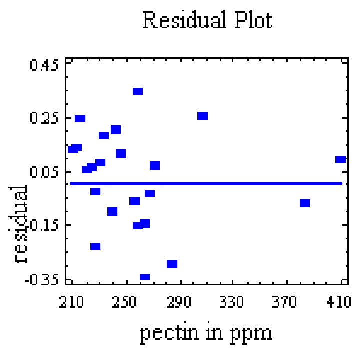
3. Identify possible outliers.

d. Other examples. Comment on each residual plot. (Is there any curved pattern among the residuals, or are they "randomly scattered"? Any other patterns or clusters that are apparent? Are there any outliers?)

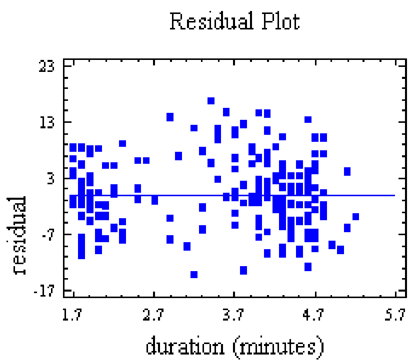
1. The baseball example (wins versus batting average)



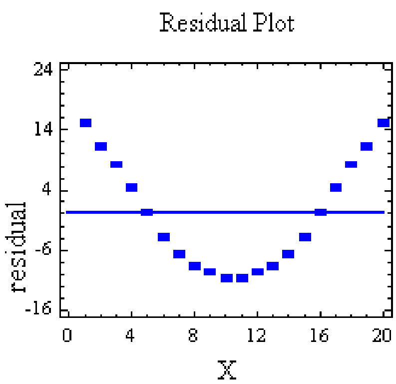
2. The orange juice example (sweetness index versus pectin)



3. The Old Faithful Geyser example (interval between eruptions versus duration of last eruption)



4. Y versus X example (the non-linear pattern)



e. Draw a Residual Plot that shows non-constant variation among the residuals.