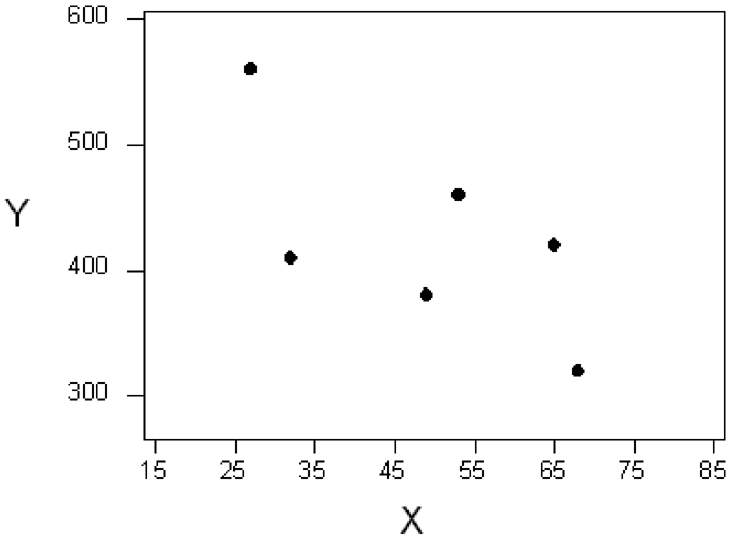
Least Squares Regression

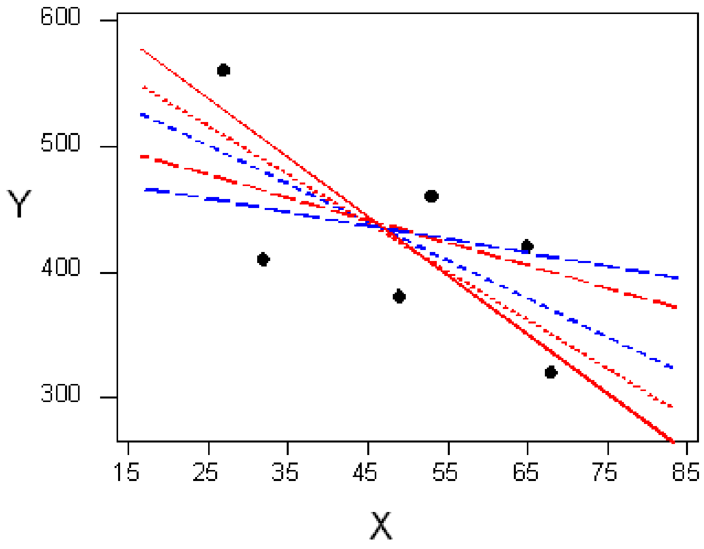
Learning Objective: In the special case of linear relationship, use the least squares regression line as a summary of the overall pattern, and use it to make predictions.

The technique that specifies the dependence of the response variable on the explanatory variable is called **regression**. When that dependence is linear (which is the case in our examples in this section), the technique is called **linear regression**. Linear regression is therefore the technique of finding the line that best fits the pattern of the linear relationship (or in other words, the line that best describes how the response variable linearly depends on the explanatory variable).

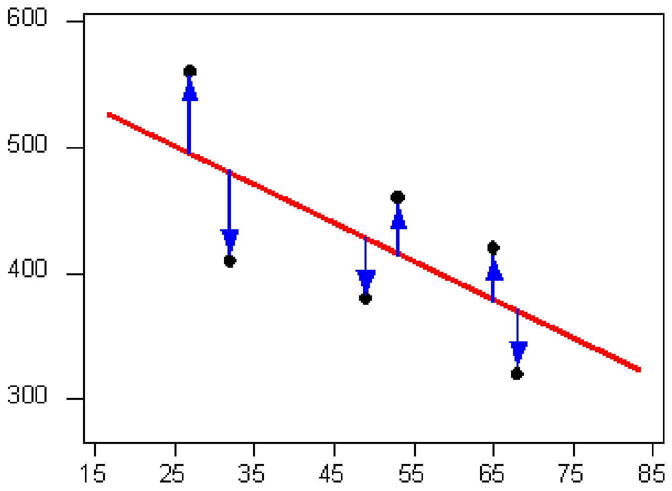
To understand how such a line is chosen, consider the following very simplified version of the age-distance example (we left just 6 of the drivers on the scatterplot):



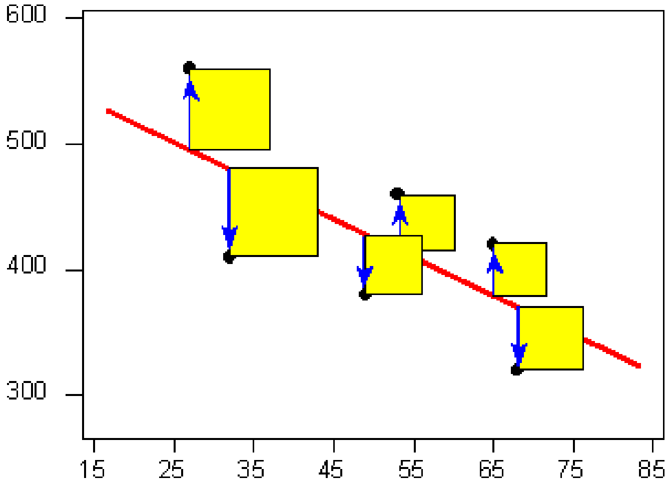
There are many lines that look like they would be good candidates to be the line that best fits the data:



It is doubtful that everyone would select the same line in the plot above. We need to agree on what we mean by "best fits the data"; in other words, we need to agree on a criterion by which we would select this line. We want the line we choose to be close to the data points. In other words, whatever criterion we choose, it had better somehow take into account the vertical deviations of the data points from the line, which are marked with blue arrows in the plot below:



The most commonly used criterion is called the **least squares** criterion. This criterion says: Among all the lines that look good on your data, choose the one that has the smallest sum of squared vertical deviations. Visually, each squared deviation is represented by the area of one of the squares in the plot below. Therefore, we are looking for the line that will have the smallest total yellow area.



This line is called the **least-squares regression line**, and, as we'll see, it fits the linear pattern of the data very well.