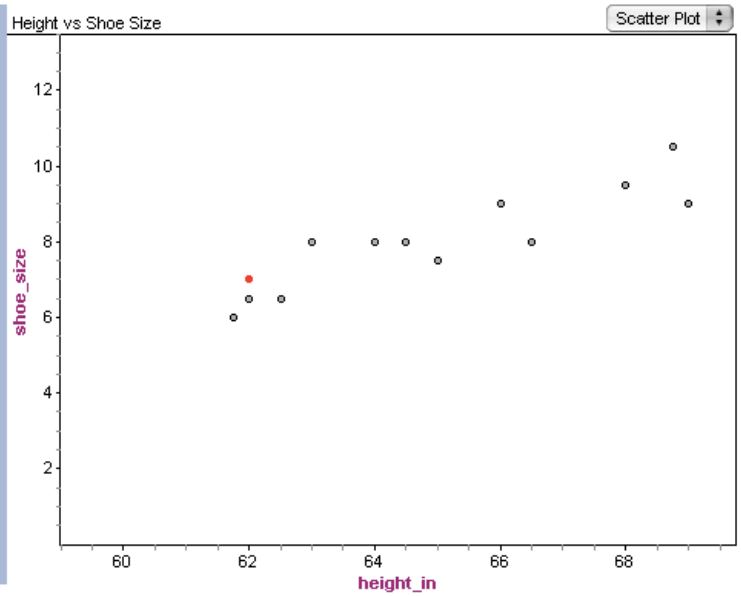
Regression
A Closer Look at the Video

After watching the *Regression* video, make sense of the mathematics by reading through the problem situation and solution. Use the comments and questions in bold to help you understand regression.

Problem: Sonya's friend, Tim, just started a new job at the bowling alley. He does all sorts of things from polishing bowling balls to handing out shoes. He was impressed by how the manager can just look at someone and tell what size shoe he/she wears. He uses his experience to guess the bowler's shoe size based on the person's height, by knowing that there is a positive correlation between the variables of height and shoe size. That is, typically taller people wear larger shoes. Sonya's friend, Jamie, needs shoes. We estimate her to be about 5'7" or 67 inches tall. Predict what size shoe she is likely to wear. Below is the scatter plot for the heights and shoe sizes of the last 15 women who rented shoes. For example, the highlighted point is a person of 62 inches with a shoe size of 7.

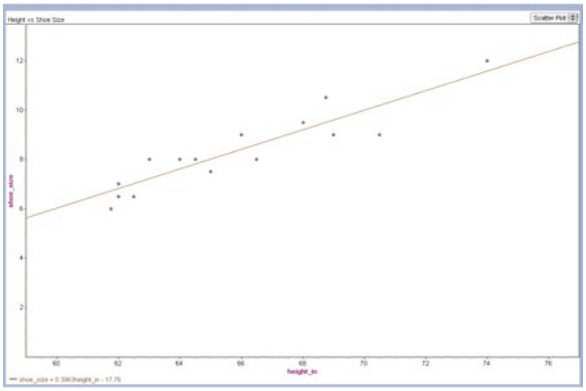
Height vs Shoe Size

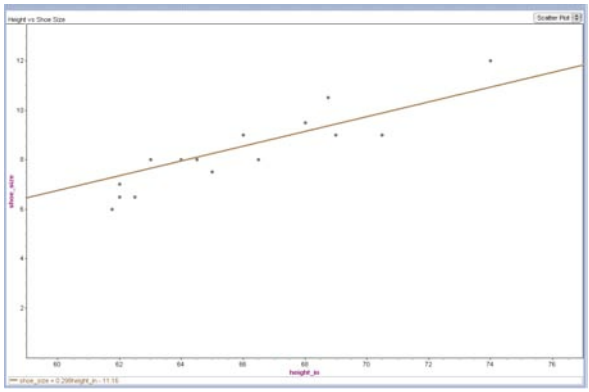
	height_in	shoe_size	<new>
1	62.00	6.50	
2	63.00	8.00	
3	62.00	7.00	
4	65.00	7.50	
5	69.00	9.00	
6	68.00	9.50	
7	74.00	12.00	
8	62.50	6.50	
9	70.50	9.00	
10	61.75	6.00	
11	66.50	8.00	
12	64.50	8.00	
13	64.00	8.00	
14	68.75	10.50	
15	66.00	9.00	



How can we use the information in the scatter plot to predict the shoe size of Jamie who is 67 inches tall?
First, we need to look at the data to see if there is a trend. The points don't quite form a line. If they did, we would be able use the equation of that line to predict the shoe size for any given height. Instead we will find a line that is close to the data and use it to predict shoe size.

Begin by drawing a line that seems close to the data points. Experiment by changing the slope of the line and by shifting it vertically. Notice, as we move the line, the equation changes.
Here are a couple possible lines.



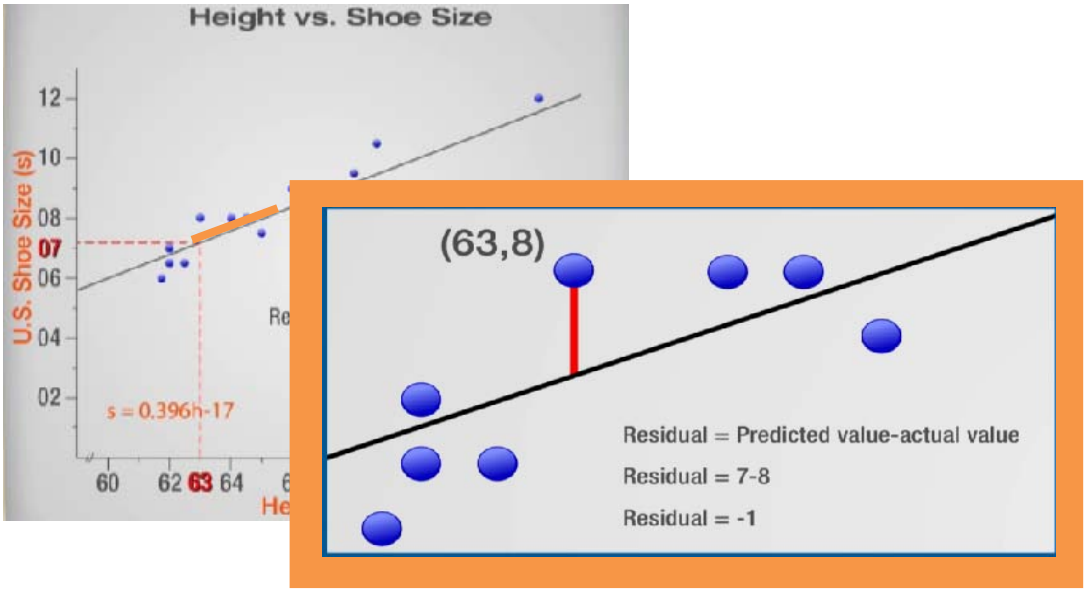


The line $s = 0.396h - 17$ **looks pretty good, but how can we measure how close is it to the actual data?**

To determine how close our line is to the actual data, we can calculate residuals.

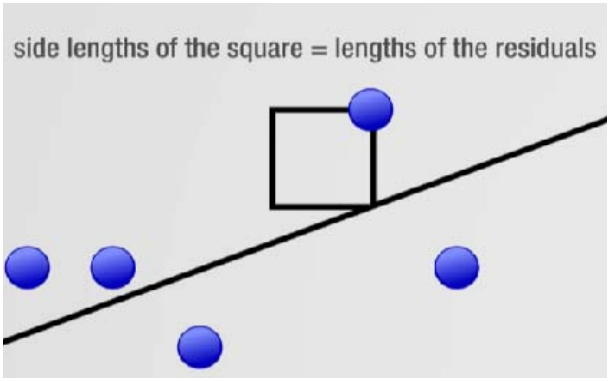
What is a residual?

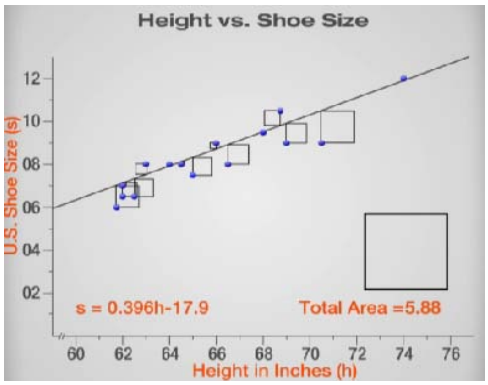
A residual is sort of like an error. It's the difference between a shoe size predicted by the line and the actual shoe size. Notice that for a woman 63 inches tall, our line predicts a size 7 shoe. However, our data set included a woman of this height, and her shoe size was 8. So, the residual at this point is $8 - 7 = 1$. We can calculate the residuals for each point in our data set.



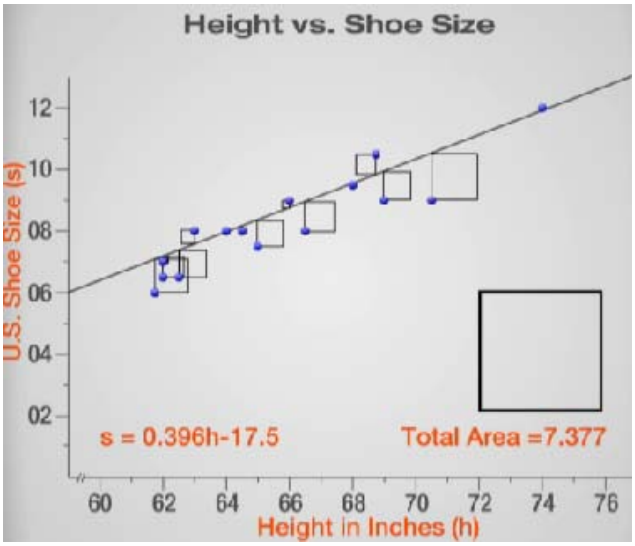
Remember, the line we are looking at isn't necessarily the best fit. We haven't found that yet.

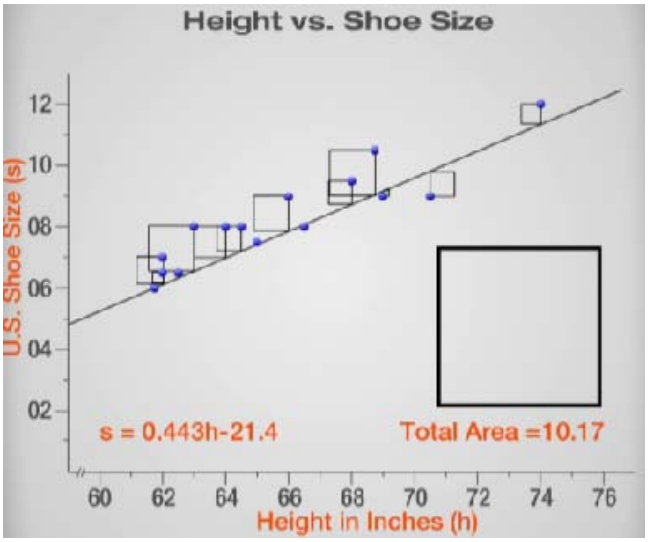
The most commonly used line of best fit is called the least-squares regression line.
Why is it called the least-squares regression line?
It is called the least-squares regression line because we are going to build squares whose sides are the lengths of the residuals. The least-squares regression line is the line for which the sum of the areas of these squares is as small as possible.





Let's experiment. When we move the line, we can see the total area or the sum of the areas of the squares changes, as the equation of the line changes.





When the total area is as small as possible, we have found the least-squares regression line.
Our least-squares regression line is $s = 0.4h - 18.4$.

What does the slope of the equation tell us?

The slope of the line is 0.4. Since we are working with shoe size, it makes sense to round 0.4 to one half. This means, if Jamie is one inch taller than her friend, she will wear about a half-size larger shoe.

What does the y-intercept tell us?
The y-intercept is approximately the point (0, -18). This would mean that if a woman is zero inches tall, she wears a size negative eighteen shoe. Of course, this does **not** make sense; the y-intercept for this regression line does not have a practical meaning.

What about our prediction for Jamie who is 67 inches tall?
We make our prediction by substituting 67 inches for the height in our equation. Calculating, we get about 8.8, or between about a size 8½ and a 9.

The least-squares regression line can be a useful tool to help make predictions, especially if the data is clustered about the line. For what values does a least-squares regression line give the best predictions?
The best predictions are for values that fall between the smallest and the largest values of the data set. If you try to make predictions beyond the data, or extrapolate, the predictions may not be as reliable.