Answer Key - Finding the Relationship Between the Regression Line and the Centroid

1. (a) Take the average value (mean) of the two *x*-coordinates and the average value of the two *y*-coordinates to get the coordinates of the centroid.
   (b) The midpoint should be the same point. The reason they are the same is that if a point is halfway between the two points, then its *x*-value is halfway between the two *x*-values and its y-value is halfway between the two *y*-values, and vice versa.
   (c) The centroid on the graph should have the same coordinates as the point you calculated in (a).

2. (a) Take the average value of the three *x*-coordinates and the average value of the three *y*-coordinates to get the coordinates of the centroid.
   (c) The centroid is on the line.

3. The centroid is always on the line.

Reflection Questions

1. Having the centroid on the regression line indicates that the mean of the output data should correspond to the mean of the input data. For example, if a person has height equal to the mean of the height measurements, then their weight should equal the mean of the weight data. In other words, a person of average height should have average weight.

2. It depends. Certainly, it is most unlikely that the weight will be exactly as predicted by the line. How far off the prediction is will depend on how spread out the data points were, how closely the data points clustered around the line. In general, the better the correlation coefficient, the better the prediction is likely to be, although we saw some anomalies in Part 3 of this i-Math. Another factor is whether the new height is within the range of the heights we measured, or whether it is outside that range. Generally, if the correlation coefficient is good and the new height is within the range of the data, we can expect the line to be a good predictor of weight. If it is outside the range of the data, we don't know what other factors may be affecting the height to weight relationship. There may be a different relationship between height and weight for excessively short or excessively tall people, whose heights are outside the range of the data.

3. It is most unlikely that the lines would be the same. Either could be used to make predictions. What is generally done if two sets of data are available, if both sets came at random from the same population at approximately the same time under similar circumstances, is to pool the data in some way. For example, one could put all the points together into one large data set and use the line that fits this big data set.

4. No. The line obtained for the switched data would not be the inverse of the original line, because the regression line is obtained by minimizing the sum of squares for the variable on the vertical axis. Suppose we switch the roles of height and weight in the above example to try to predict height given weight, making weight the input and height the output. In the original data, fitting the line involved minimizing the sum of squares for the weights. In the switched data, fitting the line involves minimizing the sum of squares for the heights. There is no reason to expect these two to give the same results, that is, to give a line that is the inverse of the original.