Ch. 10: Regression

The correlation coefficient r is a number that describes the linear relationship of two variables. _____ allows us to make predictions about one variable based on a second variable.

Ex: In the "Performance of Stats Students" data, we have

quiz-avg (x) = 44 final-avg (y) = 85
$SD_{quiz} = 12.5$ $SD_{final} = 25$ $r = 0.75$

Question: How can we predict the points in the Final based on the points in the Quiz for a student? Can we simply use the SD line?

What else could we use?

The _____ for y on x estimates the average value of y corresponding to each value of x.

This line says that associated with an increase of one SD in x, there is only an increase of r·SD's in y.

The regression line goes through the point of averages, but its slope is ...

Ex: What is the slope of the regression line for the "Performance of Stats Students" data set?

If someone scored 56.5 points in the Quiz (i.e., one SD above avg), what is their predicted score in the Final?

If someone scored 19 points in the Quiz (i.e., two SD below avg), what is their predicted score in the Final?

In general, we can predict the y-value for any given x-value as follows:

1. Convert your independent (explanatory) variable into standard units.

2. Mulitply by r. This will give the predicted value of the dependent (response) variable in standard units.

3. Convert this predicted value back into its own units.

Ex:

If someone scored 60 points in the Quiz, what is their predicted score in the Final?

If someone scored 40 points in the Quiz, what is their predicted score in the Final?

If someone scored 0 points in the Quiz, what is their predicted score in the Final?

Question: How does the regression line look in some special cases?

- If r is 0, there is no association, and the average of y doesn't change when we change x.

- If r is 1 or -1, we have perfect association, and y changes 1 SD when x changes 1 SD.

Ex: The Law School Admissions Test is a standardized test with an average score of 500 and SD 100. A large group of students took the LSAT, and then attended a LSAT-preparation class before taking it a second time. The correlation between the two scores was 0.6.

On the second test, the group of students who received a 300 on the first test improved to a 380 on average. But the students who received a 700 on the first test dropped to 620 on average. What's going on?

When correlation is less than perfect, there will be a football-shaped spread of the data points around the SD line. This spread makes the bottom group (for the first variable) rise on average, while the top group (for the first variable) drops on average. This is called the _____, and it is caused simply by the spread around the line.

The _____ is that people try to see some significance in these different outcomes.

Two Regression Lines

If there is no obvious explanatory (independent) and no obvious response (dependent) variable, then there are *two* regression lines.

Since r is symmetric, we can regress either variable on the other.

Ex: Score in Physics Quiz and Stat Quiz - which is the explanatory and which the response variable?

Ex: Age and height of children - which is the explanatory and which the response variable?