Class Worksheet

February 21 to 26

Math 125 *Kovitz* 2025

Regression

The regression line is to a scatter diagram as the average is to a list. The regression line for y on x estimates the average value for y corresponding to each value of x.

Associated with each increase of one SD in x there is an increase of only r SDs in y, on the average. (This way of using the correlation coefficient to estimate the average value of y for each value of x is called the regression method.)

A formula for the estimate of the average y for a given x is

$$\text{estimated average} = \text{average } y + (x \text{ in standard units} \times r \times \text{SD of } y)$$

The regression line is a smoothed version of the graph of averages. If the graph of averages follows a straight line, that line is the regression line.

To find y on the SD line for a given x, we would use the formula

$$y \text{ (on SD line)} = \text{average } y + (x \text{ in standard units} \times \text{SD of } y)$$

A formula predicting the value of y for an individual with a given x is

$$\text{predicted } y = \text{average } y + (x \text{ in standard units} \times r \times \text{SD of } y)$$

This is of course the same formula that was used to estimate the average y for that given x.

In this case a prediction rather than an estimate is in order because the value of y that turns up for the given x depends on which individual is chosen among all those with the given x. The actual average—on the other hand—is a certain value that we are trying to estimate using the regression line.

Percentiles and the Regression Method

When both values, x and y, follow the normal curve, regression techniques can also be used to predict percentile ranks. We need to know how many SDs a given x is above the average. The percentile rank has this information, in disguise—because the x-values follow the normal curve. The number of SDs that the y-value is predicted to be above or below average can then be translated directly into a percentile without actually using the averages and SDs of the two variables. All that will matter is r. Basically, this is because the whole problem can be worked in standard units. The percentile ranks give the standard units, but in code.

If the two variables are perfectly correlated, it makes sense to predict that the percentile of y will be equal to the percentile of x. At the other extreme, if the correlation is zero, then the percentile of x does not help at all in predicting the percentile of y. In the absence of other information, the safest guess is to put the prediction of y at the median or the 50th percentile. In fact when the correlation is somewhere between the two extremes of $r = 1$ and $r = 0$, we have to predict a value of y somewhere between the percentile of x and the median (50th percentile).

The case of r between -1 and 0 must be treated slightly differently. It is omitted here.

The Regression Fallacy

In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test—and the top group will on average fall back. This is the *regression effect.*

> The regression fallacy consists of thinking that the regression effect must be due to something important, not just the spread around the line. The uncertainty of the estimate of the regression line of y on x will lead to more uncertainty when that estimate is subsequently used to estimate x from y using the regression line of x on y; thus the result will not be the original x.

There are Two Regression Lines

For a given scatter diagram, there are two regression lines. Each line applies to the case where a particular one of the two variables is given and the other is estimated from the given. The case where x is given and y is estimated from x is called regression of y on x.

> Unless $r = +1$ or $r = -1$, the two regression lines will be different.

Problems to think about

A survey of heights and weights of a group of men led to the following results:

$$\begin{array}{rll} \text{average height} \approx 67 \text{ inches,} & \text{SD} \approx 4 \text{ inches} & \\ \text{average weight} \approx 152 \text{ pounds,} & \text{SD} \approx 30 \text{ pounds,} & r \approx 0.80 \end{array}$$

Estimate the average weight of those in the survey whose height was

(a) 67 inches (b) 68 inches (c) 75 inches (d) 60 inches

Estimate the average height of those in the survey whose weight was

(a) 208.25 pounds (b) 152 pounds (c) 143 pounds (d) 107 pounds

Suppose that a 73-inch-tall person falls on the SD line. Find his weight.

> True or false: in this example, any individual who falls on the SD line and has height greater than 67 inches must weigh more than the regression equation would predict for an individual of that height.

Predict the weight of an individual whose height is

(a) unknown (b) 71 inches (c) 73 inches (d) 65 inches

Estimate the average height of those men in the survey who weighed 188 pounds, and compare this part with the estimate of the average weight of those men in the survey who were 73 inches tall.

Suppose that the heights and weights in the survey followed the normal curve.

(a) Predict the percentile rank on weight for a man in the survey who was at an unknown percentile in height.

(b) Predict the percentile rank on weight for a man in the survey who was at the 50th percentile in height.

(c) Predict the percentile rank on weight for a man in the survey who was at the 2nd percentile in height.

(d) Predict the percentile rank on weight for a man in the survey who was at the 92nd percentile in height.

(e) Predict the percentile rank on height for a man in the survey who was at the 86th percentile in weight.

(f) True or false: the answer to part (e) could be directly inferred from the answer to part (d) without any further calculation.

Formula

$$\boldsymbol{z_y = r \cdot z_x}$$

(where z_y and z_x represent y and x in standard units)

Suppose that we are using the technique of regression of y on x to predict the value of y (or estimate the average value of all such y) given a value of x. For such an x, there will be a single point on the regression line. This formula indicates that the y-value of that point—in standard units—is r times the x-value of that point—in standard units.

The R.M.S. Error for Regression **(Chapter 11)**

The distance of a point above (+) or below (−) the regression line is

$$\text{error} = \text{actual} - \text{predicted}.$$

The r.m.s. error for regression says how far typical points are above or below the regression line.

The SD of y says how far typical points are above or below a horizontal line through the average of y. In other words, the SD of y is the r.m.s. error for predicting y by its average, just ignoring the x-values.

The r.m.s. error for the regression of y on x can be figured as

$$\sqrt{1-r^2} \times \text{the SD of } y.$$

The units for the r.m.s. error are the same as the units for the variable being predicted.

The residuals average out to 0; and the regression line for the residual plot is horizontal.

Suppose that a scatter diagram is football-shaped. Take the points in a narrow vertical strip. They will be off the regression line (up or down) by amounts similar in size to the r.m.s. error. If the diagram is heteroscedastic, the r.m.s. error should not be used for individual strips.

Suppose that a scatter diagram is football-shaped. Take the points in a narrow vertical strip. Their y-values are a new data set. The new average is estimated by the regression method. The new SD is about equal to the r.m.s. error for the regression line.

Problems to think about

A survey of heights and weights of a group of men led to the following results:

$$\begin{array}{rll} \text{average height} \approx 67 \text{ inches}, & \text{SD} \approx 4 \text{ inches} & \\ \text{average weight} \approx 152 \text{ pounds}, & \text{SD} \approx 30 \text{ pounds}, & r \approx 0.80 \end{array}$$

Find the r.m.s. error for the regression prediction of weight from height.

Find the r.m.s. error for the regression prediction of height from weight.

Assume that the scatter diagram is football-shaped.

Of those individuals that are 73 inches tall, about what percentage weigh 192 pounds or more?