3.2: Applications and Interpretations of LSRLs

Suppose that we have a bivariate quantitative dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ and we have computed its correlation coefficient r and (the coefficients of) its LSRL $\hat{y} = mx + b$. What is this information good for?

The main use of the LSRL is described in the following

[def:interpolation] Given a bivariate quantitative dataset and associated LSRL with equation $\hat{y} = mx + b$, the process of guessing that the value of the dependent variable in this relationship to have the value $mx_0 + b$, for x_0 any value for the independent variable which satisfies $x_{min} \leq x_0 \leq x_{max}$, is called **interpolation**.

The idea of interpolation is that we think the LSRL describes as well as possible the relationship between the independent and dependent variables, so that if we have a new x value, we'll use the LSRL equation to predict what would be our best guess of what would be the corresponding y. Note we might have a new value of x because we simply lost part of our dataset and are trying to fill it in as best we can. Another reason might be that a new individual came along whose value of the independent variable, x_0, was typical of the rest of the dataset – so the the very least $x_{min} \leq x_0 \leq x_{max}$ – and we want to guess what will be the value of the dependent variable for this individual before we measure it. (Or maybe we cannot measure it for some reason.)

A common (but naive) alternate approach to interpolation for a value x_0 as above might be to find two values x_i and x_j in the dataset which were as close to x_0 as possible, and on either side of it (so $x_i < x_0 < x_j$), and simply to guess that the y-value for x_0 would be the average of y_i and y_j. This is not a terrible idea, but it is not as effective as using the LSRL as described above, since we use the entire dataset when we build the coefficients of the LSRL. So the LSRL will give, by the process of interpolation, the best guess for what should be that missing y-value based on everything we know, while the "average of y_i and y_j" method only pays attention to those two nearest data points and thus may give a very bad guess for the corresponding y-value if those two points are not perfectly typical, if they have any randomness, any variation in their y-values which is not due to the variation of the x.

It is thus always best to use interpolation as described above.

EXAMPLE 3.2.2. Working with the statistics students' homework and total course points data from Example 3.1.4, suppose the gradebook of the course instructor was somewhat corrupted and the instructor lost the final course points of the student Janet. If Janet's homework points of 77 were not in the corrupted part of the gradebook, the instructor might use interpolation to guess what Janet's total course point probably were. To do this, the instructor would have plugged in $x = 77$ into the equation of the LSRL, $\hat{y} = mx + b$ to get the estimated total course points of $.754 \cdot 77 + 26.976 = 85.034$

Another important use of the (coefficients of the) LSRL is to use the underlying meanings of the slope and y-intercept. For this, recall that in the equation $y = mx + b$, the slope m tells us how much the line goes up (or down, if the slope is negative) for each increase of the x by one unit, while the y-intercept b tells us what would be the y value where the line crosses the y-axis, so when the x has the value 0. In each particular situation that we have bivariate quantitative data and compute an LSRL, we can then use these interpretations to make statements about the relationship between the independent and dependent variables.

EXAMPLE 3.2.3. Look one more time at the data on students' homework and total course points in a statistics class from Example 3.1.4, and the the LSRL computed there. We said that the slope of the LSRL was $m = .754$ and the y-intercept was $b = 26.976$. In context, what this means, is that *On average, each additional point of homework corresponded to an increase of* $.754$ *total course points.* We may hope that this is actually a causal relationship, that the extra work a student does to earn that additional point of homework score helps the student learn more statistics and therefore get $.75$ more total course points. But the mathematics here does not require that causation, it merely tells us the increase in x is *associated* with that much increase in y.

Likewise, we can also conclude from the LSRL that *In general, a student who did no homework at all would earn about* 26.976 *total course points.* Again, we cannot conclude that doing no homework *causes* that terrible final course point total, only that there is an association.