LECTURE XII. INFERENCE WITH TWO CONTINUOUS VARIABLES

Samuel DeWitt

March 27, 2019

Inference with Two Continuous Variables

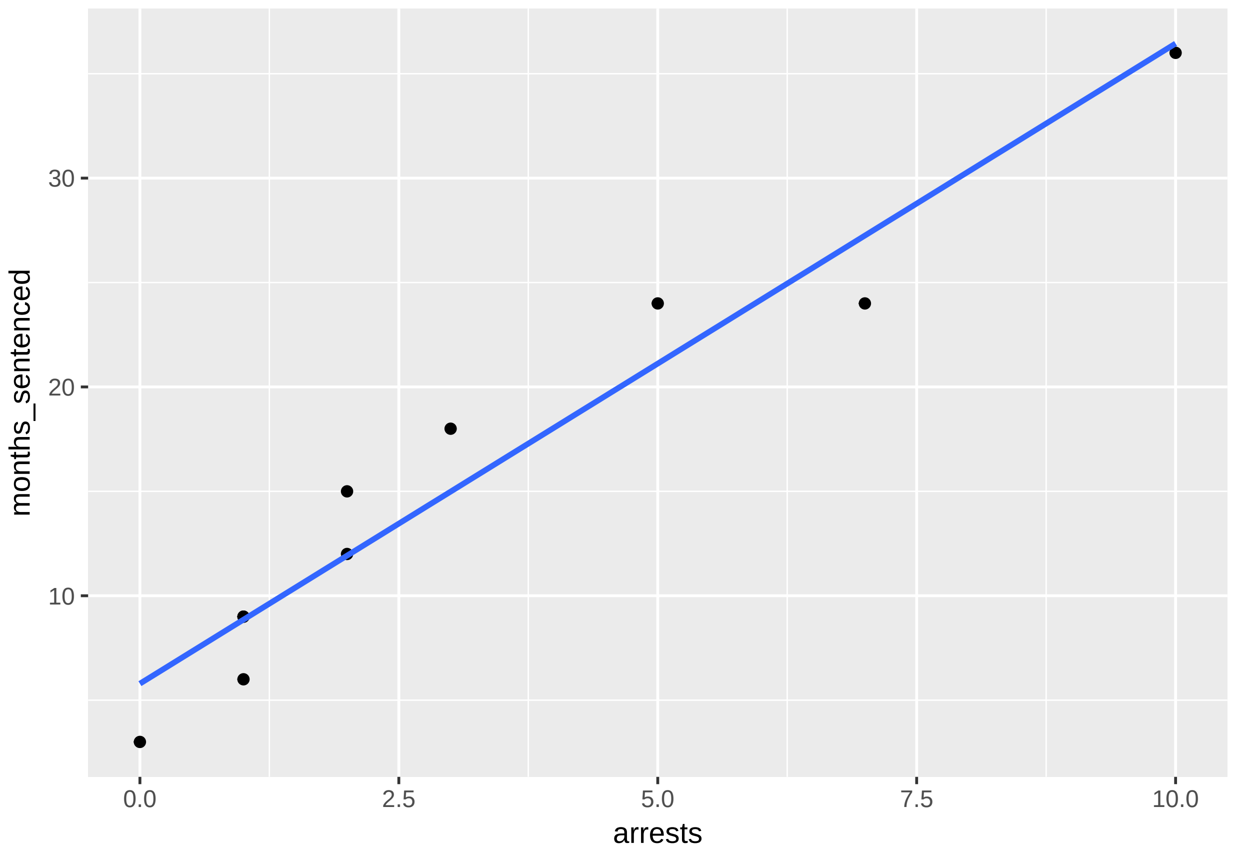
1. Scatterplots
2. Correlation coefficient
3. Regression equation
4. Comparability of correlation and regression coefficients

Scatterplots

- Correlation and regression are used to assess the relationship between two continuous variables. One variable is defined as the dependent variable (which we denote Y), and the other is defined as the independent variable (which we denote X). Let's start with an example. It is well known that one of the strongest predictors of sentence length is a defendant's prior record. We collect data from a sample of 10 inmates convicted of burglary and ask them how many months they received in their sentence (Y) and how many arrests they had prior to conviction (X). We obtain the following data.

X	Y
0	3
1	6
1	9
1	9
2	12
2	15
3	18
5	24
7	24
10	36

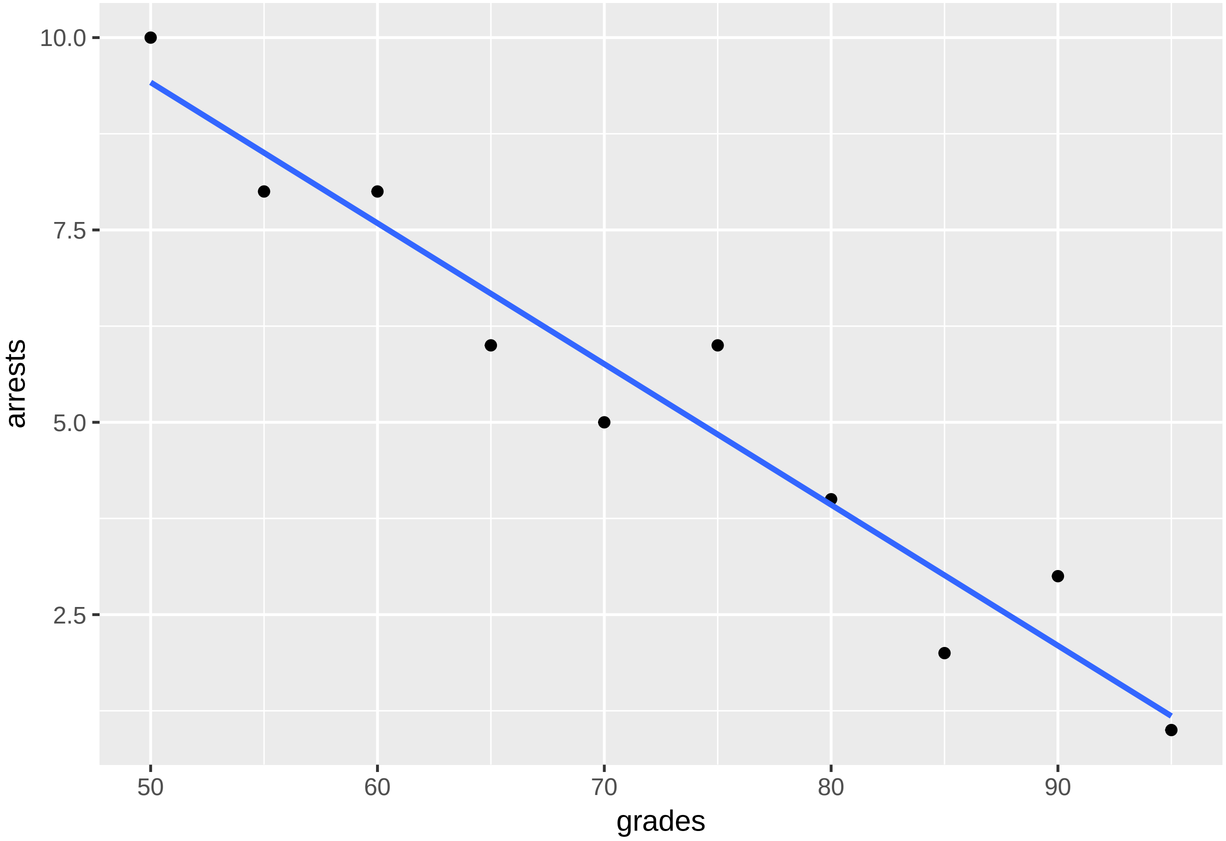
- One simple way to assess the relationship between two variables is to use a scatterplot, or graphical display that summarizes the nature of the association between an independent and dependent variable. The X-axis is the independent variable, and the Y-axis is the dependent variable. Each observation receives a dot at its respective X- and Y-values.



- Notice that as $X increases, Y also increases. We can tell with this scatterplot, without having to compute any statistics, that there is a positive relationship between prior record and sentence length. In other words, people who more prior arrests to have a longer sentence length.

- Let's say that we also gathered data from 10 youths in juvenile detention on average school performance and the number of juvenile arrests.

X	Y
50	10
55	8
60	8
65	6
70	5
75	6
80	4
85	2
90	3
95	1



- We see that, as school performances increases, juvenile arrests decrease. There is thus a negative relationship between school performance and juvenile arrest. In other words, youths with high school performance tend to have a low number of arrests.
- Scatterplots thus tell us something about the *direction* of the association between two variables. We can add a *trend line* to the scatterplot to aid in the interpretation of the direction of association, or what we will refer to later as a *regression line*. A second piece of information that we can obtain is the amount of variation there is around the regression line, which is an indication of the *strength* of the association. The closer the dots are to the regression line, the stronger the association between X and Y.
- An advantage offered by scatterplot is the ability to identify outliers. The disadvantage of using a

scatterplot to summarize the relationship between two variables is that it is not very precise. We can only determine that prior record and sentence length are positively related, and we can only eyeball the regression line. Our ultimate goal is to be a little more exact in describing the nature of the relationship between X and Y. There are two coefficients that we can compute to be more precise: a correlation coefficient or regression equation.

Correlation Coefficient

- A correlation "standardizes" the association between two variables. We need to calculate the variance of X, the variance of Y, and the crossproduct (or *covariance*) of X and Y. A correlation coefficient is represented by the r symbol, and is calculated as follows:

$$\begin{aligned} r &= \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum(X-\overline{X})^2\sum(Y-\overline{Y})^2}} \\ &= \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X^2)][n\sum Y^2 - (\sum Y^2)]}} \tag{1} \\ &= \frac{\sum XY - n\overline{XY}}{\sqrt{[\sum X^2 - n\overline{X}^2][\sum Y^2 - n\overline{Y}^2]}} \end{aligned}$$

- Let's compute the correlation between prior record and sentence length.

X	Y	XY	X^2	Y^2
0	3	0	0	9
1	6	6	1	36
1	9	9	1	81
1	9	9	1	81
2	12	24	4	144
2	15	30	4	225
3	18	54	9	324
5	24	120	25	576
7	24	168	49	576
10	36	360	100	1296
32	156	780	194	3348
$\overline{x} = 3.2$	$\overline{y} = 15.6$			

$$r = \frac{(10)(780) - (32)(156)}{\sqrt{[(10)(194) - 32^2][(10)(3348) - 156^2]}} = \frac{2808}{\sqrt{(916)(9144)}} = 0.970$$

$$r = \frac{780 - (10)(3.2)(15.6)}{\sqrt{[194 - (10)(3.2^2)][3348 - (10)(15.6^2)]}} = \frac{280.8}{\sqrt{(91.6)(914.4)}} = 0.970$$

- We can easily tell that prior record is positively related to sentence length and that it is a very strong association. Let's do the same with school performance and juvenile arrest.

- Now, let's do the same for the test scores and juvenile arrest example.

X	Y	XY	X^2	Y^2
50	10	500	2500	100
55	8	440	3025	64
60	8	480	3600	64
65	6	390	4225	36
70	5	350	4900	25
75	6	450	5625	36
80	4	320	6400	16
85	2	170	7225	4
90	3	270	8100	9
95	1	95	9025	1
725	53	3465	54625	355
$\overline{x} = 72.5$	$\overline{y} = 5.3$			

$$r = \frac{(10)(3465) - (725)(53)}{\sqrt{[(10)(54625) - 725^2][(10)(355) - 53^2]}} = \frac{-3775}{\sqrt{(20625)(741)}} = -0.966$$

$$r = \frac{3465 - (10)(72.5)(5.3)}{\sqrt{[54625 - (10)(72.5^2)][355 - (10)(5.3^2)]}} = \frac{-377.5}{\sqrt{(2062.5)(74.1)}} = -0.966$$

- The association between prior record and sentence length is thus positive and very strong while the association between test scores and juvenile arrest is negative and very strong. One nice thing about correlation coefficients is that they can be directly compared. Another useful property of the correlation coefficient is that when we square it, we can use an "explained variance" interpretation. So, prior record explains $0.970^2 = 0.941 \rightarrow 94.1\%$ of the variance in sentence length. School performance explains $-0.965^2 = 0.931 \rightarrow 93.1\%$ of the variance in juvenile arrest.

- How about if we want to conduct a hypothesis test? We want to know if a linear relationship exists between prior record and sentence length in the population, or if our estimate of the correlation is the result of sampling error. Let's go through the five steps.
    - **Step 1: State hypotheses -** Our research hypothesis is this: Does having a prior record increase sentence length? The population parameter we are trying to estimate is ??, the population correlation coefficient, and its sample analog is r. The null and alternative hypotheses are $H_0 : \rho = 0$ and $H_1 : \rho > 0$.
    - **Step 2: Obtain a probability distribution -** The probability distribution for correlation coefficients is the t-distribution, with $df = n - 2 = 8$.
    - **Step 3: Make decision rules -** Let's use $\alpha = .05$. We will reject the null hypothesis if the test statistic is greater than 1.860 (i.e., $TS > 1.860$).
    - **Step 4: Calculate the test statistic -** The test statistic for a correlation is:

$$TS = r\sqrt{\frac{n-2}{1-r^2}} = 0.970\sqrt{\frac{10-2}{1-(0.970^2)}} = 0.970\sqrt{135.36} = 11.286$$

    - **Step 5: Make a decision about the null hypothesis -** We reject the null hypothesis and conclude that having a prior record significantly increases sentence length.

- Let's do a hypothesis test for the correlation between school peformance and juvenile arrest. The research question is: Is school performance related to juvenile arrest?
    - **Step 1: State hypotheses -** $H_0 : \rho = 0$ and $H_1 : \rho \neq 0$.
    - **Step 2: Obtain a probability distribution -** t-distribution, with $df = n - 2 = 8$.
    - **Step 3: Make decision rules -** Let's use $\alpha = .01$. We will reject the null hypothesis if $|TS| > 3.355$.
    - **Step 4: Calculate the test statistic -**

$$TS = r\sqrt{\frac{n-2}{1-r^2}} = -0.966\sqrt{\frac{10-2}{1-(-0.966^2)}} = -0.966\sqrt{119.68} = -10.568$$

    - **Step 5: Make a decision about the null hypothesis -** We reject the null hypothesis and conclude that school performance is significantly related to juvenile arrest.

Regression Equation

- Another way to assess the relationship between two continuous variables is by estimating a regression equation. When a scatterplot indicates that two variables are more or less linearly related, it is convenient to draw a straight line through the middle of the data points. When we estimate the regression line (as opposed to trying to eyeball it), it can be shown that it is the "best-fitting" line, which means that the line falls as close to every data point as possible.

- A regression equation *in the population* is of the form:

  $$Y = \alpha + \beta X + \epsilon$$

  In this equation, α and β are the population parameters that summarize the association between X and Y while ϵ represents any error in using the regression equation to predict values for Y. A regression equation using sample data to estimate these parameters is of the form:

  $$Y = a + bX + e$$

  In this equation, a is the y-intercept (or constant), and b is the slope (and e is the sample data equivalent of ϵ). Once these two parameters are estimated, we can substitute any value for X to determine the "best guess" of Y for that particular value of X. The estimate b tells us the effect on Y of a one-unit increase in X. It is calculated by the formula:

  $$b = \frac{\sum(X-\overline{X})(Y-\overline{Y})}{\sum(X-\overline{X})^2} = \frac{n\sum XY - (\sum X)(\sum Y)}{n\sum X^2 - (\sum X)^2} = \frac{\sum XY - n\overline{XY}}{\sum X^2 - n\overline{X}^2}$$

  The y-intercept, a, is very simply the value of Y when $X = 0$. The purpose of the intercept is to "anchor" the regression line to the y-axis. Once we know b, we can solve for the intercept by:

  $$a = \overline{Y} - b\overline{X}$$

- Let's estimate the regression equation for our prior record and sentence length example:

  $$b = \frac{(10)(780) - (32)(156)}{(10)(194) - (32)^2} = \frac{7800 - 4992}{1940 - 1024} = \frac{2808}{916} = 3.07$$

  $$= \frac{780 - (10)(3.2)(15.6)}{194 - (10)(3.2^2)} = \frac{280.8}{91.6} = 3.07$$

  $$a = 15.6 - (3.07)(3.2) = 5.78$$

  $$Y = 5.78 + 3.07X$$

  The value for b means that an increase of one prior arrest prodices 3.07 additional months in sentence length, on average.

- This regression equation comes in handy when we want to calculate predicted values of Y for given values of X. Let's pick a few values of X to illustrate.

$Y = 5.78 + 3.07X$

X	Y
0	5.8
2	11.9
4	18.1
6	24.2
8	30.3
10	36.5

- Now let's estimate the regession equation for the school performance and juvenile arrest example:

$$b = \frac{(10)(3465) - (725)(53)}{(10)(54625) - (725)^2} = \frac{-3775}{20625} = -0.183$$

$$= \frac{3465 - (10)(72.5)(5.3)}{54625 - (10)(72.5^2)} = \frac{-377.5}{2062.5} = -0.183$$

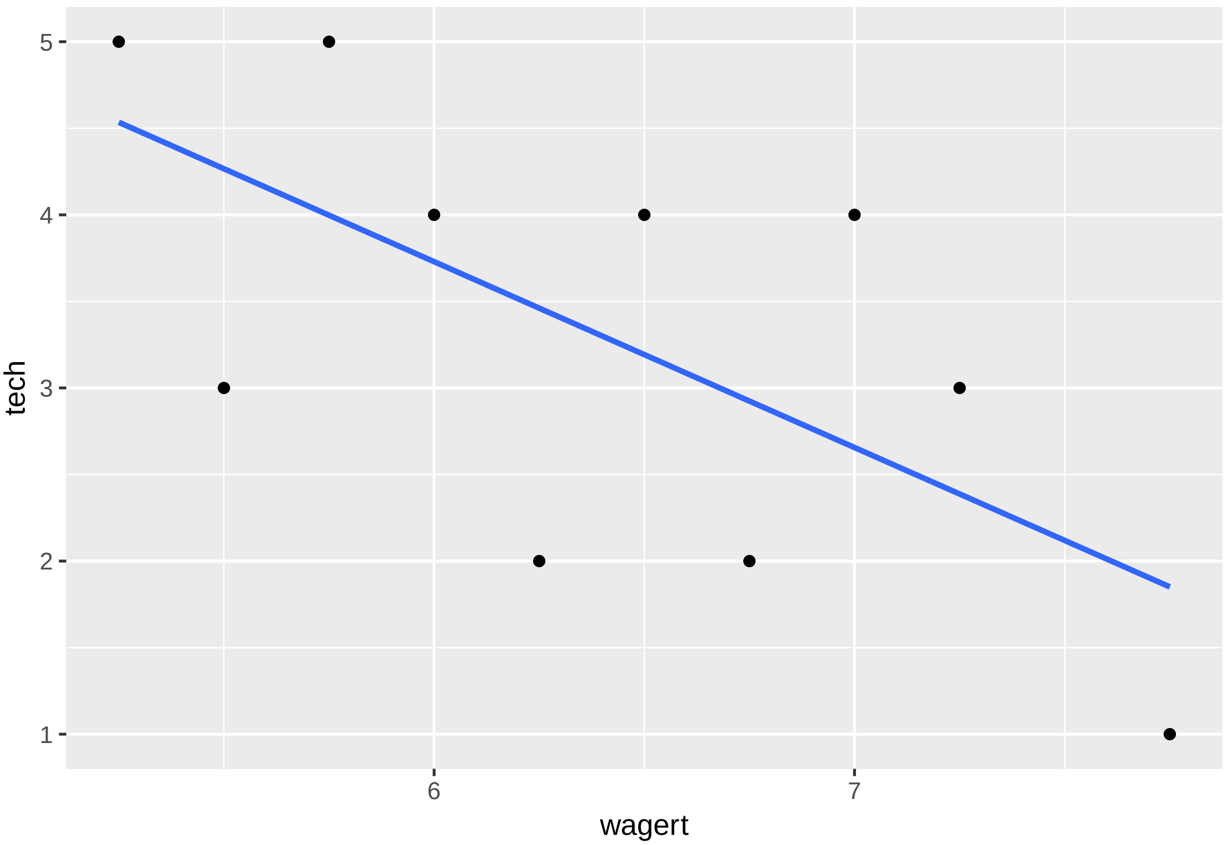
$$a = 15.6 - (-0.99)(29.9) = 45.20$$

$$Y = 45.20 - 0.99X$$

- Let's apply this approach to a new example: wages and number of technical violations on parole.

X	Y	X^2	Y^2	XY
5.25	5	27.5625	25	26.25
5.50	3	30.2500	9	16.50
5.75	5	33.0625	25	28.75
6.00	4	36.0000	16	24.00
6.25	2	39.0625	4	12.50
6.50	4	42.2500	16	26.00
6.75	2	45.5625	4	13.50
7.00	4	49.0000	16	28.00
7.25	3	52.5625	9	21.75
7.75	1	60.0625	1	7.75
64.0	33	415.375	125	205.00
$\overline{x} = 64/10 = 6.4$		$\overline{y} = 33/10 = 3.3$		

- First, let's take a look at a scatterplot of these data.



- Second, let's compute the slope and intercept:

$$b = \frac{205 - (10)(6.4)(3.3)}{415.375 - (10)(6.4^2)} = \frac{205.0 - 211.2}{415.375 - 409.6} = -1.07$$

$$a = 3.3 - (-1.07)(6.4) = 10.15$$

  - We can interpret this slope as "for every one unit increase in hourly wages, parolees experience 1.07 fewer parole violations, on average."

- Next, we should consider how well our predictions improve using this regression equation to predict parole violations as opposed to simply using the average number of parole violations to predict them. We have seen scatterplots with fit lines previously, which provide a visual indication of how well the trend line fits the data. However, a cursory visual analysis would not be direct enough evidence that a line provides a good fit to the data. Instead, we want to compute a singular measure that indicates this which can be accomplished by applying the regression equation to each observed value of y:

$$\sum e^2 = \sum (y - \hat{y})^2 = \sum (y - a - b\text{x})^2 = \text{minimum}$$

  We will use this equation shortly to compute the test statistic for a regression slope.

- We can also conduct a hypothesis test for the significance of the regression coefficient, b. This requires two additional pieces of information: 1) the standard error of the slope (s_b) and, 2) the mean squared error (s_e^2).

  - **Step 1: State hypotheses -** $H_0 : \beta \geq 0$ and $H_1 : \beta < 0$.
  - **Step 2: Obtain a probability distribution -** The probability distribution for regression slopes is the t-distribution, with df=$n - 2 = 8$.
  - **Step 3: Make decision rules -** Let's use $\alpha = .05$. We will reject the null hypothesis if the test statistic is less than -1.860 (i.e., $TS < -1.860$).
  - **Step 4: Calculate the test statistic -** The test statistic for a regression slope is:

$$TS = \frac{b - \beta}{s_b} \rightarrow \frac{b}{s_b} \text{ under the null, } \beta = 0$$

  To obtain s_b, we must use the following formula:

$$s_b = \sqrt{\frac{s_e^2}{SS_x}} = \sqrt{\frac{\sum e^2/(n-2)}{\sum (x - \overline{x})^2}} = \sqrt{\frac{\sum (y - \hat{y})^2/(n-2)}{\sum (x - \overline{x})^2}}$$

Hourly Wages	Parole Violations	$\hat{Y}$	$Y - \hat{Y}$	$(Y - \hat{Y})^2$
5.25	5	4.5305	0.4695	0.2204
5.50	3	4.2630	-1.2630	1.5952
5.75	5	3.9955	1.0045	1.0090
6.00	4	3.7280	0.2720	0.0740
6.25	2	3.4605	-1.4605	2.1331
6.50	4	3.1930	0.8070	0.6512
6.75	2	2.9255	-0.9255	0.8566
7.00	4	2.6580	1.3420	1.8010
7.25	3	2.3905	0.6095	0.3715
7.75	1	1.8555	-0.8555	0.7319
64.0	33	33.0	0	9.4439

- Now, we can estimate s_b and then the test statistic as follows:

$$s_b = \sqrt{\frac{(9.4439/(10-2)}{5.775}} = \sqrt{\frac{1.1805}{5.775}} = 0.4521$$

Therefore, the TS is then:

$$TS = \frac{-1.07}{0.4521} = -2.367$$

- **Step 5: Make a decision about the null hypothesis** - Reject H_0 and conclude that higher wages are associated with significantly fewer parole violations.

Comparing Correlation and Regression Coefficients

- b and r are very closely related, and can be easily expressed as functions of one another. In fact, within the context of bivariate regression, the standardized slopes (i.e., standardization of b) are always equivalent to the correlation coefficient.
- Here are the individual equations again for posterity:

$$b = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2}; \; r = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}}$$

- And here is how we can mathematically deomonstrate their relationship:

$$b = r\frac{s_y}{s_x} = r\left(\frac{\sqrt{\sum (y - \overline{y})^2/n}}{\sqrt{\sum (x - \overline{x})^2/n}}\right) = r\frac{\sqrt{\sum (y - \overline{y})^2}}{\sqrt{\sum (x - \overline{x})^2}}$$

$$r = b\frac{s_x}{s_y} = b\left(\frac{\sqrt{\sum (x - \overline{x})^2/n}}{\sqrt{\sum (y - \overline{y})^2/n}}\right) = b\frac{\sqrt{\sum (x - \overline{x})^2}}{\sqrt{\sum (y - \overline{y})^2}}$$