**Regression Analysis Course**

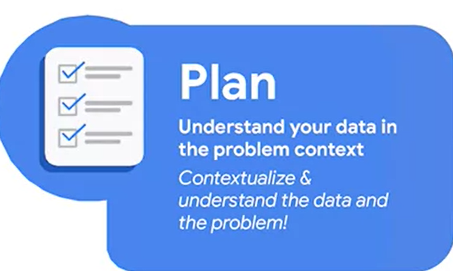
* **Course Focus**: The course will explore how data professionals use regression analysis to derive actionable insights from data.
* The course will introduce essential terms like machine learning models, regression analysis, and their statistical foundations.
* Regression models help tell a story about the relationships between variables, guiding stakeholders in making informed business decisions.
* **Pace Framework**: The course is structured around the PACE framework, which stands for Plan, Analyze, Construct, and Execute. This framework will guide the learning process.
* **Simple Linear Regression**:
  + The first regression model covered in detail.
  + The course will walk through the entire process using different scenarios and data.
* **Multiple Linear Regression**:
  + Builds on the concepts from simple linear regression.
  + Allows for solving more complex problems.
  + Topics include variable selection, model interpretation, and hypothesis testing (e.g., Chi-squared test, ANOVA).
* **Logistic Regression**:
  + The final and most complex model covered in the course.
  + Prepares learners for more advanced topics in machine learning.

 **Iterative Modelling Process**:

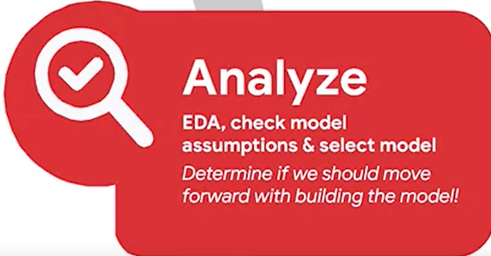
* Modelling is iterative, following frameworks like the data life cycle or exploratory data analysis (EDA).
* The course uses the PACE framework—Plan, Analyze, Construct, Execute—to structure the regression analysis process.

 **PACE Framework in Regression Analysis**:

* **Plan**:
  + Focuses on understanding data within the problem context.
  + Considers data access, collection methods, and business needs.



* **Analyze**:
  + Involves closely examining data to select appropriate models.
  + Utilizes Python for EDA and checking model assumptions.
  + Understanding statistics is crucial for validating model assumptions.



* **Construct**:
  + The phase where the model is built using Python or other coding languages.
  + Involves selecting variables, transforming data, writing code, and rechecking model assumptions.
  + The last step in this phase is evaluating model results using metrics and comparing models.



* **Execute**:
  + Involves interpreting and sharing results through formal reports and visualizations.
  + Focuses on converting model statistics into meaningful descriptions of variable relationships, considering the context and initial questions from the plan phase.



 **Role of Data and Storytelling**:

* Data is at the core of the PACE framework, ensuring insights are data-driven and contextually relevant.

 **Correlation and Regression**:

* We will explore the relationship between correlation and regression.
* Foundational regression models like linear and logistic regression will be covered in depth, providing a solid understanding for further exploration.

 **Tools and Techniques**:

* Emphasis on the importance of statistical tools and "statistical grammar" for understanding and applying regression analysis.

 **Practical Application**:

* The course will use examples to demonstrate how to navigate through the PACE stages.
* Learners will gain experience in knowing when to pivot between stages and how to refine their models based on iterative learning.

**Linear Regression**

* + Linear regression is a modeling technique used to estimate relationships between a continuous dependent variable (Y) and one or more independent variables (X).
  + The "linear" aspect refers to the straight-line relationship that can be visualized on a graph.
* **Examples of Linear Relationships**:
  + As a computer software version ages, online searches for that version may decrease.
  + As a social media personality gains follower, their book sales may increase.
* **Continuous vs. Categorical Variables**:
  + Continuous variables can take any real value within a range (e.g., product sales, vehicle speed).
  + Categorical variables have a finite number of possible values (e.g., product types, educational level).
* **Key Terms**:
  + **Dependent Variable (Y)**: The outcome or response variable that the model estimates.
  + **Independent Variable (X)**: The explanatory or predictor variable that influences the dependent variable.
  + **Slope**: The expected change in the dependent variable for each one-unit increase in the independent variable.
  + **Intercept**: The value of the dependent variable when the independent variable equals zero.
* **Correlation vs. Causation**:
  + **Positive Correlation**: Both variables tend to increase or decrease together (e.g., more coffee sales may lead to more cake sales).
  + **Negative Correlation**: One variable increases while the other decreases (e.g., as hot coffee sales increase, iced coffee sales may decrease).
  + **Correlation is not Causation**: While correlation indicates a relationship, it does not imply that one variable causes the other to change. Proving causation requires more rigorous methods.
* **Practical Applications**:
  + Linear regression helps answer questions about factors influencing sales, resource allocation, and demand in various industries.
  + The slope of the regression line provides insights into how much a dependent variable changes based on an independent variable.
* **Ethical Considerations**:
  + Data professionals must be careful not to imply causation when only correlation is present.
  + Accurately communicating the distinction between correlation and causation is essential for ethical data analysis.

**Mathematical Linear Regression Overview**

**Population vs. Sample**:

* + **Population**: The entire set of data you could possibly collect on a topic. In an ideal scenario, you'd use population data for analysis.
  + **Sample**: A subset of the population used when it's impractical to gather all data. A representative sample can provide a meaningful estimate for regression analysis.

**Observed Values**:

* + **Dependent Variable (Y)**: The outcome you're trying to predict (e.g., book sales).
  + **Independent Variable (X)**: The factor you believe influences Y (e.g., social media followers).
  + In your sample data, each data point has observed values for both X and Y.

**Mean of Y Given X**:

* + Linear regression focuses on estimating the mean of Y for each value of X. This mean is represented by μY\mu\_YμY​, the Greek letter mu.
  + The regression line is a visual representation of the mean of Y across different values of X.

**Regression Equation**:

* + The linear relationship between X and Y is represented by the equation

Y=β0​+β1​×X

*  β**0​** (Beta 0): The intercept, representing the value of Y when X is 0.
* **β1** (Beta 1): The slope, representing the change in Y for each one-unit increase in X.

**Parameter Estimation**:

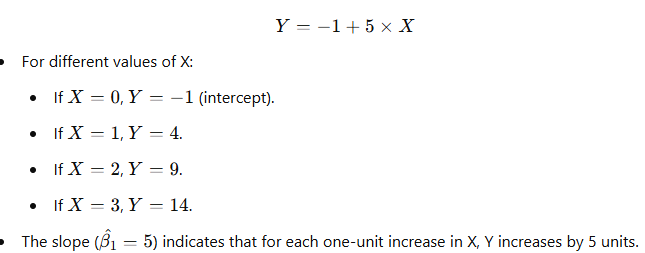
**Population Parameters**: True values of β0\beta\_0β0​ and β1\beta\_1β1​ for the entire population, which are unknown and can only be estimated.

**Parameter Estimates**: Estimates of β0\beta\_0β0​ and β1\beta\_1β1​ based on sample data, denoted as β0^\hat{\beta\_0}β0​^​ and β1^\hat{\beta\_1}β1​^​.

* The hat symbol (^\hat{}^) indicates that these are estimates.

**Example Calculation**:

* Suppose β0^=−1\hat{\beta\_0} = -1β0​^​=−1 and β1^=5\hat{\beta\_1} = 5β1​^​=5. The regression equation would be:



**Ordinary Least Squares (OLS)**:

* **OLS**: A common method to calculate the best-fit line for linear regression.
* **Loss Function**: Measures the distance between observed values (actual Y values) and estimated values (predicted Y values).
* The goal of OLS is to minimize the loss function, finding the line that best fits the data by minimizing the differences between observed and predicted values.

**Key Takeaways**:

* Linear regression provides a way to estimate the relationship between two variables.
* The slope and intercept of the regression line are calculated using sample data to estimate population parameters.
* OLS is a widely used technique to find the best-fit line by minimizing the differences between observed and estimated values.
* Understanding these concepts lays the groundwork for more advanced topics in regression analysis, including implementing OLS in Python.

#### **Recap of Linear Regression**:

* **Linear Regression**: A statistical method that estimates the relationship between a continuous dependent variable (Y) and one or more independent variables (X).
* **Independent Variable (X)**: The predictor or explanatory variable that influences the dependent variable.
* **Dependent Variable (Y)**: The outcome variable that the model aims to predict.

#### 2. **Simple Linear Regression**:

* **Definition**: A type of linear regression where there is only one independent variable (X) and one continuous dependent variable (Y).
* **Purpose**: To estimate the linear relationship between X and Y.

#### 3. **The PACE Framework**:

* **Plan**: Identify the problem, variables, and objectives.
* **Analyze**: Conduct Exploratory Data Analysis (EDA) to understand the data and check assumptions.
* **Construct**: Build the regression model using Python, ensuring the model is valid and accurate.
* **Execute**: Interpret the results and communicate them effectively to stakeholders.

#### 4. **Linear Regression Equation**:

* The equation for a simple linear regression is:

Y=β0+β1X+ϵY = \beta\_0 + \beta\_1X + \epsilonY=β0​+β1​X+ϵ

* **β0​**: The intercept (value of Y when X is 0).
* **β1:** The slope (change in Y for a one-unit change in X).
* **ϵ:** The error term, representing the difference between observed and predicted values of Y.

#### 5. **Ordinary Least Squares (OLS)**:

* **OLS**: A method to estimate the parameters (β0\beta\_0β0​ and β1\beta\_1β1​) by minimizing the sum of the squared differences between the observed and predicted values.

#### 6. **Key Assumptions of Simple Linear Regression**:

* **Linearity**: The relationship between X and Y is linear.
* **Independence**: Observations are independent of each other.
* **Homoscedasticity**: The variance of the errors is constant across all levels of X.
* **Normality**: The errors are normally distributed.

#### 7. **Analyze Stage with Python and EDA**:

* **EDA**: Use Python to explore the data and verify if it meets the assumptions of linear regression.
* **Tools**: Python libraries like Pandas, Matplotlib, and Seaborn will help in visualizing data and checking assumptions.

#### 8. **Construct Stage**:

* **Model Building**: Create a simple linear regression model using Python.
* **Practice**: You'll have the opportunity to build models and perform EDA on your own datasets.

#### 9. **Evaluation Metrics**:

* **R-squared**: Measures the proportion of variance in the dependent variable explained by the independent variable.
* **Mean Squared Error (MSE)**: The average of the squared differences between observed and predicted values.
* **Root Mean Squared Error (RMSE)**: The square root of MSE, providing a measure of the average prediction error.

#### 10. **Execute Stage**:

* **Interpretation**: Learn how to interpret the results of the model and communicate findings to non-technical audiences.
* **Stakeholder Communication**: Translate technical results into actionable insights for decision-making.

# Ordinary Least Squares

As previously mentioned, one way for finding the best fit line in regression modeling is to try different models until you find the best one. But for simple linear regression, the formulas for the best beta coefficients have been derived. In this reading, you will go through an example to gain a better understanding of how the sum of squared residuals can change as β^0*β*^​0​ and β^1*β*^​1​ change. There will be resources for further exploration if you’re interested in deriving the formulas for estimating the coefficients using ordinary least squares. In this reading, we will cover:

* Formula and notation review
* Minimizing the sum of squared residuals (SSR)
* Estimating beta coefficients

# Formula and notation review

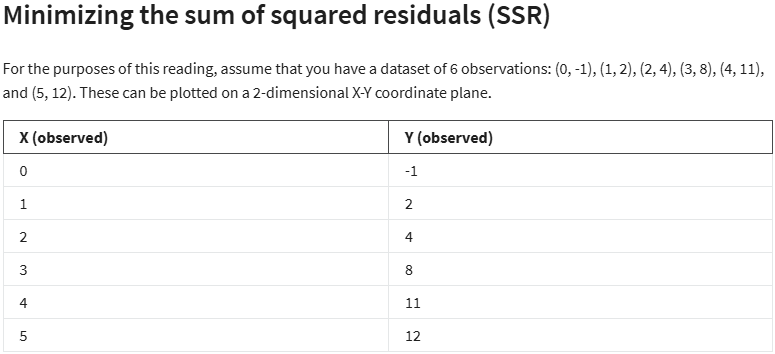
Earlier, you learned about simple linear regression as a method for estimating the linear relationship between a continuous dependent variable and one independent variable. An estimate based on simple linear regression can be represented mathematically as 

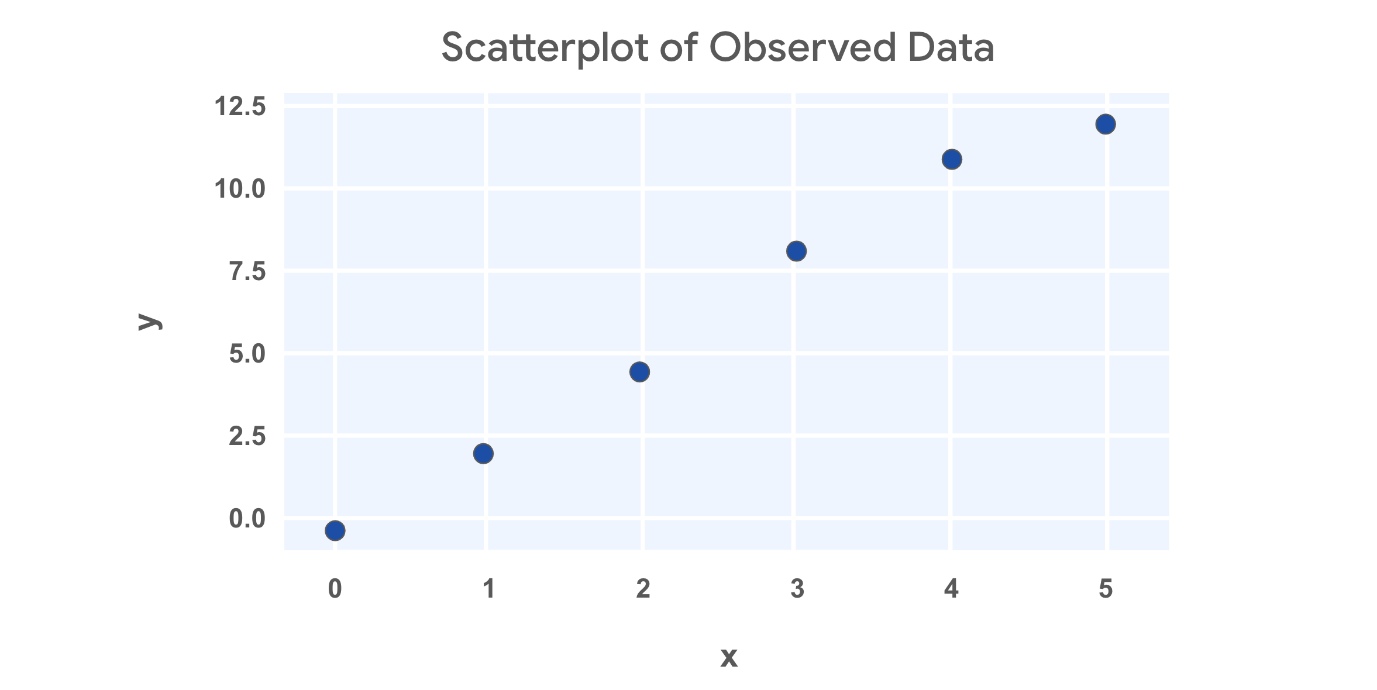
Remember that the hat symbol indicates that the beta coefficients are just estimates. As a result, the y-values derived from the regression model are also just estimates.

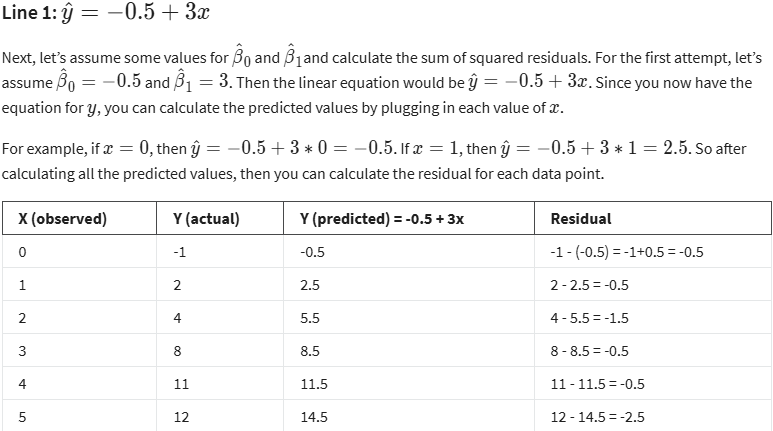
A common technique for calculating the coefficients of a linear regression model is called ordinary least squares, or OLS. Ordinary least squares estimates the beta coefficients in a linear regression model by minimizing a measure of error called the sum of squared residuals.

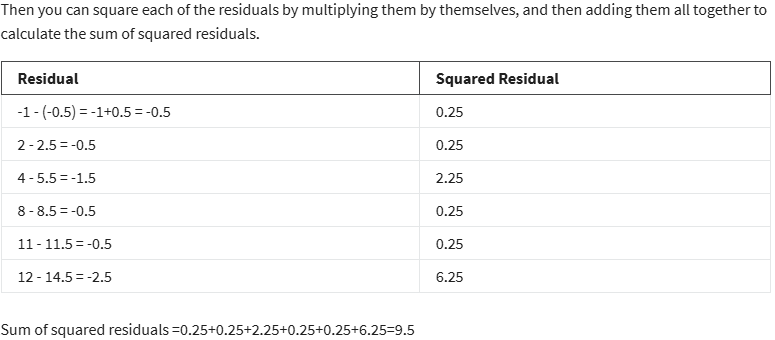


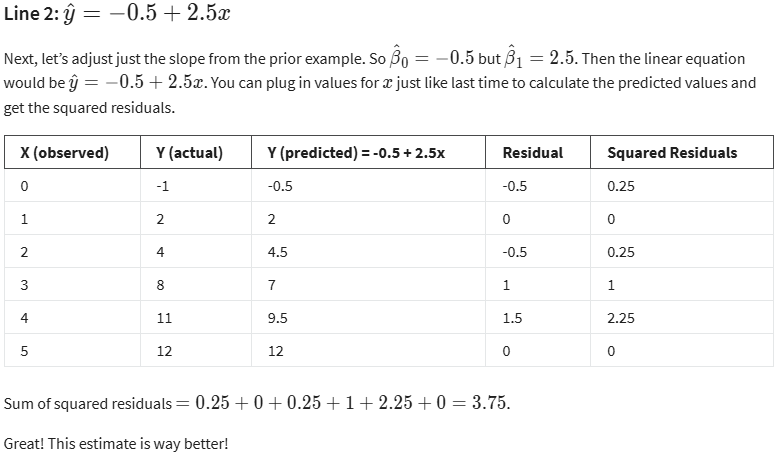
The large E shaped symbol is the capital Greek letter, sigma, and it denotes a sum. So the sum of squared residuals is the sum of the squared differences between the observed values and the values predicted by the regression model.

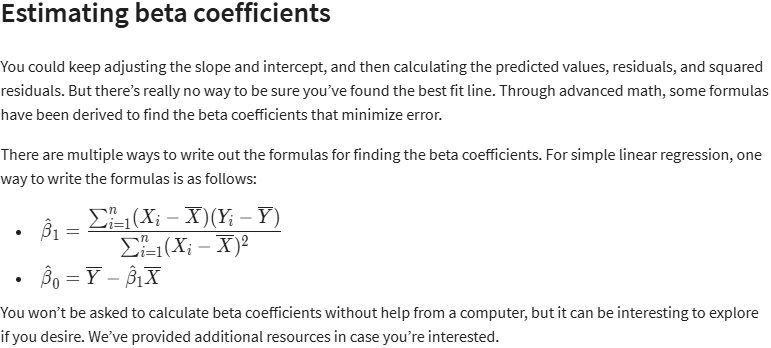












# Key takeaways

Given a sample of data, you can try out different lines that could fit your data. You could calculate the sum of squared residuals for each line to determine which fits your data best. As a data professional, it’s important to understand what the sum of squared residuals represents, and how to calculate it on your own. Thankfully, we have computers and programming languages that can calculate the sum of squared residuals and perform OLS for us. You can explore the deeper math behind OLS and SSR on your own if you wish!

# Correlation and the intuition behind simple linear regression

So far you’ve learned that simple linear regression is a technique that estimates the linear relationship between one independent variable, X, and one continuous dependent variable, Y. You’ve also learned about ordinary least squares estimation (OLS), which is a common way to determine the coefficients of the regression line—the line of “best fit” through the data. In this reading, you’ll explore the meaning of correlation; learn about r, or the “correlation coefficient;” and discover how to determine the regression equation. This knowledge will help you better understand relationships between variables, and thus how linear regression works.

## **Correlation**

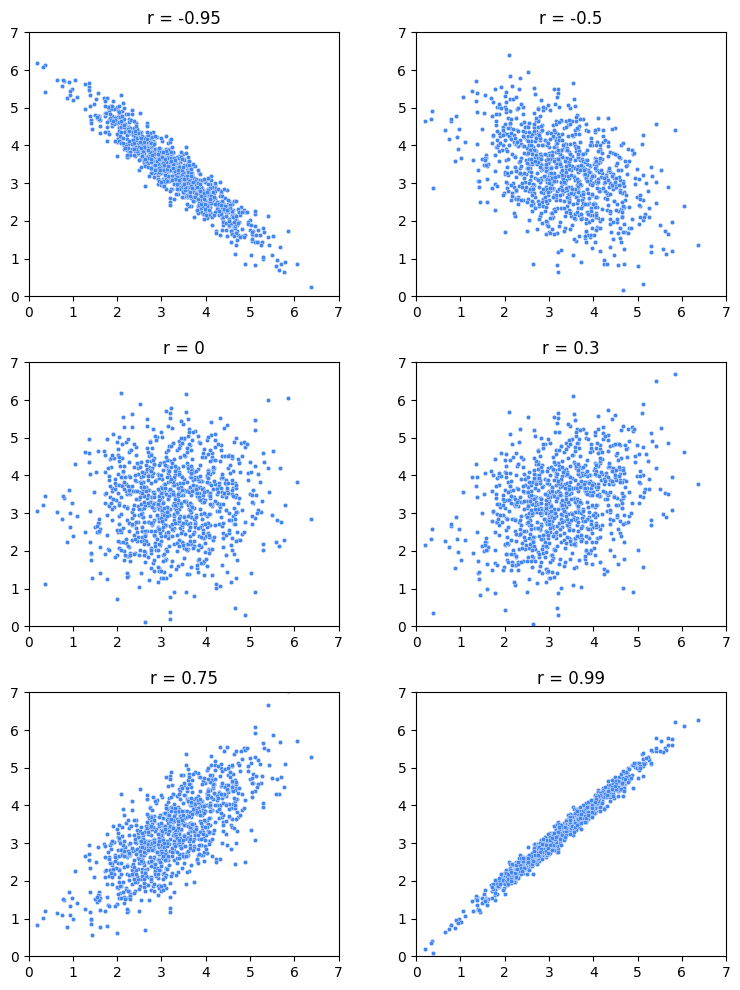
Correlation is a measurement of the way two variables move together. If there is a strong correlation between the variables, then knowing one will be very helpful to predict the other. However, if there is a weak correlation between two variables, then knowing the value of one will not tell you much about the value of the other. In the context of linear regression, correlation refers to linear correlation: as one variable changes, so does the other at a constant rate.

In the statistics course, you learned that a continuous variable can be summarized using some basic numbers. Two of these summary statistics are:

* **Average:** A measurement of central tendency (mean, median, or mode)
* **Standard deviation:** A measurement of spread

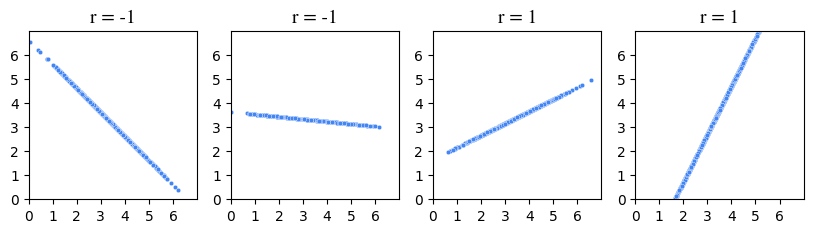
When two variables are summarized together, there is another relevant statistic called ***r***, **Pearson’s correlation coefficient** (named after the person who helped develop it), or simply the linear **correlation coefficient.** The correlation coefficient quantifies the strength of the linear relationship between two variables. It always falls in the range of [-1, 1]. When r is negative, there is a negative correlation between the variables: as one increases, the other decreases. When r is positive, there is a positive correlation between the variables: as one increases, so too does the other. When r = 0, there is no linear correlation between the variables. Note that there are cases where one variable might be precisely determined by another—like y=x2 or y=sin(x)—but the value of the linear correlation between X and Y would nonetheless be low or zero because their relationship is non-linear.

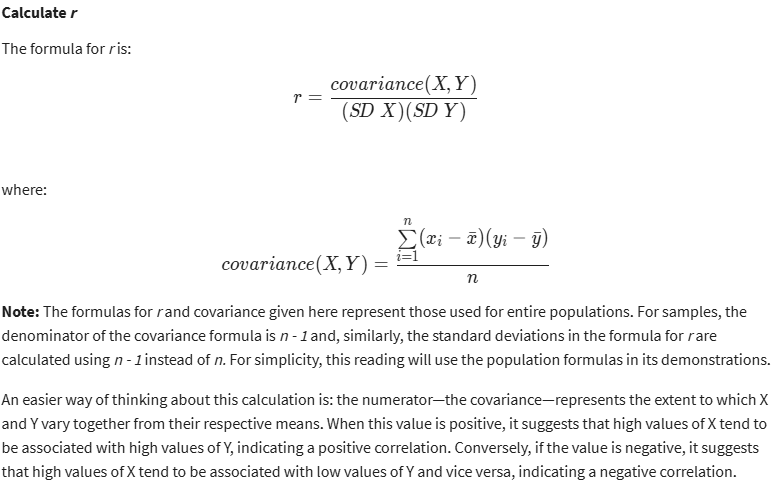
The following figure depicts scatterplots of bivariate (bi = “two”, variate = “variables”) data where each variable has the same mean and standard deviation and only the correlation coefficient varies.



Notice that the closer to -1 or 1 *r* is, the more linear the data appears. When *r* is exactly 1 or exactly -1, then the variables are perfectly correlated, and their graph is a line. When *r* is zero, there is no correlation between the variables, and, in this example, the data appears as a shapeless cloud of points.

However, *r* only tells you the strength of the linear correlation between the variables; it does not tell you anything about the magnitude of the slope of the relationship between the variables aside from its sign. For example, variables with *r*=1 wouldn't tell you if increasing X by one would lead to Y increasing by 10, 100, 0.1, or something else. It would only tell you that you can be sure that it *would* increase. This fact is illustrated in the following figure, where even though the slopes of the lines are all different, *r* is only either -1 or 1. If the line is perfectly horizontal or perfectly vertical, then *r* is undefined. (If you’re wondering why, refer to the equation below. One of the terms in the denominator would equal zero, which would make the whole denominator equal zero, which would result in an undefined solution.)

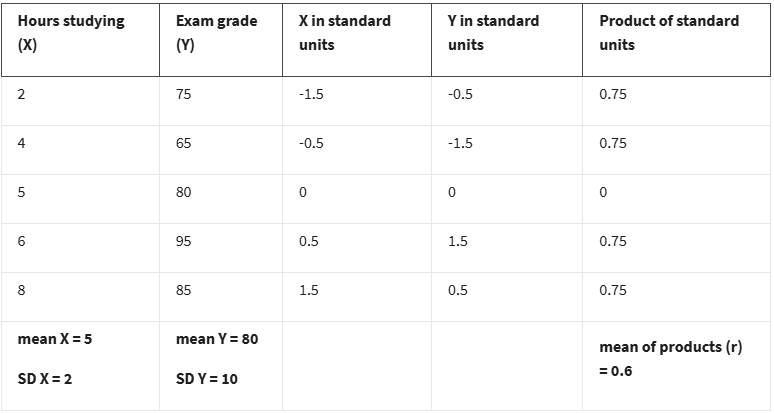


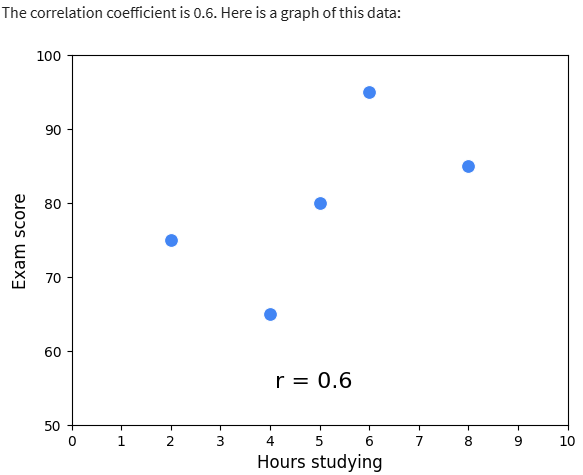
****

The denominator—the product of the standard deviations—standardizes the units of the numerator. It adjusts for the inherent variability of the individual variables. This makes r a statistic without a unit. It is a pure number, without dimension.

An equivalent way to calculate r is to convert each data point in each variable to standard units (subtract the mean, divide by the standard deviation), then take the average of the products.

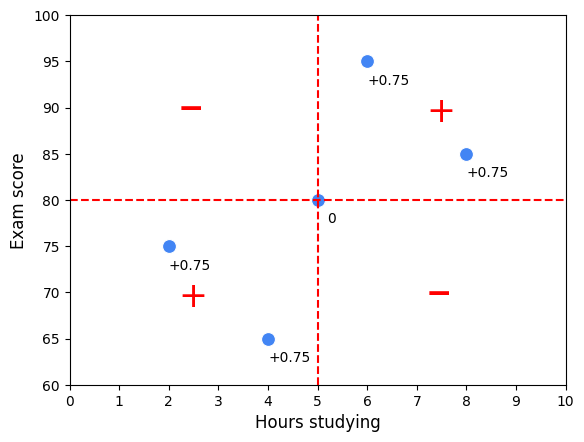
Here’s an example. Suppose five students took an exam and you recorded how many hours they spent studying and also their grade. The following table breaks out the calculation of r.





Notice that the cloud of points slopes upwards. This corresponds with r being positive. The correlation coefficient works as an indicator of association because it uses the product of each variable’s deviation from its mean. When the product is positive, it means both the X and the Y values are either below their respective means (negative standard units) or above their respective means (positive standard units). They vary together. However, when this product is negative, it means one of the values is above its mean and the other is below it. They vary in opposing directions relative to their respective means.

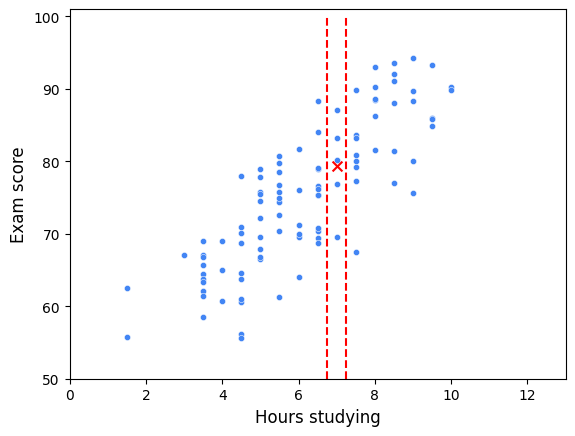
The following figure illustrates this idea. The figure is divided into quadrants. The vertical line represents the mean X value and the horizontal line represents the mean Y value. Each point is labeled with the product of its standardized scores (refer to the table above). The average of these scores is r. When r is positive, more points will tend to be in the positive quadrants, and vice versa.



## **Regression**

In the absence of any other information, if you had to guess a randomly selected student’s exam score, the best way for you to minimize your error would be to guess the average of all the students’ scores. But what if you also knew how many hours that student studied? Now, your best guess might be the average score of only the students who studied for that many hours.

Here is an example using a sample of 100 students with study times rounded to the nearest half hour. Suppose you were told a student studied for seven hours. To guess their exam score, one way to minimize error is to guess the average of only the students who studied for seven hours.



In this scatterplot, all of the students who studied for seven hours fall between the two vertical lines. Their mean exam score is represented by an X. Linear regression expands on this concept. A regression line represents the estimated average value of Y for every value of X, given the assumptions and limitations of a linear model. In other words, the actual average Y values for each X might not lie exactly on the regression line if the relationship between X and Y is not perfectly linear or if there are other factors influencing Y that are not included in the model. The regression line attempts to balance out these influences to find a straight-line relationship that best fits the data as a whole. It’s an estimation of the central tendency of Y, given X.

### The regression equation

Now that you know about r and you better understand the concept of regression, you’re ready to put everything together to find the line of best fit through the data. The formula for this line is known as the regression equation. There are two keys to this step.

The first is:

* The mean value of X and the mean value of Y (i.e., point (x̄, ȳ)) will always fall on the regression line.

The second is to understand what r means:

* For each increase of one standard deviation in X, there is an expected increase of r standard deviations in Y, on average over X.

The following figure illustrates how these concepts work together to determine the regression line.

