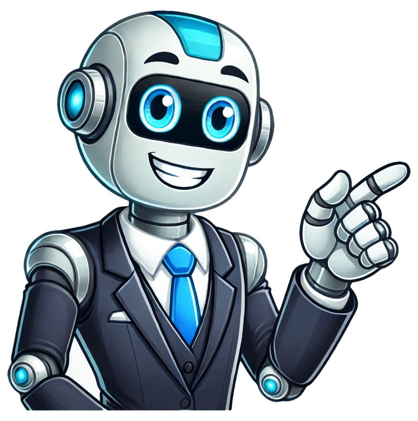

Regression is a statistical method that analyzes the relationship between a dependent variable and one or more independent variables. It helps predict or understand how changes in the independent variable(s) are associated with changes in the dependent variable. Linear regression is the most common form of this technique. It establishes the linear relationship between two variables and is also referred to as simple regression or ordinary least squares (OLS) regression. Linear regression is graphically depicted using a straight line of best fit, with the slope defining how the change in one variable impacts a change in the other. The y-intercept of a linear regression relationship represents the value of the dependent variable when the value of the independent variable is zero. Nonlinear regression models also exist but are far more complex. Regression is a statistical technique that relates a dependent variable to one or more independent variables.A regression model shows whether changes observed in the dependent variable are associated with changes in one or more of the independent variables.It does this by determining a best-fit line to see how the data is dispersed around this line.Regression helps economists and financial analysts with challenges ranging from asset valuation to making predictions.Several assumptions about the data and the model itself must hold for regression results to be properly interpreted. Regression is used in economics to help investment managers value assets and understand the relationships between factors such as commodity prices and the stocks of businesses dealing in those commodities. It's a powerful tool for uncovering the associations between variables observed in data, but it can't easily indicate causation. Regression as a statistical technique shouldn't be confused with the concept of regression to the mean, also known as mean reversion. Joules Garcia / Investopedia Regression captures the correlation between variables observed in a data set and quantifies whether those correlations are statistically significant. The two basic types of regression are simple linear regression and multiple linear regression, but there are nonlinear regression methods for more complicated data and analysis. Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y. Multiple linear regression uses two or more independent variables to predict the outcome. Analysts can use stepwise regression to examine each independent variable contained in the linear regression model. Regression can help finance and investment professionals. A company might use it to predict sales based on weather, previous sales, gross domestic product (GDP) growth, or other types of conditions. The capital asset pricing model (CAPM) is a regression model that's often used in finance for pricing assets and discovering the costs of capital. Econometrics is a set of statistical techniques that are used to analyze data in finance and economics. It effectively studies the income effect using observable data. An economist might hypothesize that a consumer's spending will increase as they increase their income. A regression analysis can then be conducted to understand the strength of the relationship between income and consumption if the data show that such an association is present. It can indicate whether that relationship is statistically significant. You can have several independent variables in an analysis, such as changes to GDP and inflation in addition to unemployment in explaining stock market prices. It's referred to as multiple linear regression when more than one independent variable is used. This is the most commonly used tool in econometrics. Econometrics is sometimes criticized for relying too heavily on the interpretation of regression output without linking it to economic theory or looking for causal mechanisms. It's crucial that the findings revealed in the data can be adequately explained by a theory. Linear regression models often use a least-squares approach to determine the line of best fit. The least-squares technique is determined by minimizing the sum of squares created by a mathematical function. A square is then determined by squaring the distance between a data point and the regression line or mean value of the dataset. A regression model is constructed when this process has been completed. It's usually accomplished with software. The general form of each type of regression model is: Simple linear regression: Y = a + b X + u \begin{aligned}&Y = a + bX + u \\\end{aligned} Y=a+bX+u Multiple linear regression: Y = a + b 1 X 1 + b 2 X 2 + b 3 X 3 + . . . + b t X t + u where: Y = The dependent variable you are trying to predict or explain X = The explanatory (independent) variable(s) you are using to predict or associate with Y a = The y-intercept b = (beta coefficient) is the slope of the explanatory variable(s) u = The regression residual or error term \begin{aligned}&Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_tX_t + u \\&\textbf{where:} \\&Y = \text{The dependent variable you are trying to predict} \\&\text{or explain} \\&X = \text{The explanatory (independent) variable(s) you are} \\&\text{using to predict or associate with Y} \\&a = \text{The y-intercept} \\&b = \text{(beta coefficient) is the slope of the explanatory} \\&\text{variable(s)} \\&u = \text{The regression residual or error term} \\\end{aligned} Y=a+b1X1+b2X2+b3X3+...+btXt+uwhere:Y=The dependent variable you are trying to predictor explainX=The explanatory (independent) variable(s) you are using to predict or associate with Ya=The y-interceptb=(beta coefficient) is the slope of the explanatoryvariable(s)u=The regression residual or error term Regression is often used to determine how specific factors such as the price of a commodity, interest rates, particular industries, or sectors influence the price movement of an asset. The CAPM is based on regression and is used to project the expected returns for stocks and generate costs of capital. A stock's returns are regressed against the returns of a broader index such as the S&P 500 to generate a beta for the particular stock. Beta is the stock's risk in relation to the market or index, and it's reflected as the slope in the CAPM. The return for the stock in question would be the dependent variable Y. The independent variable X would be the market risk premium. Additional variables such as the market capitalization of a stock, valuation ratios, and recent returns can be added to the CAPM to get better estimates for returns. These additional factors are known as the Fama-French factors. They're named after the professors who developed the multiple linear regression model to better explain asset returns. Regression tries to see if there's a relationship between two things, such as whether there's a link between how you do one thing and how you do one or more other things. Regression helps you make educated guesses, or predictions, based on past information. It's about finding a pattern between two or more things and using that pattern to make a good guess about what might happen in the future. There's some debate about the origins of the name, but this statistical technique was most likely termed "regression" by Sir Francis Galton in the 19th century. It described the statistical feature of biological data, such as the heights of people in a population, to regress to a mean level. There are shorter and taller people, but only outliers are very tall or short, and most people cluster somewhere around or "regress" to the average. Regression in statistics is a statistical technique used to identify the relationship between variables occurring in some data. It can show the magnitude of such an association and determine its statistical significance. Regression is a powerful tool for statistical inference and has been used to try to predict future outcomes based on past observations. A regression model output may be in the form of Y = 1.0 + (3.2)X1 - 2.0(X2) + 0.21.Here we have a multiple linear regression that relates some variable Y with two explanatory variables X1 and X2. We would interpret the model as the value of Y changes by 3.2× for every one-unit change in X1. If X1 goes up by 2, Y goes up by 6.4, holding all else constant.This means that when controlling for X2, X1 has this observed relationship. Every one-unit increase in X2 is associated with a 2× decrease in Y if X1 holds constant. We can also note the y-intercept of 1.0, indicating that Y = 1 when X1 and X2 are both zero. The error term or residual is 0.21. Four main assumptions about the underlying data process of what you're analyzing must hold to properly interpret the output of a regression model:The relationship between variables is linear.There must be homoskedasticity, or the variance of the variables and the error term must remain constant.All explanatory variables are independent of each other.All variables are normally distributed. Regression is a statistical method that tries to determine the strength and character of the relationship between one dependent variable and a series of other variables. It's used in finance, investing, and other disciplines. Regression analysis uncovers the associations between variables observed in data, but it can't easily indicate causation. Regression coefficients are fundamental components in statistical modeling, particularly in linear regression analysis. They represent the relationship between independent variables and the dependent variable, indicating how much the dependent variable is expected to change when an independent variable changes. Each coefficient quantifies the change in the dependent variable for a one-unit change in the corresponding independent variable, assuming all other variables remain constant. This concept is crucial for interpreting the results of regression analyses and understanding how different factors influence outcomes. Ad description. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Types of Regression Coefficients There are primarily two types of regression coefficients: the intercept and the slope coefficients. The intercept, often denoted as β0, indicates the expected value of the dependent variable when all independent variables are set to zero. Slope coefficients, denoted as β1, β2, etc., represent the change in the dependent variable for each unit increase in the respective independent variable. Understanding these coefficients helps in making predictions and analyzing the impact of various factors. Interpreting Regression Coefficients Interpreting regression coefficients requires a solid grasp of the context of the data. A positive coefficient suggests a direct relationship, meaning that as the independent variable increases, the dependent variable also tends to increase. Conversely, a negative coefficient indicates an inverse relationship. The magnitude of the coefficient provides insight into the strength of this relationship, allowing analysts to prioritize which variables have the most significant impact on the outcome. Standardized vs. Unstandardized Coefficients In regression analysis, coefficients can be presented in two forms: standardized and unstandardized. Unstandardized coefficients are the raw coefficients obtained from the regression equation, reflecting the actual units of measurement. Standardized coefficients, on the other hand, are dimensionless and allow for comparison across different variables by converting them to a common scale. This distinction is essential for understanding the relative importance of predictors in a regression model. Significance of Regression Coefficients The significance of regression coefficients is assessed through hypothesis testing, typically using t-tests. A coefficient is considered statistically significant if the p-value associated with it is below a predetermined threshold (commonly 0.05). This significance indicates that there is a meaningful relationship between the independent variable and the dependent variable, providing confidence in the predictive power of the model. Ad description. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Multicollinearity and Its Impact Multicollinearity occurs when independent variables in a regression model are highly correlated, which can distort the estimation of regression coefficients. This phenomenon can lead to inflated standard errors, making it difficult to assess the individual effect of each variable. Detecting multicollinearity is crucial, as it can undermine the reliability of the regression analysis and lead to misleading conclusions. Coefficient of Determination (R²) The coefficient of determination, commonly denoted as R², is a key statistic that indicates the proportion of variance in the dependent variable that can be explained by the independent variables in the model. While R² does not directly relate to individual regression coefficients, it provides context for understanding the overall effectiveness of the model. A higher R² value suggests that the regression coefficients collectively explain a larger portion of the variability in the dependent variable. Applications of Regression Coefficients Regression coefficients are widely used across various fields, including economics, social sciences, and health research. They enable researchers and analysts to make informed predictions, assess risk factors, and evaluate the effectiveness of interventions. By quantifying relationships between variables, regression coefficients serve as powerful tools for decision-making and strategic planning. Limitations of Regression Coefficients Despite their usefulness, regression coefficients have limitations. They assume a linear relationship between variables, which may not always hold true in real-world scenarios. Additionally, coefficients do not imply causation; a significant coefficient does not necessarily mean that changes in the independent variable cause changes in the dependent variable. Analysts must exercise caution and consider other factors when interpreting regression results. Conclusion on Regression Coefficients In summary, regression coefficients are vital for understanding the dynamics of relationships between variables in statistical modeling. Their interpretation, significance, and application are essential for effective data analysis and decision-making. By grasping the nuances of regression coefficients, analysts can leverage their insights to drive meaningful conclusions and actions based on data. Ad description. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Coefficients are the numbers by which the variables in an equation are multiplied. For example, in the equation y = -3.6 + 5.0X1 - 1.8X2, the variables X1 and X2 are multiplied by 5.0 and -1.8, respectively, so the coefficients are 5.0 and -1.8. The size and sign of a coefficient in an equation affect its graph. In a simple linear equation (contains only one x variable), the coefficient (and slope) is the slope of the line. Minitab displays the coefficients for each predictor variable and the constant. When calculating a regression equation to model data, Minitab estimates the coefficients for each predictor variable based on your sample and displays these estimates in a coefficients table. For example, the following coefficients table is shown in the output for this regression equation: Regression Equation Heat Flux = 325.4 + 2.55 East + 3.80 South - 22.95 North + 0.0675 Insolation + 2.42 Time of Day This equation predicts the heat flux in a home based on the position of its focal points, the insolation, and the time of day. Minitab displays the coefficient values for the equation in the second column: Coefficients Term Coef SE Coef T-Value P-Value VIF Constant 325.4 96.1 3.39 0.003 East 2.55 1.25 2.04 0.053 1.36 South 3.80 1.46 2.60 0.016 3.18 North -22.95 2.70 -8.49 0.000 2.61 Insolation 0.0675 0.0290 2.33 0.029 2.32 Time of Day 2.42 1.81 1.34 0.194 5.37 Each coefficient estimates the change in the mean response per unit increase in X when all other predictors are held constant. For example, in the regression equation, if the North variable increases by 1 and the other variables remain the same, heat flux decreases by about 22.95 on average. If the p-value of a coefficient is less than the chosen significance level, such as 0.05, the relationship between the predictor and the response is statistically significant. Minitab also includes a value for the constant in the equation in the Coef column. The term coefficient can also be used to denote a calculated numerical value used as an index, such as a coefficient of correlation, a coefficient of determination, or Kendall's coefficient. Data rarely fit a straight line exactly. Usually, you must be satisfied with rough predictions. Typically, you have a set of data whose scatter plot appears to "fit" a straight line. This is called a Line of Best Fit or Least-Squares Line. If you know a person's pinky (smallest) finger length, do you think you could predict that person's height? Collect data from your class (pinky finger length, in inches). The independent variable, \(x\), is pinky finger length and the dependent variable, \(y\), is height. For each set of data, plot the points on graph paper. Make your graph big enough and use a ruler. Then "by eye" draw a line that appears to "fit" the data. For your line, pick two convenient points and use them to find the slope of the line. Find the \(y\)-intercept of the line by extending your line so it crosses the \(y\)-axis. Using the slopes and the \(y\)-intercepts, write your equation of "best fit." Do you think everyone will have the same equation? Why or why not? According to your equation, what is the predicted height for a pinky length of 2.5 inches? A random sample of 11 statistics students produced the following data, where \(x\) is the third exam score out of 80, and \(y\) is the final exam score out of 200. Can you predict the final exam score of a random student if you know the third exam score? 1a: Table showing the scores on the final exam based on scores from the third exam. \(x\) (third exam score) \(y\) (final exam score) 65 175 67 133 71 185 71 163 66 126 75 198 67 153 70 163 71 159 69 151 69 159 Figure \(\PageIndex{1}\): Scatter plot showing the scores on the final exam based on scores from the third exam. SCUBA divers have maximum dive times they cannot exceed when going to different depths. The data in Table show different depths with the maximum dive times in minutes. Use your calculator to find the least squares regression line and predict the maximum dive time for 110 feet. \(X\) (depth in feet) \(Y\) (maximum dive time) 50 80 60 55 70 45 80 35 90 25 100 22 Answer \(\hat{y} = 127.24 - 1.11x\) At 110 feet, a diver could dive for only five minutes. The third exam score, \(x\), is the independent variable and the final exam score, \(y\), is the dependent variable. We will plot a regression line that best "fits" the data. If each of you were to fit a line "by eye," you would draw different lines. We can use what is called a least-squares regression line to obtain the best fit line. Consider the following diagram. Each point of data is of the the form (\(x, y\)) and each point of the line of best fit using least-squares linear regression has the form (\(x, \hat{y}\)). The \(\hat{y}\) is read "\(y\) hat" and is the estimated value of \(y\). It is the value of \(y\) obtained using the regression line. It is not generally equal to \(y\) from data. Figure \(\PageIndex{2}\) The term \(y_{0} – \hat{y}_{0} = \varepsilon_{0}\) is called the "error" or residual. It is not an error in the sense of a mistake. The absolute value of a residual measures the vertical distance between the actual value of \(y\) and the estimated value of \(y\). In other words, it measures the vertical distance between the actual data point and the predicted point on the line. If the observed data point lies above the line, the residual is positive, and the line underestimates the actual data value for \(y\). If the observed data point lies below the line, the residual is negative, and the line overestimates that actual data value for \(y\). In the diagram in Figure, \(y_{0} – \hat{y}_{0} = \varepsilon_{0}\) is the residual for the point shown. Here the point lies above the line and the residual is positive. \(\varepsilon =\) the Greek letter epsilon For each data point, you can calculate the residuals or errors, \(y_{i} - \hat{y}_{i} = \varepsilon_{i}\) for \(i = 1, 2, 3, ..., 11\). Each \(|\varepsilon|\) is a vertical distance. For the example about the third exam scores and the final exam scores for the 11 statistics students, there are 11 data points. Therefore, there are 11 \(\varepsilon\) values. If you square each \(\varepsilon\) and add, you get \[(\varepsilon_{1})^{2} + (\varepsilon_{2})^{2} + \dotso + (\varepsilon_{11})^{2} = \sum^{11}_{i = 1} \varepsilon^{2} \label{SSE}\] Equation\ref{SSE} is called the Sum of Squared Errors (SSE). Using calculus, you can determine the values of \(a\) and \(b\) that make the SSE a minimum. When you make the SSE a minimum, you have determined the points that are on the line of best fit. It turns out that the line of best fit has the equation: \[\hat{y} = a + bx\] where \(a = \bar{y} - b\bar{x}\) and \(b = \dfrac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^{2}}\). The sample means of the \(x\) values and the \(x\) values are \(\bar{x}\) and \(\bar{y}\), respectively. The best fit line always passes through the point \((\bar{x}, \bar{y})\). The slope \(b\) can be written as \(b = r\left(\dfrac{s_{y}}{s_{x}}\right)\) where \(s_{y} =\) the standard deviation of the \(y\) values and \(s_{x} =\) the standard deviation of the \(x\) values. \(r\) is the correlation coefficient, which is discussed in the next section. The process of fitting the best-fit line is called linear regression. The idea behind finding the best-fit line is based on the assumption that the data are scattered about a straight line. The criteria for the best fit line is that the sum of the squared errors (SSE) is minimized, that is, made as small as possible. Any other line you might choose would have a higher SSE than the best fit line. This best fit line is called the least-squares regression line . Computer spreadsheets, statistical software, and many calculators can quickly calculate the best-fit line and create the graphs. The calculations tend to be tedious if done by hand. Instructions to use the TI-83, TI-83+, and TI-84+ calculators to find the best-fit line and create a scatterplot are shown at the end of this section. THIRD EXAM vs FINAL EXAM EXAMPLE: The graph of the line of best fit for the third-exam/final-exam example is as follows: Figure \(\PageIndex{3}\) The least squares regression line (best-fit line) for the third-exam/final-exam example has the equation: \[\hat{y} = -173.51 + 4.83x\] Remember, it is always important to plot a scatter diagram first. If the scatter plot indicates that there is a linear relationship between the variables, then it is reasonable to use a best fit line to make predictions for \(y\) given \(x\) within the domain of \(x\)-values in the sample data, but not necessarily for x-values outside that domain. You could use the line to predict the final exam score for a student who earned a grade of 73 on the third exam. You should NOT use the line to predict the final exam score for a student who earned a grade of 50 on the third exam, because 50 is not within the domain of the \(x\)-values in the sample data, which are between 65 and 75. The slope of the line, \(b\), describes how changes in the variables are related. It is important to interpret the slope of the line in the context of the situation represented by the data. You should be able to write a sentence interpreting the slope in plain English. INTERPRETATION OF THE SLOPE: The slope of the best-fit line tells us how the dependent variable (\(y\)) changes for every one unit increase in the independent (\(x\)) variable, on average. THIRD EXAM vs FINAL EXAM EXAMPLE Slope: The slope of the line is \(b = 4.83\). Interpretation: For a one-point increase in the score on the third exam, the final exam score increases by 4.83 points, on average. Using the Linear Regression T Test: LinRegTTest In the STAT list editor, enter the \(X\) data in list L1 and the Y data in list L2, paired so that the corresponding (\(x,y\)) values are next to each other in the lists. (If a particular pair of values is repeated, enter it as many times as it appears in the data.) On the STAT TESTS menu, scroll down with the cursor to select the LinRegTTest. (Be careful to select LinRegTTest, as some calculators may also have a different item called LinRegTInt.) On the LinRegTTest input screen enter: Xlist: L1 ; Ylist: L2 ; Freq: 1 On the next line, at the prompt \(\beta\) or \(\rho\), highlight "\(\neq 0\)" and press ENTER Leave the line for "RegEq:" blank Highlight Calculate and press ENTER. Figure \(\PageIndex{4}\) The output screen contains a lot of information. For now we will focus on a few items from the output, and will return later to the other items. The second line says \(y = a + bx\). Scroll down to find the values \(a = -173.513\), and \(b = 4.8273\); the equation of the best fit line is \(\hat{y} = -173.51 + 4.83x\) The two items at the bottom are \(r_{2} = 0.43969\) and \(r = 0.663\). For now, just note where to find these values; we will discuss them in the next two sections. Graphing the Scatterplot and Regression Line We are assuming your \(X\) data is already entered in list L1 and your \(Y\) data is in list L2 Press 2nd STATPLOT ENTER to use Plot 1 On the input screen for PLOT 1, highlight On, and press ENTER For TYPE: highlight the very first icon which is the scatterplot and press ENTER Indicate Xlist: L1 and Ylist: L2 For Mark: it does not matter which symbol you highlight. Press the ZOOM key and then the number 9 (for menu item "ZoomStat") ; the calculator will fit the window to the data To graph the best-fit line, press the "\(Y =\)" key and type the equation \(-173.5 + 4.83X\) into equation Y1. (The \(X\) key is immediately left of the STAT key). Press ZOOM 9 again to graph it. Optional: If you want to change the viewing window, press the WINDOW key. Enter your desired window using Xmin, Xmax, Ymin, Ymax Another way to graph the line after you create a scatter plot is to use LinRegTTest. Make sure you have done the scatter plot. Check it on your screen. Go to LinRegTTest and enter the lists. At RegEq: press VARS and arrow over to Y-VARS. Press 1 for 1:Function. Press 1 for 1:Y1. Then arrow down to Calculate and do the calculation for the line of best fit. Press \(Y = \text{you will see the regression equation}\)). Press GRAPH. The line will be drawn." Besides looking at the scatter plot and seeing that a line seems reasonable, how can you tell if the line is a good predictor? Use the correlation coefficient as another indicator (besides the scatterplot) of the strength of the relationship between \(x\) and \(y\). The correlation coefficient, \(r\), developed by Karl Pearson in the early 1900s, is numerical and provides a measure of strength and direction of the linear association between the independent variable \(x\) and the dependent variable \(y\). The correlation coefficient is calculated as \[r = \dfrac{n \sum(xy) - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n \sum x^{2} - \left(\sum x\right)^{2}\right] \left[n \sum y^{2} - \left(\sum y\right)^{2}\right]}}\] where \(n =\) the number of data points. If you suspect a linear relationship between \(x\) and \(y\), then \(r\) can measure how strong the linear relationship is. What the VALUE of \(r\) tells us: The value of \(r\) is always between –1 and +1: –1 ≤ r ≤ 1. The size of the correlation \(r\) indicates the strength of the linear relationship between \(x\) and \(y\). Values of \(r\) close to –1 or to +1 indicate a stronger linear relationship between \(x\) and \(y\). If \(r = 0\) there is absolutely no linear relationship between \(x\) and \(y\) (no linear correlation). If \(r = 1\), there is perfect positive correlation. If \(r = -1\), there is perfect negative correlation. In both these cases, all of the original data points lie on a straight line. Of course,in the real world, this will not generally happen. What the SIGN of \(r\) tells us: A positive value of \(r\) means that when \(x\) increases, \(y\) tends to increase and when \(x\) decreases, \(y\) tends to decrease (positive correlation). A negative value of \(r\) means that when \(x\) increases, \(y\) tends to decrease and when \(x\) decreases, \(y\) tends to increase (negative correlation). The sign of \(r\) is the same as the sign of the slope, \(b\), of the best-fit line. Strong correlation does not suggest that \(x\) causes \(y\) or \(y\) causes \(x\). We say "correlation does not imply causation." Figure \(\PageIndex{5}\): (a) A scatter plot showing data with a positive correlation. \(0 < r < 1\) (b) A scatter plot showing data with a negative correlation. \(-1 < r < 0\) (c) A scatter plot showing data with zero correlation. \(r = 0\) The formula for \(r\) looks formidable. However, computer spreadsheets, statistical software, and many calculators can quickly calculate \(r\). The correlation coefficient \(r\) is the bottom item in the output screens for the LinRegTTest on the TI-83, TI-83+, or TI-84+ calculator (see previous section for instructions). The variable \(r^{2}\) is called the coefficient of determination and is the square of the correlation coefficient, but is usually stated as a percent, rather than in decimal form. It has an interpretation in the context of the data: \(r^{2}\), when expressed as a percent, represents the percent of variation in the dependent (predicted) variable \(y\) that can be explained by variation in the independent (explanatory) variable \(x\) using the regression (best-fit) line. \(1 - r^{2}\), when expressed as a percentage, represents the percent of variation in \(y\) that is NOT explained by variation in \(x\) using the regression line. This can be seen as the scattering of the observed data points about the regression line. Consider the third exam/final exam example introduced in the previous section The line of best fit is: \(\hat{y} = -173.51 + 4.83x\) The correlation coefficient is \(r = 0.6631\) The coefficient of determination is \(r^{2} = 0.6631^{2} = 0.4397\) Interpretation of \(r^{2}\) in the context of this example: Approximately 44% of the variation (0.4397 is approximately 0.44) in the final-exam grades can be explained by the variation in the grades on the third exam, using the best-fit regression line. Therefore, approximately 56% of the variation (\(1 - 0.44 = 0.56\)) in the final exam grades can NOT be explained by the variation in the grades on the third exam, using the best-fit regression line. (This is seen as the scattering of the points about the line.) A regression line, or a line of best fit, can be drawn on a scatter plot and used to predict outcomes for the \(x\) and \(y\) variables in a given data set or sample data. There are several ways to find a regression line, but usually the least-squares regression line is used because it creates a uniform line. Residuals, also called “errors,” measure the distance from the actual value of \(y\) and the estimated value of \(y\). The Sum of Squared Errors, when set to its minimum, calculates the points on the line of best fit. Regression lines can be used to predict values within the given set of data, but should not be used to make predictions for values outside the set of data. The correlation coefficient \(r\) measures the strength of the linear association between \(x\) and \(y\). The variable \(r\) has to be between –1 and +1. When \(r\) is positive, the \(x\) and \(y\) will tend to increase and decrease together. When \(r\) is negative, \(x\) will increase and \(y\) will decrease, or the opposite, \(x\) will decrease and \(y\) will increase. The coefficient of determination \(r^{2}\), is equal to the square of the correlation coefficient. When expressed as a percent, \(r^{2}\) represents the percent of variation in the dependent variable \(y\) that can be explained by variation in the independent variable \(x\) using the regression line. Coefficient of Correlation a measure developed by Karl Pearson (early 1900s) that gives the strength of association between the independent variable and the dependent variable; the formula is: \[r = \dfrac{n \sum xy - \left(\sum x\right) \left(\sum y\right)}{\sqrt{\left[n \sum x^{2} - \left(\sum x\right)^{2}\right] \left[n \sum y^{2} - \left(\sum y\right)^{2}\right]}}\] where \(n\) is the number of data points. The coefficient cannot be more than 1 or less than –1. The closer the coefficient is to ±1, the stronger the evidence of a significant linear relationship between \(x\) and \(y\). Linear regression strives to show the relationship between two variables by applying a linear equation to observed data. One variable is supposed to be an independent variable, and the other is to be a dependent variable. For example, the weight of the person is linearly related to his height. Hence this shows a linear relationship between the height and weight of the person. As the height is increased, the weight of the person also gets increased. It is not necessary that here one variable is dependent on others, or one causes the other, but there is some critical relationship between the two variables. In such cases, we use a scatter plot to imply the strength of the relationship between the variables. If there is no relation or linking between the variables, the scatter plot does not indicate any increasing or decreasing pattern. For such cases, the linear regression design is not beneficial to the given data. Also, read: Linear Regression Equation The measure of the extent of the relationship between two variables is shown by the correlation coefficient. The range of this coefficient lies between -1 to +1. This coefficient shows the strength of the association of the observed data for two variables. A linear regression line equation is written in the form of: Y = a + bX where X is the independent variable and plotted along the x-axis Y is the dependent variable and plotted along the y-axis The slope of the line is b, and a is the intercept (the value of y when x = 0). Linear Regression Formula Linear regression shows the linear relationship between two variables. The equation of linear regression is similar to the slope formula what we have learned before in earlier classes such as linear equations in two variables. It is given by; Y= a + bX Now, here we need to find the value of the slope of the line, b, plotted in scatter plot and the intercept, a. Simple Linear Regression The very most straightforward case of a single scalar predictor variable x and a single scalar response variable y is known as simple linear regression. The equation for this regression is represented by; y=a+bx The expansion to multiple and vector-valued predictor variables is known as multiple linear regression, also known as multivariable linear regression. The equation for this regression is represented by; Y = a+bX Almost all real-world regression patterns include multiple predictors, and basic explanations of linear regression are often explained in terms of the multiple regression form. Note that, though, in these cases, the dependent variable y is yet a scalar. Least Square Regression Line or Linear Regression Line The most popular method to fit a regression line in the XY plot is the method of least-squares. This process determines the best-fitting line for the noted data by reducing the sum of the squares of the vertical deviations from each data point to the line. If a point rests on the fitted line accurately, then its perpendicular deviation is 0. Because the variations are first squared, then added, their positive and negative values will not be cancelled. Linear regression determines the straight line, called the least-squares regression line or LSRL, that best expresses observations in a bivariate analysis of data set. Suppose Y is a dependent variable, and X is an independent variable, then the population regression line is given by; Y = B0+B1X Where B0 is a constant B1 is the regression coefficient If a random sample of observations is given, then the regression line is expressed by; ŷ = b0 + b1x where b0 is a constant, b1 is the regression coefficient, x is the independent variable, and ŷ is the predicted value of the dependent variable. Properties of Linear Regression For the regression line where the regression parameters b0 and b1 are defined, the properties are given as: The line reduces the sum of squared differences between observed values and predicted values. The regression line passes through the mean of X and Y variable values The regression constant (b0) is equal to y-intercept the linear regression The regression coefficient (b1) is the slope of the regression line which is equal to the average change in the dependent variable (Y) for a unit change in the independent variable (X). Regression Coefficient In the linear regression line, we have seen the equation is given by; Y = B0+B1X Where B0 is a constant B1 is the regression coefficient Now, let us see the formula to find the value of the regression coefficient. B1 = b1 = Σ [(xi – x)(yi – y)] / Σ [(xi – x)2] Where xi and yi are the observed data sets. And x and y are the mean value. The estimation of relationships between a dependent variable and one or more independent variables Over 2 million + professionals use CFI to learn accounting, financial analysis, modeling and more. Unlock the essentials of corporate finance with our free resources and get an exclusive sneak peek at the first module of each course. Start Free Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modeling the future relationship between them. Regression analysis includes several variations, such as linear, multiple linear, and nonlinear. The most common models are simple linear and multiple linear. Nonlinear regression analysis is commonly used for more complicated data sets in which the dependent and independent variables show a nonlinear relationship. Regression analysis offers numerous applications in various disciplines, including finance. Regression Analysis – Linear Model Assumptions Linear regression analysis is based on six fundamental assumptions: The dependent and independent variables show a linear relationship between the slope and the intercept. The independent variable is not random. The value of the residual (error) is zero. The value of the residual (error) is constant across all observations. The value of the residual (error) is not correlated across all observations. The residual (error) values follow the normal distribution. Regression Analysis – Simple Linear Regression Simple linear regression is a model that assesses the relationship between a dependent variable and an independent variable. The simple linear model is expressed using the following equation: Y = a + bX + ϵ Where: Y – Dependent variable X – Independent (explanatory) variable a – Intercept b – Slope ϵ – Residual (error) Check out the following video to learn more about simple linear regression: Regression Analysis – Multiple Linear Regression Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is: Y = a + bX1 + cX2 + dX3 + ϵ Where: Y – Dependent variable X1, X2, X3 – Independent (explanatory) variables a – Intercept b, c, d – Slopes ϵ – Residual (error) Multiple linear regression follows the same conditions as the simple linear model. However, since there are several independent variables in multiple linear analysis, there is another mandatory condition for the model: Non-collinearity: Independent variables should show a minimum correlation with each other. If the independent variables are highly correlated with each other, it will be difficult to assess the true relationships between the dependent and independent variables. Regression Analysis in Finance Regression analysis comes with several applications in finance. For example, the statistical method is fundamental to the Capital Asset Pricing Model (CAPM). Essentially, the CAPM equation is a model that determines the relationship between the expected return of an asset and the market risk premium. The analysis is also used to forecast the returns of securities, based on different factors, or to forecast the performance of a business. Learn more forecasting methods in CFI’s Budgeting and Forecasting Course! 1. Beta and CAPM In finance, regression analysis is used to calculate the Beta (volatility of returns relative to the overall market) for a stock. It can be done in Excel using the Slope function. Download CFI’s free beta calculator! 2. Forecasting Revenues and Expenses When forecasting financial statements for a company, it may be useful to do a multiple regression analysis to determine how changes in certain assumptions or drivers of the business will impact revenue or expenses in the future. For example, there may be a very high correlation between the number of salespeople employed by a company, the number of stores they operate, and the revenue the business generates. The above example shows how to use the Forecast function in Excel to calculate a company’s revenue, based on the number of ads it runs. Learn more forecasting methods in CFI’s Budgeting and Forecasting Course! Excel remains a popular tool to conduct basic regression analysis in finance, however, there are many more advanced statistical tools that can be used. Python and R are both powerful coding languages that have become popular for all types of financial modeling, including regression. These techniques form a core part of data science and machine learning, where models are trained to detect these relationships in data. Learn more about regression analysis, Python, and Machine Learning in CFI’s Business Intelligence & Data Analysis certification. Additional Resources To learn more about related topics, check out the following free CFI resources: Regression coefficients are the quantities by which the variables in a regression equation are multiplied. The most commonly used type of regression is linear regression. The aim of linear regression is to find the regression coefficients that produce the best-fitted line. The regression coefficients in linear regression help in predicting the value of an unknown variable using a known variable. In this article, we will learn more about regression coefficients, their formulas as well as see certain associated examples so as to find the best-fitted regression line. What are Regression Coefficients? Regression coefficients can be defined as estimates of some unknown parameters to describe the relationship between a predictor variable and the corresponding response. In other words, regression coefficients are used to predict the value of an unknown variable using a known variable. Linear regression is used to quantify how a unit change in an independent variable causes an effect in the dependent variable by determining the equation of the best-fitted straight line. This process is known as regression analysis. Formula for Regression Coefficients The goal of linear regression is to find the equation of the straight line that best describes the relationship between two or more variables. For example, suppose a simple regression equation is given by y = 7x - 3, then 7 is the coefficient, x is the predictor and -3 is the constant term. Suppose the equation of the best-fitted line is given by Y = aX + b then, the regression coefficients formula is given as follows: a = \(\frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^{2})-(\sum x)^{2}}\) b = \(\frac{(\sum y)(\sum x^{2})-(\sum x)(\sum xy)}{n(\sum x^{2})-(\sum x)^{2}}\) here, n refers to the number of data points in the given data sets. How to Find Regression Coefficients? Before determining the regression coefficients to find the best-fitted line, it is necessary to check whether the variables follow a linear relationship or not. This can be done by using the correlation coefficient and interpreting the corresponding value. Given below are the steps to find the regression coefficients for regression analysis. To find the coefficient of X use the formula a = \(\frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^{2})-(\sum x)^{2}}\). To find the constant term the formula is b = \(\frac{(\sum y)(\sum x^{2})-(\sum x)(\sum xy)}{n(\sum x^{2})-(\sum x)^{2}}\). Now input the regression coefficients in the equation Y = aX + b. A scatter plot can also be made so as to visually depict the regression line as shown below. Regression Coefficients Interpretation It is necessary to understand the nature of the regression coefficient as this helps to make certain predictions about the unknown variable. It helps to check to what extent a dependent variable will change with a unit change in the independent variable. Given below are the regression coefficients interpretation. If the sign of the coefficients is positive it implies that there is a direct relationship between the variables. This means that if the independent variable increases (or decreases) then the dependent variable also increases (or decreases). If the sign of the coefficients is negative it means that if the independent variable increases then the dependent variable decreases and vice versa. This means it is an indirect relationship. Related Articles: Probability and Statistics Data Handling Introduction to Graphing Important Notes on Regression Coefficients Regression coefficients are values that are used in a regression equation to estimate the predictor variable and its response. The most commonly used type of regression is linear regression. Equation of the best-fitted line is given by Y = aX + b. By using formulas, the values of the regression coefficient can be determined so as to get the regression line for the given variables. Example 1: Find the regression coefficients for the following data: Age Glucose Level 43 99 21 65 25 79 42 75 57 87 59 81 Solution: Age (x) Glucose Level (y) xy x2 y2 43 99 4257 1849 9801 21 65 1365 441 4225 25 79 1975 625 6241 42 75 3150 1764 5625 57 87 4959 3249 7569 59 81 4779 3481 6561 Total = 247 486 20485 11409 40022 The formula for finding the regression coefficients are as follows: a = \(\frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^{2})-(\sum x)^{2}}\) = 0.39 b = \(\frac{(\sum y)(\sum x^{2})-(\sum x)(\sum xy)}{n(\sum x^{2})-(\sum x)^{2}}\) = 65.14 The regression equation is Y = 0.39X + 65.14 Answer: a = 0.39 and b = 65.14 Example 2: Find the regression line for the following data. A B 6.25 4.03 6.5 4.02 6.5 4.02 6 4.04 6.25 4.03 6.25 4.03 Solution: X Y XY X2 Y2 6.25 4.03 25.19 39.06 16.24 6.5 4.02 26.13 42.25 16.16 6.5 4.02 26.13 42.25 16.16 6 4.04 24.24 36 16.32 6.25 4.03 25.19 39.06 16.24 6.25 4.03 25.19 39.06 16.24 Total = 37.75 24.17 152.06 237.69 97.37 The formula for finding the regression coefficients are as follows: a = \(\frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^{2})-(\sum x)^{2}}\) = -0.04 b = \(\frac{(\sum y)(\sum x^{2})-(\sum x)(\sum xy)}{n(\sum x^{2})-(\sum x)^{2}}\) = 4.28 The regression equation is Y = -0.04X + 4.28 Answer: Regression equation is Y = -0.04X + 4.28 Example 3: Plot the graph for the following data if the regression coefficients are given as a = -0.07 and b = 68.63 X Y 130 55 135 56 140 62 142 63 147 63 156 51 Solution: The regression coefficients are given as a = -0.07 and b = 68.63 Thus, the regression line is Y = -0.07X + 68.63 Thus, the scatter plot can be drawn as follows: Show more > go to slidego to slidego to slide Breakdown tough concepts through simple visuals. Math will no longer be a tough subject, especially when you understand the concepts through visualizations. Book a Free Trial Class FAQs on Regression Coefficients In statistics, regression coefficients can be defined as multipliers for variables. They are used in regression equations to estimate the value of the unknown parameters using the known parameters. What are Regression Coefficients Independent of? Regression coefficients are independent of the change of scale as well as the origin of the plot. What is the Formula for Regression Coefficients? The formula for regression coefficients is given as a = \(\frac{n(\sum xy)-(\sum x)(\sum y)}{n(\sum x^{2})-(\sum x)^{2}}\) and b = \(\frac{(\sum y)(\sum x^{2})-(\sum x)(\sum xy)}{n(\sum x^{2})-(\sum x)^{2}}\). How are Regression Coefficients used in a Linear Regression Line? The equation of a linear regression line is given as Y = aX + b, where a and b are the regression coefficients. How to Interpret Regression Coefficients? If the value of the regression coefficients is positive then it means that the variables have a direct relationship while negative regression coefficients imply that the variables have an indirect relationship. How to Calculate Regression Coefficients? The steps to calculate the regression coefficients are as follows: Substitute values to find a (coefficient of X). Substitute values for b (constant term). Put the values of these regression coefficients in the linear equation Y = aX + b. What Do Regression Coefficients Tell Us? Regression Coefficients tell us how much a dependent variable changes with a unit change in the independent variables. In This ArticleRegression analysis is a statistical method used in many fields including data science, machine learning, economics, and medicine. In this article, we’ll explain one of the key elements of a regression: the regression coefficient. What Is a Regression Coefficient? A regression coefficient is the quantity that sits in front of an independent variable in your regression equation. It is a parameter estimate describing the relationship between one of the independent variables in your model and the dependent variable. In the simple linear regression below, the quantity 0.5, which sits in front of the variable X, is a regression coefficient. The intercept—in this case 2—is also a coefficient, but you’ll hear it referred to, instead, as the “intercept,” “constant,” or "β0\beta_0β0". For the sake of this article, we will leave the intercept out of our discussion. Y^=2+0.5X\hat{Y}= 2 + 0.5XY^=2+0.5XRegression coefficients tell us about the line of best fit and the estimated relationship between an independent variable and the dependent variable in our model. In a simple linear regression with only one independent variable, the coefficient determines the slope of the regression line; it tells you whether the regression line is upward or downward-sloping and how steep the line is. Regression Coefficient in Multiple Regression Regressions can have more than one dependent variable, and, therefore, more than one regression coefficient. In the multivariate regression below, there are two independent variables (X1X_1X1 and X2X_2X2). This means you have two regression coefficients: 0.7 and -3.2. Each coefficient gives you information about the relationship between one of the independent variables and the dependent (or response) variable, Y. Y^=−4+0.7X1−3.2X2\hat{Y}= -4 + 0.7X_1 -3.2X_2Y^=−4+0.7X1−3.2X2In the regression here, the coefficient 0.7 suggests a positive linear relationship between X1X_1X1 and Y. If all other independent variables in the model are held constant, as X1X_1X1 increases by 1 unit, we estimate that Y increases by 0.7 units. The coefficient in front of X2X_2X2 is negative, indicating a negative correlation between X2X_2X2 and Y. If X2X_2X2 were to increase by one unit, and all other variables were held constant, we would predict Y to decrease by -3.2. Applying the Regression Coefficient Simple Linear Regression Model The figure below is a scatterplot showing the relationship between an independent variable—also called a predictor variable—plotted along the x-axis and the dependent variable plotted along the y-axis. Each point on the scatter plot represents an observation from a dataset. In linear regression, we estimate the relationship between the independent variable (X) and the dependent variable (Y) using a straight line. There are a few different ways to fit this line, but the most common method is called the Ordinary Least Squares Method (or OLS). In OLS, you can find the regression line by minimizing the sum of squared errors. Here the errors—or residuals—are the vertical distances between each point on the scatter plot and the regression line. The regression coefficient on X tells you the slope of the regression line. Regression Output Tables We typically perform our regression calculations using statistical software like R or Stata. When we do this, we not only create scatter plots and lines but also create a regression output table like the one below. A regression output table is a table summarizing the regression line, the errors of your model, and the statistical significance of each parameter estimated by your model. In the table here, the independent variables (X1X_1X1 and X2X_2X2) are listed in the first column of the table, and the coefficients on these variables are listed in the second column of the table in rows 3 and 4. Regression Coefficient Interpretation In linear regression, your regression coefficients will be constants that are either positive or negative. Here is how you can interpret the coefficients. 1. Non-Zero Regression Coefficient A non-zero regression coefficient indicates a relationship between the independent variable and the dependent variable. 2. Positive Coefficient If the regression coefficient is positive, there is a positive relationship between the independent variable and the dependent variable. As X increases, Y tends to increase, and as X decreases, Y tends to decrease. 3. Negative Coefficient If the regression coefficient is negative, there is a negative (or inverse) relationship between the independent variable and the dependent variable. As X increases, Y tends to decrease, and as X decreases, Y tends to increase. Remember, you still cannot say with certainty that X causes Y. In regression, you can estimate how much of the variation in your independent variable can be explained by the dependent variable with R2R^2R2, and what the exact relationship between the variables is. In regression, you can estimate how much of the variation in your independent variable can be explained by the variation in your dependent variable... How to Find Regression Coefficients In a simple OLS linear regression of the form Y = B0+B1X, the regression coefficient B1 is calculated using the following formula: B1=Cov(X,Y)Var(X)... Regression Coefficients in Different Types of Regression Models In this article, we've mainly discussed the simplest form of regression: a linear regression with one independent variable. As you continue to study statistics, you'll encounter many more complex forms of regression. In these other regression models, the regression coefficients might take on slightly different forms and may need to be interpreted differently. Here's a list of some commonly used regression models. Linear Regression Linear regression is one of the most basic forms of regressions. As you saw earlier, in linear regression, you find a line of best fit (a regression line) that minimizes the sum of squared errors. Logistic Regression (or Logit Regression) We use logistic regression when you want to study a binary outcome and you want to estimate the likelihood of one of the two possible outcomes occurring. In logistic regression, your outcome variable is binary and can take on only one of two values. In logistic regression, you interpret the regression coefficients differently than you would in a linear model. In linear regression, a coefficient of 2 means that as your independent variable increases by one unit, your dependent variable is expected to increase by 2 units. In logistic regression, a coefficient of 2 means that as your independent variable increases by one unit, the log odds of your dependent variable increase by 2. Regression Models with Non-linear Terms In a non-linear regression, you might estimate the relationship between variables using a curve rather than a line. For example, if the relationship between X and Y cannot be expressed by a line, but rather by a curve, you might want to include X1X_1X1 but also its quadratic version X12X_1^2X12. In this case, we will get two coefficients related to X1X_1X1; one for X1X_1X1 and one for X12X_1^2X12. Something like this: Y^=−8.2+1.5X1−0.5(X1)2... In other words, we no longer can say that if X1 changes by one unit, Y changes by 0.4 units since X1X_1X1 appears twice in the regression. Instead, the relationship between Y and X1X_1X1 is non-linear. If the level of X1X_1X1 is 1 and we increase it by 1 unit, then Y increases by (1.5 - 1) units. However, if the level of X1X_1X1 is 2 and we increase it by 1 unit, then Y increases by (1.5 - 2). This is because the partial derivative of Y with respect to X1X_1X1 is no longer a constant and in this case it is 1.5 - 2 × 0.5 X1. Ridge Regression Ridge regression is a technique used in machine learning. Statisticians and data scientists use ridge regression to adjust linear regressions to avoid overfitting the model to training data. In ridge regression, the parameters of the model (including the regression coefficients) are found by minimizing the sum of squared errors plus a value called the ridge regression penalty. As a result of the adjustment, the dependent variables become less sensitive to changes in the independent variable. In other words, the coefficients in a ridge regression tend to be smaller in absolute value than the coefficients in an OLS regression. Lasso Regression Lasso regression is similar to ridge regression. It is an adjustment method used with OLS to adjust for the risk of overfitting a model to training data. In a Lasso regression, you adjust your OLS regression line by using a value known as the Lasso regression penalty. Similar to the Ridge regression, the lasso regression penalty shrinks the coefficients in a regression model. Explore Outlier's Award-Winning For-Credit Courses... Linear regression is a powerful tool for statistical analysis... Linear regression aims to determine the regression coefficients that result in the best-fitting line. These coefficients are helpful when estimating the value of an unknown variable using a known variable. This article explains regression coefficients and provides related examples. What are Regression Coefficients?Regression coefficients are estimations of unknown parameters that describe the connection between a predictor variable and its associated response. In other words, regression coefficients are used to estimate the value of an unknown variable using a known variable. Linear regression is used to measure how a unit change in the independent variable affects the dependent variable by calculating the equation of the best-fitted straight line. This method is referred to as regression analysis. Regression LineLinear regression models aim to find a line equation that best represents the relationship between dependent (y) and independent (x) variables.y = a + bx y is the dependent variable, also known as the response or explained variable.x is the independent variable, also known as the predictor or explanatory variable. a is the y-intercept, which represents the value of y when x is 0.b is the slope of the line, indicating the change in y for one-unit change in x. It represents the strength and direction of the relationship between x and y.Formula for Regression CoefficientsThe formula for regression coefficients lies at the heart of linear regression analysis, a powerful statistical technique used to model the relationship between variables. At its core, linear regression seeks to find the best-fitting straight line that describes the relationship between a predictor variable (often denoted as X) and a response variable (often denoted as Y).In the formula for regression coefficients:a = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}b = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}Each term plays a crucial role in determining the slope (a) and intercept (b) of the best-fitted line:n: Represents the number of data points in the dataset. It ensures that the calculations are representative of the entire dataset. By computing a and b using these formulas, analysts can derive the equation of the best-fitted line: Y = aX + b. This equation enables predictions and insights into the relationship between the variables, empowering decision-making processes across various domains.Regression Coefficients InterpretationUnderstanding the nature of regression coefficients helps in predicting unknown variables... Regression coefficients provide key insights into these relationships.The interpretation of regression coefficients depends on their sign.A positive coefficient indicates a direct relationship between the variables, where an increase in the independent variable leads to an increase in the dependent variable.Conversely, a negative coefficient signifies an inverse relationship, where an increase in the independent variable results in a decrease in the dependent variable.Steps to Calculate the Regression CoefficientsBefore calculating regression coefficients for finding the best-fitted line, it is important to determine if the variables have a linear relationship. This can be done by interpreting the value and using correlation coefficient.Apply the formula a to determine the coefficient of X... To obtain the constant term the formula is... Calculate regression coefficients by entering them into the Y = aX + b equation and visually represent the regression line with a scatter plot.Regression Coefficients in Different Types of Regression ModelsRegression coefficients in different types of regression models:Linear Regression: It's like drawing a straight line through data points... Logistic Regression: This is used when we're dealing with yes/no or true/false outcomes, like predicting if an email is spam or not... Polynomial Regression: Instead of a straight line, this is like fitting a curve to our data... Ridge and Lasso Regression: These are methods to prevent our model from becoming too complicated and fitting the data too closely... Time Series Regression: When we're looking at data over time, like stock prices or temperature changes, we use this... Types of Regression Coefficients... Simple Regression Coefficient... Partial Regression Coefficient... Multiple Regression Coefficient... Positive regression coefficient... Negative regression coefficient... Linear regression coefficient... Non-Linear regression coefficient... Solved Examples on Regression Coefficients1. Find the regression coefficients for the following data:AgeGlucose Level... Linear regression is a statistical method that is used in various machine learning models to predict the value of unknown data using other related data values. Linear regression is used to study the relationship between a dependent variable and an independent variable.In this article, we will learn about, Linear Regression, Linear Regression Equation, Linear Equation Formulas, and others in detail.What is Linear Regression?Linear regression is a very common formula used in various machine learning models that perform a predictive analysis. In linear regression, we have two variables and they are considered as independent variable and dependent variable. In Linear Regression we assumes a linear relationship between the variables, which means that changes in the independent variables are associated with proportional changes in the dependent variable.Linear Regression FormulaVarious linear regression that are commonly used are,1) Simple Linear Regression: This is the simplest form, where we have one thing we're trying to predict and one thing we think might influence it. For example, We are perform a predictive analysis where are trying to predict someone's weight based on their height.2) Multiple Linear Regression: Here, things get a bit more complex. We're still predicting one thing, but now we're considering multiple factors that might influence it. For instance, we might predict a person's weight based on their height, age, and maybe even their diet habits.3) Logistic Regression: This one comes into play when we're dealing with binary outcomes, like whether someone will click on an ad or not. We're still looking at multiple factors that might play a role.4) Ordinal Regression: Sometimes, what we're trying to predict isn't exactly numerical, but it has an order. Think of rating something from 1 to 5 stars. This kind of regression helps us predict such ordinal outcomes.5) Multinomial Regression: When our outcome has several categories but no inherent order, like predicting someone's favorite color among several options, we turn to multinomial regression.6) Discriminant Analysis: Similar to multinomial regression, this helps us when we have multiple categories for our outcome variable, but here, we're specifically focused on classifying cases into those categories based on the predictor variables.Each of these methods has its own strengths and best-use scenarios.Linear Regression EquationLinear regression line equation is written in the form:y = a + bxwhere,x is an Independent Variable, Plotted along the X-axisy is the Dependent Variable, Plotted along the Y-axisThe slope of the regression line is "b", and the intercept value of the regression line is "a"(the value of y when x = 0).Linear Regression FormulaThe formula used for linear regressions is, y = a + bxIntercept value, a, and slope of the line, b, are evaluated using the formulas given below:\begin{array}{l}\large a~=~\frac{\sum y \sum x^{2} ~-~ \sum x \sum xy} {n(\sum x^{2}) ~-~ (\sum x)^{2}}\end{array} \\\begin{array}{l}\large b~=~\frac{n\sum xy~-~\left(\sum x\right)\left(\sum y\right)}{n\sum x^{2}~-~\left(\sum x\right)^{2}}\end{array}where,y is the Dependent Variable that Lies along the Y-axisa is Y-Interceptb is the Slope of the Regression Linex is an Independent Variable that Lies along the X-axisProperties of Linear RegressionIn the linear regression line, if the regression parameters a0 and a1 are defined, the properties are given below:Linear regression line reduces the sum of squared differences between observed values and predicted values.Linear regression line always passes through the mean of X and Y variable values.The linear regression constant (b0) is equal to the y-intercept of the linear regression.Linear regression coefficient (b0) is the slope of the regression line.Linear Regression Line Least square method is the most common method used to fit a regression line, in the X-Y graph. In this process, we determines the line of best fit by reducing the sum of the squares of the vertical deviations from each data point to the line.For any point that is fitted accurately, its perpendicular deviation is zero. The linear regression line is shown in the image added below,Regression CoefficientLinear regression line, equation:Y = B0 + B1Xwhere,B0 is a ConstantB1 is the Regression CoefficientHere, B1 is the regression coefficient and its formula is,B1 = b1 = Σ [(xi – x)(yi – y)] / Σ [(xi – x)2]where,xi and yi are Observed Data Setsx and y are Mean ValueWhat is Linear Regression Used for?Various uses of Linear Regression are,It is used in market research and the study of customer survey results.It is used for studying the performance of the engines of automobiles.It is used in deciding the effective price of any goods.It is used in astronomy.Error in Linear Regression FormulaStandard error about the regression line is defined as the measure of the average proportion that the regression equation predicts. Standard error in this case is denoted by 'SE'. Higher the coefficient of the determination involved, the lower the standard error and hence, a more accurate result is generated.Solved Example Questions on Linear RegressionQuestion 1: Find the linear regression equation for the given data:Solution:Calculating intercept and slope value.xyx2xy389249681545425203296∑x = 20∑y = 20∑x2 = 124∑xy = 104Using formula,\begin{array}{l}\large a~=~\frac{\sum y \sum x^{2} ~-~ \sum x \sum xy} {n(\sum x^{2}) ~-~ (\sum x)^{2}}\end{array}\\a = {20 (124) - 20 (104)} / {4 (124) - 400}a = 400/96 = 4.17\begin{array}{l}\large b~=~\frac{n\sum xy~-~\left(\sum x\right)\left(\sum y\right)}{n\sum x^{2}~-~\left(\sum x\right)^{2}}\end{array}b = {4 (104) - 20 (20)} / {4 (124) - 400}b = 16/96 = 0.166 So, linear regression equation is, y=a+bx → y = 4.17 + 0.167xQuestion 2: Find the linear regression equation for the given data:Solution:Calculating intercept and slope value.xyx2xy461624754935389241313∑x = 15∑y = 22∑x2 = 75∑xy = 86Using formula,\begin{array}{l}\large a~=~\frac{\sum y \sum x^{2} ~-~ \sum x \sum xy} {n(\sum x^{2}) ~-~ (\sum x)^{2}}\end{array}\\= (22 (75) - 15 (86)) / (4 (75) - 225)= 360/75= 4.8\begin{array}{l}\large b~=~\frac{n\sum xy~-~\left(\sum x\right)\left(\sum y\right)}{n\sum x^{2}~-~\left(\sum x\right)^{2}}\end{array}= (4 (86) - 15 (22)) / (4 (75) - 225)= 14/75= 0.1867So, the linear regression equation is, y = 4.8 + 0.187x.Question 3: Find the intercept of the linear regression line if ∑x = 25, ∑y = 20, ∑x2 = 90, ∑xy = 150, and n = 5.Solution:Using formula,\begin{array}{l}\large a=\frac{\sum y \sum x^{2} – \sum x \sum xy} {n(\sum x^{2}) – (\sum x)^{2}}\end{array}\\= (20 (90) - 25 (150)) / (5 (90) - 625)= -1950/-175= 11.14Question 4: Find the intercept of the linear regression line if ∑x = 30, ∑y = 27, ∑x2 = 110, ∑xy = 190 and n = 4.Solution:Using formula,\begin{array}{l}\large a=\frac{\sum y \sum x^{2} – \sum x \sum xy} {n(\sum x^{2}) – (\sum x)^{2}}\end{array}\\= (27 (110) - 30 (190)) / (4 (110) - 900)= -2730/-460= 5.93Question 5: Find slope of linear regression line if ∑x = 10, ∑y = 16, ∑x2 = 60, ∑xy = 120 and n = 4.Solution:Using formula,\begin{array}{l}\large b=\frac{n\sum xy-\left(\sum x\right)\left(\sum y\right)}{n\sum x^{2}-\left(\sum x\right)^{2}}\end{array}= (4 (120) - 10 (16)) / (4 (60) - 100)= 320/140= 2.29Question 6: Find slope of linear regression line if ∑x = 40, ∑y = 32, ∑x2 = 130, ∑xy = 210 and n = 4.Solution:Using formula,\begin{array}{l}\large b=\frac{n\sum xy-\left(\sum x\right)\left(\sum y\right)}{n\sum x^{2}-\left(\sum x\right)^{2}}\end{array}= (4 (210) - 40 (32)) / (4 (130) - 1600)= -440/-1080= 0.407Question 7: Find slope of linear regression line if ∑x = 50, ∑y = 44, ∑x2 = 150, ∑xy = 230 and n = 4.Solution:Using formula,\begin{array}{l}\large a=\frac{\sum y \sum x^{2} – \sum x \sum xy} {n(\sum x^{2}) – (\sum x)^{2}}\end{array}\\= (44 (150) - 50 (230)) / (4 (150) - 2500)= -4900/-1900= 2.57\begin{array}{l}\large b=\frac{n\sum xy-\left(\sum x\right)\left(\sum y\right)}{n\sum x^{2}-\left(\sum x\right)^{2}}\end{array}= (4 (230) - 50 (44)) / (4 (150) - 2500)= -1280/-1900= 0.673Conclusion Linear regression is an essential and widely used statistical method in predictive modeling and data analysis. By leveraging the linear regression formula and understanding its components such as the slope, intercept, and regression coefficients, we can effectively model the relationship between independent and dependent variables. Mastering the linear regression formula will provide us with the ability to analyze data trends, forecast outcomes, and derive meaningful insights, enhancing our decision-making capabilities in various applications including market research, financial analysis, sports analytics, and more. Basic Math Formulas What is the Division Formula? LCM Formula | Examples & Practice Questions Distributive Property | Definition and Examples Exponents Consecutive Integers Scientific Notation Formula Binary Formula Convert Binary fraction to Decimal Fibonacci Sequence Formula Direct Variation: Definition, Formula and Examples What is Celsius Formula? Fahrenheit to Celsius (Â°F to Â°C) | Formula , Conversion and Examples Revenue Formula Selling Price Formula Simple Interest Compound Interest Monthly Compound Interest Formula Daily Compound Interest Formula with Examples Double Time Formula Perpendicular Lines Right Angle What is Parallel Lines Formula? Angles Formula Degrees to Radian Converter Area of 2D Shapes Area of Quadrilateral Area of Square What is the Diameter Formula? Arc Length Formula Central Angle of Circle Formula with Solved Examples Asymptote Formula Axis of Symmetry of a Parabola Centroid of a Trapezoid Formula Area of a Circle: Formula, Derivation, Examples Parallelogram Formulas Perimeter Formulas for Geometric Shapes Perimeter of Triangle Equilateral Triangle Scalene Triangle: Definition, Properties, Formula, Examples Right Angled Triangle | Properties and Formula Perimeter of Rectangle What is the Formula for Perimeter of a Square? Circumference Formula Perimeter of a Parallelogram Rhombus Formula Perimeter of Rhombus Formula Diagonal Formula Diagonal of a Polygon Formula Diagonal of a Square Formula Diagonal of Parallelogram Formula Diagonal of a Cube Formula Euclid Euler Theorem What is Side Angle Side Formula? Polygon Formula - Definition, Symbol, Examples Annulus Area Formula Volume Formulas for 3D Shapes Volume of a Cube Volume of a Cylinder| Formula, Definition and Examples Volume of Cone- Formula, Derivation and Examples Volume of a Sphere Surface Area Formulas Surface Area of a Cone Surface Area of Sphere | Formula, Derivation and Solved Examples Surface Area of a Square Pyramid Volume of a Pyramid Formula Frustum of Cone Volume of a Square Pyramid Formula Surface Area of a Prism Frustum of a Regular Pyramid Formula Basic Math Formulas Algebra Formulas - List of all Algebra Formulas Polynomial Formula Factorization of Polynomial What is Factoring Trinomials Formula? a2 - b2 Formula Difference of Cubes Discriminant Formula in Quadratic Equations Sum of Arithmetic Sequence Formula Function Notation Formula Binomial Distribution in Probability Binomial Expansion Formulas Binomial Theorem FOIL Method Exponential Decay Formula Factorial Formula Combinations Formula with Examples Fourier Series Formula Maclaurin series 30-60-90 Formula Cofunction Formulas What is Cos Square theta Formula? What are Cosine Formulas? Cosecant Formula Cotangent Formula Tangent Formulas Cot Half Angle Formula 2cosA cosB Formula Multiple Angle Formulas Double Angle Formula for Cosine Inverse Trigonometric Functions | Definition, Formula, Types and Examples Limit Formula Average and Instantaneous Rate of Change Calculus | Differential and Integral Calculus Total Derivative Difference Quotient Formula Chain Rule: Theorem, Formula and Solved Examples Implicit Differentiation Antiderivative: Integration as Inverse Process of Differentiation Integration Formulas Integration by Parts Integration by Substitution Formula Definite Integral | Definition, Formula & How to Calculate Area Under Curve Differentiation and Integration Formula Differential Equations In statistics, regression analysis is a technique that can be used to analyze the relationship between predictor variables and a response variable. When you use software (like R, Stata, SPSS, etc.) to perform a regression analysis, you will receive a regression table as output that summarize the results of the regression. Arguably the most important numbers in the output of the regression table are the regression coefficients. Yet, despite their importance, many people have a hard time correctly interpreting these numbers. This tutorial walks through an example of a regression analysis and provides an in-depth explanation of how to interpret the regression coefficients that result from the regression. Related: How to Read and Interpret an Entire Regression Table A Regression Analysis Example Suppose we are interested in running a regression analysis using the following variables: Predictor Variables Total number of hours studied (continuous variable – between 0 and 20) Whether or not a student used a tutor (categorical variable – "yes" or "no") Response Variable Exam score (continuous variable – between 1 and 100) We are interested in examining the relationship between the predictor variables and the response variable to find out if hours studied and whether or not a student used a tutor actually have a meaningful impact on their exam score. Suppose we run a regression analysis and get the following output: Term Coefficient Standard Error t Stat P-value Intercept 48.56 14.32 3.39 0.002 Hours studied 2.03 0.67 3.03 0.009 Tutor 8.34 5.68 1.47 0.138 Let's take a look at how to interpret each regression coefficient. Interpreting the Intercept The intercept term in a regression table tells us the average expected value for the response variable when all of the predictor variables are equal to zero. In this example, the regression coefficient for the intercept is equal to 48.56. This means that for a student who studied for zero hours (Hours studied = 0) and did not use a tutor (Tutor = 0), the average expected exam score is 48.56. It's important to note that the regression coefficient for the intercept is only meaningful if it's reasonable that all of the predictor variables in the model can actually be equal to zero. In this example, it's certainly possible for a student to have studied for zero hours (Hours studied = 0) and to have also not used a tutor (Tutor = 0). Thus, the interpretation for the regression coefficient of the intercept is meaningful in this example. In some cases, though, the regression coefficient for the intercept is not meaningful. For example, suppose we ran a regression analysis using square footage as a predictor variable and house value as a response variable. In the output regression table, the regression coefficient for the intercept term would not have a meaningful interpretation since square footage of a house can never actually be equal to zero. In that case, the regression coefficient for the intercept term simply anchors the regression line in the right place. Interpreting the Coefficient of a Continuous Predictor Variable For a continuous predictor variable, the regression coefficient represents the difference in the predicted value of the response variable for each one-unit change in the predictor variable, assuming all other predictor variables are held constant. In this example, Hours studied is a continuous predictor variable that ranges from 0 to 20 hours. In some cases, a student studied as few as zero hours and in other cases a student studied as much as 20 hours. From the regression output, we can see that the regression coefficient for Hours studied is 2.03. This means that, on average, each additional hour studied is associated with an increase of 2.03 points on the final exam, assuming the predictor variable Tutor is held constant. For example, consider student A who studies for 10 hours and uses a tutor. Also consider student B who studies for 11 hours and also uses a tutor. According to our regression output, student B is expected to receive an exam score that is 2.03 points higher than student A. The p-value from the regression table tells us whether or not this regression coefficient is actually statistically significant. We can see that the p-value for Hours studied is 0.009, which is statistically significant at an alpha level of 0.05. Note: The alpha level should be chosen before the regression analysis is conducted – common choices for the alpha level are 0.01, 0.05, and 0.10. Related post: An Explanation of P-Values and Statistical Significance Interpreting the Coefficient of a Categorical Predictor Variable For a categorical predictor variable, the regression coefficient represents the difference in the predicted value of the response variable between the category for which the predictor variable = 0 and the category for which the predictor variable = 1. In this example, Tutor is a categorical predictor variable that can take on two different values: 1 = the student used a tutor to prepare for the exam 0 = the student did not used a tutor to prepare for the exam From the regression output, we can see that the regression coefficient for Tutor is 8.34. This means that, on average, a student who used a tutor scored 8.34 points higher on the exam compared to a student who did not used a tutor, assuming the predictor variable Hours studied is held constant. For example, consider student A who studies for 10 hours and uses a tutor. Also consider student B who studies for 10 hours and does not use a tutor. According to our regression output, student A is expected to receive an exam score that is 8.34 points higher than student B. The p-value from the regression table tells us whether or not this regression coefficient is actually statistically significant. We can see that the p-value for Tutor is 0.138, which is not statistically significant at an alpha level of 0.05. This indicates that although students who used a tutor scored higher on the exam, this difference could have been due to random chance. Interpreting All of the Coefficients At Once We can use all of the coefficients in the regression table to create the following estimated regression equation: Expected exam score = 48.56 + 2.03*(Hours studied) + 8.34*(Tutor) Note: Keep in mind that the predictor variable "Tutor" was not statistically significant at alpha level 0.05, so you may choose to remove this predictor from the model and not use it in the final estimated regression equation. Using this estimated regression equation, we can predict the final exam score of a student based on their total hours studied and whether or not they used a tutor. For example, a student who studied for 10 hours and used a tutor is expected to receive an exam score of: Expected exam score = 48.56 + 2.03*(10) + 8.34*(1) = 77.2 Considering Correlation When Interpreting Regression Coefficients It's important to keep in mind that predictor variables can influence each other in a regression model. For example, most predictor variables will be at least somewhat related to one another (e.g. perhaps a student who studies more is also more likely to use a tutor). This means that regression coefficients will change when different predict variables are added or removed from the model. One good way to see whether or not the correlation between predictor variables is severe enough to influence the regression model in a serious way is to check the VIF between the predictor variables. This will tell you whether or not the correlation between predictor variables is a problem that should be addressed before you decide to interpret the regression coefficients. If you are running a simple linear regression model with only one predictor, then correlated predictor variables will not be a problem.