Click to verify

A fundamental component of statistical technique, regression analysis is essential for examining and measuring connections between variables. Its uses are numerous and diverse, from forecasting financial trends to evaluating medical results. This in-depth manual explores the essence of regression analysis, explaining its various kinds, applications, and underlying concepts. Discover the various regression techniques, their uses, and the underlying mathematics. Acquire knowledge of fundamental ideas, including the regression equation, coefficient evaluation, and fit metrics quality. Examine the fundamental presumptions of regression analysis and their significance for trustworthy outcomes. Recognize the many ways that regression analysis may be used in various contexts. Analyze the benefits and drawbacks of regression analysis, considering its diagnostic capabilities, quantification of correlations, ability to account for confounding factors, predictive strength, and limits. Table of contents Regression analysis is a reliable statistical method for ascertaining the relationship between a dependent variable and one or more independent variables. It clarifies how changes in the independent components impact the dependent variable, making it a basic idea in finance, economics, and the social sciences. Simple Linear Regression: As simple linear regression shows, a line that passes through the displayed data points represents the association between one predictor variable and one responder variable. The objective is to determine sales using, for example, the amount of money spent on advertising or to approximate the level of the dependent variable numerically such that it corresponds to the level of the independent variable. Multiple Linear Regression: Multiple linear regression incorporates two or more independent variables to predict a single dependent variable, extending the capabilities of basic linear regression. This method estimates property values based on size, location, and age and reflects the cumulative influence of several factors on the dependent variable. Logistic Regression: Logistic regression is used when the dependent variable is categorical or binary (e.g., true or false, yes/no). Instead of fitting a straight line to forecast the likelihood of a specific result, it utilizes a logistic function (sigmoid curve). For example, it can predict if a consumer will make a purchase (yes or no). Polynomial Regression: Polynomial regression uses an nth-degree polynomial to express the relationship between the independent and dependent variables. By changing the predictors, it can now fit more intricate, nonlinear connections. Also Read: 7 Regression Techniques You Should Know! The fundamental idea is to fit a mathematical equation to observed data. In simple linear regression, the equation is: The coefficients represent the intercept and slope. They show that the amount of y varies when x increases by one unit. Every independent variable in multiple regression has a coefficient representing its influence on the dependent variable. R-squared (R^2): The percentage of the dependent variable's volatility can be predicted based on the independent variables. Higher R^2 values suggest a better match. Adjusted R-squared: This approach provides a more accurate estimate in multiple regression settings by adjusting R^2 for the number of predictors in the model. P-values: Evaluate the coefficients' significance. Low P-values, usually less than 0.05, indicate that the association is statistically significant. Linearity: The relationship between dependent and independent variables should be linear. Independence: Observations should be independent of each other. Homoscedasticity: The variance of errors should be consistent across all levels of the independent variables. Normality: One should normally distribute the errors. Business and Economics: Regression analysis helps businesses forecast sales, adapt prices, and search for market indicators. It is also used to understand economic factors like GDP and unemployment. Finance: They enable evaluation of the threats facing a certain investment and portfolio management by demonstrating dependency between asset prices and other variables, such as interest rates or profits. Healthcare: It employs information on patients' clinical and demographic data to identify factors associated with ill health. It also evaluates the effectiveness of therapeutic interventions and predicts patient outcomes. Marketing: Regression analysis is a method marketers use to predict sales, evaluate advertising campaigns, and analyze consumer behavior. Social Sciences: Sociologists and psychologists use regression analysis to comprehend the relationship between variables and results, such as education and income patterns. Predictive Power: This research's data were analyzed through regression analysis to arrive at the findings related to future results. When the dependency between the variables is understood, future prospects, sales, and other factors can be calculated in specific detail. Quantification of Relationships: It offers a precise mathematical framework for calculating the direction and intensity of correlations between different variables. This aids in comprehending how modifications to one variable impact those to another. Control for Confounding Variables: Multiple regression can include several independent variables, which aids in determining one variable's influence while accounting for others. This is especially helpful in challenging real-world situations. Diagnostic Tools: Regression analysis helps with model validation and improvement by offering diagnostic tools (such as R-squared, p-values, and residual plots) to evaluate the model's fit and the importance of predictors. Versatility: Regression analysis works with various data kinds and scenarios, including continuous, categorical, and binary outcomes. Moreover, it applies to multiple professions, including economics, engineering, and social sciences. Ease of Implementation: Thanks to modern statistical software and tools, regression analysis is now easier to apply, even for those without extensive statistical knowledge. Process simplification is achieved using automated tools in Python, R, and other platforms. Hypothesis Testing: Regression analysis aids in testing theories about the correlations between variables. It offers a structure for determining whether specific predictors significantly impact the dependent variable. Assumption Dependencies: Several presumptions, including linearity, independence, homoscedasticity, and error normalcy, underpin regression models. Breaking these presumptions may lead to erroneous or deceptive results. Multicollinearity: Many independent variables may affect outcomes, making it difficult to determine the influence of specific predictors. Overfitting: When trained on training data, a model overfits the training data and performs much worse predicting new data. This happens when the model collects noise in addition to the signal. Sensitivity to Outliers: Outliers can significantly alter the model's coefficients and outcomes in regression analysis. Limited by Linear Relationships: The assumption of a linear connection between variables in simple linear regression may not hold in all cases. Researchers need advanced methods like polynomial regression or machine learning models for more complicated interactions. Interpretability Issues: It can be challenging to determine how each predictor affects the results of a model with many predictors, particularly in multiple regression. This difficulty increases if there are interactions between the variables. Sample Size Requirements: Regression analysis requires a substantial sample size to yield accurate estimations. Tiny sample sizes may result in unstable estimates and inadequate generalization. As a fundamental tool for data analysis, regression analysis continues to provide insights and predictive power for a wide range of applications. However, reliability depends on paying close attention to assumptions, model selection, and validation, even if it offers valuable tools for forecasting and relationship comprehension. Q1. What is a regression analysis in simple terms? A. Regression analysis is a statistical method used to understand the relationship between one dependent variable and one or more independent variables. Q2. What does a regression analysis tell you? A. It tells you how changes in the independent variables are associated with changes in the dependent variable, helping to predict or explain the dependent variable. Q3. What is the main purpose of regression analysis? A. The main purpose is to model the relationship between variables, allowing for predictions, insights into causal relationships, and understanding the strength of these relationships. Q4. What is an example of a regression analysis? A. An example is predicting a person's salary (dependent variable) based on their years of experience and education level (independent variables).

ALibaba.comAmazon.comAppleInsider.comBenzingaBusinessInsiderCNBlogsBitBucketCoinBaseCornellCrunchBaseDecryptDigitalJournalForbesEntreprenerGeeksForGeeksGithubGitlabGoogleHackermoonHubspotHuschBlackwellInvestingInvestopediaJefferiesJhuMitNasdaqNewsweekNYTimesShopifyStackExchangeTechCrunchUmichUMNWeirdWSPALibaba.comAmazon.comAppleInsider.comBenzingaBusinessInsiderCNBlogsBitBucketCoinBaseCornellCrunchBaseDecryptDigitalJournalForbesEntreprenerGeeksForGeeksGithubGitlabGoogleHackermoonHubspotH
Share — copy and redistribute the material in any medium or format for any purpose, even commercially. Adapt — remix, transform, and build upon the material for any purpose, even commercially. The licensor cannot revoke these freedoms as long as you follow the license terms. Attribution — You must give appropriate credit , provide a link to the license, and indicate if changes were made . You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original. No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits. You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable exception or limitation . No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as publicity, privacy, or moral rights may limit how you use the material. Jobs Companies Articles Tracker uptonpark/iStock/Getty ImagesYou probably know by now that whenever possible you should be making data-driven decisions at work. But do you know how to parse through all the data available to you? The good news is that you probably don't need to do the number crunching yourself (hallelujah!) but you do need to correctly understand and interpret the analysis created by your colleagues. One of the most important types of data analysis is called regression analysis. Regression analysis is a statistical technique used for modeling the relationship between the dependent variable and one or more independent variables, enabling prediction, decision-making, and insights across various fields. Highlights Regression analysis models relationships between dependent and independent variables for prediction and decision-making. Linear, logistic, and polynomial are key types of regression, each suited to different data and goals. Goodness-of-fit metrics, like R-squared and adjusted R-squared, assess model performance and explainability. Regression assumptions include linearity, normality, independence of errors, and homoscedasticity, which must be validated. Common pitfalls include confusing correlation with causation, overfitting, multicollinearity, omitted variable bias, and extrapolation. Ad description. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Regression analysis is a cornerstone technique in statistics and data science that allows us to explore and quantify the relationships between variables. It is used to predict outcomes, identify trends, and make data-driven decisions across various fields, from business and finance to healthcare and engineering. At its core, regression analysis seeks to model the relationship between a dependent variable (the variable we are trying to predict or explain) and one or more independent variables (the factors that influence the dependent variable). By doing so, we can gain insights into our data's underlying patterns and causal relationships, enabling us to make better predictions and more informed decisions. The fundamental idea behind regression analysis is to find the best-fitting model that accurately represents the relationship between the dependent and independent variables. This often involves fitting a line or curve to the data points to minimize the differences between the observed and predicted values, known as the residuals. In practice, regression analysis can take many forms, from simple linear regression, which models the relationship between one dependent and one independent variable, to more advanced techniques, such as multiple polynomial regression, which allow for the analysis of more complex relationships. As a versatile and powerful statistical tool, regression analysis is essential for anyone looking to make sense of data and draw meaningful conclusions. In this guide, you will learn about the various types of regression analysis, their fundamental concepts, key assumptions and limitations, and practical applications across multiple industries. By understanding what's regression analysis and mastering its techniques, you will be well-equipped to tackle complex data challenges and make data-driven decisions with confidence. Linear Regression: This is the most basic form of regression analysis to model the relationship between the dependent and independent variables. It assumes a linear relationship between variables, represented by a straight line. The objective is to find the best-fitting line that minimizes the squared differences sum between the observed and predicted values. Linear regression is widely used for forecasting, trend analysis, and identifying the impact of one variable on another. Logistic Regression: This is a type of regression analysis designed for problems with categorical dependent variables, including binary (two possible outcomes), nominal (unordered categories), and ordinal (ordered categories). Rather than predicting the actual value of the dependent variable, logistic regression estimates the probability of an event occurring based on the independent variable. It employs the logistic function, which transforms the regression model output into a probability value between 0 and 1. Commonly used in fields such as medical research, marketing, and finance, logistic regression is versatile for predicting the likelihood of specific events or outcomes and understanding the impact of various factors on those outcomes. Polynomial Regression: This is an extension of the linear regression used when the relationship between the dependent and independent variables is nonlinear. Instead of fitting a straight line, a polynomial function is used to model the curvature in the relationship. Polynomial regression can capture more complex patterns in the data, providing a better fit for situations where linear models are inadequate. However, it is essential to be cautious with the choice of polynomial degrees, as overly complex models can lead to overfitting and poor generalization of new data. *Multiple Regression: Multiple regression is an advanced technique that allows for the inclusion of multiple independent variables, enabling the analysis of more complex relationships between variables and controlling for confounding factors. It is not only a generalization of linear regression. Still, it can also be extended to other types of regression, such as logistic and polynomial regression. Multiple regression aims to find the best-fitting model explaining the relationship between dependent and independent variables. Dependent and Independent Variables: In regression analysis, the dependent variable (the response or target variable) is the outcome we try to predict or explain. At the same time, independent variables (also known as predictors, features, or explanatory variables) are the factors that influence the dependent variable. Regression analysis aims to model the relationship between these variables, allowing us to understand the effects of the independent variables on the dependent variable and make predictions. Coefficients and Intercepts: Coefficients and intercepts are essential components of a regression model. Coefficients represent the effect of the independent variables on the dependent variable, indicating how much the dependent variable changes for each unit increase in the independent variables, assuming all other variables are held constant. The intercept represents the expected value of the dependent variable when all independent variables are equal to zero. These values are estimated using ordinary least squares or maximum likelihood estimation techniques during the model fitting. The Goodness of Fit and R-squared: The goodness of fit measures how well the regression model fits the observed data. Several metrics can be used to assess the goodness of fit. Still, one of the most common is R-squared (also known as the coefficient of determination). R-squared represents the variance proportion in the dependent variable explained by the model's independent variables. It ranges between 0 and 1, with higher values indicating a better fit. In multiple regression, it is essential to consider the adjusted R-squared, which accounts for the number of independent variables in the model. The R-squared value can increase as more variables are added, even if the additional variables do not contribute significantly to the model's predictive power. The adjusted R-squared corrects this issue by penalizing the R-squared value for the inclusion of unnecessary variables. This results in a more accurate model's performance assessment and helps to prevent overfitting. However, both R-squared and adjusted R-squared should be interpreted cautiously, as a high value does not guarantee a good model. The model may still be susceptible to issues such as multicollinearity, omitted variable bias, or other violations of regression assumptions. Ad description. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Regression analysis relies on several key assumptions common across different types of regression, including linear, logistic, polynomial, and multiple regression extensions. If these assumptions are not met, the accuracy and validity of the model's predictions and conclusions may be limited. Therefore, it is crucial to understand and assess these assumptions when performing regression analysis. Appropriate functional form: The relationship between the dependent and independent variables should follow the functional form of the regression model. This means a linear relationship for linear regression, while a polynomial relationship is assumed for polynomial regression. Checking scatterplots or residual plots can help identify violations of this assumption. Independence of observations: The observations in a dataset should be independent of each other. In autocorrelation (when consecutive observations are related), the regression model may underestimate the true variability in the data, leading to unreliable coefficient estimates and hypothesis tests. Independence of errors: The residuals (errors) should be independent, meaning that the error for one observation should not influence the error for another. If errors are correlated, the standard errors of the coefficients may be underestimated, leading to incorrect inferences and confidence intervals. Homoscedasticity (for linear and polynomial regression): The residual variance should be constant across all levels of the independent variables. If the variance changes (a phenomenon called heteroscedasticity), the standard errors of the coefficients may be biased, affecting hypothesis testing and confidence intervals. Normality of residuals (for linear and polynomial regression): The residuals (i.e., the differences between the observed and predicted values) should follow a normal distribution. Violations of normality can affect hypothesis testing and the validity of confidence intervals. No multicollinearity: Independent variables should not be highly correlated with one another. Multicollinearity can cause unstable coefficient estimates and make it difficult to interpret the individual contributions of each variable. Correctly specified link function (for logistic regression): The link function must be correctly defined in logistic regression to transform the predicted values into probabilities. Using an incorrect link function can lead to biased coefficient estimates and inaccurate predictions. Understanding and interpreting regression analysis results is crucial for making better decisions and drawing meaningful conclusions. Here are some key aspects to consider when interpreting the results of any regression model, including linear, logistic, polynomial, and multiple regression: Coefficient estimates: The coefficients represent the effect of each independent variable on the dependent variable, holding all other variables constant. In linear and polynomial regression, the coefficients indicate the change in the dependent variable for a unit increase in the independent variable. In logistic regression, the coefficients represent the change in the log-odds of the outcome for a unit increase in the independent variable. Significance of coefficients: Hypothesis tests, such as t-tests or z-tests, are performed to determine the statistical significance of the coefficients. A statistically significant coefficient suggests that the independent variable has a meaningful impact on the dependent variable. A non-significant coefficient implies that the independent variable may not contribute significantly to the model. Confidence intervals: Confidence intervals estimate the range within which the population coefficient will likely fall. Narrower intervals suggest more precise estimates, while wider intervals indicate greater uncertainty. Model fit statistics: Goodness-of-fit metrics, such as R-squared, adjusted R-squared, or the Akaike Information Criterion (AIC), can help assess the overall performance of the model. These metrics should be considered alongside other diagnostic measures and plots to evaluate the model's adequacy. Residual analysis: Examining the residuals can reveal patterns or trends that suggest violations of regression assumptions or areas where the model does not fit the data well. Residual plots, normal probability plots, and autocorrelation plots can be used to diagnose potential issues and guide model improvement. Outliers and influential points: Outliers and influential points can significantly impact the regression model. Identifying and addressing these observations by excluding them or using robust regression techniques can help improve the model's performance. Validation and generalization: Cross-validation or hold-out validation can be used to assess the model's performance on new data, helping to gauge its generalizability and prevent overfitting. Ad description. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Regression analysis is a powerful statistical tool with many practical applications across diverse industries. The regression analysis can help inform decision-making, optimize processes, and predict future outcomes by modeling the relationship between the dependent and independent variables. Here are some examples of how regression analysis is used in various industries: Finance and Economics: In finance, regression analysis can be used to model the relationship between stock prices and economic indicators, such as interest rates or unemployment rates. This can help investors make better decisions about portfolio allocation and risk management. Additionally, economists may use regression to study the monetary and fiscal policies' impact on economic growth and inflation. Marketing and Sales: Regression models can be employed to analyze the effectiveness of marketing campaigns, understand consumer behavior, and predict sales. For example, businesses can use regression to determine the impact of advertising spending on sales, allowing them to optimize their marketing budget for maximum return on investment. Healthcare: Regression analysis explores the relationship between patient characteristics and health outcomes in the healthcare industry. This can help identify disease risk factors, inform treatment decisions, and optimize patient care. For instance, logistic regression can predict the likelihood of a patient developing a particular condition based on demographic and clinical variables. Manufacturing and Quality Control: Regression analysis can optimize manufacturing processes, improve product quality, and reduce production costs. By modeling the relationship between process variables and product characteristics, companies can identify the optimal conditions for achieving desired product specifications while minimizing waste and resource consumption. Human Resources: In HR management, regression analysis can be used to understand the factors influencing employee performance, retention, and job satisfaction. This can help organizations develop targeted recruitment, training, and employee engagement strategies, ultimately improving productivity and reducing turnover. Sports Analytics: Regression models are increasingly used in sports analytics to evaluate player performance, inform coaching decisions, and optimize team strategies. For example, multiple regression analysis can quantify individual player statistics' contribution to team success, helping coaches and managers make more informed roster decisions. Environmental Science: In ecological research, regression analysis can be employed to model the relationship between environmental factors, such as temperature or precipitation, and ecological outcomes, such as species distribution or ecosystem productivity. This can inform conservation efforts, natural resource management, and policy development. While regression analysis is a powerful and widely-used statistical tool, it has challenges and potential pitfalls. Being aware of these common misconceptions and issues can help practitioners avoid mistakes and improve the quality of their analysis: Correlation vs. causation: A common misconception in regression analysis is that correlation implies causation. Although regression can identify relationships between variables, it does not necessarily prove a causal link. Establishing causality requires a deeper understanding of the underlying mechanisms and often involves experimental designs or additional analyses. Overfitting: Building a model that is too complex or includes too many independent variables can lead to overfitting, where the model captures noise in the data rather than the underlying relationships. Overfit models perform poorly on new data and can lead to misleading conclusions. To prevent overfitting, consider using cross-validation, regularization, or model selection based on information criteria like AIC or BIC. Multicollinearity: When independent variables are highly correlated, it becomes difficult to interpret the individual contribution of each variable to the model. Multicollinearity can lead to unstable estimates and inflated standard errors. Detecting multicollinearity through variance inflation factors (VIFs) or correlation matrices, and addressing it through techniques like variable selection or dimensionality reduction, can help improve model interpretation and performance. Omitted variable bias: Excluding important variables from the regression model can result in biased coefficient estimates and misleading conclusions. To avoid omitted variable bias, ensure that all relevant variables are included in the analysis. Consider using stepwise regression or model selection techniques to identify the most important predictors. Violation of assumptions: Ignoring or failing to test the underlying assumptions of regression analysis can lead to unreliable results. It is essential to assess the validity of the assumptions, such as linearity, independence of errors, and homoscedasticity, and to employ alternative techniques or transformations if necessary. Extrapolation: Using regression models to make predictions beyond the range of the observed data can be risky, as the relationships between variables may not hold in unobserved regions. Exercise caution when extrapolating predictions, and consider the model's limitations and the potential for unforeseen factors to influence the outcome. Misinterpreting coefficients: Interpreting regression coefficients without considering the scale of the variables or the link function (in the case of logistic regression) can lead to confusion and incorrect conclusions. Ensure that the interpretation of coefficients is context-appropriate and considers the units of measurement, the direction of the effect, and the magnitude of the relationship. Ad description. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Regression analysis is a powerful and versatile statistical tool enabling practitioners to model the relationships between variables, make predictions, and inform decision-making across various industries. By understanding the fundamental concepts, such as dependent and independent variables, coefficients, and goodness of fit, analysts can select the appropriate type of regression model for their data, including linear, logistic, polynomial, and multiple regression. However, it is crucial to be aware of regression analysis's assumptions and limitations and carefully interpret the results to avoid common pitfalls and misconceptions. By considering the assumptions, addressing issues such as multicollinearity and overfitting, and using techniques like cross-validation and regularization, practitioners can build more accurate and generalizable models that contribute valuable insights and drive data-driven decision-making. In summary, regression analysis is an invaluable tool for understanding complex relationships and solving real-world problems, as long as practitioners are mindful of its potential challenges and limitations and committed to using the method responsibly and effectively. Recommended Articles Ready to expand your knowledge? Check out our blog for more relevant articles and deepen your understanding of these crucial statistical tests. Take advantage of the opportunity to enhance your data analysis skills and drive more informed decision-making. Read now! Q1: What's Regression Analysis? Regression analysis is a statistical approach to model relationships between dependent and independent variables for prediction and decision-making. Q2: What are the different types of regression analysis? The main types are linear, logistic, polynomial, and multiple regression, each suited to different data types and goals. Q3: What's the difference between correlation and causation? Correlation measures the strength and direction of a relationship between variables. At the same time, causation implies that one variable directly influences the other. Q4: What are the key assumptions of regression analysis? Assumptions include linearity, normality, independence of errors, and homoscedasticity, which must be validated for reliable results. Q5: What is multicollinearity, and how can it be addressed? When independent variables are highly correlated, multicollinearity occurs, making it difficult to interpret individual contributions. It can be addressed using variable selection or dimensionality reduction techniques. Q6: What is overfitting, and how can it be prevented? Overfitting occurs when a model captures noise rather than underlying relationships, resulting in poor generalizability. It can be prevented using cross-validation, regularization, or model selection. Q7: How do you interpret regression coefficients? Coefficients represent the effect of independent variables on the dependent variable while holding other variables constant. Interpretation depends on the type of regression and variable scales. Q8: What is residual analysis? The residual analysis examines residuals (differences between observed and predicted values) to identify patterns, trends, or violations of regression assumptions, guiding model improvement. Q9: In which industries are regression analysis commonly used? Regression analysis is widely used in finance, marketing, healthcare, manufacturing, HR, sports analytics, and environmental science. Q10: What are some common pitfalls in regression analysis? Common pitfalls include confusing correlation with causation, overfitting, multicollinearity, omitted variable bias, extrapolation, and misinterpreting coefficients. Regression analysis is a statistical technique for studying linear relationships. [1] It begins by supposing a general form for the relationship, known as the regression model: $Y = \alpha + \beta 1X1 + \ldots + \beta kXk + \varepsilon$. Example: In the motorpool case, the manager of the motorpool considers the model Cost = $\alpha + \beta 1$Mileage + $\beta 2$Age + $\beta 3$Make + ε . Y is the dependent variable, representing a quantity that varies from individual to individual throughout the population, and is the primary focus of interest. X1, ... Xk are the explanatory variables (the so-called "independent variables"), which also vary from one individual to the next, and are thought to be related to Y. Finally, ε is the residual term, which represents the composite effect of all other types of individual differences not explicitly identified in the model. [2] Beside the model, the other input into a regression analysis is some relevant sample data, consisting of the observed values of the dependent and explanatory variables for a sample of members of the population. The primary result of a regression analysis is a set of estimates of the regression coefficients α, $\beta 1$,..., βk. These estimates are made by finding values for the coefficients that make the average residual 0, and the standard deviation of the residual term as small as possible. The result is summarized in the prediction equation: Ypred = a + b1X1 +...+ bkXk . Example: Fitting the model above to the motorpool data, we obtain: Costpred = 107.34 + 29.65 Mileage + 73.96 Age + 47.43 Make . (Dive down for further discussion of the assumptions underlying regression analysis, or examine a workbook which illustrates some of the underlying computations.) Why do a regression analysis? Typically, a regression analysis is done for one of two purposes: In order to predict the value of the dependent variable for individuals for whom some information concerning the explanatory variables is available, or in order to estimate the effect of some explanatory variable on the dependent variable. Making individual predictions If we know the value of several explanatory variables for an individual, but do not know the value of that individual's dependent variable, we can use the prediction equation (based on a model using the known variables as its explanatory variables) to estimate the value of the dependent variable for that individual. In order to see how much our prediction can be trusted, we use the standard error of the prediction [3] to construct confidence intervals for the prediction. (Examine a workbook that provides a detailed discussion of the standard error of the prediction.) Example: In order to predict the next twelve-month's maintenance and repair expenses for a specific one-year-old Ford currently in the motorpool, we'd first perform a regression analysis using age and make as the explanatory variables: Costpred = 705.66 + 8.53 Age - 54.27 Make . Our prediction will then be $714.19, and the margin of error (at the 95%-confidence level) for the prediction is 2.1788 × 124.0141 = $270.20 . If our goal is not to make a prediction for an individual, but rather to estimate the mean value of the dependent variable across a large pool of similar individuals, we use the standard error of the estimated mean instead when computing confidence intervals. Example: Our estimate of the average cost of keeping one-year-old Fords working is $714.19, with a margin of error of 2.1788 × 41.573 = $90.58 . Estimating the effect of an explanatory variable on the dependent variable In order to estimate the "pure" effect of some explanatory variable on the dependent variable, we want to control for as many other effects as possible. That is, we'd like to see how our prediction would change for an individual if this explanatory variable were different, while all others aspects of the individual were kept the same. In order to do this, we should always use the most complete model available, i.e., we should include all other relevant factors as additional explanatory variables. (Dive down for further discussion.) Our estimate of impact of a unit difference in the targeted explanatory variable is its coefficient in the prediction equation. The extent to which our estimate can be trusted is measured by the standard error of the coefficient. Example: Using the full regression model, we estimate that the mean marginal maintenance and repair cost associated with driving one of the cars in the motorpool an additional 1000 miles is $29.65, with a margin of error in the estimate of 2.2010 × 3.915 = $8.62 . To better understand why we use the most complete model available, note that any "one of the cars" has a particular age and make, and we want to hold those constant while considering the incremental effect of another 1000 miles of driving. Determining whether there is evidence that an explanatory variable belongs in a regression model Given a specific model, one might wonder whether a particular one of the explanatory variables really "belongs" in the model; equivalently, one might ask if this variable has a true regression coefficient different from 0 (and therefore would affect predictions). We take the standard approach of classical hypothesis testing: In order to see if there is evidence supporting the inclusion of the variable in the model, we start by hypothesizing that it does not belong, i.e., that its true regression coefficient is 0. Dividing the estimated coefficient by the standard error of the coefficient yields the t-ratio of the variable, which simply shows how many standard-deviations-worth of sampling error would have to have occurred in order to yield an estimated coefficient so different from the hypothesized true value of 0. We then ask how likely it is to have experienced so much sampling error: This yields the significance level of the sample data with respect to the null hypothesis that 0 is the true value of the coefficient. The closer this significance level is to 0%, the stronger is the evidence against the null hypothesis, and therefore the stronger the evidence is that the true coefficient is indeed different from 0, i.e., that the variable does belong in the model. Example: In the full model, the significance level of the t-ratio of mileage is 0.0011%. We have overwhelmingly strong evidence that mileage has a true non-zero effect in the model. On the other hand, the significance level of the t-ratio of make is only 12.998%. We have here only a little bit of evidence that the true difference between Fords and Hondas is nonzero. (If we really wish to make a case against Hondas, we'd require that the estimated difference persist as the sample size is increased, i.e., as more evidence is collected.) Measuring the explanatory power of a model Why does the dependent variable take different values for different members of the population? There are two possible answers: "Because the explanatory variables vary." "Because things still sitting in the residual term vary." The total variation seen in the dependent variable can be broken down into these two components, and the coefficient of determination [4] is the fraction of the total variation that is explained by the model, i.e., the fraction explained by variation in the explanatory variables. Subtracting the coefficient of determination from 100% indicates the fraction of variation in the dependent variable that the model fails to explain. Example: Looking at mileage alone, it can explain 56% of the observed car-to-car variation in annual maintenance costs. Looking at age alone, it can't explain much of anything. But variations in mileage and age together can explain over 78% of the variation in costs. The reason they can explain more together than the sum of what they can explain separately is that mileage masks the effect of age in our data. When both are included in the regression model, the effect of mileage is separated from the effect of age, and the latter effect then can be seen. A natural follow-up is to ask what the relative importance of variation in the explanatory variables is in explaining observed variation in the dependent variable. The beta-weights [5] of the explanatory variables can be compared to answer this question. (Dive down for a discussion of the distinction between t-ratios and beta-weights.) Example: In the full model, the beta-weight of mileage is roughly twice that of age, which in turn is more than twice that of make. If asked, "Why does the annual maintenance cost vary from car to car?" one would answer, "Primarily because the cars vary in how far they're driven. Of secondary explanatory importance is that they vary in age. Trailing both is the fact that some are Fords and others Hondas, i.e., that make varies across the fleet." Summary The six "steps" to interpreting the result of a regression analysis are: Look at the prediction equation to see an estimate of the relationship. Refer to the standard error of the prediction (in the appropriate model) when making predictions for individuals, and the standard error of the estimated mean when estimating the average value of the dependent variable across a large pool of similar individuals. Refer to the standard errors of the coefficients (in the most complete model) to see how much you can trust the estimates of the effects of the explanatory variables. Look at the significance levels of the t-ratios to see how strong is the evidence in support of including each of the explanatory variables in the model. Use the "adjectived" coefficient of determination to measure the potential explanatory power of the explanatory variables in order to rank them in order of explanatory importance. [1] Why is it valuable to be able to unravel linear relationships? Some interesting relationships are linear, essentially all managerial relationships are at least locally linear, and several modeling tricks help to transform the most commonly-encountered nonlinear relationships into linear relationships. [2] The dependent and explanatory variables, as well as the residual term, can be thought of as random variables resulting from the random selection of a single member of the population, i.e., as quantities that vary from one individual to the next. [3] The standard error of the prediction takes into account both our exposure to error in using a value of 0 for the individual's residual when making the prediction (measured by the standard error of the regression), and our exposure to sampling error in estimating the regression coefficients (measured by the standard error of the estimated mean). [4] The coefficient of determination is sometimes called the "R-square" of the model. Some computer packages will offer two coefficients of determination, one with an adjective – "adjusted", "corrected", or "unbiased" – in front. Given the choice, use the one with the adjective. If it is somewhat less than zero, read it as 0%. [5] The beta-weight of an explanatory variable has the same sign as the estimated coefficient of that variable. It is the magnitude, i.e., absolute value, of the beta-weight that is of relevance. Regression analysis is a statistical technique that develops a relationship between explanatory (independent) variables and response (dependent) variable. It measures the dependence of one (dependent) variable on one or more than one other (independent) variable. Linear regressionIntroduction of regression analysisThe term regression was first introduced in nineteenth century to describe a biological phenomena, namely that the progeny of exceptional individuals tends on average to be less exceptional than their parents and more alike their more distant ancestors. Francis Galton, cousin of Charles Darwin studied this phenomena. He said that the mean value of a child's characteristic (such as height, weight etc) was not equal to his /her parents height/weight but rather was between this value and the average value of the entire population. Thus, for instance , the height of the offspring of very tall people (called by Galton, people "taller than mediocrity") would tend to be shorter than their parents. Galton called this phenomenon 'regression to mediocrity'.Objectives of Regression analysisEstimate the relationship between explanatory and response variable.Determine the effect of each of the explanatory variables on the response variable.Predict the value of the response variable for a given value of explanatory variable.Types of regression analysisThere are various types of regression analysis based on the regression model. Such as,Linear RegressionLogistic RegressionPolynomial RegressionStepwise RegressionRidge RegressionLasso RegressionElasticNet RegressionThere are three types of linear regression, i.eLinear regression modelMultiple linear regression modelPolynomial regression modelLinear regression modelA regression model where exists one dependent (response) variable and one independent (explanatory) variable. Multiple linear regression modelA regression model where exists one dependent (response) variable and more than one independent (explanatory) variable. Polynomial regression modelA regression model where exists one dependent (response) variable and one independent (explanatory) variable but there is a polynomial function exist in the explanatory variable.Application of regression modelA company might wish to improve its marketing process. After collecting data on the demand for a product, the product's price and the advertising expenditure incurred in promoting the product, the company might use regression analysis to develop an equation to predict the future demand on the basis of price and advertising.A real state company fixes the selling price of its apartments, as it claims, on the basis of the size of the apartments measured in terms of square footage of living space. A sample of 20 apartments was chosen and the apartment size and selling price were recorded. To see if there is any such relationship, the company fits a regression model.Consider the following data on the selling price of a specified floor space.A physician collected blood sample from 50 infants on pulmonary blood flow (PBF) and pulmonary blood volume (PBV) to examine if there is any relationship between PBF and PBV. A linear regression analysis seems appropriate for purpose to see if there is any such relationship. What is standard deviation in regression model?This is common question that everybody asked about regression. In regression model the differences between the regression line and the data at each value of the independent variable is called standard deviation.And more simplystandard deviation= Root of Mean Square Error(RMSE)Data ScienceBivariate analysis| How to analyze data using spss (part-10)Normality check| how to analyze data using spss (part-11) Let us understand the concept with some regression formula example.Example #1Consider the following two variables x and y, you are required to do the calculation of the regression.Solution:Using the above formula, we can calculate linear regression in excel as follows.We have all the values in the above table with n = 5.Now, first, calculate the intercept and slope for the regression.Calculation of Intercept is as follows,a = (628.33 * 88,017.46) – (519.89 * 106,206.14) / 5* 88,017.46 – (519.89)2a = 0.52Calculation of Slope is as follows,b = (5 * 106,206.14) – (519.89 * 628.33) / (5 * 88,017.46) – (519,89)2b = 1.20Let's now input the values in the regression formula to get regression.Hence the regression line Y = 0.52 + 1.20 * X. Alternatively, instead of calculating manually or to verify your result after using Excel, you can use a linear regression calculator to quickly find the values and confirm the regression equation.Example #2Let us take another regression formula example, where the State Bank of India recently established a new policy linking savings account interest rates to Repo rates. Therefore, the auditor of the State Bank of India wants to conduct an independent analysis of the decisions taken by the bank regarding interest rate changes and whether those have been changed whenever there have been changes in the Repo rate. Therefore, the following is the summary of the Repo rate and Bank's savings account interest rate that prevailed in those months are below.The State Bank of India auditor has approached you to conduct an analysis and provide a presentation on the same in the next meeting. Use the regression formula and determine whether the bank's rate changed as and when it changed the Repo rate.Solution:Using the formula discussed above, we can calculate linear regression in Excel. Treating the Repo rate as an independent variable, i.e., X, and treating Bank's rate as the dependent variable as Y.We have all the values in the above table with n = 6Now, first, calculate the intercept and slope for the regression.Calculation of Intercept is as follows,a = (24.17 * 237.69) – (37.75 * 152.06) / 6 * 237.69 – (37.75)2a = 4.28Calculation of Slope is as follows,b = (6 * 152.06) – (37.75 *24.17) / 6 * 237.69 – (37.75)2b= -0.04Let's now input the formulas' values to arrive at the figure.Hence, the regression line Y = 4.28 – 0.04 * X.Analysis: The State Bank of India is indeed following the rule of linking its saving rate to the repo rate, as some slope value signals a relationship between the repo rate and the bank's saving account rate.Example #3ABC laboratory is researching height and weight and wanted to know if there is any relationship, like as the height increases, the weight will also increase. So, they gathered a sample of 1,000 people for each category and found an average height in that group.Below are the details that they have gathered.You are required to do the calculation of regression and come up with the conclusion that any such relationship exists.Solution:Using the formula discussed above, we can calculate linear regression in Excel. Treating height as an independent variable, i.e., X, and weight as the dependent variable as Y.We have all the values in the above table with n = 6Now, first, calculate the intercept and slope for the regression.Calculation of Intercept is as follows,a = (350 * 120,834) – (850 * 49,553) / 6 * 120,834 – (850)2a = 68.63Calculation of Slope is as follows,b = (6 * 49,553) – (850 *350) / 6 * 120,834 – (850)2b = -0.07Let's now input the values in the formula to arrive at the figure.Hence the regression line Y = 68.63 – 0.07 * XAnalysis: There is a significant, less relationship between height and weight, as the slope is very low.