Linear regression

What Is Linear Regression?

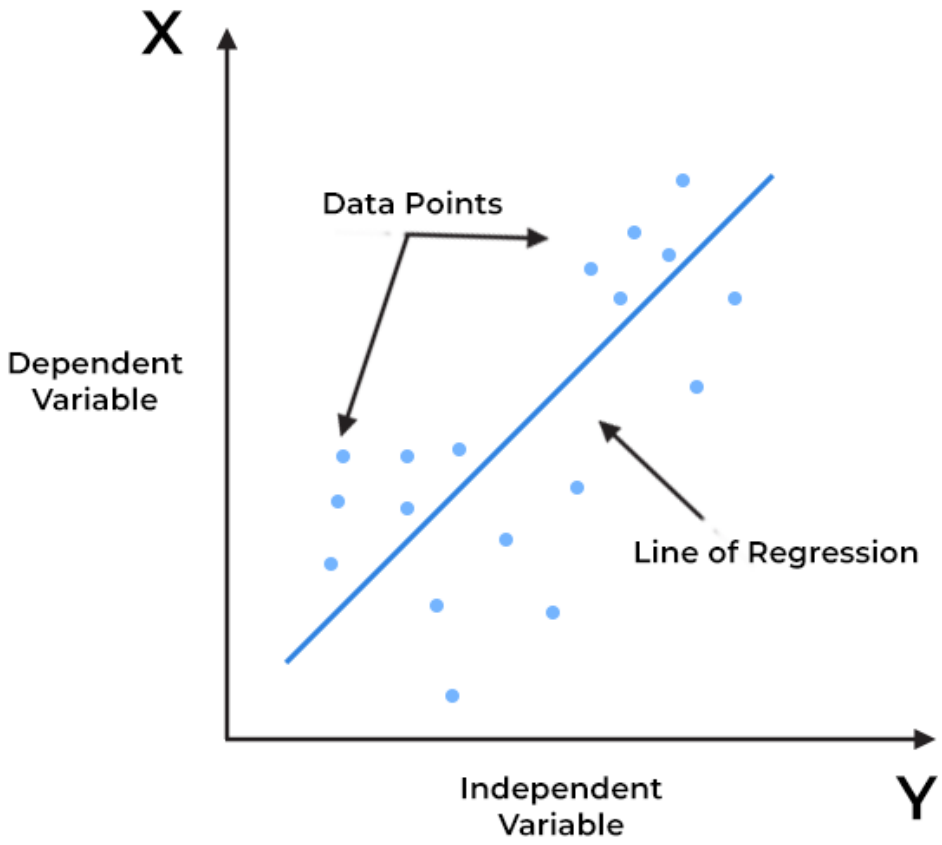
Linear regression is an algorithm that provides a linear relationship between an independent variable and a dependent variable to predict the outcome of future events. It is a statistical method used in data science and machine learning for predictive analysis.

The independent variable is also the predictor or explanatory variable that remains unchanged due to the change in other variables. However, the dependent variable changes with fluctuations in the independent variable. The regression model predicts the value of the dependent variable, which is the response or outcome variable being analyzed or studied.

Thus, linear regression is a supervised learning algorithm that simulates a mathematical relationship between variables and makes predictions for continuous or numeric variables such as sales, salary, age, product price, etc.

This analysis method is advantageous when at least two variables are available in the data, as observed in stock market forecasting, portfolio management, scientific analysis, etc.

A sloped straight line represents the linear regression model.



Best Fit Line for a Linear Regression Model

In the above figure,

X-axis = Independent variable

Y-axis = Output / dependent variable

Line of regression = Best fit line for a model

Here, a line is plotted for the given data points that suitably fit all the issues. Hence, it is called the 'best fit line.' The goal of the linear regression algorithm is to find this best fit line seen in the above figure.

Key benefits of linear regression

Linear regression is a popular statistical tool used in data science, thanks to the several benefits it offers, such as:

1. Easy implementation

The linear regression model is computationally simple to implement as it does not demand a lot of engineering overheads, neither before the model launch nor during its maintenance.

2. Interpretability

Unlike other deep learning models (neural networks), linear regression is relatively straightforward. As a result, this algorithm stands ahead of black-box models that fall short in justifying which input variable causes the output variable to change.

3. Scalability

Linear regression is not computationally heavy and, therefore, fits well in cases where scaling is essential. For example, the model can scale well regarding increased data volume (big data).

4. Optimal for online settings

The ease of computation of these algorithms allows them to be used in online settings. The model can be trained and retrained with each new example to generate predictions in real-time, unlike the neural networks or support vector machines that are computationally heavy and require plenty of computing resources and substantial waiting time to retrain on a new dataset. All these factors make such compute-intensive models expensive and unsuitable for real-time applications.

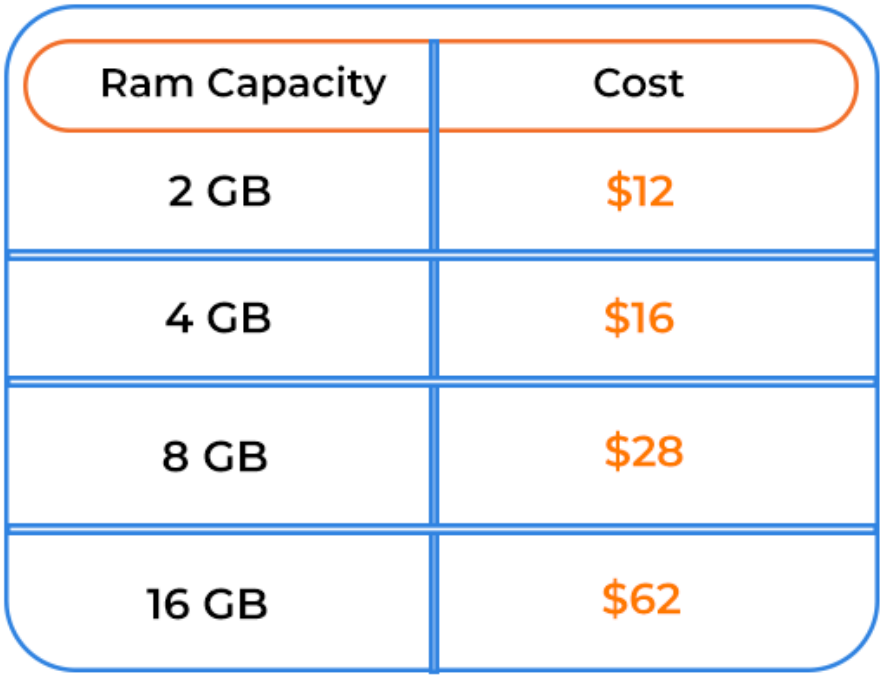
The above features highlight why linear regression is a popular model to solve real-life machine learning problems.

See More: [What Is Super Artificial Intelligence (AI)? Definition, Threats, and Trends]

Linear Regression Equation

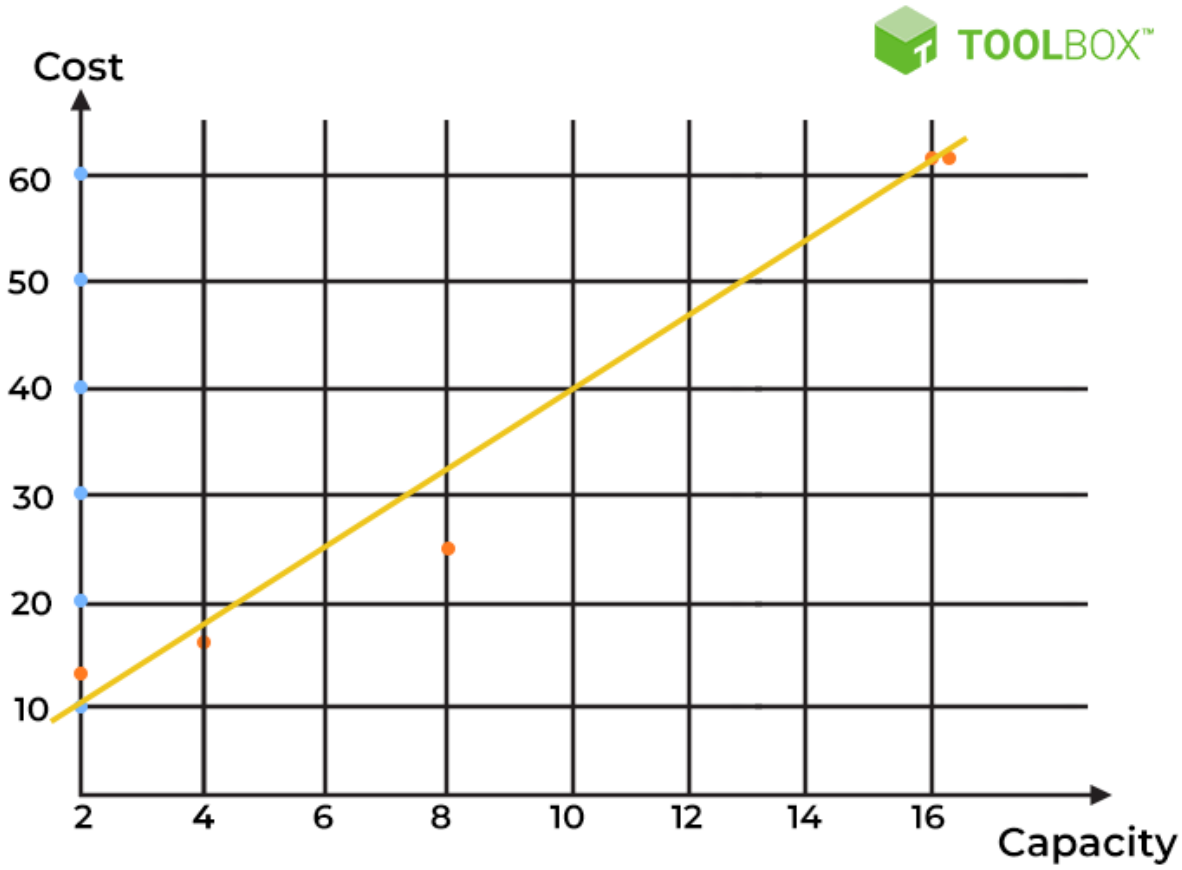
Let's consider a dataset that covers RAM sizes and their corresponding costs.

In this case, the dataset comprises two distinct features: memory (capacity) and cost. The more RAM, the more the purchase cost of RAMs.



Dataset: RAM Capacity vs. Cost

If we plot RAM on the X-axis and its cost on the Y-axis, a line from the lower-left corner of the graph to the upper right represents the relationship between X and Y. On plotting these data points on a scatter plot, we get the following graph:



Scatter Plot: PAM Capacity vs. Cost

The memory-to-cost ratio may vary according to different manufacturers and RAM versions, but the data trend shows a pattern. The data on the bottom left shows cheaper RAMs with smaller memory, and the line continues to the upper right corner of the graph, where the RAMs are of higher capacity and are costly).

The regression model defines a linear function between the X and Y variables that best showcases the relationship between the two. It is represented by the slant line seen in the above figure, where the objective is to determine an optimal 'regression line' that best fits all the individual data points.

Mathematically these slant lines follow the following equation,
Y = m*X + b
Where X = dependent variable (target)
Y = independent variable
m = slope of the line (slope is defined as the 'rise' over the 'run')
However, machine learning experts have a different notation to the above slope-line equation,
y(x) = p0 + p1 * x
where,

- y = output variable. Variable y represents the continuous value that the model tries to predict.
- x = input variable. In machine learning, x is the feature, while it is termed the independent variable in statistics. Variable x represents the input information provided to the model at any given time.
- p0 = y-axis intercept (or the bias term).
- p1 = the regression coefficient or scale factor. In classical statistics, p1 is the equivalent of the slope of the best-fit straight line of the linear regression model.
- pi = weights (in general).

Thus, regression modeling is all about finding the values for the unknown parameters of the equation, i.e., values for p0 and p1 (weights).

The equation for multiple linear regression

The above process applies to simple linear regression having a single feature or independent variable. However, a regression model can be used for multiple features by extending the equation for the number of variables available within the dataset.

The equation for multiple linear regression is similar to the equation for a simple linear equation, i.e., y(x) = p0 + p1x1 plus the additional weights and inputs for the different features which are represented by p(n)x(n). The formula for multiple linear regression would look like, y(x) = p0 + p1x1 + p2x2 + … + p(n)x(n)

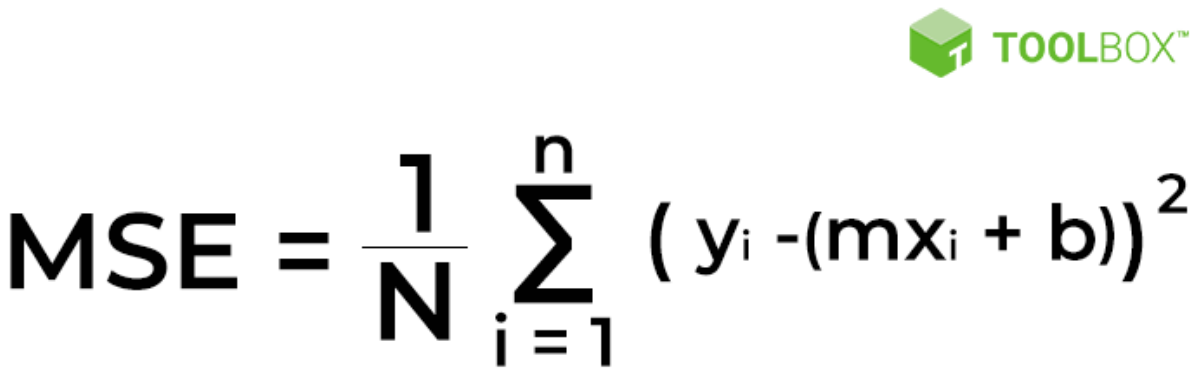
The machine learning model uses the above formula and different weight values to draw lines to fit. Moreover, to determine the line best fits the data, the model evaluates different weight combinations that best fit the data and establishes a strong relationship between the variables.

Furthermore, along with the prediction function, the regression model uses a cost function to optimize the weights (pi). The cost function of linear regression is the root mean squared error or mean squared error (MSE).

Fundamentally, MSE measures the average squared difference between the observation's actual and predicted values. The output is the cost or score associated with the current set of weights and is generally a single number. The objective here is to minimize MSE to boost the accuracy of the regression model.

Math

Given the simple linear equation y=mx+b, we can calculate the MSE values:



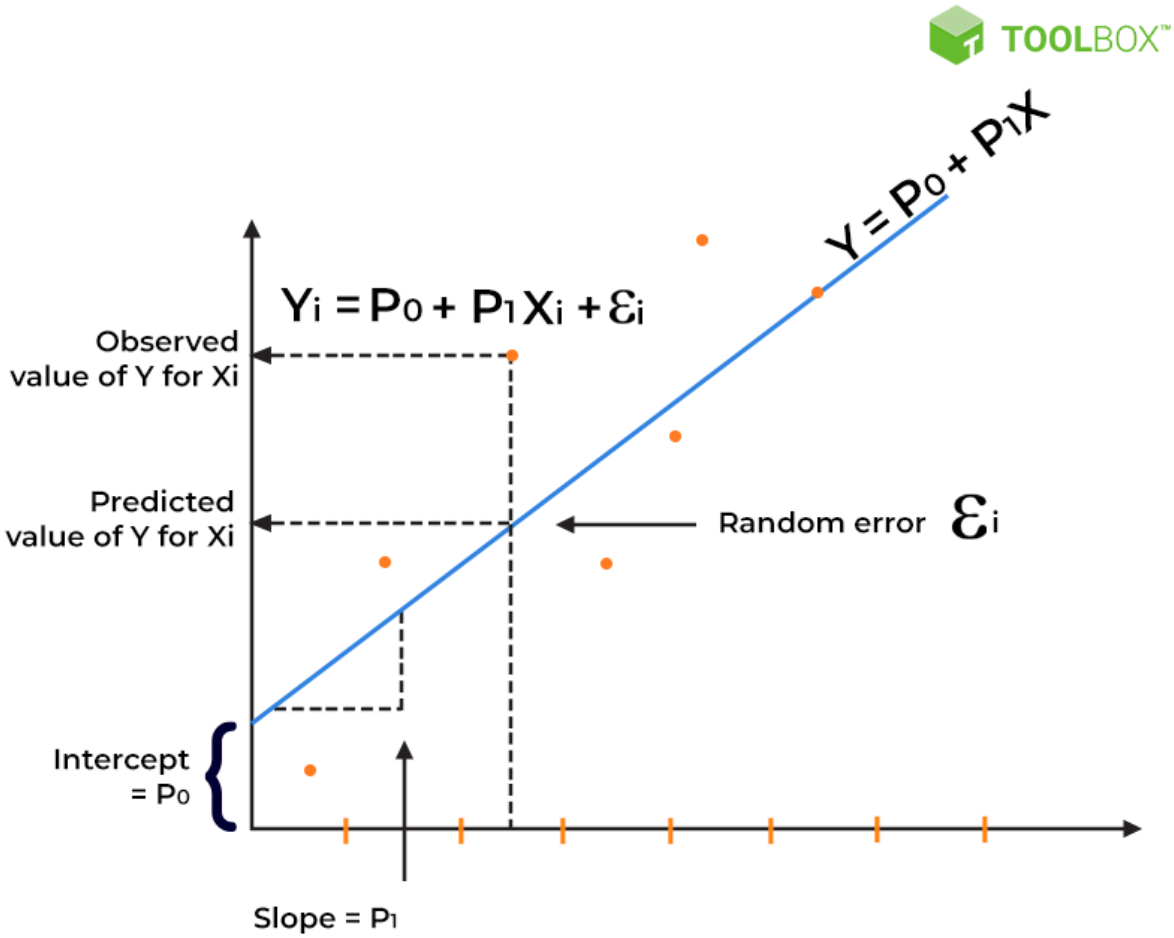
Equation to Calculate MSE Values

Where,

- N = total number of observations (data points)
- 1/N∑ni=1 = mean
- yi = actual value of an observation
- mxi+b = prediction

Along with the cost function, a 'Gradient Descent' algorithm is used to minimize MSE and find the best-fit line for a given training dataset in fewer iterations, thereby improving the overall efficiency of the regression model.

The equation for linear regression can be visualized as:



Visualization of Equation for Linear Regression

Types of Linear Regression with Examples

Linear regression has been a critical driving force behind many AI and data science applications. This statistical technique is beneficial for businesses as it is a simple, interpretable, and efficient method to evaluate trends and make future estimates or forecasts.

The types of linear regression models include:

1. Simple linear regression

Simple linear regression reveals the correlation between a dependent variable (input) and an independent variable (output). Primarily, this regression type describes the following:

- Relationship strength between the given variables.

Example: The relationship between pollution levels and rising temperatures.

- The value of the dependent variable is based on the value of the independent variable.

Example: The value of pollution level at a specific temperature.

2. Multiple linear regression

Multiple linear regression establishes the relationship between independent variables (two or more) and the corresponding dependent variable. Here, the independent variables can be either continuous or categorical. This regression type helps foresee trends, determine future values, and predict the impacts of changes.

Example: Consider the task of calculating blood pressure. In this case, height, weight, and amount of exercise can be considered independent variables. Here, we can use multiple linear regression to analyze the relationship between the three independent variables and one dependent variable, as all the variables considered are quantitative.

3. Logistic regression

Logistic regression—also referred to as the logit model—is applicable in cases where there is one dependent variable and more independent variables. The fundamental difference between multiple and logistic

regression is that the target variable in the logistic approach is discrete (binary or an ordinal value). Implying, the dependent variable is finite or categorical–either P or Q (binary regression) or a range of limited options P, Q, R, or S.

The variable value is limited to just two possible outcomes in linear regression. However, logistic regression addresses this issue as it can return a probability score that shows the chances of any particular event.

Example: One can determine the likelihood of choosing an offer on your website (dependent variable). For analysis purposes, you can look at various visitor characteristics such as the sites they came from, count of visits to your site, and activity on your site (independent variables). This can help determine the probability of certain visitors who are more likely to accept the offer. As a result, it allows you to make better decisions on whether to promote the offer on your site or not.

Furthermore, logistic regression is extensively used in machine learning algorithms in cases such as spam email detection, predicting a loan amount for a customer, and more.

4. Ordinal regression

Ordinal regression involves one dependent dichotomous variable and one independent variable, which can either be ordinal or nominal. It facilitates the interaction between dependent variables with multiple ordered levels with one or more independent variables.

For a dependent variable with m categories, (m -1) equations will be created. Each equation has a different intercept but the same slope coefficients for the predictor variables. Thus, ordinal regression creates multiple prediction equations for various categories. In

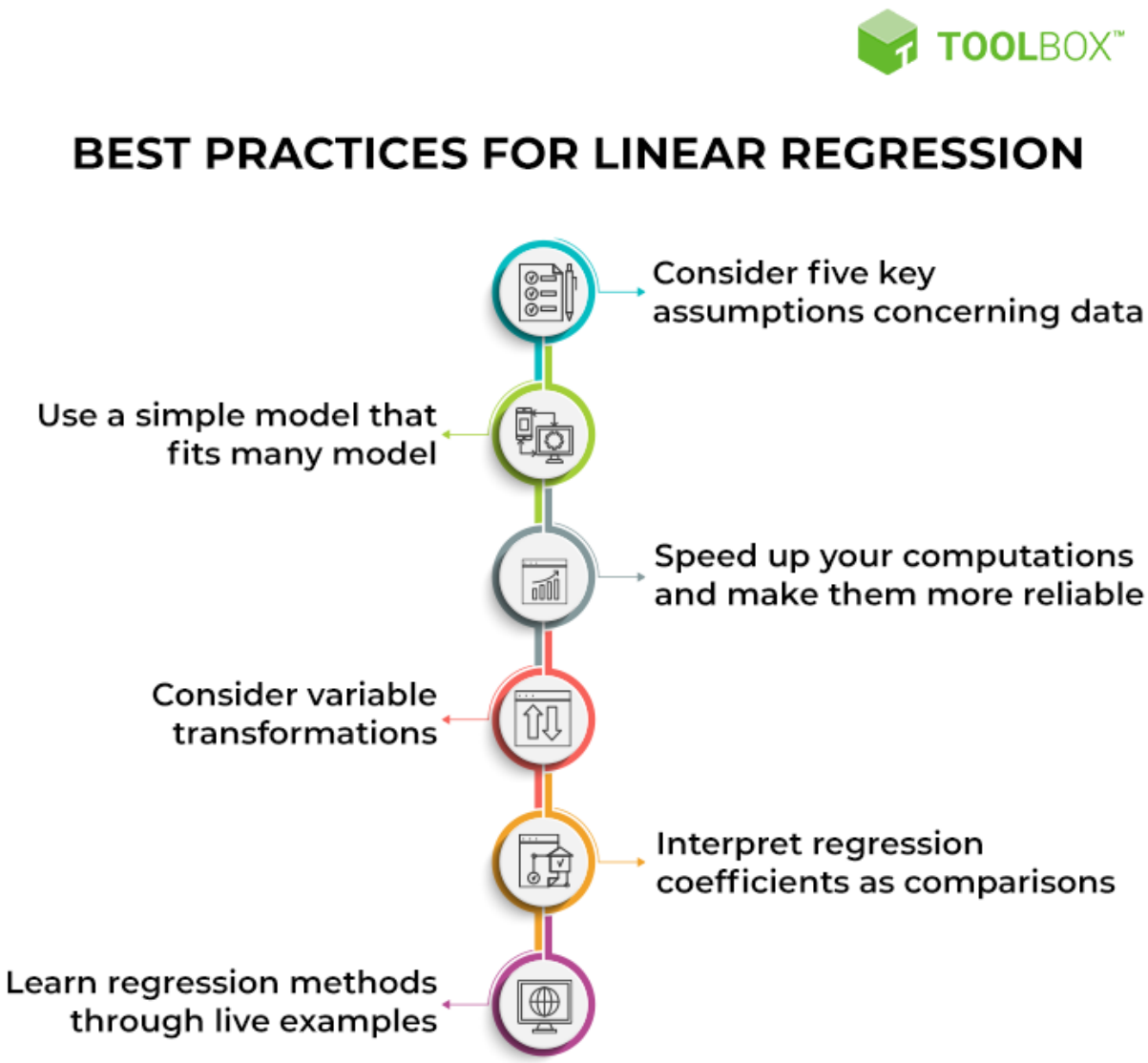
machine learning, ordinal regression refers to ranking learning or ranking analysis computed using a generalized linear model (GLM).
Example: Consider a survey where the respondents are supposed to answer as ‘agree’ or ‘disagree.’ In some cases, such responses are of no help as one cannot derive a definitive conclusion, complicating the generalized results. However, you can observe a natural order in the categories by adding levels to responses, such as agree, strongly agree, disagree, and strongly disagree. Ordinal regression thus helps in predicting the dependent variable having ‘ordered’ multiple categories using independent variables.

5. Multinomial logistic regression

Multinomial logistic regression (MLR) is performed when the dependent variable is nominal with more than two levels. It specifies the relationship between one dependent nominal variable and one or more continuous-level (interval, ratio, or dichotomous) independent variables. Here, the nominal variable refers to a variable with no intrinsic ordering.
Example: Multinomial logit can be used to model the program choices made by school students. The program choices, in this case, refer to a vocational program, sports program, and academic program. The choice of type of program can be predicted by considering a variety of attributes, such as how well the students can read and write on the subjects given, gender, and awards received by them.
Here, the dependent variable is the choice of programs with multiple levels (unordered). The multinomial logistic regression technique is used to make predictions in such a case.
See More: 5 Ways To Avoid Bias in Machine Learning Models

Linear Regression Best Practices for 2022

Today, linear regression models are widely used by data scientists across industries to make a variety of observations. These models help evaluate trends, make sales estimates, analyze pricing elasticity, and assess risks to companies. Experts can adopt specific best practices to ensure the smooth implementation and functioning of linear regression models.



Best Practices for Linear Regression

Here, we list down the linear regression best practices for 2022.

1. Consider five key assumptions concerning data

Linear regression analysis can be practical and be performed accurately when the data abides by a set of rules. These are referred to as key assumptions concerning the data.

- **Linear relationship**

The first important assumption of linear regression is that the dependent and independent variables should be linearly related. The relationship can be determined with the help of scatter plots that help in visualization. Also, one needs to check for outliers as linear regression is sensitive to them.

- **Normal distribution of residuals**

The second assumption relates to the normal distribution of residuals or error terms, i.e., if residuals are non-normally distributed, the model-based estimation may become too wide or narrow. The non-normal distribution also underscores that you need to closely observe some unusual data points to make a good model.

- **Multicollinearity**

The third assumption relates to multicollinearity, where several independent variables in a model are highly correlated. More correlated variables make it difficult to determine which variable contributes to predicting the target variable. Also, standard errors inevitably increase due to correlated variables.
Moreover, with such a robust variable correlation, the predicted regression coefficient of a correlated variable further depends on the other variables available in the model, leading to wrong conclusions

and poor performance. The goal, therefore, is to have minimal or lesser multicollinearity.

- **Autocorrelation**

One fundamental assumption of linear regression specifies that the given dataset should not be autocorrelated. This mostly happens when residuals or error terms are not independent of each other.
In other words, the situation arises when the value of f(a+1) is not independent of the value of f(a). For example, in the case of stock prices, the price of one stock depends on the cost of the previous one.

- **Homoscedasticity**

Another assumption of linear regression analysis is referred to as homoscedasticity. Homoscedasticity relates to cases where the residuals (error terms) between the independent and dependent variables remain the same for all independent variable values.
In simple words, the residuals or error terms must have 'constant variance.' If not, it leads to an unbalanced scatter of residuals, known as heteroscedasticity. With heteroscedasticity, you cannot trust the results of the regression analysis.

2. Use a simple model that fits many models

The common misconception is that complex problems require complex regression models. However, research has identified that simpler models offer precise predictions as they effectively show how well the data fit with the models.
Furthermore, as most models have similar explanatory abilities, simple linear regression models are likely the best choice. Start with a simple regression model and make it complex as per the need. This

has its implications, as the more complex the model is, the more tailored the model will be to the specific dataset. As a result, generalizability suffers.
One must verify, validate, and ensure that the added complexity produces narrower prediction intervals. Also, keep checking the predicted R-squared value rather than chasing the high R-squared range. R-squared is a statistical measure, also termed a coefficient of determination, which evaluates how close the data (data points) are to the fitted regression line.

3. Speed up your computations and make them more reliable

The practice improves how quickly computing systems process the data while using a linear regression model. You may incur additional startup costs associated with data preparation or model complexity by speeding up the computations and making the model run faster. Some methods to boost model speed include:

- **Data subsetting**

Subsetting allows you to explore data with potentially more models, .and separate analysis reveals variations across these subsets.

- **Predictive simulation (Fake data)**

You can resolve glitches in the code or the model fit by employing predictive simulation. This can be achieved using Bayesian generalized linear models that help to make probabilistic predictions via simulations. Also, predictive simulation helps in comparing the data to the fitted model’s prediction.
Fake-data simulation enables you to verify the correctness of the code.

- **Graph the relevant and not the irrelevant**

Graphing is a crucial tool used for visualization while performing regression analysis. The objective of these graphs is to communicate the information to oneself or a wider audience. Displaying the raw data (exploratory data analysis [EDA]) is the first graphing step.

- **Graph the model**

Fitted model graphs have the following representation:

1. The data plots overlay show the model fit.
2. Graphed sets of estimated parameters.
3. Plot predicted datasets and compare them to actual data.

- **Graph the data**

Real-world data is complex as it has multiple dimensions. Hence, making different graphs and observing the model from different vantage points is essential. In other words, use a series of graphs for better data visualizations instead of depending on a single image.

- **Avoid graphing irrelevant data**

Experts in the field know the importance of plotting raw data and regression diagnostics (residual variance). Although such plots help determine the usability of the model in predicting individual data points, the plots do not satisfy the assumptions of linearity, normality, etc., in regression stated earlier.

Implying, focus on plotting graphs that you are capable of explaining rather than graphing irrelevant data that are unexplainable.

4. Consider variable transformations

Consider transforming every variable under consideration in the regression model. Known transformation ways include:

- **Log transformations:** Derive logarithms of positive variables that enable you to consider multiplicative models.
- **Standardization**: Standardizing allows straightforward interpretation and scaling of all the statistics or coefficients in a model. It ensures whether the model data is within a specific range or scale.
- **Transform first, model later (multilevel modeling):** This practice makes model coefficients comparable, and the model makes better sense.

Additionally, consider plotting raw data and residuals while performing the transformations.

The above transformations are univariate. However, interactions and predictors formed by combining inputs can be transformed too; for example, combining all the survey responses creates a total score. Transformations aim to create fit models that include relevant information and can be compared to data.

5. Interpret regression coefficients as comparisons

Interpreting regression coefficients is critical to understanding the model. Let's consider a sample linear regression equation,

*Monthly wage = – 20 + 0.7 * height + error*

(Where wage = per $1k and height = inches)

The model tells us that taller people in this sample earn more on average. In other words, the model reveals the average difference in earnings between two people who have some height difference.

Interpreting regression in the context of comparisons has the following benefits:

- **Interpretation as a comparison**: Comparisons explain the model without the need for any causal assumptions.
- **Complicated regressions can be built using simpler models**: For complex regressions, start with simpler models and make/add adjustments as needed.
- **Comparisons help in causal inference**: Deriving the causal impact on an outcome (causal inference) becomes possible with comparative interpretation.

6. Learn regression methods through live examples

Linear regression is best learned when you apply complex statistical methods to real-life problems that you care about.
The process begins with proper data collection procedures related to the population of interest. Then, you need to determine the objective of data collection and analysis. This implies identifying what you want to achieve and is it achievable with the data at hand. Finally, gain a complete statistical understanding of your data through simulations and visualizations.

Takeaway

Linear regression models are based on a simple and easy-to-interpret mathematical formula that helps in generating accurate predictions. They find applications across business areas and academic fields such as social sciences, management, environmental and computational science.

With a scientific base, linear regression has proven to predict future trends reliably. It has been widely adopted as these models are easy to interpret, comprehend and can be trained quickly.