Linear Regression

Two major types of problems that machine learning algorithms try to solve are:

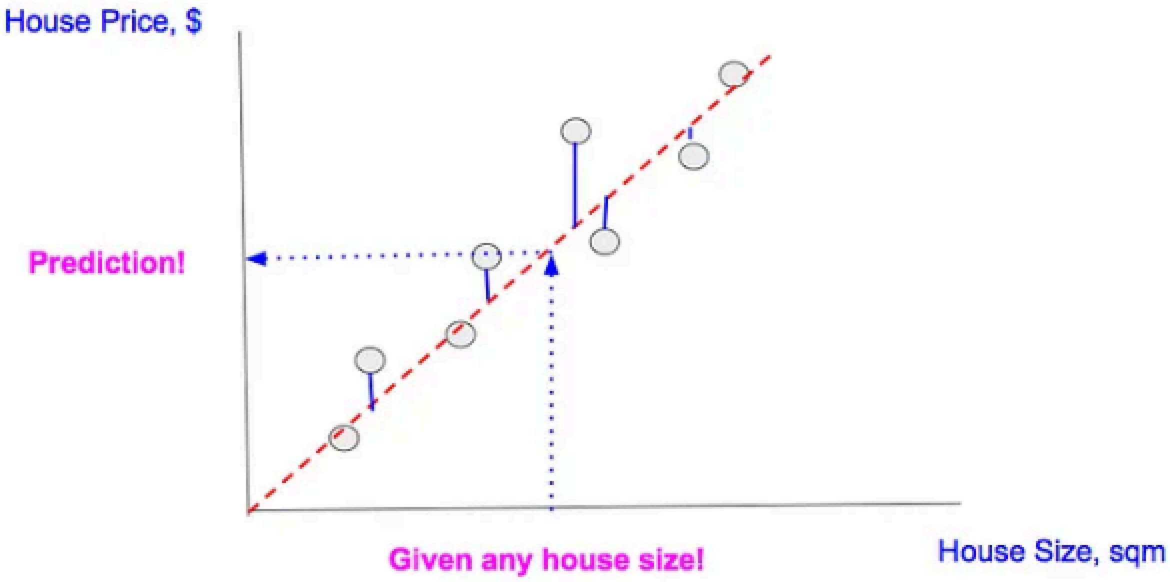
Regression — Predict continuous value of a given data point.

Classification — Predict the class of the given data point.

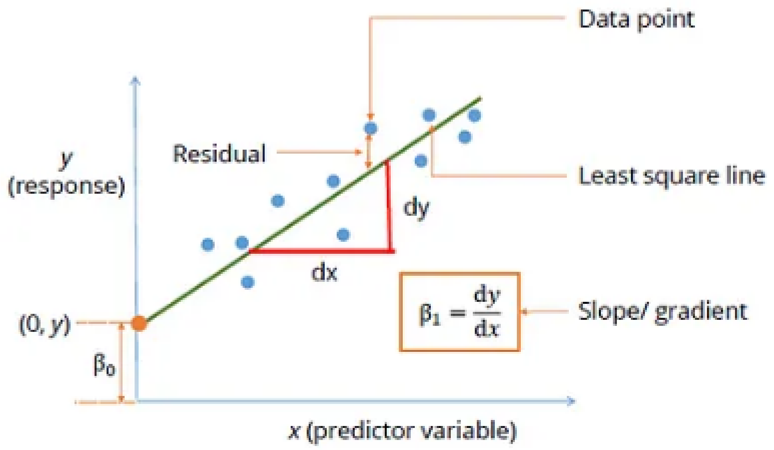
Linear Regression, intuitively is a regression algorithm with a Linear approach. We try to predict a continuous value of a given data point by generalising on the data that we have in hand. The line part indicates that we are using a linear approach in generalising over the data.

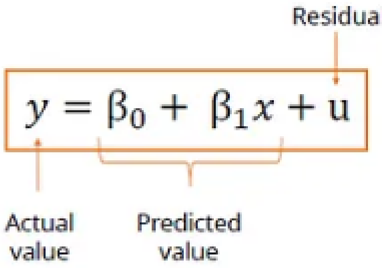
Example:

Examples makes it easy to understand, so suppose you want to predict the price of a house by knowing its size. You have the data which has some house prices and corresponding sizes. Charting the data and fitting a line among them will look something like this:



To generalize, you draw a straight line such that it crosses through the maximum points. Once you get that line, for house of any size you just project that data point over the line which gives you the house price.





We will go over the elements of the equation one by one:

y — The value that you want to predict

β_0 — The y-intercept of the line means where the line intersects the Y-axis

β_1 — The slope or gradient of the line means how steep is the line

x — The value of the data point

u — The residual or noise that are caused by unexplained factors

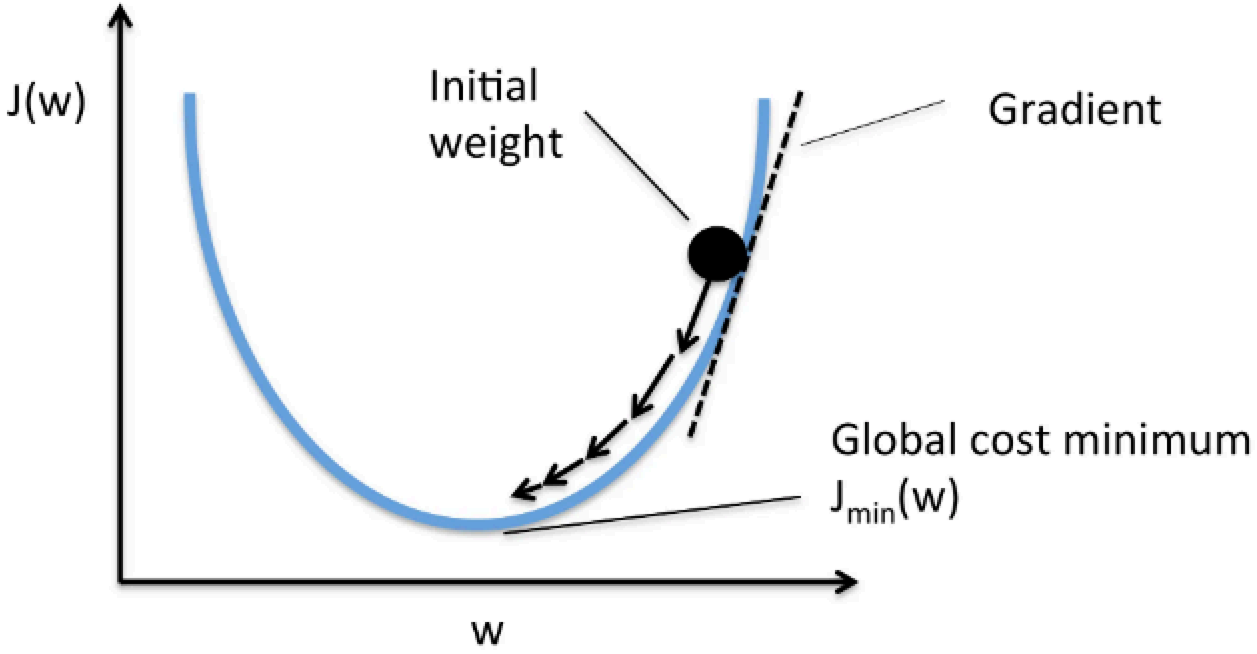
Cost Function:

The Cost Function is a mathematical construct which is calculated by adding up the squared error terms:

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

Gradient Descent:

Gradient Descent is a simple optimisation technique for finding minimum of any function, in this case we want to find the minima of our MSE function.



To know the slope of a function at any point you differentiate that point with respect to its parameters, thus Gradient Descent differentiates the above Cost function and comes to know the slope of that point.

To go to the bottom most point it has to go in the opposite direction of the slope, i.e. where the slope is decreasing.

It has to take small steps to move towards the bottom point and thus learning rate decides the length of step that gradient descent will take.

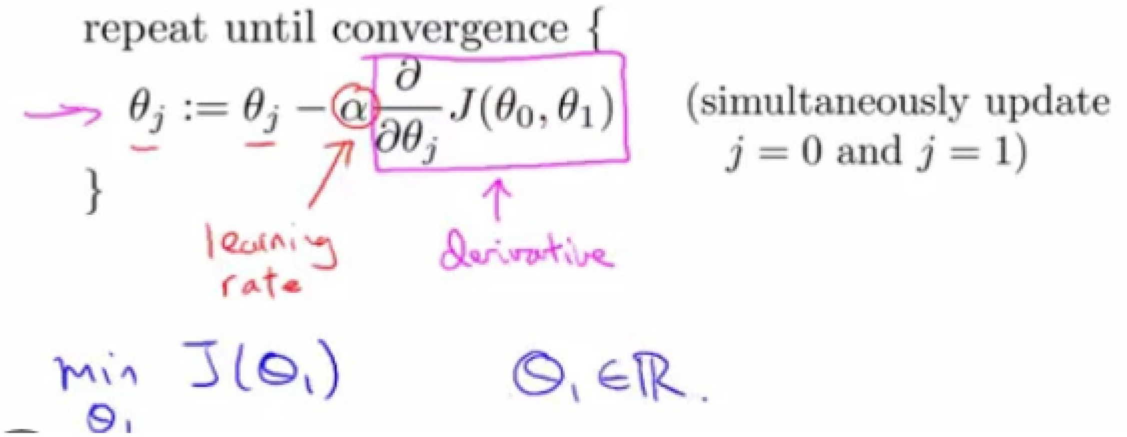
After every move it validates that the current position is global minima or not. This is validated by the slope of that point, if the slope is zero then the algorithm has reached the bottom most point.

After every step, it updates the parameter (or weights) and by doing the above step repeatedly it reaches to the bottom most point.

If we consider J function as our cost function:

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

The formula to update the weights would be:



```
repeat until convergence {
    $\theta_j := \theta_j - \alpha \frac{\partial}{\partial\theta_j} J(\theta_0, \theta_1)$    (simultaneously update $j = 0$ and $j = 1$)
}
```

R-Squered:

It is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. R-squared explai to what extent the variance of one variable explains the variance of the second variable. So, if the R2 of a model is 0.50, then approximately half of the observed variation can be explained by the model's inputs.

If the mean of the observed data is:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

then the variability of the data set can be measured with two sums of squares formulas:

$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

The most general definition of the coefficient of determination is:

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Adjusted R-Squared:

Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases when the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected. Typically, the adjusted R-squared is positive, not negative. It is always lower than the R-squared.

$$\bar{R}^2 = 1 - \frac{SS_{\text{res}}/\text{df}_{\text{res}}}{SS_{\text{tot}}/\text{df}_{\text{tot}}}$$

Inserting the degrees of freedom and using the definition of R2, it can be rewritten as:

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

where p is the total number of explanatory variables in the model, and n is the sample size.

Bias and Variance

What is bias?

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

Double-click (or enter) to edit

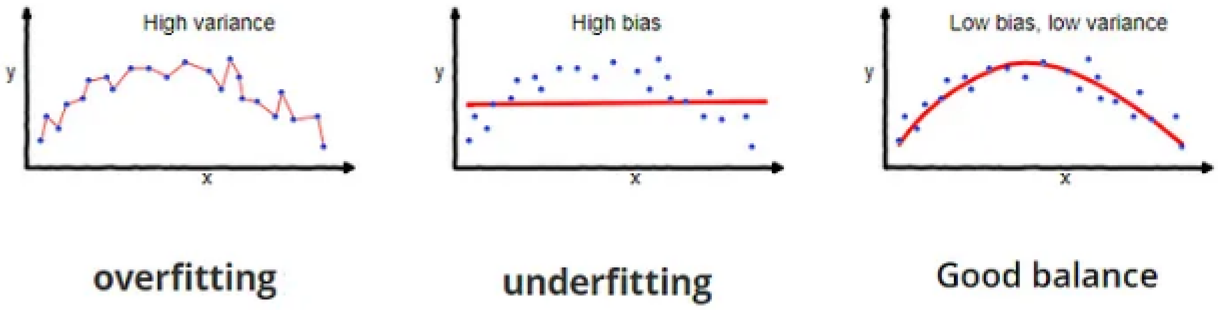
What is variance?

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn’t seen before. As a result, such models perform very well on training data but has high error rates on test data.

Overfitting and Underfitting

In supervised learning, **underfitting** happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance. It happens when we have very less amount of data to build an accurate model or when we try to build a linear model with a nonlinear data. Also, these kind of models are very simple to capture the complex patterns in data like Linear and logistic regression.

In supervised learning, **overfitting** happens when our model captures the noise along with the underlying pattern in data. It happens when we train our model a lot over noisy dataset. These models have low bias and high variance. These models are very complex like Decision trees wh are prone to overfitting.



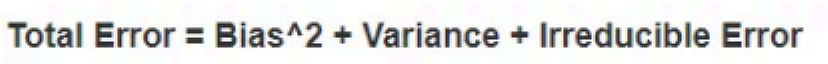
Why is Bias Variance Tradeoff?

If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.

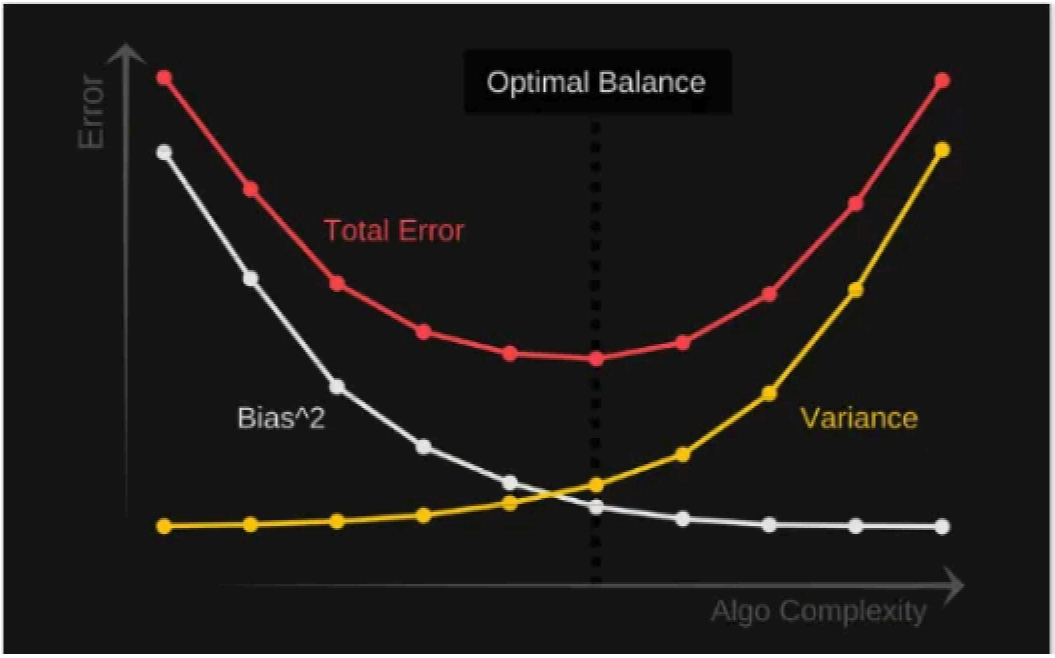
This tradeoff in complexity is why there is a tradeoff between bias and variance. An algorithm can't be more complex and less complex at the same time.

Total Error

To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.



Total Error = Bias^2 + Variance + Irreducible Error



Multicollinearity

Multicollinearity is a statistical concept where several independent variables in a model are correlated.

Two variables are considered perfectly collinear if their correlation coefficient is +/- 1.0.

Multicollinearity among independent variables will result in less reliable statistical inferences.

When you're analyzing an investment, it is better to use different types of indicators rather than multiple indicators of the same type to avoid multicollinearity.

Multicollinearity can lead to less reliable results because the results you're comparing are generally the same.

Detecting Multicollinearity

A statistical technique called the variance inflation factor (VIF) can detect and measure the amount of collinearity in a multiple regression model. VIF measures how much the variance of the estimated regression coefficients is inflated as compared to when the predictor variables are not linearly related. A VIF of 1 will mean that the variables are not correlated; a VIF between 1 and 5 shows that variables are moderately correlated, and a VIF between 5 and 10 will mean that variables are highly correlated.

What are the basic assumptions for linear regression?

1. Linear Relationship: There should be a linear relationship between the independent and dependent variables.
2. Multivariate Normality: The residuals (errors) of the regression should be normally distributed.
3. No or Little Multicollinearity: Independent variables should not be too highly correlated with each other.
4. No Auto-correlation: The residuals should not be correlated with each other. This is especially important for time series data.
5. Homoscedasticity: The variance of residual errors should be constant across all levels of t' independent variables.

If these assumptions are not met, there are several ways to satisfy them:

1. Linear Relationship: If the relationship is not linear, consider transforming the variables (e.g., using logarithmic or square root transformations) or using polynomial regression.
2. Multivariate Normality: If the residuals are not normally distributed, consider transforming the dependent variable or using non-parametric regression methods.
3. Multicollinearity: If there is multicollinearity, consider removing some of the highly correlated independent variables, or use regularization techniques like Ridge or Lasso regression.
4. Auto-correlation: If there is auto-correlation, consider using time series analysis techniques like ARIMA models or adding a time lag variable to the regression.
5. Homoscedasticity: If the variance of residuals is not constant, consider transforming the dependent variable or using weighted least squares regression.

What are the advantages of linear regression?

1. Simplicity and Interpretability: One of the main strengths of linear regression is its simplicity. The relationship between independent and dependent variables is expressed through a linear equation, making the results easy to interpret. This simplicity also makes the model easy to explain to stakeholders.
2. Less Data Required: Compared to more complex models, linear regression can perform well with a smaller amount of data. This is especially useful in situations where data collection is expensive or difficult.
3. Efficiency: Linear regression models are computationally inexpensive to run, which means they require less computational resources and time, making them suitable for situations with limited computational power or when quick decisions are needed.

4. Predictive Performance: When the relationship between variables is approximately linear and the data is not overly complex, linear regression can provide accurate and reliable predictions.
5. Basis for Other Techniques: Linear regression serves as the foundation for many other statistical modeling techniques. It introduces many of the principles used in more complex models, such as regularization in ridge and lasso regression, and it forms the basis for logistic regression used in classification tasks.
6. Quantifiable: Linear regression provides coefficients for each explanatory variable, allowing for quantitative assessment of the strength and type (positive or negative) of the relationship between the dependent variable and each independent variable.
7. Versatility: It can be applied to a wide range of real-world problems, from predicting housi prices to estimating sales or evaluating trends over time.
8. Assessment of Relationships: It allows you to understand the relationships and influences among variables, which is valuable for hypothesis testing in scientific research, where you might want to understand the impact of variables on a particular outcome.
9. Handling Overfitting: With methods like ridge and lasso, which extend linear regression, you can manage overfitting by penalizing large coefficients, thus enhancing the model’s ability to generalize.

What are the disadvantages of linear regression?

1. Assumption of Linearity: Linear regression assumes that there is a linear relationship between the independent and dependent variables. This can be a major limitation if the true relationship is non-linear; in such cases, linear regression might not effectively model the data or predict outcomes accurately.
2. Sensitivity to Outliers: Linear regression models are highly sensitive to outliers. A single outlier can significantly affect the slope of the regression line and the overall results, leading to poor model performance, especially in predicting future outcomes.
3. Assumption of Homoscedasticity: The model assumes that the residuals (errors) have constant variance at all levels of the independent variables. If the variance of the residuals varies, a condition known as heteroscedasticity, the predictive performance and the reliability of the model's inferences can be adversely affected.
4. Independence of Observations: Linear regression assumes that all observations are independent of each other. In cases where there is correlation between observations (as often found in time series data where data points naturally relate to preceding points), the model’s performance can degrade.
5. Multicollinearity: If there is high correlation between independent variables, it can lead to multicollinearity, which reduces the precision of the estimate coefficients, thereby making the statistical tests less reliable. This can also make it difficult to determine the effect of each independent variable on the dependent variable.

6. Normality of Residuals: For proper inference, linear regression assumes that the residuals are normally distributed. This assumption is critical for constructing confidence intervals and conducting hypothesis tests. If the residuals are not normally distributed, the confidence intervals and hypothesis tests may not be valid.
7. Limited Predictive Power for Complex Relationships: Since linear regression is constrained to linear relationships, it often does not perform well with more complex data patterns or interactions that are better modeled by non-linear or multilevel hierarchical models.
8. Can’t Fit Count and Binary Outcomes Well: Linear regression is not suitable for count data or binary outcomes. Other types of regression such as Poisson regression and logistic regression are better suited for these data types.
9. Model Overfitting: Especially in cases where the number of predictors is close to the number of observations, linear regression can overfit the data, meaning it performs well on training data but poorly on unseen data.

Whether feature scaling is required in linear regression?

Feature scaling is not strictly required for linear regression to function, but it can be very beneficial in several contexts, especially when using techniques that involve regularization or when the features differ significantly in scales. Here are some points to consider about the importance of feature scaling in linear regression: When Feature Scaling is Beneficial:

1. Gradient Descent Optimization: If you are using gradient descent to find the regression coefficients, feature scaling can speed up the convergence of the algorithm. Features with vastly different scales can result in a skewed optimization landscape, making it harder for gradient descent to find the optimal solution efficiently.
2. Regularization: Techniques like Ridge (L2 regularization) and Lasso (L1 regularization) penalize the magnitude of the coefficients in the regression model. Without feature scaling, the penalty might disproportionately affect smaller scale features. By scaling all features to a comparable range, you ensure that the regularization term treats all features equally, improving the effectiveness of the model.
3. Interpretability: When features are on the same scale, it becomes easier to compare the magnitude of coefficients to see which features have more influence on the response variable. This can be crucial for understanding your model in business or scientific settings.

What is the impact of missing values on linear regression?

Missing values in datasets can significantly impact linear regression models, both in terms of model training and the validity of the results. Here’s how missing values affect linear regression:

1. Inability to Perform Calculations Linear regression requires complete cases (rows of data) to compute the regression coefficients. If any values are missing in the rows used for fitting the model, those rows cannot be used in the computation without appropriate handling of the missing data. Most statistical software or libraries will automatically exclude any row with a missing value in any variable used in the model, which can reduce the sample size and potentially bias the results if the missing data are not missing completely at random (MCAR).
2. Biased Estimates If the missing values are not random (Missing Not at Random - MNAR or Missing at Random - MAR), excluding cases with missing data can lead to biased estimates of the coefficients. For instance: • MNAR (Missing Not at Random): The missingness of the data is related to its hypothetical value. For example, higher values are more likely to be missing because they are not reported. • MAR (Missing at Random): The missingness is related to some other observed data in the dataset but not to the values of the missing data themselves.
3. Reduced Statistical Power By automatically excluding rows with missing values, the effect sample size used in the model is reduced. This reduction in sample size decreases the statistical power of the regression analysis, potentially leading to a failure in detecting significant relationships between variables.
4. Impact on Model Accuracy Missing data can lead to an underfitting model if significant predictors have missing values that aren't adequately imputed or handled. This can make the model less accurate and reliable. Solutions to Handle Missing Data in Linear Regression:
5. Deletion: • Listwise Deletion: Removing entire records where any single value is missing. Simple but can lead to a significant reduction in data size and potential biases. • Pairwise Deletion: Uses all available data for each calculation. This can lead to using different data sets for different parts of the analysis, which can be problematic for consistency.
6. Imputation: • Mean/Median/Mode Imputation: Replacing missing values with the mean, median, or mode of the respective column. Easy to implement but can introduce bias if the data are not MCAR. • Regression Imputation: Estimating missing values using regression techniques based on other available variables. • Advanced Methods: Techniques such as multiple imputation or using algorithms like k-nearest neighbors (KNN) or machine learning-based approaches (e.g., predictive mean matching).
7. Using Indicator Variables: Adding a binary indicator variable that flags data as missing can sometimes help the model account for the nature of the missingness if it’s systematic. The choice of method depends on the extent and nature of missingness, as well as on the specific analysis goals and assumptions about the data. Proper handling of missing data is crucial for building reliable and valid linear regression models.

what is the impact of outliers on linear regression?

Outliers can significantly impact linear regression models because they can influence the slope of the regression line and potentially lead to misleading results. Here’s a detailed look at the impact of outliers on linear regression:

1. Impact on the Fit of the Model Slope and Intercept: Outliers, especially those in the X (predictor) or Y (response) variable, can dramatically change the slope and intercept of the regression line. An outlier can pull the regression line toward itself, leading to a less accurate estimate of the relationship for the majority of the data. Influence: Outliers may exert an inordinate influence on the model fit, meaning a single or a few points could control the slope and intercept of the regression line. This can distort the true relationship between variables in the rest of the data.
2. Inflated Error Metrics Increased Residuals: Outliers can cause larger residuals (differences between observed and predicted values), which can inflate measures of the model's error, such as the mean squared error (MSE) or the root mean squared error (RMSE). Reduced Accuracy: The presence of outliers often results in a less accurate model because the statistical assumptions required for linear regression (like homoscedasticity and normality residuals) are violated. This reduction in model accuracy can compromise the reliability of its predictions.
3. Biased Regression Coefficients Parameter Estimates: Outliers can lead to biased or skewed estimates of the regression coefficients. This bias can affect the interpretation and the significance of the predictor variables, leading to incorrect conclusions about the data.
4. Assumption Violations Normality of Residuals: Linear regression assumes that the residuals are normally distributed. Outliers can cause the distribution of residuals to become skewed or heavy-tailed, violating this assumption. Homoscedasticity (Equal Variance): Another assumption of linear regression is that the residuals have constant variance (homoscedasticity). Outliers can lead to heteroscedasticity, where variances are unequal across the range of data, affecting the efficiency of the estimates. Methods to Handle Outliers in Linear Regression: Detection: Graphical Methods: Plots like scatter plots, box plots, or residual plots can help visualize outliers. Statistical Tests: Use tests like Z-scores, IQR (Interquartile Range), or influence measures (Cook’s distance, leverage) to identify outliers. Removal: If an outlier is a result of data entry or measurement error, and if its removal can be justified, it might be removed from the dataset to improve the model’s performance. Transformation: Applying transformations such as logarithmic, square root, or Box-Cox can reduce the effect of outliers by making the data more homogeneous and closer to meeting the assumptions of linear regression. Robust Regression Methods: Utilizing robust regression techniques that are less sensitive to outliers, such as RANSAC (Random Sample Consensus), Theil-Sen estimator, or Huber regression. Weighting Schemes: Implementing methods that give less weight to outlier observations during the regression analysis. Understanding the source and nature of outliers is crucial before deciding on the method for dealing with them. In some cases, outliers may contain valuable information about the data set, or they may

represent a critical part of the population you are studying. In such cases, it's essential to incorporate them appropriately in the analysis rather than simply removing them.

Refrence:

https://towardsdatascience.com/an-intuitive-perspective-to-linear-regression-32bb9885b312

https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

https://www.investopedia.com/terms/m/multicollinearity.asp

Start coding or generate with AI.