# Regression Analysis

## Definition of Regression Analysis

Regression analysis is a statistical method used to estimate the relationship between a dependent variable (also known as the response variable) and one or more independent variables (also known as predictors or explanatory variables). The goal of regression analysis is to identify the strength and direction of the relationship between the variables, and to make predictions or draw inferences about the dependent variable based on the independent variables.

## Types of Regression Analysis:

There are several types of regression analysis, including:

- Simple linear regression: a regression analysis that involves a single independent variable and a linear relationship between the independent and dependent variables.
- Multiple linear regression: a regression analysis that involves two or more independent variables and a linear relationship between the independent and dependent variables.
- Polynomial regression: a regression analysis that involves a curvilinear relationship between the independent and dependent variables.
- Logistic regression: a regression analysis that is used when the dependent variable is binary (e.g., yes/no, true/false) and the relationship between the independent and dependent variables is nonlinear.

## Assumptions of Regression Analysis:

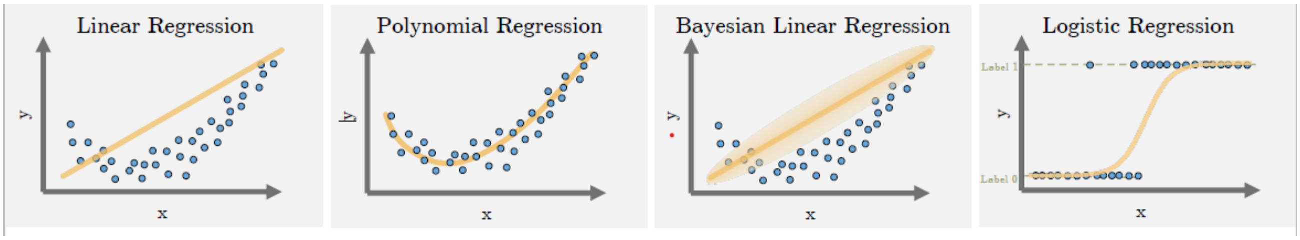
Regression analysis is based on several assumptions, including:

- Linearity: the relationship between the independent and dependent variables is linear.
- Independence: the observations are independent of each other.
- Homoscedasticity: the variance of the dependent variable is constant across different levels of the independent variable(s).
- Normality: the residuals (the differences between the predicted and actual values) are normally distributed.

## Interpretation of Regression Analysis:

The output of a regression analysis typically includes several statistics, such as the coefficients, standard errors, and R-squared value. The coefficients represent the estimated change in the dependent variable for each unit change in the independent variable, while the standard errors reflect the degree of uncertainty in the coefficient estimates. The R-squared value indicates the proportion of variance in the dependent variable that is explained by the independent variables.

## Visual Representation:



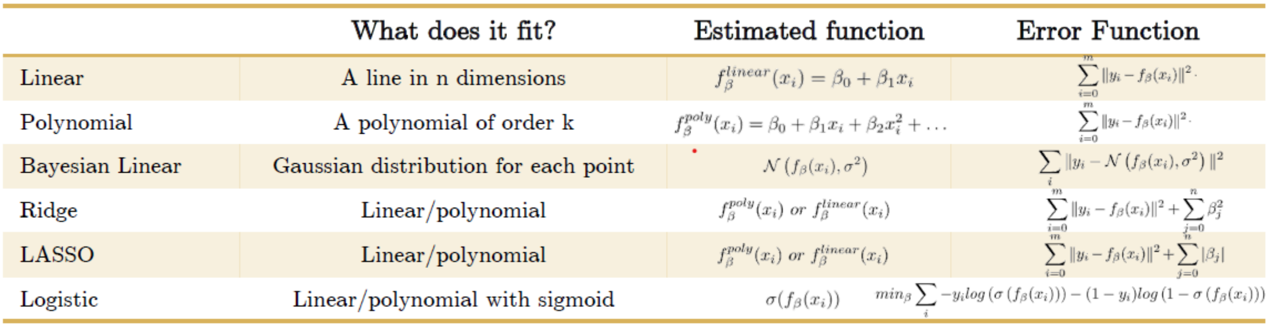
Linear Regression

Polynomial Regression

Bayesian Linear Regression

Logistic Regression

## Summary:



| | What does it fit? | Estimated function | Error Function |
|---|---|---|---|
| Linear | A line in n dimensions | $f_\beta^{linear}(x_i) = \beta_0 + \beta_1 x_i$ | $\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2.$ |
| Polynomial | A polynomial of order k | $f_\beta^{poly}(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \ldots$ | $\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2.$ |
| Bayesian Linear | Gaussian distribution for each point | $\mathcal{N}\left(f_\beta(x_i), \sigma^2\right)$ | $\sum_i \|y_i - \mathcal{N}\left(f_\beta(x_i), \sigma^2\right)\|^2$ |
| Ridge | Linear/polynomial | $f_\beta^{poly}(x_i)$ $or$ $f_\beta^{linear}(x_i)$ | $\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2 + \sum_{j=0}^{n} \beta_j^2$ |
| LASSO | Linear/polynomial | $f_\beta^{poly}(x_i)$ $or$ $f_\beta^{linear}(x_i)$ | $\sum_{i=0}^{m} \|y_i - f_\beta(x_i)\|^2 + \sum_{j=0}^{n} \|\beta_j\|$ |
| Logistic | Linear/polynomial with sigmoid | $\sigma(f_\beta(x_i))$ | $min_\beta \sum_i -y_i log\left(\sigma\left(f_\beta(x_i)\right)\right) - (1-y_i) log\left(1-\sigma\left(f_\beta(x_i)\right)\right)$ |

# Regularization in ML

## Definition of Regularization:

Regularization is a technique used in machine learning to prevent over fitting of a model by adding a penalty term to the objective function that the model is trying to optimize. The penalty term encourages the model to have smaller coefficients, which reduces the complexity of the model and helps to prevent it from fitting the noise in the data.

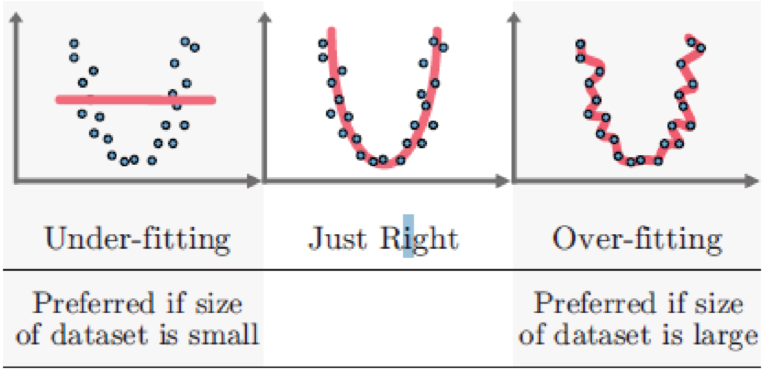

Under-fitting
Just Right
Over-fitting
Preferred if size of dataset is small
Preferred if size of dataset is large

Figure 1. Overfitting

## Types of Regularization:

There are two main types of regularization:

- L1 regularization (also known as Lasso regularization): a technique that adds a penalty term proportional to the absolute value of the coefficients to the objective function. This type of regularization is useful when the goal is to obtain a sparse model, where many of the coefficients are zero.
- L2 regularization (also known as Ridge regularization): a technique that adds a penalty term proportional to the square of the coefficients to the objective function. This type of regularization is useful when the goal is to obtain a model with small coefficients that are all non-zero.

## Tuning the Regularization Parameter:

The amount of regularization applied to a model is controlled by a hyperparameter known as the regularization parameter. This parameter is typically set using a validation set or a cross-validation procedure, where different values of the regularization parameter are tested and the one that leads to the best performance on the validation set is chosen.

## Advantages of Regularization:

Regularization has several advantages, including:

- It helps to prevent over fitting of a model by reducing the complexity of the model and making it less sensitive to noise in the data.
- It can improve the interpretability of a model by reducing the number of non-zero coefficients and highlighting the most important features.
- It can help to handle multi collinearity (i.e., high correlation between the independent variables) by reducing the impact of correlated features on the model.

# Quiz Questions:

**1.In linear regression, what is the purpose of the residual plot?**

A) To visualize the relationship between the dependent variable and the independent variable.

B) To check for outliers in the dataset.

C) To assess the distribution of the residuals.

D) To check for heteroscedasticity in the model.

**Answer: C) To assess the distribution of the residuals.**

Explanation: Residuals are the differences between the actual values of the dependent variable and the predicted values based on the regression model. A residual plot is a graphical representation of these differences, and it is used to assess the assumptions of the linear regression model. Specifically, the residual plot is used to check for the normality and homoscedasticity of the residuals. If the residuals are normally distributed and have constant variance across different levels of the independent variable(s), then the linear regression model is considered to be valid. Therefore, option C is the correct answer. Option A is incorrect because visualizing the relationship between the dependent and independent variables is typically done through scatterplots or other graphical methods. Option B is incorrect because checking for outliers can be done using other methods such as boxplots or z-scores. Option D is incorrect because checking for heteroscedasticity can also be done using other methods such as the Breusch-Pagan test or the White test.

**2.In linear regression, what is the purpose of the coefficient of determination (R-squared)?**

A) To measure the strength of the relationship between the dependent and independent variables.

B) To measure the amount of variance in the dependent variable that is explained by the independent variable(s).

C) To measure the amount of variance in the independent variable(s) that is explained by the dependent variable.

D) To measure the accuracy of the predictions made by the model.

**Answer: B) To measure the amount of variance in the dependent variable that is explained by the independent variable(s).**

Explanation: The coefficient of determination, also known as R-squared, is a statistical measure that represents the proportion of variance in the dependent variable that is explained by the independent variable(s) in a linear regression model. It ranges from 0 to 1, with higher values indicating a better fit between the model and the data. Therefore, option B is the correct answer. Option A is incorrect because the correlation coefficient (r) is used to measure the strength of the relationship between the dependent and independent variables. Option C is incorrect because there is no equivalent measure of the amount of variance in the independent variable(s) that is explained by the dependent variable. Option D is incorrect because the accuracy of the predictions made by the model can be assessed using other metrics such as the mean squared error (MSE) or the root mean squared error (RMSE).

**3.Which of the following statements are true about regularization in machine learning?**

A) Regularization is used to prevent overfitting in a model

B) Regularization adds a penalty term to the cost function

C) L1 regularization encourages sparse models

D) L2 regularization encourages sparse models

E) Regularization can be applied to all types of models

F) Regularization always improves the performance of a model

Choose all that apply:

**Answer: A, B, C.**

Explanation:

A) Regularization is commonly used to prevent overfitting by adding a penalty term to the cost function to discourage the model from learning complex relationships in the training data that may not generalize well to new data.

B) Regularization adds a penalty term to the cost function that penalizes large weights, forcing the model to choose smaller weights, which can help avoid overfitting.

C) L1 regularization can encourage sparse models by shrinking the weights of some features to zero, effectively removing them from the model, while L2 regularization doesn't necessarily lead to sparsity, but it can still help reduce the impact of outliers in the data.

D) This statement is false, L2 regularization doesn't encourage sparsity, instead it shrinks the weights towards zero but doesn't eliminate them.

E) Regularization can be applied to a wide range of models including linear regression, logistic regression, neural networks, and support vector machines among others.

F) This statement is false, regularization doesn't always improve the performance of a model, it depends on the data and the specific model being used. In some cases, regularization can hurt performance if the penalty term is too high, or if the model is too simple or complex for the given data.

**4.Which of the following statements are true about bias and variance in machine learning?**

A) Bias measures how well a model fits the training data

B) Variance measures how well a model generalizes to new data

C) Bias is caused by underfitting

D) Variance is caused by overfitting

E) Increasing the complexity of a model can increase bias

F) Increasing the complexity of a model can increase variance

Choose all that apply:

**Answer: A, B, C, D, F.**

Explanation:

A) Bias measures how well a model fits the training data and how well it captures the underlying relationships in the data. A high bias model is too simple and may miss important relationships in the data.

B) Variance measures how well a model generalizes to new data, by quantifying how much the model's predictions change when trained on different subsets of the data. A high variance model is too complex and may overfit the training data, leading to poor performance on new data.

C) Bias is caused by underfitting, where the model is too simple and can't capture the underlying relationships in the data, resulting in high training error and poor generalization to new data.

D) Variance is caused by overfitting, where the model is too complex and fits the training data too well, resulting in low training error but high generalization error.

E) Increasing the complexity of a model can increase bias if the model is not expressive enough to capture the underlying relationships in the data, but this is not a universal truth as sometimes increasing model complexity can actually decrease bias.

F) Increasing the complexity of a model can increase variance as the model becomes more flexible and prone to overfitting, leading to high variance and poor generalization to new data.

**5.Which of the following statements are true about AutoML?**

A) AutoML is a set of techniques and tools for automating the machine learning process

B) AutoML can automatically select the best model architecture for a given task

C) AutoML can automatically preprocess and transform the data

D) AutoML can automatically optimize hyperparameters

E) AutoML can replace the need for human expertise in machine learning

F) AutoML always produces better models than those created manually

Choose all that apply:

**Answer: A, B, C, D**

Explanation:

A) AutoML refers to a set of techniques and tools that automate various parts of the machine learning process, such as data preprocessing, feature engineering, model selection, hyperparameter tuning, and model deployment, among others.

B) AutoML can automatically select the best model architecture for a given task by trying out different model families or architectures and selecting the one that performs best on a validation set or using Bayesian optimization.

C) AutoML can automatically preprocess and transform the data, such as handling missing values, scaling the features, encoding categorical variables, among others.

D) AutoML can automatically optimize hyperparameters, such as learning rate, regularization strength, batch size, and others, using various techniques such as grid search, random search, and Bayesian optimization.

E) AutoML can help automate some parts of the machine learning process, but it doesn't replace the need for human expertise in machine learning, as domain knowledge and critical thinking are still important for interpreting the results and making decisions.

F) This statement is false, AutoML doesn't always produce better models than those created manually, as it depends on various factors such as the quality of the data, the complexity of the task, and the expertise of the human in charge of the machine learning process.