Regression in statistics

Python

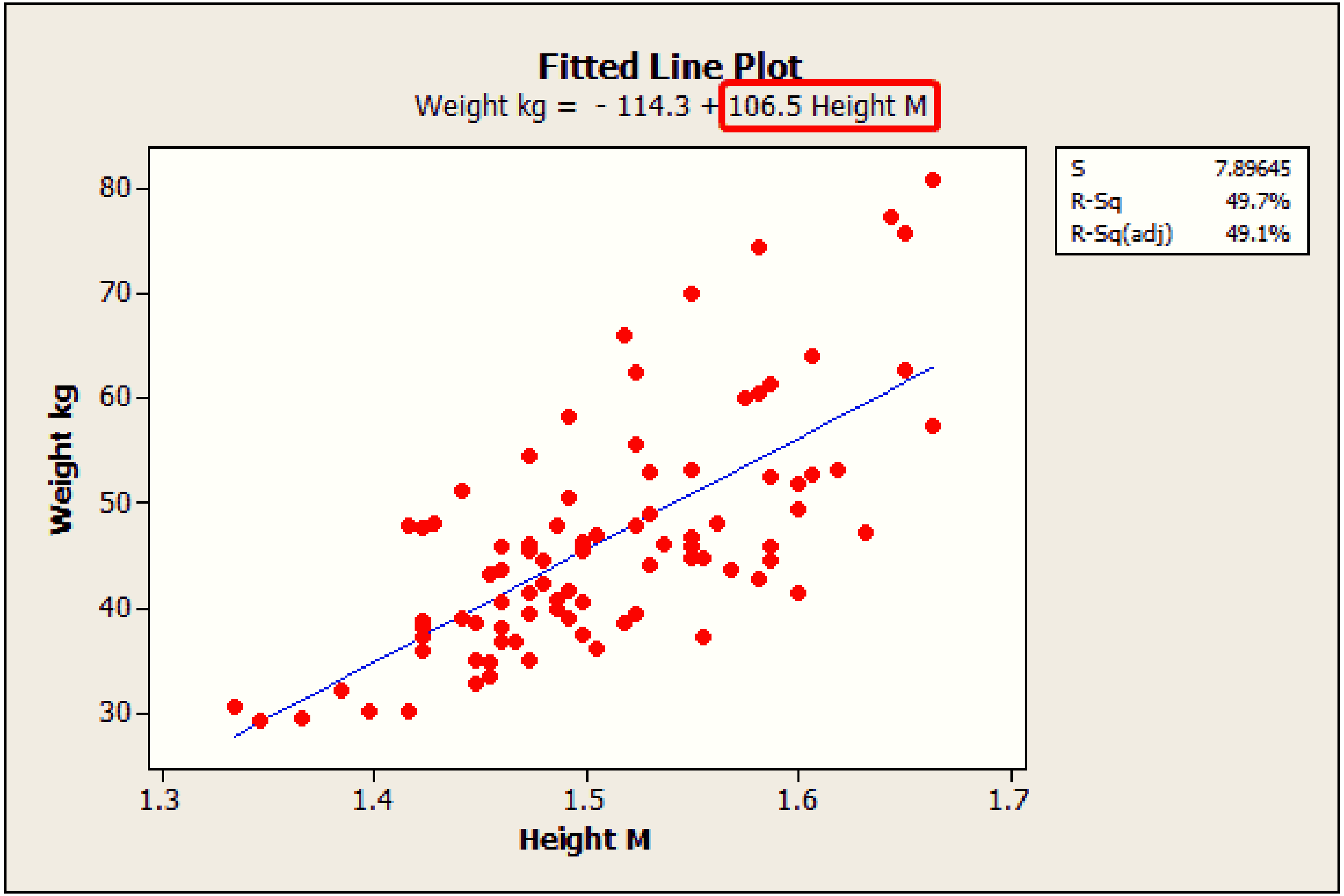
Regression in statistics

- Regression in statistics refers to a set of techniques used to model the relationship between a dependent variable and one or more independent variables. It is a widely used statistical method for understanding and predicting the behaviour of a dependent variable based on the values of independent variables.
- In regression analysis, the dependent variable is often referred to as the target variable or the response variable, while the independent variables are called predictor variables or features. The goal is to find the best-fitting regression model that describes the relationship between the predictors and the target variable.
- The most common type of regression is linear regression, where the relationship between the variables is assumed to be linear. In linear regression, the goal is to find a linear equation that minimizes the difference between the predicted values and the actual values of the target variable.

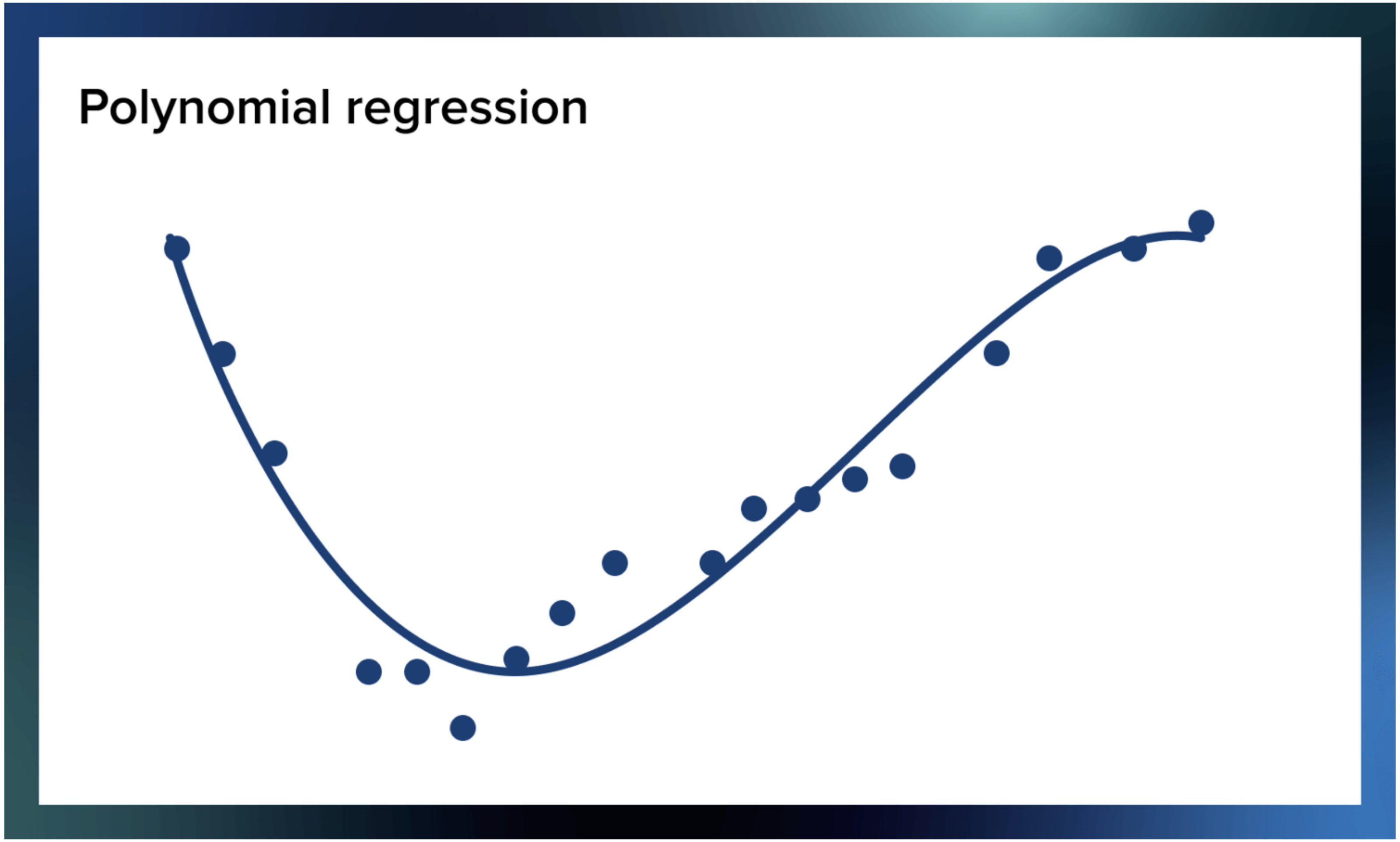
Regression in statistics

- There are various types of regression techniques that extend beyond simple linear regression, including:
  - 1. Multiple Linear Regression: It involves multiple independent variables to predict the target variable using a linear equation.
  - 2. Polynomial Regression: It allows for modeling relationships that are not strictly linear by using polynomial functions of the predictors.
  - 3. Logistic Regression: It is used when the target variable is binary or categorical. It models the probability of a binary outcome based on the values of the predictors.
  - 4. Ridge Regression and Lasso Regression: These are regularization techniques used to prevent overfitting in regression models by introducing a penalty term for the coefficients.
  - 5. Time Series Regression: It involves analyzing time-dependent data and predicting future values based on historical patterns.
- Regression analysis provides valuable insights into the relationship between variables, helps in understanding the impact of predictors on the target variable, and enables predictions and forecasting. It is widely used in various fields such as economics, finance, social sciences, marketing, and machine learning.

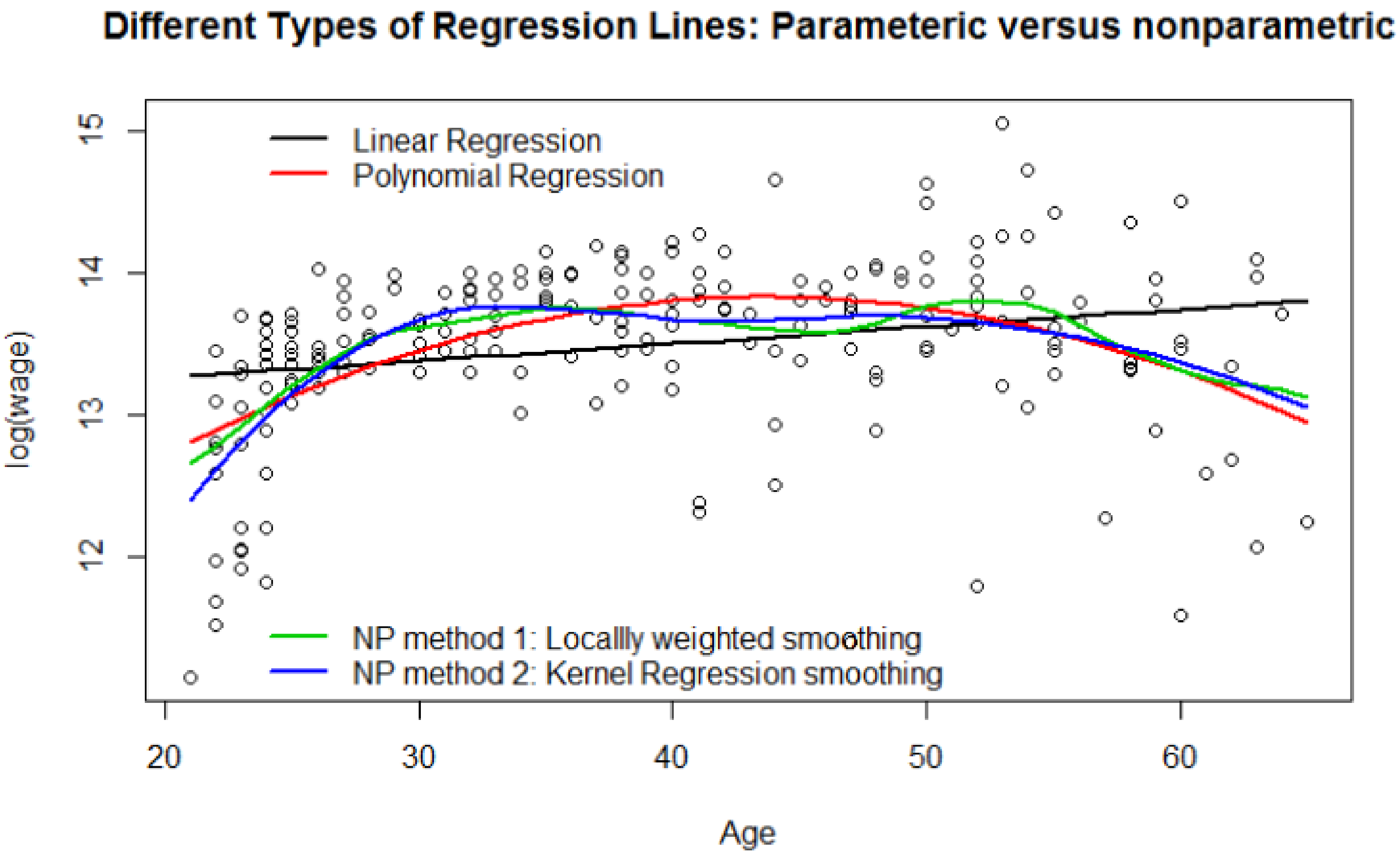
Regression in statistics



Regression in statistics



Regression in statistics



Linear regression

Python

Linear regression

- Linear regression is a statistical technique used to model the relationship between a dependent variable and one or more independent variables. It assumes a linear relationship between the predictors and the target variable. The goal of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that minimizes the difference between the predicted values and the actual values of the dependent variable.
- In simple linear regression, there is a single independent variable (predictor) and a single dependent variable (target). The relationship between the two is represented by the equation of a straight line:

$$y = mx + b$$

- where y is the dependent variable, x is the independent variable, m is the slope (regression coefficient), and b is the y-intercept. The slope represents the change in the dependent variable for a unit change in the independent variable.
- The parameters (slope and intercept) of the linear regression model are estimated using a method called ordinary least squares (OLS). OLS minimizes the sum of the squared differences between the predicted values and the actual values, resulting in the line that best fits the data.
- Multiple linear regression extends the concept to include multiple independent variables. The relationship between the predictors and the target variable is represented by a linear equation:

$$y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots + \beta_n * x_n,$$

where y is the dependent variable, β_0 is the intercept, $x_1 . x_2, \ldots, x_n$ are the independent variables, and $\beta_1, \beta_2, \ldots, \beta_n$are the regression coefficients for each predictor.

OLS - assumptions

- Linear regression relies on several assumptions to ensure the validity of the model and the interpretation of the results. Violations of these assumptions can affect the reliability and accuracy of the regression analysis. Here are the key assumptions of linear regression:
  - 1. Linearity: The relationship between the independent variables and the dependent variable is assumed to be linear. This means that the regression model represents a straight-line relationship between the variables.
  - 2. Independence: The observations in the dataset are assumed to be independent of each other. This assumption implies that the values of the dependent variable for one observation do not depend on or influence the values of the dependent variable for other observations.
  - 3. Homoscedasticity: The variability (or dispersion) of the residuals (the differences between the observed and predicted values of the dependent variable) should be constant across all levels of the independent variables. In other words, the spread of the residuals should be the same throughout the range of the independent variables.
  - 4. Normality: The residuals are assumed to follow a normal distribution. This assumption allows for reliable hypothesis testing, confidence interval estimation, and model inference. It is particularly important when using statistical tests or constructing confidence intervals for the regression coefficients.
  - 5. No multicollinearity: The independent variables should not be highly correlated with each other. High multicollinearity can lead to unstable estimates of the regression coefficients and difficulties in interpreting the individual effects of the independent variables.
  - 6. No endogeneity: There should be no endogeneity or reverse causality, meaning that the dependent variable should not cause the independent variables. Endogeneity can lead to biased coefficient estimates and undermine the interpretation of the regression results.
- It is important to assess these assumptions when conducting linear regression analysis. Various techniques and diagnostic tools, such as residual analysis, normality tests, and checking for multicollinearity, can help evaluate whether the assumptions hold. If violations are detected, appropriate corrective measures or alternative regression approaches may be necessary.

Linear regression

- Linear regression analysis involves several key steps:
  - 1. Data Preparation: Gather and prepare the data, ensuring it meets the assumptions of linear regression, such as linearity, independence, homoscedasticity, and absence of multicollinearity.
  - 2. Model Building: Select the appropriate variables and specify the form of the linear regression model (simple or multiple). Fit the model to the data using the OLS method.
  - 3. Model Evaluation: Assess the quality of the model by examining statistical measures such as the coefficient of determination (R-squared), p-values of the coefficients, standard error, and residual analysis.
  - 4. Predictions and Inferences: Use the fitted model to make predictions on new data and draw inferences about the relationship between the predictors and the dependent variable.
- Linear regression is widely used in various fields, including economics, social sciences, finance, marketing, and machine learning. It provides insights into the impact of independent variables on the dependent variable and can be utilized for prediction, forecasting, and understanding relationships in the data.

Linear regression in Python

In Python, you can perform linear regression using the statsmodels or scikit-learn libraries. Here's an example of performing linear regression using both libraries:

Linear regression – most important measures

- When performing linear regression, there are several important metrics that can be used to evaluate the quality of the model and assess its performance. Here are some of the most commonly used metrics:
  - 1. R-squared (R^2): R-squared measures the proportion of the variance in the dependent variable that is explained by the independent variables. It ranges from 0 to 1, where a value of 1 indicates a perfect fit. However, R-squared alone does not provide information about the goodness of fit or the predictive power of the model.
  - 2. Adjusted R-squared: Adjusted R-squared adjusts the R-squared value by the number of predictors in the model. It penalizes the addition of unnecessary variables and provides a more reliable measure of the model's goodness of fit.
  - 3. Mean Squared Error (MSE): MSE measures the average squared difference between the predicted and actual values of the dependent variable. It provides an overall measure of the model's accuracy, with lower values indicating better fit. However, MSE is not easily interpretable as it is on a different scale than the dependent variable.
  - 4. Root Mean Squared Error (RMSE): RMSE is the square root of MSE, which gives the error metric in the same unit as the dependent variable. It provides a more interpretable measure of the average prediction error.
  - 5. Mean Absolute Error (MAE): MAE measures the average absolute difference between the predicted and actual values of the dependent variable. It is less sensitive to outliers compared to MSE and provides a more intuitive measure of the model's performance.
  - 6. Residuals: Residuals are the differences between the observed and predicted values of the dependent variable. Analyzing the distribution of residuals can help assess the assumptions of linear regression, such as linearity, normality, and homoscedasticity.
  - 7. p-values: The p-values associated with the regression coefficients indicate the statistical significance of each predictor. Lower p-values suggest a stronger relationship between the predictor and the dependent variable. It is important to consider these values to determine the significance of the predictors in the model.
- These metrics provide valuable insights into the performance and quality of a linear regression model. It is recommended to assess multiple metrics to gain a comprehensive understanding of the model's effectiveness and to validate its assumptions.

Linear Regression with statsmodels

```
import statsmodels.api as sm
import pandas as pd

# Prepare the data
data = pd.read_csv('data.csv')
X = data[['x1', 'x2', 'x3']]  # Independent variables
y = data['y']  # Dependent variable

# Add a constant column to the independent variables matrix
X = sm.add_constant(X)

# Fit the model
model = sm.OLS(y, X)
results = model.fit()

# Print the model summary
print(results.summary())

# Get the estimated coefficients
coefficients = results.params
print(coefficients)
```

Linear Regression with statsmodels

- In this example, we first import the necessary libraries and load the data into a Pandas DataFrame. We then specify the independent variables X and the dependent variable y. The add_constant function is used to add a constant column to the independent variables matrix, which accounts for the intercept term in the linear regression model.
- Next, we create an Ordinary Least Squares (OLS) model using sm.OLS and fit it to the data using the fit method. We can then print the summary of the model using results.summary(), which provides detailed statistics about the regression.

Linear Regression with scikit-learn

- Finally, we can access the estimated coefficients of the model using results.params and perform further analysis or predictions using the obtained coefficients.

Linear Regression with scikit-learn

```
from sklearn.linear_model import LinearRegression
import pandas as pd

# Prepare the data
data = pd.read_csv('data.csv')
X = data[['x1', 'x2', 'x3']]  # Independent variables
y = data['y']  # Dependent variable

# Create and fit the model
model = LinearRegression()
model.fit(X, y)

# Get the coefficients and intercept
coefficients = model.coef_
intercept = model.intercept_

# Print the coefficients and intercept
print('Coefficients:', coefficients)
print('Intercept:', intercept)
```

Linear Regression with scikit-learn

- In this example, we import the LinearRegression class from scikit-learn. After loading the data into a Pandas DataFrame, we define the independent variables X and the dependent variable y.
- We create an instance of the LinearRegression model and fit it to the data using the fit method. The coef_ attribute provides the coefficients of the linear regression model, and the intercept_ attribute gives the intercept term.
- You can use these coefficients and the intercept for making predictions or further analysis.
- Both statsmodels and scikit-learn offer various additional functionalities for linear regression, such as handling categorical variables, evaluating model performance, and dealing with multicollinearity. You can refer to their documentation for more details on advanced usage and customization.