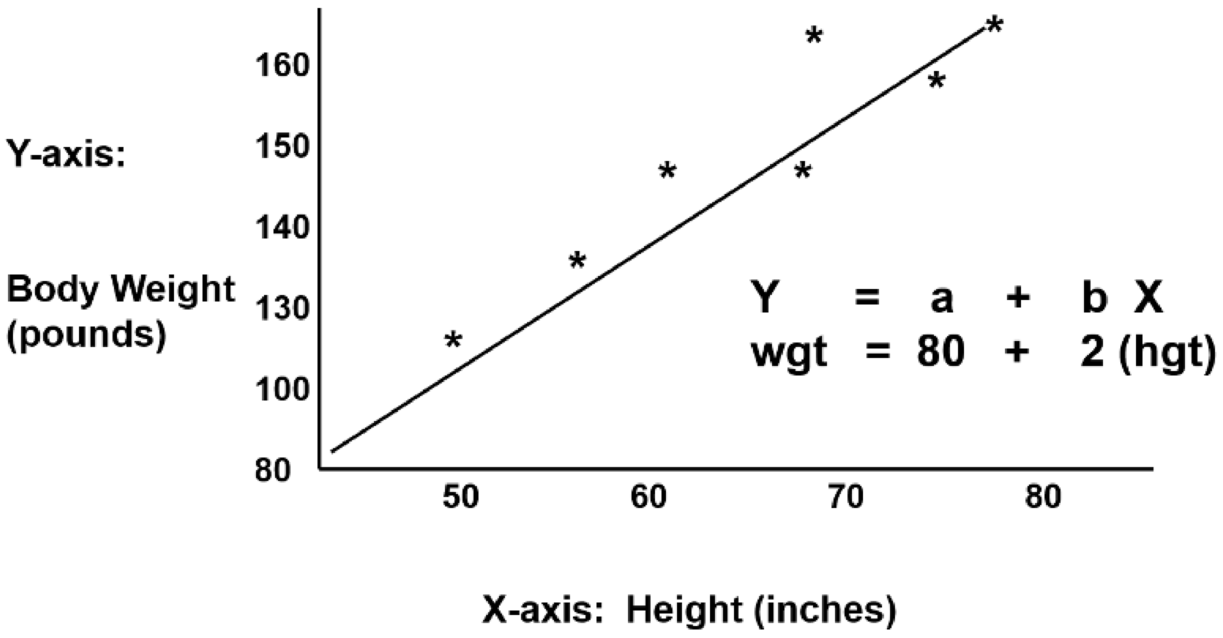
Lesson 2: Linear Regression

Linear regression is a fundamental concept in machine learning and statistics. It is used to model the relationship between a dependent variable and one or more independent variables. In this chapter, we will explore the basic concepts of linear regression. Linear regression is a simple yet powerful tool that can be used in a wide range of machine learning and statistical applications, making it an essential technique for any data scientist or machine learning practitioner.

Simple Linear Regression

Simple linear regression is a statistical technique used to model the relationship between a dependent variable and a single independent variable. The goal of this technique is to find the best linear relationship between the variables, which can then be used to make predictions about the dependent variable.

The assumptions of linear regression include that the relationship between the variables is linear, the errors are normally distributed, and the variance of the errors is constant across all levels of the independent variable. These assumptions should be checked before fitting the linear regression model.



To fit a simple linear regression model, we first need to collect data on both the dependent and independent variables. We then use a method called ordinary least squares to estimate the coefficients of the linear equation that best fits the data. This involves minimizing the sum of the squared errors between the predicted values and the actual values.

Once we have fitted the model, we can interpret the results by examining the coefficients of the equation. The intercept represents the predicted value of the dependent variable when the independent variable is zero, while the slope represents the change in the dependent variable for each one-unit increase in the independent variable.

Simple linear regression has a wide range of real-world applications, including in finance, economics, and engineering. For example, it can be used to predict the price of a house based on its size or to estimate the amount of rainfall based on the temperature.

EXAMPLE CODE

Here is an example code for implementing simple linear regression in Python using the **LinearRegression** class from the **sklearn** library. This code fits a linear regression model to sample data with one independent variable **x** and one dependent variable **y**. It retrieves the intercept and slope of the linear equation and makes a prediction for a new value of **x**.

```
import numpy as np
from sklearn.linear_model import LinearRegression


# Sample data
x = np.array([5, 10, 15, 20, 25]).reshape((-1, 1))
y = np.array([10, 20, 30, 40, 50])


# Create a linear regression model and fit the data
```

```
model = LinearRegression().fit(x, y)


# Print the coefficients of the linear equation
print('Intercept:', model.intercept_)
print('Slope:', model.coef_[0])


# Predict the value of y for a new value of x
new_x = [[30]]
print('Predicted y for x = 30:', model.predict(new_x))
```

Multiple Linear Regression

Multiple linear regression is a powerful tool for modeling the relationship between a dependent variable and multiple independent variables. To use this technique, it is important to consider the assumptions of multiple linear regression, including linearity, independence, homoscedasticity, and normality. Violations of these assumptions can affect the accuracy of the model, so it is important to diagnose and address these issues.

To fit a multiple linear regression model, we use least squares regression to estimate the coefficients of the model. The coefficient of determination (R-squared) measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. The coefficients can be interpreted to understand the relationship between each independent variable and the dependent variable and can be used to make predictions.

Multiple linear regression has a wide range of real-world applications, such as predicting housing prices based on factors like location, square footage, and number of bedrooms, or predicting sales based on factors like advertising spend, seasonality, and pricing strategies. By understanding the concepts and techniques of multiple linear regression, we can apply this powerful tool to solve problems in various industries.

EXAMPLE CODE

The following Python code example demonstrates how to fit a multiple linear regression model using the statsmodels library in Python. This example assumes that the data is stored in a CSV file, and demonstrates how to load the data, define the dependent and independent variables, and fit the model using the ordinary least squares (OLS) method. The example also shows how to add a constant column to the independent variables using the add_constant function, and how to print a summary of the model using the summary method.

```python
import pandas as pd
import numpy as np
import statsmodels.api as sm


# load data
data = pd.read_csv('data.csv')

# define dependent and independent variables
y = data['sales']
X = data[['TV', 'radio', 'newspaper']]

# add constant column to independent variables
X = sm.add_constant(X)

# fit multiple linear regression model
model = sm.OLS(y, X).fit()

# print model summary
print(model.summary())
```