Course Outline: Regression Analysis

Muhammed Sezer & Şevval Belkıs Dikkaya

1 Introduction to Regression

Regression analysis is a powerful statistical technique used to model the relationship between a dependent variable and one or more independent variables. It allows us to make predictions about the dependent variable based on the values of the independent variables.

In this course, we will cover the basics of regression analysis, including simple linear regression and multiple linear regression. We will also explore the assumptions and limitations of regression analysis, and learn how to interpret the results of a regression model.

By the end of this course, you will be able to build and interpret regression models, understand the strengths and limitations of regression analysis, and be able to apply these techniques to real-world problems. So, let's get started!

1.1 Definition of Regression

Regression analysis is a statistical method used to analyze the relationship between one or more independent variables and a dependent variable. It is used to predict the value of a dependent variable based on the value of one or more independent variables.

1.2 Importance of Regression

Regression analysis is important because it can help identify the strength and direction of the relationship between variables. This information can be used to make predictions and inform decision-making. Additionally, regression analysis can be used to test hypotheses about the relationship between variables.

1.3 Brief History of Regression

Regression analysis has a long history, with roots in the work of Francis Galton in the late 1800s. In the early 1900s, Karl Pearson developed the concept of correlation and expanded on Galton's work. In the 1930s, Ronald Fisher developed the method of least squares, which is still widely used in regression analysis today.

1.4 Types of Problems that can be Solved using Regression Analysis

Regression analysis can be used to solve a wide range of problems, including:

- Predictive modeling: predicting the value of a dependent variable based on the values of one or more independent variables
- Causal inference: determining the causal relationship between variables
- Forecasting: predicting future values of a dependent variable
- Trend analysis: analyzing changes in a dependent variable over time
- Quality control: identifying factors that contribute to variation in a dependent variable

1.5 Applications of Regression in Various Fields

Regression analysis is used in a variety of fields, including:

- Economics: analyzing the relationship between economic variables, such as GDP and unemployment
- Finance: predicting stock prices based on market trends and other variables
- Marketing: predicting consumer behavior based on demographic and behavioral variables
- Healthcare: analyzing the relationship between patient characteristics and health outcomes
- Education: predicting academic performance based on demographic and educational variables

Examples of specific applications of regression analysis include:

- Linear regression: predicting the price of a house based on its size, location, and other variables
- Logistic regression: predicting whether a customer will churn based on their purchase history and other variables
- Poisson regression: predicting the number of accidents on a given stretch of road based on traffic volume and other variables

2 Theoretical Background of Regression

In regression analysis, the goal is to estimate the relationship between a dependent variable (Y) and one or more independent variables ($X_1, X_2, ..., X_p$). This relationship is usually expressed as an equation, which is called the regression equation. The regression equation is used to predict the value of the dependent variable based on the values of the independent variables.

The general form of the regression equation is given as follows:

$$Y = f(X_1, X_2, ..., X_p) + \epsilon \tag{1}$$

where $f(X_1, X_2, ..., X_p)$ is the functional form of the relationship between the dependent variable and the independent variables, and ϵ is the error term, which captures the random variation in the dependent variable that is not explained by the independent variables.

2.1 Linear Regression Models and Assumptions

Linear regression is a type of regression analysis that models the relationship between a dependent variable and one or more independent variables using a linear equation. The linear regression equation is given as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{2}$$

where β_0 is the intercept, $\beta_1, \beta_2, ..., \beta_p$ are the coefficients, and $X_1, X_2, ..., X_p$ are the independent variables.

The assumptions of linear regression models include:

- Linearity: The relationship between the dependent variable and the independent variables is linear.
- Independence: The error terms are independent of each other.
- Homoscedasticity: The variance of the error terms is constant across all levels of the independent variables.
- Normality: The error terms are normally distributed.

2.2 Polynomial Regression Models and Assumptions

Polynomial regression is a type of regression analysis that models the relationship between a dependent variable and one or more independent variables using a polynomial equation. The polynomial regression equation is given as follows:

$$\begin{aligned} Y = \beta_0 &+ \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \\ &+ \beta_{p+1} X_1^2 + \beta_{p+2} X_2^2 + ... + \beta_{2p} X_p^2 \\ &+ \beta_{2p+1} X_1 X_2 + ... + \beta_{(k-1)p+1} X_1^{k-1} + ... \\ &+ \beta_{kp} X_p^{k-1} + \epsilon \end{aligned} \tag{3}$$

where β_0 is the intercept, $\beta_1, \beta_2, ..., \beta_p$ are the coefficients of the first-order terms, $\beta_{p+1}, \beta_{p+2}, ..., \beta_{2p}$ are the coefficients of the second-order terms, $\beta_{2p+1}, ..., \beta_{(k-1)p+1}$ are the coefficients of the interaction terms, and β_{kp} are the coefficients of the kth-order terms. $X_1, X_2, ..., X_p$ are the independent variables, and ϵ is the error term.

The assumptions of polynomial regression models are similar to those of linear regression models and include:

- Linearity: The relationship between the dependent variable and the independent variables is linear.
- Independence: The error terms are independent of each other.
- Homoscedasticity: The variance of the error terms is constant across all levels of the independent variables.
- Normality: The error terms are normally distributed.

However, in polynomial regression models, the assumption of linearity requires that the relationship between the dependent variable and the independent variables be linear in terms of the transformed variables, not necessarily in terms of the original variables. Additionally, the assumption of independence applies to the error terms after the polynomial terms have been added to the model.

2.3 Logistic Regression Models and Assumptions

Logistic regression is a type of regression analysis used to model the relationship between a binary dependent variable and one or more independent variables using a logistic function. The logistic regression equation is given as follows:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}} \tag{4}$$

where β_0 is the intercept, $\beta_1, \beta_2, ..., \beta_p$ are the coefficients, and $X_1, X_2, ..., X_p$ are the independent variables.

The logistic function is a sigmoidal function that ranges from 0 to 1, representing the probability of the dependent variable being in the "success" category. One advantage of logistic regression is that its derivative is easy to calculate and is defined, which is useful for optimization algorithms that require gradient information.

The assumptions of logistic regression models include:

- Independence: The observations are independent of each other.
- Linearity of independent variables and log odds: The relationship between the independent variables and the log odds of the dependent variable is linear.

- Absence of multicollinearity: The independent variables are not highly correlated with each other.
- Large sample size: The sample size is large enough to ensure stable parameter estimates.

Note that normality and homoscedasticity are not assumptions of logistic regression models since the dependent variable is binary.

2.4 Ridge Regression Models and Assumptions

Ridge regression is a type of linear regression that incorporates regularization to address the problem of multicollinearity in the independent variables. The ridge regression equation is given as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{5}$$

where β_0 is the intercept, $\beta_1, \beta_2, ..., \beta_p$ are the coefficients, and $X_1, X_2, ..., X_p$ are the independent variables.

The ridge regression model introduces a regularization term, which adds a penalty for large coefficients. The objective function to be minimized is:

$$\sum_{i=1}^{n}(y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}))^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{6}$$

where λ is the regularization parameter, which determines the strength of the penalty term.

The assumptions of ridge regression models are similar to those of linear regression models, with some differences:

- Linearity: The relationship between the dependent variable and the independent variables is linear.
- Independence: The error terms are independent of each other.
- Homoscedasticity: The variance of the error terms is constant across all levels of the independent variables.
- Normality: The error terms are normally distributed.
- Multicollinearity: Ridge regression addresses multicollinearity by introducing the regularization term, which shrinks the coefficients of highly correlated variables toward each other.

Ridge regression is particularly useful when the independent variables are highly correlated, as it helps to stabilize the parameter estimates and prevent overfitting.

2.5 Time Series Regression Models and Assumptions

Time series regression is a type of regression analysis that models the relationship between a dependent variable and one or more independent variables, considering the temporal ordering of the observations. The time series regression equation is given as follows:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_p X_{pt} + \epsilon_t \tag{7}$$

where Y_t is the dependent variable at time t, β_0 is the intercept, $\beta_1, \beta_2, ..., \beta_p$ are the coefficients, $X_{1t}, X_{2t}, ..., X_{pt}$ are the independent variables at time t, and ϵ_t is the error term at time t.

The assumptions of time series regression models include:

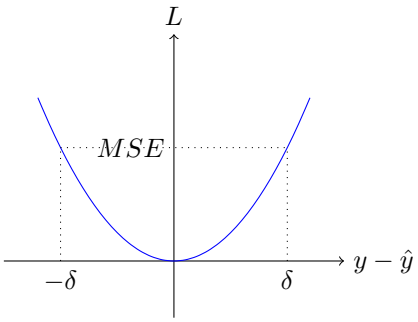
- Linearity: The relationship between the dependent variable and the independent variables is linear.
- Stationarity: The time series data should be stationary, i.e., the mean, variance, and autocorrelation structure do not change over time.
- No multicollinearity: The independent variables should not be highly correlated with each other.
- No autocorrelation: The error terms should not exhibit autocorrelation, meaning that the error term at time t should not be correlated with the error term at any other time.
- Homoscedasticity: The variance of the error terms is constant across all levels of the independent variables.
- Normality: The error terms are normally distributed.

Time series regression models often incorporate lags of the dependent variable and/or the independent variables to account for temporal dependencies in the data. These models can be useful for understanding the effects of past values on the current value of the dependent variable and for forecasting future values of the dependent variable.

2.6 Error Functions in Regression Models

Error functions are used to measure the difference between the predicted values and the actual values in a regression model. The goal is to minimize this difference to obtain the best-fit line. There are several types of error functions, each with its own advantages and disadvantages. Here are some commonly used error functions in regression models:

2.6.1 Mean Squared Error (MSE)



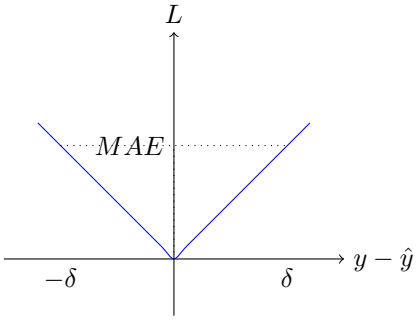
The mean squared error (MSE) is a commonly used error function in regression models. It measures the average of the squared differences between the predicted values and the actual values:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

where n is the number of observations, y_i is the actual value of the dependent variable for observation i, and $\hat{y}_i$ is the predicted value of the dependent variable for observation i.

The advantage of MSE is that it penalizes large errors more than small errors, making it useful in cases where we want to reduce the impact of outliers on the model.

2.6.2 Mean Absolute Error (MAE)



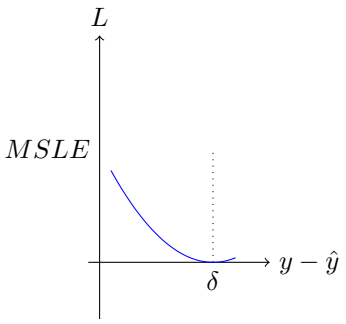
The mean absolute error (MAE) is another commonly used error function in regression models. It measures the average of the absolute differences between the predicted values and the actual values:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{9}$$

where n is the number of observations, y_i is the actual value of the dependent variable for observation i, and $\hat{y}_i$ is the predicted value of the dependent variable for observation i.

One advantage of MAE over MSE is that it is less sensitive to outliers, since it penalizes all errors linearly. This can be useful in cases where the dataset has many outliers or the model should not be heavily influenced by them.

2.6.3 Mean Squared Logarithmic Error (MSLE)



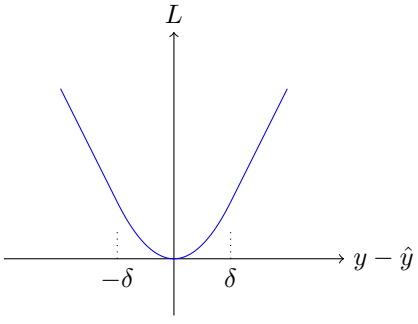
The mean squared logarithmic error (MSLE) is an error function used in regression models when dealing with exponential growth or predicting multiplicative factors. It measures the average of the squared differences between the logarithms of the predicted values and the actual values:

$$MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(1+y_i) - \log(1+\hat{y}_i))^2 \tag{10}$$

where n is the number of observations, y_i is the actual value of the dependent variable for observation i, and $\hat{y}_i$ is the predicted value of the dependent variable for observation i.

The advantage of MSLE over MSE is that it is less sensitive to large differences between the predicted and actual values when both are large. This can be useful when modeling growth rates or multiplicative factors.

2.6.4 Huber Loss



The Huber loss is a combination of the mean squared error (MSE) and the mean absolute error (MAE). It is quadratic for errors smaller than a certain threshold (usually represented as δ) and linear for errors larger than that threshold. This loss function is less sensitive to outliers than the MSE:

$$HuberLoss(y, \hat{y}) = \begin{cases} \frac{1}{2}(y - \hat{y})^2, & \text{if } |y - \hat{y}| \leq \delta \\ \delta(|y - \hat{y}| - \frac{1}{2}\delta), & \text{otherwise} \end{cases} \tag{11}$$

The advantages of the Huber loss function include: - It is less sensitive to outliers than the MSE, as it uses a linear function for large errors. - It retains the differentiability property of the MSE for small errors, which is useful in optimization algorithms that require gradient information. - By adjusting the δ parameter, we can control the transition point between the quadratic and linear regions of the loss function, allowing us to fine-tune the model's sensitivity to outliers.

2.7 Model Selection and Loss Function Evaluation

Selecting the appropriate model and loss function is crucial for obtaining accurate and reliable predictions. This section will discuss how to choose a suitable model, select the right loss function, and evaluate the goodness of fit.

2.7.1 Model Selection

When selecting a model, one should consider the following aspects:

1. The nature of the data: Consider the type of data you are working with (e.g., continuous, binary, count data) and the relationships between variables (linear, nonlinear, time-dependent).

2. Model complexity: Choose a model that is neither too simple nor too complex. A simple model may not capture the underlying relationships in the data, while a complex model may overfit the data.

3. Interpretability: In some cases, it is essential to have a model that is easy to interpret and understand.

4. Cross-validation: Use techniques like k-fold cross-validation to estimate the model's performance on unseen data.

2.7.2 Loss Function Selection

When selecting a loss function, consider the following aspects:

1. The type of problem: Different loss functions are suitable for different types of problems, such as regression, classification, or ranking.

2. The distribution of the data: Some loss functions may be more suitable for data with specific characteristics, such as skewed or heavy-tailed distributions.

3. The impact of outliers: If your data contains outliers, you may want to choose a robust loss function that is less sensitive to extreme values.

2.7.3 Goodness of Fit Evaluation

To evaluate the goodness of fit of a model, consider the following criteria:

1. R-squared and adjusted R-squared: These measures indicate the proportion of the variance in the dependent variable explained by the independent variables in the model.

2. Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC): These criteria balance the goodness of fit against the complexity of the model.

3. Residual analysis: Examine the residuals (errors) of the model to check for patterns or deviations from the assumptions (e.g., linearity, independence, homoscedasticity, and normality).

4. Cross-validation: Use techniques like k-fold cross-validation to estimate the model's performance on unseen data.

3 Practical Use of Regression (30 minutes)

In this section, we will discuss the practical aspects of using regression, starting with data preparation and preprocessing.

3.1 Data Preparation and Preprocessing

3.1.1 Cleaning the Data

The first step in using regression is to clean the data. This involves:

1. Identifying and correcting data entry errors, inconsistencies, and duplicate records.

2. Ensuring that the data is in a suitable format for analysis (e.g., numbers stored as text should be converted to numeric format).

3.1.2 Handling Missing Values

Handling missing values in the dataset is crucial for the reliability of the regression model. Common strategies for dealing with missing values include:

1. Removing records with missing values: This approach can be used if the number of records with missing values is relatively small and their removal does not significantly impact the representativeness of the dataset.

2. Imputation: Replace missing values with a reasonable estimate, such as the mean, median, or mode of the variable. More sophisticated imputation methods, such as k-Nearest Neighbors or multiple imputation, can also be used.

3. Using models that can handle missing values directly: Some regression models, such as decision trees and random forests, can handle missing values without the need for imputation.

3.1.3 Transforming Variables

In some cases, transforming variables may be necessary to meet the assumptions of the regression model or to improve its performance. Common variable transformations include:

1. Logarithmic transformation: Applying the natural logarithm to a variable can help stabilize variance and make the data more normally distributed.

2. Standardization: Scaling the variables to have a mean of 0 and a standard deviation of 1 can help with the interpretation of coefficients and improve the performance of some regression models.

3. Categorical encoding: Converting categorical variables into numerical format using techniques such as one-hot encoding or ordinal encoding.

3.1.4 Variable Selection

Selecting the appropriate variables for the regression model involves:

1. Identifying relevant independent variables based on domain knowledge, literature review, or exploratory data analysis.

2. Checking for multicollinearity between independent variables, which can cause instability in the regression coefficients and make the model difficult to interpret. Methods to detect multicollinearity include calculating the correlation matrix, variance inflation factors (VIF), or using dimensionality reduction techniques such as principal component analysis (PCA).

3. Using variable selection techniques such as stepwise selection, Lasso regression, or recursive feature elimination to identify the most important variables for the model.

3.2 Model Training and Evaluation

3.2.1 Training the Regression Model

After preparing the data and selecting the variables, the next step is to train the regression model on the data. This involves selecting the appropriate regression algorithm, such as linear regression, ridge regression, or LASSO, depending on the problem at hand and the assumptions made about the data.

To improve the model's performance, it's essential to tune the model's hyperparameters. Hyperparameters are external factors that influence the training process, such as learning rate, regularization strength, or the degree of a polynomial in polynomial regression. Grid search, random search, or Bayesian optimization are common methods for hyperparameter tuning.

3.2.2 Evaluating the Model on a Test Set

Once the model is trained, it's essential to evaluate its performance on a test set. The test set should be separate from the training set to ensure the model's performance is assessed on unseen data. This helps to understand how well the model generalizes to new data and prevents overfitting.

3.2.3 Model Evaluation Metrics

Several evaluation metrics can be used to measure the performance of the regression model. Some common metrics include:

- Mean Squared Error (MSE): The average of the squared differences between the predicted values and the actual values. MSE emphasizes larger errors over smaller ones.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{12}$$

- Root Mean Squared Error (RMSE): The square root of the MSE. RMSE has the same units as the dependent variable and provides a more interpretable measure of the model's performance.

$$RMSE = \sqrt{MSE} \tag{13}$$

- Mean Absolute Error (MAE): The average of the absolute differences between the predicted values and the actual values. MAE is less sensitive to outliers than MSE or RMSE and provides a linear penalty for all errors.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{14}$$

When evaluating the model, it's essential to compare its performance to a baseline, such as the performance of a simple linear regression model or the mean of the dependent variable. This helps to understand if the model is adding value beyond simple heuristics.

3.3 Feature Selection and Engineering

3.3.1 Feature Selection

Feature selection is the process of selecting the most relevant variables to include in the regression model. By reducing the number of input features, the model becomes more interpretable, less prone to overfitting, and computationally less expensive.

Some common feature selection techniques include:

- Filter methods: Techniques like correlation analysis and mutual information analysis are used to rank the importance of each feature based on their relationship with the dependent variable.
- Wrapper methods: Recursive feature elimination (RFE) is an example of a wrapper method that iteratively removes the least important features and evaluates the model's performance.
- Embedded methods: Techniques like LASSO regression perform feature selection as part of the model training process by applying a penalty to the model coefficients.

3.3.2 Feature Engineering

Feature engineering is the process of creating new features from existing ones to improve the model's performance. This can involve:

- Transformations: Applying mathematical transformations such as logarithmic, square root, or power transformations to the original features.
- Interaction terms: Creating new features by multiplying two or more existing features to capture their interactions.
- Polynomial features: Generating higher-degree terms of the input features to model more complex relationships.
- Domain-specific features: Creating new features based on domain knowledge and expertise to better capture the underlying pattern in the data.

3.4 Regularization and Overfitting

3.4.1 Regularization

Regularization is a technique used to prevent overfitting of the regression model by adding penalties to the model coefficients. This discourages the model from becoming too complex and fitting the noise in the data instead of the underlying pattern.

Two common regularization techniques are:

- L1 regularization (LASSO): Adds the sum of the absolute values of the coefficients as a penalty term. This can result in some coefficients being set to zero, effectively performing feature selection.
- L2 regularization (Ridge): Adds the sum of the squared values of the coefficients as a penalty term. This can shrink the coefficients towards zero, but typically does not set them to exactly zero.

3.4.2 Overfitting

Overfitting occurs when the model is too complex and captures the noise in the data instead of the underlying pattern. This results in a model that performs well on the training data but poorly on new, unseen data.

To prevent overfitting:

- Regularization techniques can be applied to discourage complex models.
- Feature selection can be used to reduce the number of input features and the model's complexity.
- Cross-validation can be used to evaluate the model's performance on different subsets of the data, ensuring that the model generalizes well to unseen data.

3.5 Practical Applications of Regression

3.5.1 Finance and Economics

Regression analysis is widely used in finance and economics to model relationships between variables and make predictions. Some examples include:

- Predicting stock prices: Regression models can be used to predict the future prices of stocks based on factors such as past prices, trading volume, and market sentiment.
- Estimating customer lifetime value: Regression models can help estimate the total value a customer brings to a business over their entire relationship, based on factors like demographics, purchase history, and customer interactions.
- Analyzing the impact of advertising on sales: Regression models can be used to determine the relationship between advertising spend and sales revenue, allowing businesses to optimize their advertising strategies for maximum return on investment.

3.5.2 Engineering and Natural Sciences

Regression models are also used in engineering and natural sciences to model complex systems and predict outcomes based on input variables. Some examples include:

- Modeling the relationship between weather and crop yield: Regression models can be used to understand the impact of weather variables such as temperature, precipitation, and solar radiation on crop yields, enabling more effective agricultural planning.
- Predicting equipment failure: Regression models can help predict the failure of equipment or machinery based on factors such as age, usage, and maintenance history, allowing for proactive maintenance and reduced downtime.
- Estimating pollutant emissions: Regression models can be used to estimate the emissions of pollutants from industrial processes, transportation, and other sources based on factors like production levels and fuel consumption.

3.5.3 Machine Learning Applications

Regression is a fundamental concept in machine learning, with numerous practical applications across various domains:

- Predicting housing prices: Regression models can be used to predict housing prices based on features such as square footage, location, and the number of bedrooms and bathrooms.
- Credit risk assessment: Regression models can help assess the creditworthiness of borrowers based on factors like income, credit history, and debt-to-income ratio, enabling better lending decisions.
- Image and speech recognition: Regression models can be used in combination with deep learning techniques to recognize objects in images or transcribe spoken words into text.

In summary, regression analysis is a versatile and powerful tool used across various fields to model relationships between variables, make predictions, and inform decision-making.

4 Intuition for Regression (30 minutes)

4.1 Identifying Regression Problems

4.1.1 When to Apply Regression

The first step in using regression is to identify when it can be applied to a problem. Regression is most suited for problems where there is a relationship

between two or more variables, and we want to predict the value of one variable (the dependent variable) based on the values of the other variables (the independent variables).

4.1.2 Examples of Regression Problems

To understand when to apply regression, let's look at some examples of problems where regression can be applied:

- **Predicting housing prices:** In the real estate industry, we might want to predict housing prices based on features such as square footage, number of bedrooms, location, and other factors. In this case, the dependent variable is the housing price, and the independent variables are the various features of the house.
- **Predicting the weight of a person:** In the field of health and wellness, we may want to predict a person's weight based on their height and age. In this example, the dependent variable is the person's weight, while the independent variables are their height and age.
- **Forecasting product demand:** In supply chain management, we might want to predict product demand based on factors such as seasonality, promotional activities, and historical sales data. Here, the dependent variable is product demand, and the independent variables are the factors affecting demand.

In each of these examples, we can see that regression is a suitable approach because there is a relationship between the dependent variable and one or more independent variables. By understanding the nature of the problem and the relationships between variables, we can determine when it is appropriate to use regression for prediction and analysis.

4.2 Types of Regression Problems

4.2.1 Simple Linear Regression

Simple linear regression is used when there is a linear relationship between one independent variable and the dependent variable. In this case, the goal is to find the best-fitting straight line that represents the relationship between the two variables. The equation for simple linear regression is:

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{15}$$

where Y is the dependent variable, X is the independent variable, β_0 is the intercept, β_1 is the coefficient, and ϵ is the error term.

4.2.2 Multiple Linear Regression

Multiple linear regression is used when there is a linear relationship between multiple independent variables and the dependent variable. The goal is to find the best-fitting hyperplane that represents the relationship between the variables. The equation for multiple linear regression is:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{16}$$

where Y is the dependent variable, $X_1, X_2, ..., X_p$ are the independent variables, β_0 is the intercept, $\beta_1, \beta_2, ..., \beta_p$ are the coefficients, and ϵ is the error term.

4.2.3 Polynomial Regression

Polynomial regression is used when there is a curved relationship between the independent and dependent variables. This type of regression extends simple linear regression by adding higher-degree terms of the independent variable. The equation for polynomial regression is:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + ... + \beta_p X^p + \epsilon \tag{17}$$

where Y is the dependent variable, X is the independent variable, β_0 is the intercept, $\beta_1, \beta_2, ..., \beta_p$ are the coefficients, and ϵ is the error term.

4.2.4 Logistic Regression

Logistic regression is used when the dependent variable is categorical, typically binary (e.g., success/failure, yes/no). The goal is to estimate the probability of an event occurring based on the values of the independent variables. The equation for logistic regression is:

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{18}$$

where $P(Y = 1)$ is the probability of the event occurring (i.e., $Y = 1$), $X_1, X_2, ..., X_p$ are the independent variables, and $\beta_0, \beta_1, \beta_2, ..., \beta_p$ are the coefficients.

Understanding the different types of regression problems and their respective models is crucial for selecting the appropriate method to solve a particular problem based on the nature of the independent and dependent variables.