Course Outline: Regression Analysis
METU EEE A block 209

Muhammed Sezer & Şevval Belkıs Dikkaya

March 27, 2023

1 Introduction to Regression (15 minutes)

a) Definition of regression

- Regression is a statistical method used to establish a relationship between a dependent variable and one or more independent variables.
- The method seeks to find the best-fitting line (or curve) that represents this relationship.

b) Importance of regression

- Regression analysis is important in a variety of fields such as finance, marketing, engineering, and social sciences.
- It helps to understand the relationship between variables and can be used to make predictions or to inform decision-making.

c) Brief history of regression

- The concept of regression was first introduced by Sir Francis Galton in the late 19th century.
- He used the method to study the relationship between the heights of fathers and sons.
- Since then, regression analysis has been extensively studied and developed, and is now widely used in various fields.

d) Types of problems that can be solved using regression analysis

- Regression analysis can be used to solve a variety of problems, such as predicting stock prices, forecasting sales, analyzing the relationship between advertising spend and sales, and understanding the effect of education on income.

e) Applications of regression in various fields

- Regression analysis has wide applications in fields such as finance, marketing, healthcare, social sciences, and engineering.

- It is used to understand and predict customer behavior, to forecast sales and revenues, to optimize manufacturing processes, and to analyze the relationship between environmental factors and health outcomes, among other things.

2 Theoretical Background of Regression (30 minutes)

a) Basic concepts of regression

- Regression analysis is a statistical method used to model the relationship between a dependent variable and one or more independent variables.
- The method is used to estimate the parameters of a mathematical function that describes this relationship.
- The function is used to make predictions or to understand the relationship between the variables.

b) Linear regression models and assumptions

- Linear regression models assume that the relationship between the dependent variable and the independent variables is linear.
- The model is represented by a straight line equation, where the slope and intercept are estimated from the data.
- The assumptions of linear regression include normality of errors, constant variance, and linearity of the relationship.

c) Nonlinear regression models and when to use them

- Nonlinear regression models allow for more complex relationships between the dependent variable and the independent variables.
- They can be used when the relationship between the variables is not linear or when the data exhibits a pattern that cannot be modeled using a linear equation.
- Nonlinear regression models are used to fit curves, surfaces, or other complex functions to the data.

d) Different types of regression models

- There are various types of regression models, each with its own assumptions and applications.
- Simple linear regression is used to model the relationship between two variables.
- Multiple linear regression is used to model the relationship between a dependent variable and two or more independent variables.

- Polynomial regression is used to model nonlinear relationships by fitting a polynomial function to the data.
- Logistic regression is used to model binary or categorical outcomes.
- Time series regression is used to model the relationship between a dependent variable and time.

e) Model selection criteria

- Model selection criteria are used to evaluate the performance of different regression models and to select the best model.
- Some common model selection criteria include R-squared, adjusted R-squared, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC).
- These criteria evaluate the goodness of fit of the model and balance the trade-off between model complexity and accuracy.

3 Practical Use of Regression (30 minutes)

a) Data preparation and preprocessing

- The first step in using regression is to prepare and preprocess the data.
- This involves cleaning the data, handling missing values, and transforming the variables if necessary.
- Data preparation also involves selecting the variables that will be used in the regression model and checking for multicollinearity.

b) Model training and evaluation

- The next step is to train the regression model on the data.
- This involves selecting the appropriate regression algorithm and tuning the model hyperparameters.
- Once the model is trained, it needs to be evaluated on a test set to check its performance.
- Model evaluation metrics include mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE).

c) Feature selection and engineering

- Feature selection involves selecting the most relevant variables to include in the regression model.
- Feature engineering involves creating new features from existing ones to improve the model's performance.

- Some techniques for feature selection and engineering include principal component analysis (PCA), recursive feature elimination (RFE), and polynomial features.

d) Regularization and overfitting

- Regularization is a technique used to prevent overfitting of the regression model.
- Overfitting occurs when the model is too complex and fits the noise in the data instead of the underlying pattern.
- Regularization techniques include L1 and L2 regularization, which add penalties to the model coefficients to discourage overfitting.
- Cross-validation is also used to evaluate the model's performance and to prevent overfitting.

e) Practical applications of regression

- Regression is widely used in various fields, including finance, economics, engineering, and social sciences.
- Some practical applications of regression include predicting stock prices, estimating customer lifetime value, analyzing the impact of advertising on sales, and modeling the relationship between weather and crop yield.
- Regression is also used in machine learning applications such as predicting housing prices, credit risk assessment, and image and speech recognition.

4 Intuition for Regression (30 minutes)

a) Identifying regression problems

- The first step in using regression is to identify when it can be applied to a problem.
- Regression is used when 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).
- Some examples of problems where regression can be applied include predicting housing prices based on features such as square footage and number of bedrooms, and predicting the weight of a person based on their height and age.

b) Types of regression problems

- There are different types of regression problems depending on the nature of the independent and dependent variables.

- Simple linear regression is used when there is a linear relationship between one independent variable and the dependent variable.

- Multiple linear regression is used when there is a linear relationship between multiple independent variables and the dependent variable.

- Polynomial regression is used when there is a curved relationship between the independent and dependent variables.

- Logistic regression is used when the dependent variable is categorical.

c) The role of regression coefficients

- The regression coefficients represent the relationship between the independent variables and the dependent variable.

- In simple linear regression, the coefficient represents the slope of the line.

- In multiple linear regression, the coefficients represent the change in the dependent variable for a one-unit change in the independent variable, holding all other independent variables constant.

- The coefficients can be used to interpret the relationship between the independent and dependent variables and to make predictions.

d) Limitations of regression

- While regression is a powerful tool, it has some limitations.

- One limitation is that it assumes a linear relationship between the independent and dependent variables, which may not be the case in all situations.

- Another limitation is that it assumes that the relationship between the independent and dependent variables is constant across the range of the independent variables, which may not be true in all cases.

- It is important to be aware of these limitations when using regression and to use other tools such as data visualization and exploratory data analysis to assess the relationship between variables.

e) Tips for using regression effectively

- To use regression effectively, it is important to have a good understanding of the data and the problem at hand.

- It is also important to carefully select the independent variables and to preprocess the data appropriately.

- Regularization techniques can be used to prevent overfitting and to improve the generalization of the model.

- Finally, it is important to interpret the results of the regression analysis carefully and to use other tools such as data visualization and exploratory data analysis to validate the results.

5 Conclusion (15 minutes)

a) Recap of key concepts

- In this course, we have covered the basic concepts of regression analysis, including the definition of regression, the theoretical background of regression, practical use cases for regression, and intuition for identifying regression problems and using regression effectively.
- We have also discussed the limitations of regression and the importance of careful data preprocessing and model selection to ensure accurate and reliable results.

b) Importance of regression

- Regression is an important tool for data analysis and machine learning, with applications in a wide range of fields including finance, healthcare, and marketing.
- By understanding the basic concepts of regression and how to apply it effectively, participants will be better equipped to analyze and interpret data, make informed decisions, and develop predictive models.

c) Next steps

- To further develop their skills in regression analysis, participants may want to explore advanced topics such as non-linear regression, time series analysis, and ensemble methods.
- Additionally, participants may want to practice applying regression to real-world datasets and use cases to gain hands-on experience and develop their intuition for identifying regression problems and selecting appropriate models.
- Finally, participants are encouraged to continue learning and staying up-to-date with the latest developments in regression analysis and related fields.

6 Bonus: Applying Regression to Real-World Problems (Never Ends)

a) Introduction

- In this section, we will briefly discuss two real-world problems where regression analysis can be applied: predicting stock prices and predicting housing prices.
- Please note that this section is optional and not part of the main course material. If you are interested in learning more about these topics, we recommend further study and research.

Example 1: Predicting stock prices

- Stock prices are notoriously difficult to predict, but regression analysis can be a useful tool for identifying trends and making informed predictions.
- One approach to predicting stock prices using regression is to analyze historical price data and identify patterns and correlations that can be used to develop a predictive model.
- Factors that may influence stock prices include company financials, industry trends, macroeconomic conditions, and market sentiment, among others.

Example 2: Predicting housing prices

- Another common application of regression analysis is predicting housing prices, which can be useful for homebuyers, real estate investors, and other stakeholders.
- Factors that may influence housing prices include location, property size and condition, local market conditions, and economic trends, among others.
- Regression models for predicting housing prices may incorporate multiple variables and may require careful data preprocessing and model selection to ensure accurate and reliable results.

Conclusion

- In this section, we briefly discussed two real-world problems where regression analysis can be applied: predicting stock prices and predicting housing prices.
- While regression can be a useful tool for making predictions and identifying trends in complex datasets, it is important to be aware of its limitations and to carefully consider data quality, model selection, and other factors that can affect the accuracy and reliability of results.
- If you are interested in learning more about these topics or applying regression to other real-world problems, we recommend further study and research to develop your skills and intuition.