

# Exam SRM Summary Sheet

Last Updated: April 2025

|   |           |
|---|-----------|
| <b>0. Review</b>  | <b>7</b>  |
| 0.1 Sampling Assumptions  | 8         |
| 0.2 Pearson Correlation Coefficient                             | 8         |
| 0.3 Matrices  | 9         |
| 0.3 Maximum Likelihood Estimation                               | 10        |
| <b>1. Basics of Statistical Learning (Learning Objective 1)</b> | <b>12</b> |
| 1.1 Types of Variables  | 13        |
| 1.2 Prediction and Inference                                    | 15        |
| 1.3 Decomposition of the Expected Squared Error                 | 15        |
| 1.4 Parametric and Non-Parametric Methods                       | 16        |
| 1.5 Supervised vs Unsupervised Learning                         | 16        |
| 1.6 Regression vs Classification                                | 17        |
| 1.7 Mean Squared Error and Error Rate                           | 17        |
| 1.8 Bias-Variance Tradeoff                                      | 18        |
| 1.8.1 Definitions   | 18        |
| 1.8.2 Tradeoff Table  | 18        |
| 1.9 Data Collection   | 19        |
| 1.10 Bayes Classifier   | 20        |
| 1.11 K-Nearest Neighbors  | 21        |
| 1.11.1 Algorithm  | 21        |
| 1.11.2 Bias-Variance Tradeoff in KNN                            | 21        |
| 1.12 The Validation Set Approach                                | 22        |
| 1.12.1 Algorithm  | 22        |
| 1.12.2 Pros and Cons  | 22        |
| 1.13 Leave-One-Out Cross-Validation                             | 23        |
| 1.13.1 Algorithm  | 23        |
| 1.13.2 Pros and Cons  | 23        |
| 1.14 K-Fold Cross-Validation                                    | 24        |
| 1.14.1 Algorithm  | 24        |
| 1.14.2 Model Comparison   | 24        |
| <b>2. Linear Models (Learning Objective 2)</b>                  | <b>25</b> |
| 2.1 Simple Linear Regression                                    | 26        |
| 2.1.1 Theoretical Representation of a Linear Model              | 26        |
| 2.1.2 Observations vs Predictions (Actuals vs Estimates)        | 26        |
| 2.1.3 Ordinary Least Squares                                    | 27        |
| 2.1.4 Error Term  | 28        |
| 2.2 Mean Squared Error and Standard Error                       | 29        |
| 2.3 Sum of Squares and R-Squared                                | 30        |
| 2.4 The t-Test  | 31        |
| 2.5 Intervals and Partial Correlations                          | 33        |

|   |    |
|---|----|
| 2.5.1 Confidence Interval                                       | 33 |
| 2.5.2 Prediction  | 34 |
| 2.5.3 Prediction Interval                                       | 34 |
| 2.5.4 Partial Correlations                                      | 35 |
| 2.6 Multiple Linear Regression                                  | 36 |
| 2.6.1 Concepts  | 36 |
| 2.6.2 Matrix Notation   | 37 |
| 2.7 The F-Test  | 38 |
| 2.7.1 Definitions   | 38 |
| 2.7.2 Partial F-Test  | 39 |
| 2.8 ANOVA Table   | 39 |
| 2.9 Subset Selection  | 40 |
| 2.10 Choosing the Best Model from Subset Selection              | 42 |
| 2.11 Residual Analysis  | 42 |
| 2.11.1 Information  | 43 |
| 2.11.2 Standardized Residuals                                   | 44 |
| 2.12 Influential Points   | 45 |
| 2.13 Collinearity   | 46 |
| 2.14 Homoscedasticity and Heteroscedasticity                    | 47 |
| 2.14.1 Definitions  | 47 |
| 2.14.2 Breusch-Pagan Test                                       | 48 |
| 2.15 Ridge Regression   | 49 |
| 2.16 Lasso  | 51 |
| 2.16.1 Definitions  | 51 |
| 2.16.2 A Geometric Interpretation of Ridge Regression and Lasso | 52 |
| 2.17 Binary Dependent Variables                                 | 52 |
| 2.18 Logit and Probit Models                                    | 54 |
| 2.18.1 Logit Models   | 54 |
| 2.18.2 Probit Models  | 55 |
| 2.18.3 Threshold Interpretation                                 | 55 |
| 2.18.4 Parameter Estimation                                     | 56 |
| 2.19 Nominal Dependent Variables                                | 57 |
| 2.19.1 Generalized Logit Model                                  | 57 |
| 2.19.2 Other Models   | 58 |
| 2.20 Ordinal Dependent Variables                                | 59 |
| 2.21 Poisson Regression   | 60 |
| 2.22 Other Count Models   | 61 |
| 2.23 Generalized Linear Models                                  | 62 |
| 2.23.1 Definitions  | 62 |
| 2.23.2 Canonical Link Function                                  | 63 |
| 2.23.3 Linear Regression Sampling Assumptions on GLMs           | 63 |

|   |           |
|---|-----------|
| 2.23.4 Variance Function  | 64        |
| 2.23.5 The Tweedie Distribution                                   | 64        |
| 2.24 Estimation in GLMs   | 65        |
| 2.24.1 Maximum Likelihood Estimation for Canonical Links          | 65        |
| 2.24.2 Goodness-of-Fit Statistics for GLMs                        | 66        |
| 2.24.3 Residual Analysis for GLMs                                 | 67        |
| <b>3. Time Series Models (Learning Objective 3)</b>               | <b>68</b> |
| 3.1 Introduction to Time Series                                   | 69        |
| 3.1.1 Key Terms   | 69        |
| 3.1.2 Time Series Models  | 70        |
| 3.2 Stationarity  | 71        |
| 3.2.1 Stationarity  | 71        |
| 3.2.2 White Noise   | 71        |
| 3.2.3 Random Walk   | 72        |
| 3.3 Forecast Evaluation   | 73        |
| 3.3.1 Out-of-Sample Validation Process                            | 73        |
| 3.3.2 Statistics for Comparing Forecasts                          | 74        |
| 3.4 Autoregressive Models   | 75        |
| 3.4.1 Autocorrelation   | 75        |
| 3.4.2 AR(1) Model   | 76        |
| 3.5 Smoothing   | 78        |
| 3.6 Exponential Smoothing   | 79        |
| 3.7 Seasonal Adjustments  | 80        |
| 3.8 Unit Root Test  | 81        |
| 3.9 ARCH and GARCH Models   | 82        |
| <b>4. Decision Trees (Learning Objective 4)</b>                   | <b>83</b> |
| 4.1 Introduction to Decision Trees                                | 84        |
| 4.2 Regression Trees  | 84        |
| 4.3 Recursive Binary Splitting                                    | 85        |
| 4.4 Pruning   | 86        |
| 4.5 Classification Trees  | 87        |
| 4.6 Trees vs Linear Models  | 88        |
| 4.7 Bagging   | 89        |
| 4.8 Random Forests  | 90        |
| 4.9 Boosting  | 91        |
| <b>5. Unsupervised Learning Techniques (Learning Objective 5)</b> | <b>92</b> |
| 5.1 Introduction to Unsupervised Learning                         | 93        |
| 5.2 Principal Components Regression                               | 94        |
| 5.2.1 Linear Combinations of Predictors                           | 94        |
| 5.2.2 Principal Components Regression                             | 95        |
| 5.2.3 Partial Least Squares                                       | 95        |

|   |     |
|---|-----|
| 5.3 Principal Component Analysis                  | 96  |
| 5.3.1 Definitions                                 | 96  |
| 5.3.2 Methodology                                 | 97  |
| 5.3.3 Proportion of Variance Explained            | 99  |
| 5.4 K-Means Clustering                            | 100 |
| 5.4.1 Definitions                                 | 100 |
| 5.4.2 K-Means Clustering Algorithm                | 101 |
| 5.4.3 Algorithm Visual Example                    | 101 |
| 5.5 Hierarchical Clustering                       | 102 |
| 5.5.1 Definitions                                 | 102 |
| 5.5.2 Hierarchical Clustering Algorithm           | 103 |
| 5.5.3 Calculating Dissimilarity - Linkage Methods | 103 |
| 5.5.4 Visual Example                              | 104 |

This document is provided as a free resource for public use. You are welcome to use it for personal consumption. The following restrictions apply:

1. **Non-Commercial Use Only**

Any reproduction, distribution, or use for profit or commercial purposes is prohibited.

2. **No Unauthorized Modifications**

You may not alter or edit the content and present it as your own without proper attribution. Please contact [admin@theactuarialnexus.com](mailto:admin@theactuarialnexus.com) if you would like to distribute this document for non-commercial purposes.

While every effort has been made to ensure the accuracy of this document, it may contain typos or errors. Please email [admin@theactuarialnexus.com](mailto:admin@theactuarialnexus.com) if you encounter any typos.

This document was authored in association with The Actuarial Nexus, and is not endorsed by or affiliated with the Society of Actuaries.

Most questions on Exam SRM test a conceptual understanding of the material. As such, this document includes definitions and explanations to accompany some of the formulas. The formulas and notation in this document have been kept as close as possible to those in the source material.

Memorizing the information in this document alone is not sufficient preparation to pass Exam SRM.

Visit [The Actuarial Nexus](https://theactuarialnexus.com) for a comprehensive study program, including 70+ written chapters, 800+ practice questions, and powerful analytic tools designed to help you pass the exam.

## 0. Review

## 0.1 Sampling Assumptions

| Concept                            | Description  |
|------------------------------------|--|
| $E(y_i) = \mu$                     | The expected value (mean) of $y_i$ converges to the population mean, $\mu$ . |
| $\text{Var}(y_i) = \sigma^2$       | The variance of $y_i$ converges to the population variance, $\sigma^2$ .     |
| $\{y_i\}$ are independent          | The set $\{y_i\}$ consists of independent variables.                         |
| $\{y_i\}$ are normally distributed | The set $\{y_i\}$ follows a normal distribution.                             |

$\mu$  and  $\sigma^2$  are parameters. The goal is to use statistics, such as  $\bar{y}$  and  $s_y^2$  to infer information about parameters.

## 0.2 Pearson Correlation Coefficient

| Concept    | Description  |
|------------|--|
| Definition | The <b>Pearson (ordinary) correlation coefficient</b> measures the strength and direction of the linear relationship between two continuous variables. |
| Formula    | $r = \frac{1}{(n-1)s_x s_y} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$   |
| Values     | $-1 \leq r < 0$ : Negative linear relationship.<br>$r = 0$ : No linear relationship.<br>$0 < r \leq 1$ : Positive linear relationship.                 |
| Properties | It is a dimensionless measure (units of measurement removed) and location and scale invariant.   |



### 0.3 Matrices

| Concept                        | Formula/Notation   | Notes   |
|--------------------------------|--|---|
| Matrix                         | <p>Matrices are typically denoted by uppercase letters such as <b>A</b>, <b>B</b>, and <b>C</b>.</p> <p>If a matrix <b>A</b> has <math>m</math> rows and <math>n</math> columns, it is referred to as an <math>m \times n</math> matrix.</p>   | $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$ <p><b>A</b> is a <math>3 \times 3</math> matrix with 3 rows and 3 columns.</p>   |
| Inverse                        | $\mathbf{A}^{-1}$  | <p>Only square matrices can be inverted.</p> $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ <p>Inverting matrices by hand is beyond the scope of the exam.</p>                        |
| Transpose                      | $\mathbf{X}^{\top}, \mathbf{X}^T$ or $\mathbf{X}'$   | $\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \end{pmatrix}$ $\mathbf{X}^{\top} = \begin{pmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \\ x_{13} & x_{23} \end{pmatrix}$ |
| The Variance-Covariance Matrix | $\Sigma = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_n) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_n, X_1) & \text{Cov}(X_n, X_2) & \cdots & \text{Var}(X_n) \end{bmatrix}$ $\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$ | <ul style="list-style-type: none"> <li>- Covariance measures how much two variables change together.</li> <li>- The matrix is symmetric.</li> <li>- The variances are always non-negative.</li> </ul>       |

Matrices are commonly used in [multiple linear regression](#).

### 0.3 Maximum Likelihood Estimation

| Terminology   | Formula / Description  |
|---|--|
| Maximum Likelihood Estimation   | <b>Maximum likelihood estimation</b> is a method for estimating the parameters of a statistical model by maximizing the likelihood that the observed data occurred under the model.  |
| Likelihood Function<br>$L(\theta; \mathbf{x})$                                | For a random sample $\mathbf{x} = (x_1, x_2, \dots, x_n)$ drawn from a distribution with PDF $f(x; \theta)$ , the <b>likelihood function</b> is:<br>$L(\theta; \mathbf{x}) = \prod_{i=1}^n f(x_i; \theta)$   |
| Log-Likelihood Function<br>$\ell(\theta; \mathbf{x})$                         | The <b>log-likelihood function</b> is:<br>$\ell(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}) = \sum_{i=1}^n \log f(x_i; \theta)$<br>where $\log(\cdot)$ is the natural logarithm function (by convention).   |
| Score Function<br>$\frac{\partial \ell(\theta; \mathbf{x})}{\partial \theta}$ | The <b>score function</b> is the partial derivative of the log-likelihood function with respect to the parameter(s):<br>$\frac{\partial \ell(\theta; \mathbf{x})}{\partial \theta}$ <hr/> For a parameter vector, $\theta = (\theta_1, \theta_2, \dots, \theta_p)$ , the score function is:<br>$\left( \frac{\partial \ell(\theta; \mathbf{x})}{\partial \theta_1}, \frac{\partial \ell(\theta; \mathbf{x})}{\partial \theta_2}, \dots, \frac{\partial \ell(\theta; \mathbf{x})}{\partial \theta_p} \right)$ |
| Maximum Likelihood Estimator<br>$\hat{\theta}$                                | $\hat{\theta} = \arg \max_{\theta} L(\theta; \mathbf{x}) = \arg \max_{\theta} \ell(\theta; \mathbf{x})$<br>The <b>maximum likelihood estimator</b> (MLE) is found by setting the score function to zero and solving for $\theta$ . The argmax is the value of $\theta$ where the function is maximized.  |
| (Fisher) Information Matrix<br>$I(\theta)$                                    | $I(\theta) = -E \left[ \frac{\partial^2}{\partial \theta \partial \theta^T} \ell(\theta) \right]$<br>A larger value of $I(\theta)$ indicates that the data provides more information about $\theta$ , leading to more precise estimates.   |

|   |  |
|---|--|
| <p>Covariance Matrix</p> $I(\theta)^{-1}$ | <p>The asymptotic variance-covariance matrix (covariance matrix) is the inverse of the information matrix: <math>I(\theta)^{-1}</math>.</p> <p>As the sample size grows, the distribution of the MLE <math>\hat{\theta}</math> approaches a normal distribution with mean <math>\theta</math> and covariance matrix <math>I(\theta)^{-1}</math>.</p> |
|---|--|

# **1. Basics of Statistical Learning (Learning Objective 1)**

## 1.1 Types of Variables

| Terminology          | Definition  | Example   |
|----------------------|---|---|
| Input Variable       | A variable used to predict the output variable (a.k.a. predictors, explanatory variables, exogenous variables, independent variables, features, regressors).  | In a house price prediction model, input variables could include the size of the house, the number of bedrooms, and the location.   |
| Output Variable      | The variable that you are trying to predict or explain (a.k.a. response, outcome of interest, endogenous variable, explained variable, outcome, regressand, or dependent variable).   | In a house price prediction model, the output variable is the price of the house.   |
| Confounding Variable | A variable that affects both the independent variable and the dependent variable, potentially leading to a false association between them.  | <p>In a study examining the relationship between coffee consumption and heart disease, smoking can act as a confounding variable.</p> <p>People who drink more coffee might also be more likely to smoke, and smoking is a known risk factor for heart disease.</p> |
| Binary Variable      | A variable that captures the presence or absence of a particular attribute, event, or condition within a dataset.   | Gender (male/female), yes/no responses, or the occurrence of an event.  |
| Dummy Variable       | <p>A binary (0/1) indicator used in a regression model to represent the presence or absence of a categorical attribute.</p> <p>Several dummy variables can be created to represent variables with more than two categories.</p> | <p>For a color variable with red and blue, use 1 if blue and 0 if red.</p> <p>For a color variable with red, blue, and green, use 1 if blue and 0 otherwise, 1 if green and 0 otherwise. Red is the reference when both are 0.</p>                                  |
| Nominal Variable     | A variable that categorizes data without any intrinsic order.   | Blood type (A, B, AB, O) or eye color (blue, green, brown).   |

|                      |  |   |
|----------------------|--|---|
| Ordinal Variable     | A variable that categorizes and ranks data in a specific order.  | Satisfaction ratings (poor, fair, good, excellent) or education levels (high school, bachelor's, master's, doctorate).                  |
| Count Variable       | A variable that quantifies the number of occurrences of an event within a fixed period or space.   | The number of customer visits to a store or the number of emails received per day.  |
| Interaction Variable | A variable that captures the combined effect of two variables when their joint impact on the outcome is different from their individual effects. | In a model with "Education" and "Gender," adding an Education × Gender variable lets the effect of education differ between genders.    |
| Omitted Variable     | A relevant variable left out of a model, which can bias the results if it's correlated with included variables.                                  | Leaving out experience in a wage model that includes education may overstate education's effect if experience is also related to wages. |
| Suppressor Variable  | A variable that increases the predictive validity of other variables when included in the model.   | Party hours may act as a suppressor when added to a model predicting GPA from SAT scores.   |

## 1.2 Prediction and Inference

**Prediction** aims to forecast the output variable based on the input variables. The focus is on the accuracy of the predictions.

**Inference** aims to understand the relationship between the input variables and the output variable. The focus is on interpretability.

In some situations, it is necessary to model for **both** inference and prediction simultaneously.

## 1.3 Decomposition of the Expected Squared Error

The **expected squared error**, or squared expected difference, is a theoretical measure of how far predictions,  $\hat{Y}$ , are from actual outcomes,  $Y$ :

$$\mathbb{E}[(Y - \hat{Y})^2] = \mathbb{E}[(Y - \hat{f}(X))^2] = (f(X) - \hat{f}(X))^2 + \text{Var}[\epsilon]$$

| Terminology | Reducible Error  | Irreducible Error (Noise)  |
|-------------|--|--|
| Formula     | $(f(X) - \hat{f}(X))^2$  | $\text{Var}[\epsilon]$   |
| Description | <b>Reducible error</b> is the error introduced by the model's approximation.                                   | <b>Irreducible error</b> (noise) captures the effects of unobserved variables or inherent randomness in the data.                                  |
| Comments    | Decreased by improving the model, choosing better predictors, or using more sophisticated modeling techniques. | In practice, the irreducible error can never be zero. There are always factors affecting the response variable that are not included in the model. |

## 1.4 Parametric and Non-Parametric Methods

A **parametric method** assumes the data can be modeled by a particular distribution characterized by a fixed number of parameters. The most popular parametric models take a linear form.

A **non-parametric method** does not assume a fixed form or structure for the underlying data distribution. Instead of having a predetermined number of parameters, non-parametric models allow the data to dictate the model complexity.

**Flexibility** refers to the model's capacity to fit a wide variety of shapes and patterns in the data. Flexible models can model complex and non-linear relationships, but have a higher risk of overfitting.

**Overfitting** occurs when a model is too closely fit to the specific features of the training data, including noise, leading to poor generalization to new, unseen data. An overfit model has low training error but high test error.

## 1.5 Supervised vs Unsupervised Learning

| Method     | Supervised Learning  | Unsupervised Learning   |
|------------|--|---|
| Definition | <b>Supervised learning</b> involves using labeled data, where each data point is associated with a response measurement.   | <b>Unsupervised learning</b> deals with data that has no labeled responses.   |
| Goal       | Predict or classify the response based on inputs.  | Discover patterns, relationships, and groupings in the data.  |
| Examples   | <ul style="list-style-type: none"><li>- <a href="#">Linear Regression</a></li><li>- <a href="#">Logistic Regression</a></li><li>- <a href="#">Decision Trees</a></li></ul> | <ul style="list-style-type: none"><li>- <a href="#">Principal Component Analysis</a></li><li>- <a href="#">K-means Clustering</a></li><li>- <a href="#">Hierarchical Clustering</a></li></ul> |
| Exam SRM   | Learning Objectives 1-4  | Learning Objective 5  |



## 1.6 Regression vs Classification

The response variable plays an important role in selecting the learning method.

| Method            | Regression  | Classification  |
|-------------------|---|---|
| Response Variable | Quantitative (numerical)  | Qualitative (categorical)   |
| Goal              | Predict a quantitative outcome.   | Classify data into categories.  |
| Examples          | <ul style="list-style-type: none"> <li>- <a href="#">Linear Regression</a> (continuous)</li> <li>- <a href="#">Time Series</a> (continuous)</li> <li>- <a href="#">Poisson Regression</a> (discrete)</li> </ul> | <ul style="list-style-type: none"> <li>- <a href="#">Logistic Regression</a></li> <li>- <a href="#">Decision Trees</a></li> </ul> |

## 1.7 Mean Squared Error and Error Rate

| Data  | Training Data   | Test Data  |
|---|---|--|
| <b>Mean Squared Error Regression</b><br>Difference between observed and predicted values. | $\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2$ <a href="#">Example</a> | $\text{Ave}(y_0 - \hat{f}(x_0))^2$                           |
| <b>Error Rate Classification</b><br>Proportion of misclassified training observations.    | $\frac{1}{n} \sum_{i=1}^n I(y_i \neq \hat{y}_i)$                          | $\text{Ave}(I(y_i \neq \hat{y}_i))$ <a href="#">Example</a>  |
| Purpose   | Measures how well a model fits the training data.                         | Measures prediction accuracy on unseen data.                 |
| Notes   | Lower values indicate better fit on training data.                        | Estimated with cross-validation if test data is unavailable. |

## 1.8 Bias-Variance Tradeoff

### 1.8.1 Definitions

| Terminology       | Formula/Description   |
|-------------------|---|
| Expected Test MSE | $\mathbb{E}[(y_0 - \hat{f}(x_0))^2] = \text{Var}[\hat{f}(x_0)] + \text{Bias}[\hat{f}(x_0)]^2 + \text{Var}(\epsilon)$ <hr/> Measures model generalization to new data. |
| Variance          | $\text{Var}[\hat{f}(x_0)] = \mathbb{E}[(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)])^2]$ <hr/> Measures model sensitivity to fluctuations in the training data.           |
| Bias              | $\text{Bias}[\hat{f}(x_0)] = \mathbb{E}[\hat{f}(x_0)] - f(x_0)$ <hr/> Measures the error introduced by approximating a complex reality with a simpler model.          |
| Irreducible Error | $\text{Var}(\epsilon)$ <hr/> Represents noise in the data which cannot be explained by any model.   |

### 1.8.2 Tradeoff Table

| Model Flexibility | Bias   | Variance | Expected Test MSE           | Notes   |
|-------------------|--------|----------|-----------------------------|---|
| Low               | High   | Low      | High (due to high bias)     | - Simple models<br>- Leads to underfitting    |
| Medium            | Medium | Medium   | Low (optimal trade-off)     | - Optimal trade-off between bias and variance |
| High              | Low    | High     | High (due to high variance) | - Complex models<br>- Leads to overfitting    |

## 1.9 Data Collection

**Sampling frame error** occurs when the sampling frame, or the list from which the sample is drawn, does not adequately approximate the population of interest. Limited sampling regions can introduce bias when attempting to extrapolate beyond the sampled area.

When an omitted variable influences both the dependent variable  $y$  and the explanatory variable  $x$ , it can create a **spurious** relationship. The effects of the omitted variable may be incorrectly attributed to other included variables, potentially creating a misleading or false association between those variables and the outcome.

Techniques to handle data that are **missing at random**:

1. Ignore the problem.
2. If only a few data points are missing, remove the observations with missing data.
3. If missing data are concentrated in one variable, omit the variable.
4. Impute the data by substituting missing values with reasonable estimates.

Traditional statistical techniques are intended for the **low-dimensional** setting in which the number of observations is much greater than the number of features ( $n \gg p$ ). New technologies have enabled the collection of an almost unlimited number of features, enabling analysis around **high-dimensional** ( $p > n$ ) data sets. In practice, while  $p$  can be very large,  $n$  can be limited due to cost or sample availability. Applying least squares regression in a high-dimensional setting can lead to overfitting. Methods such as [forward stepwise selection](#), [lasso](#), and [principal components regression](#) are particularly useful in high-dimensional settings.

## 1.10 Bayes Classifier

| Concept          | Description   |
|------------------|---|
| Bayes Classifier | The <b>Bayes classifier</b> is a hypothetical optimal classifier that assigns the most probable class given a predictor vector. |
| Formula          | $\Pr(Y = j   X = x_0)$  |
| Goal             | Minimizes test error by selecting the class with the highest conditional probability.   |
| Limitations      | Requires full knowledge of the true conditional probability distributions, which is rarely available in practice.               |
| Approximation    | Approximated using models like logistic regression, decision trees, and neural networks.  |

| Concept          | Description  |
|------------------|--|
| Bayes Error Rate | The <b>Bayes error rate</b> is the lowest possible test error rate achievable, even with the optimal classifier. It is similar to the irreducible error. |
| Formula          | $1 - \mathbb{E} \left( \max_j \Pr(Y = j   X) \right)$  |

## 1.11 K-Nearest Neighbors

### 1.11.1 Algorithm

For a new point  $x_0$ :

| Step   | Description   |
|--|---|
| 1. Choose $K$  | Select the number of neighbors.   |
| 2. Compute Distances                                   | Measure the distance from $x_0$ to all other observations (e.g., Euclidean distance). |
| 3. Find $K$ Nearest Neighbors                          | Select the $K$ closest points to $x_0$ .  |
| 4. Classify the Point                                  | Assign the class based on the majority vote among the $K$ neighbors.                  |
| The conditional probability $x_0$ belongs to class $j$ | $\Pr(Y = j \mid X = x_0) \approx \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j)$   |

### 1.11.2 Bias-Variance Tradeoff in KNN

| $K$ Value | Bias | Variance | Behavior                                 |
|-----------|------|----------|--|
| Small $K$ | Low  | High     | Sensitive to noise; may overfit.         |
| Large $K$ | High | Low      | May miss patterns; smoother predictions. |

## 1.12 The Validation Set Approach

### 1.12.1 Algorithm

| Step                | Description  |
|---------------------|--|
| 1. Data Splitting   | Randomly divide data into training (70–80%) and validation (20–30%) sets.  |
| 2. Model Training   | Fit the model using the training set.  |
| 3. Model Validation | Evaluate the model on the validation set using metrics like mean squared error (MSE) or the sum of squared prediction errors (SSPE). |

### 1.12.2 Pros and Cons

| Pros                   | Cons                              |
|------------------------|-----------------------------------|
| Simple                 | High variance due to single split |
| Fast                   | Not all data is used for training |
| Low computational cost | May overestimate test error       |

## 1.13 Leave-One-Out Cross-Validation

### 1.13.1 Algorithm

| Step                | Description   |
|---------------------|---|
| 1. Data Splitting   | <p>For each observation <math>i</math> in the dataset:</p> <ul style="list-style-type: none"> <li>- Remove the <math>i</math>-th observation, resulting in the training set <math>D_{-i}</math>.</li> <li>- Train the model on <math>D_{-i}</math>.</li> <li>- Test the model on the removed <math>i</math>-th observation <math>(x_i, y_i)</math>, obtaining the prediction <math>\hat{y}_i</math>.</li> </ul> |
| 2. Model Training   | Compute the performance metric (e.g., mean squared error or error rate) for each iteration.   |
| 3. Model Validation | <p>Average the performance metrics over all <math>n</math> iterations to obtain the final estimate.</p> <p>For regression, use the predicted residual sum of squares (PRESS):</p> $CV_{(n)} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n MSE_i$ <p>For classification, use:</p> $CV_{(n)} = \frac{1}{n} \sum_{i=1}^n Err_i$  |

### 1.13.2 Pros and Cons

| Pros   | Cons                                       |
|--|--|
| Nearly unbiased estimate of test error                 | Requires training the model $n$ times      |
| Uses almost all data for training                      | High variance due to similar training sets |
| No randomness in splits => leads to consistent results | Computationally expensive for large $n$    |

## 1.14 K-Fold Cross-Validation

### 1.14.1 Algorithm

| Step                | Description  |
|---------------------|--|
| 1. Data Splitting   | Divide the data set into $k$ folds. When $k = n$ , k-fold CV is equivalent to LOOCV. Popular choices for $k$ are $k = 5$ or $k = 10$ .   |
| 2. Model Training   | For each fold: <ul style="list-style-type: none"><li>- Train the model on <math>k - 1</math> folds.</li><li>- Validate the model on the remaining fold.</li><li>- Compute the MSE for the predictions on the validation fold.</li></ul>  |
| 3. Model Validation | <p>Calculate the mean of the MSE values obtained from each of the <math>k</math> folds.</p> <p>For regression, use:</p> $CV_{(k)} = \frac{1}{k} \sum_{i=1}^k \text{MSE}_i$ <p><a href="#">Example</a></p> <p>For classification, use:</p> $CV_{(k)} = \frac{1}{k} \sum_{i=1}^k \text{Err}_i$ |

### 1.14.2 Model Comparison

LOOCV has higher variance and lower bias than K-fold CV for  $k < n$  since LOOCV averages the outputs of  $n$  fitted models. These models are highly correlated with each other.

In terms of the bias-variance tradeoff, K-fold can be considered between the validation set approach and LOOCV.



## **2. Linear Models (Learning Objective 2)**

## 2.1 Simple Linear Regression

### 2.1.1 Theoretical Representation of a Linear Model

| Concept                      | Description  |
|------------------------------|--|
| Equation                     | $y \approx \beta_0 + \beta_1 x$ (approximation)    |
| Dependent Variable ( $y$ )   | The outcome variable we aim to predict.            |
| Independent Variable ( $x$ ) | The input variable used for prediction.            |
| Intercept ( $\beta_0$ )      | Expected value of $y$ when $x = 0$ .               |
| Slope ( $\beta_1$ )          | The change in $y$ for a one-unit increase in $x$ . |
| Coefficients (or Parameters) | $\beta_0, \beta_1$                                 |

### 2.1.2 Observations vs Predictions (Actuals vs Estimates)

| Concept  | Description   |
|--|---|
| Observed Values ( $x_i, y_i$ )                   | Actual data points collected from empirical data for observations $i = 1, \dots, n$ .   |
| Predicted Value ( $\hat{y}_i$ )                  | The value of $y$ predicted by the regression equation for a given observation, $x_i$ .  |
| Regression Equation                              | $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$   |
| Estimated Intercept ( $\hat{\beta}_0$ or $b_0$ ) | Estimate from ordinary least squares:<br>$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$  |
| Estimated Slope ( $\hat{\beta}_1$ or $b_1$ )     | Estimate from ordinary least squares:<br>$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = r \frac{s_y}{s_x}$ |

### 2.1.3 Ordinary Least Squares

| Concept                                   | Description  |
|---|--|
| Terminology                               | <p>The following refer to the same concept:</p> <ul style="list-style-type: none"> <li>- Ordinary least squares (OLS)</li> <li>- Method of least squares</li> <li>- Least squares method</li> <li>- Least squares regression</li> </ul> <hr/> <ul style="list-style-type: none"> <li>- Simple linear regression: One independent variable</li> <li>- Multiple linear regression: More than one independent variable</li> </ul> |
| Goal                                      | The goal of OLS is to estimate regression coefficients by minimizing the residual sum of squares (RSS).  |
| Residual ( $e_i$ )                        | <p>A <b>residual</b> measures the difference between the observed value <math>y_i</math> and the predicted value <math>\hat{y}_i</math>.</p> $e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$   |
| Residual Sum of Squares (RSS)             | $\begin{aligned} \text{RSS} &= e_1^2 + e_2^2 + \dots + e_n^2 \\ &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \end{aligned}$  |
| Fitted Regression Line (Line of Best Fit) | $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$  |

### 2.1.4 Error Term

The **true model** is defined as  $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ .

The **error term** (disturbance term) is  $\epsilon_i = \varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$ . It acknowledges that observed data points do not perfectly fit the underlying theoretical model, and accounts for random variation (noise).

Assumptions about the error terms,  $\{\epsilon_i\}$ , are made to ensure OLS produces unbiased estimates. An estimator  $\hat{\mu}$  of a parameter  $\mu$  is **unbiased** if  $E(\hat{\mu}) = \mu$ .

The following assumptions apply to observations,  $\{y_i\}$ , and errors,  $\{\epsilon_i\}$ :

| Observation Assumptions                                | Error Assumptions                                       |
|--|---|
| 1. $E[y_i] = \beta_0 + \beta_1 x_i$                    | 1. $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$           |
| 2. $\{x_1, \dots, x_n\}$ are non stochastic variables. | 2. $\{x_1, \dots, x_n\}$ are non stochastic variables.  |
| 3. $Var(y_i) = \sigma^2$                               | 3. $E[\epsilon_i] = 0$ and $Var(\epsilon_i) = \sigma^2$ |
| 4. $\{y_i\}$ are independent random variables.         | 4. $\{\epsilon_i\}$ are independent random variables.   |

From the central limit theorem, these assumptions imply that  $\{y_i\}$  and  $\{\epsilon_i\}$  are approximately normally distributed.

## 2.2 Mean Squared Error and Standard Error

The formulas below only apply to simple linear regression (one independent variable).

| Concept  | Description   |
|--|---|
| $Var(\epsilon_i) = \sigma^2$                                 | The variance of the error term provides a measure of how much the observed values deviate from the true values due to random noise.   |
| Mean Squared Error (MSE)<br><a href="#">Example</a>          | <p>To obtain an unbiased estimator of <math>\sigma^2</math>:</p> $MSE = s^2 = \hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ $= \frac{1}{n-2} \sum_{i=1}^n e_i^2 = \frac{RSS}{n-2}$ |
| Residual Standard Error (RSE)<br><a href="#">Example</a>     | <p>The RSE is an estimate of <math>\sigma</math>:</p> $RSE = \sqrt{MSE} = \sqrt{s^2} = s = \hat{\sigma} = \sqrt{\frac{RSS}{n-2}}$   |
| Standard Error of $\hat{\beta}_0$<br><a href="#">Example</a> | $SE(\hat{\beta}_0) = \sqrt{MSE \left[ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$  |
| Standard Error of $\hat{\beta}_1$<br><a href="#">Example</a> | $SE(\hat{\beta}_1) = \sqrt{\frac{MSE}{\sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{RSE}{s_x \sqrt{n-1}}$  |

## 2.3 Sum of Squares and R-Squared

| Concept   | Description   |
|---|---|
| <p>Total Sum of Squares (<i>TSS</i>)</p> <p><b>Total deviation:</b><br/><math>y_i - \bar{y}</math></p>  | <p>Measures the total variation in the dependent variable around its mean. <a href="#">Example</a>.</p> <hr/> $TSS = \sum_{i=1}^n (y_i - \bar{y})^2$  |
| <p>Residual Sum of Squares (<i>RSS</i>, <i>ErrorSS</i>, <i>SSE</i>)</p> <p><b>Unexplained deviation:</b><br/><math>y_i - \hat{y}_i</math></p> | <p>Measures the discrepancy between the observed data and the values predicted by the model. <a href="#">Example</a>.</p> <hr/> $RSS = \sum_{i=1}^n (y_i - \hat{y}_i)^2$  |
| <p>Regression Sum of Squares (Regression SS)</p> <p><b>Explained deviation:</b><br/><math>\hat{y}_i - \bar{y}</math></p>                      | <p>Measures the variation explained by the regression model. <a href="#">Example</a>.</p> <hr/> $\text{Regression SS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$   |
| <p>Decomposition of TSS (Linear Regression Only)</p>  | $TSS = RSS + \text{Regression SS}$ $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ <p>The cross-product term equals zero in linear regression. A nonzero cross-product term exists for <a href="#">nonlinear models</a>.</p> |
| <p>Coefficient of Determination or R-Squared (<math>R^2</math>)</p>   | <p><b>R-squared</b> measures the proportion of the variance in the dependent variable that is explained by the independent variable.</p> <hr/> $R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} = \frac{\text{Regression SS}}{TSS}$ <p>Value ranges from 0 to 1.</p>                  |

## 2.4 The t-Test

For Exam SRM, the **t-test** is mainly used to test the significance of a regression coefficient, such as the slope,  $\beta_1$ .

| Concept                          | Description   |
|----------------------------------|---|
| Null Hypothesis ( $H_0$ )        | <p><math>H_0: \beta_1 = d</math>, represents the status quo, that <math>\beta_1</math> is equal to a specific value <math>d</math>.</p> <p>In regression, often <math>H_0: \beta_1 = 0</math>, indicating no relationship between the predictor and the response.</p>   |
| Alternative Hypothesis ( $H_a$ ) | <p>Competes with <math>H_0</math>, representing a difference from <math>d</math>.</p> <ul style="list-style-type: none"> <li>- <math>\beta_1 \neq d</math> (two-tailed)</li> <li>- <math>\beta_1 &lt; d</math> or <math>\beta_1 &gt; d</math> (one-tailed)</li> </ul>   |
| t-Statistic (t-ratio)            | $t = \frac{\hat{\beta}_1 - d}{SE(\hat{\beta}_1)}$ <p>The <b>t-statistic</b> measures how many standard errors the estimate <math>\hat{\beta}_1</math> is from the hypothesized value <math>d</math>.</p>  |
| Degrees of Freedom (df)          | <p><math>df = n - k</math></p> <p>For simple regression, <math>df = n - 2</math>, since two parameters, the intercept and slope are estimated.</p>  |
| Significance Level ( $\alpha$ )  | <p>Probability threshold for rejecting <math>H_0</math>, such as <math>\alpha = 0.05</math>.</p> <ul style="list-style-type: none"> <li>- One-tailed: critical value at <math>\alpha</math>.</li> <li>- Two-tailed: critical value at <math>\alpha/2</math> in each tail.</li> </ul> <p>See the table on the next page for details.</p> |
| Decision Rule                    | <p>Reject <math>H_0</math> if <math> t  &gt;</math> critical value from the t-distribution table or p-value <math>&lt; \alpha</math>. Otherwise fail to reject <math>H_0</math>. See the table on the next page for details.</p>  |
| Interpretation                   | <ul style="list-style-type: none"> <li>- Reject <math>H_0 \Rightarrow</math> variable likely significant; keep in model</li> <li>- Fail to reject <math>H_0 \Rightarrow</math> variable may be excluded from the model</li> </ul>   |

| Alternative Hypothesis | Reject Null Hypothesis Criteria |
|------------------------|---------------------------------|
| $H_a : \beta_1 \neq d$ | $\ t\  > t_{df,\alpha/2}$       |
| $H_a : \beta_1 > d$    | $t > t_{df,\alpha}$             |
| $H_a : \beta_1 < d$    | $t < -t_{df,\alpha}$            |





## 2.5 Intervals and Partial Correlations

### 2.5.1 Confidence Interval

A **confidence interval** is a range of values, derived from sample data, that is likely to contain the true value of an unknown population parameter.

A 95% confidence interval suggests that if the same population were sampled multiple times, approximately 95% of the calculated confidence intervals from those samples would contain the true parameter value.

For simple linear regression with parameters  $\beta_0$  and  $\beta_1$ , a  $100(1 - \alpha)\%$  confidence interval for the slope  $\beta_1$  is given by:

$$\hat{\beta}_1 \pm t_{n-2, \alpha/2} \cdot \text{se}(\hat{\beta}_1)$$

For multiple linear regression with  $p$  predictors, a  $100(1 - \alpha)\%$  confidence interval for the slope  $\beta_j$  is given by:

$$\hat{\beta}_j \pm t_{n-(p+1), \alpha/2} \cdot \text{se}(\hat{\beta}_j)$$

### 2.5.2 Prediction

**Prediction** (or forecasting) is the process of estimating future values based on historical data. Let the response variable from a series of known explanatory variables,  $\mathbf{x}_* = (1, x_{*1}, \dots, x_{*p})$ , be denoted as:

$$y_* = \beta_0 + \beta_1 x_{*1} + \dots + \beta_p x_{*p} + \epsilon_*$$

The least squares point predictor for  $y_*$  is:

$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_{*1} + \dots + \hat{\beta}_p x_{*p}$$

For simple linear regression, the least squares point predictor for  $y_*$  is denoted:

$$\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$$

We can decompose the prediction error into the estimation error and the random error:

$$\underbrace{y_* - \hat{y}_*}_{\text{prediction error}} = \underbrace{(\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1)x_{*1} + \dots + (\beta_p - \hat{\beta}_p)x_{*p}}_{\text{regression estimation error at } x_{*1}, \dots, x_{*p}} + \underbrace{\epsilon_*}_{\text{deviation error}}$$

This decomposition allows us to model the distribution of the prediction error, and construct a prediction interval for  $y_*$ .

### 2.5.3 Prediction Interval

A  $100(1 - \alpha)\%$  **prediction interval** (forecast interval) for the dependent variable  $y$  at a given  $\mathbf{x}_*$  is:

$$\hat{y}_* \pm t_{n-(p+1), \alpha/2} \cdot \text{se}(\text{pred}) \text{ where } \text{se}(\text{pred}) = s \sqrt{1 + \mathbf{x}_*'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_*}$$

The prediction interval is generally wider than the confidence interval for the mean response because it includes the variability of the individual observations.

For simple linear regression where,  $\hat{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$ , the standard error of the prediction at  $x_*$  is:

$$\text{se}(\text{pred}) = \sqrt{\text{MSE} \times \left( 1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)}$$

### 2.5.4 Partial Correlations

The **partial correlation coefficient** measures the strength and direction of the linear relationship between two variables, while controlling for the effect of one or more other variables:

$$r(y, x_j \mid x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_p) = \frac{t(b_j)}{\sqrt{t(b_j)^2 + n - (p + 1)}}$$

Calculating partial correlation coefficients using the t-ratio is efficient and allows all partial correlation coefficients to be computed from a single regression, though it might miss nonlinear relationships.

Create **added variable plots** (partial regression plots) by plotting  $y \mid x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_p$  vs  $x_j \mid x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_p$  to visualize nonlinear relationships.

## 2.6 Multiple Linear Regression

### 2.6.1 Concepts

| Concept                                       | Description  |
|---|--|
| Definition                                    | Multiple linear regression extends from simple linear regression and models the relationship between a dependent variable $y$ and multiple independent variables $x_1, x_2, \dots, x_p$ .  |
| Model Form                                    | $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$   |
| Least Squares Method                          | Estimate regression coefficients by minimizing the RSS.  |
| Prediction                                    | $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_p x_p$  |
| Residual Sum of Squares<br>(RSS)              | $\begin{aligned} \text{RSS} &= e_1^2 + e_2^2 + \dots + e_n^2 \\ &= \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 \end{aligned}$  |
| Mean Squared Error<br><a href="#">Example</a> | $\begin{aligned} \text{MSE} = s^2 = \hat{\sigma}^2 &= \frac{1}{n - (p + 1)} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \frac{1}{n - (p + 1)} \sum_{i=1}^n e_i^2 = \frac{\text{RSS}}{n - (p + 1)} \end{aligned}$ <p><math>p + 1</math> is the number of parameters including the intercept.</p> |
| Residual Standard Error                       | $\text{RSE} = \sqrt{\text{MSE}} = \sqrt{s^2} = s = \hat{\sigma} = \sqrt{\frac{\text{RSS}}{n - (p + 1)}}$   |

## 2.6.2 Matrix Notation

| Concept                                 | Description   |
|---|---|
| Model Form                              | $y = \mathbf{X}\beta + \epsilon$  |
| Dependent Variables ( $\mathbf{y}$ )    | $\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$  |
| Design Matrix ( $\mathbf{X}$ )          | $\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$  |
| Coefficients ( $\beta$ )                | $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}$  |
| Errors ( $\epsilon$ )                   | $\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$   |
| Least Squares Method                    | Estimate regression coefficients, $\beta$ , by minimizing the RSS.  |
| Coefficients Estimate ( $\hat{\beta}$ ) | $\mathbf{b} = \hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$ <hr/> $\mathbb{E}[\hat{\beta}] = \beta \text{ (unbiased) \& Var}[\hat{\beta}] = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$ <p>where <math>\sigma^2</math> is the variance-covariance matrix.</p> <hr/> <p>Alternatively,</p> $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \sum_{i=1}^n \mathbf{w}_i y_i$ $\mathbf{w}_i = (\mathbf{X}^\top \mathbf{X})^{-1} (1, x_{i1}, \dots, x_{ip})^\top$ |
| Prediction Estimate                     | $\hat{y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$  |

## 2.7 The F-Test

### 2.7.1 Definitions

While the t-test assesses the significance of individual regression coefficients, the **F-test** measures the overall significance of a regression model.

| Concept                                | Description   |
|--|---|
| Null Hypothesis<br>( $H_0$ )           | $H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$<br>None of the predictor variables have any effect on the response variable.  |
| Alternative Hypothesis<br>( $H_a$ )    | $H_a : \text{at least one } \beta_j \text{ is non-zero}$<br>At least one predictor variable is significantly associated with the response variable.                         |
| F-Statistic<br><a href="#">Example</a> | $F = \frac{(TSS - RSS)/p}{RSS/(n - p - 1)}$<br>The <b>F-statistic</b> measures the ratio of explained variance per predictor to unexplained variance per degree of freedom. |
| Decision Rule                          | Reject $H_0$ if $F > \text{critical value}$ from the F-distribution table.<br>Otherwise fail to reject $H_0$ . A F-distribution table is <b>not</b> provided on the exam.   |
| Interpretation                         | Reject $H_0 \Rightarrow$ at least one predictor variable is significant.<br>Fail to reject $H_0 \Rightarrow$ no predictor variables are significant.                        |

### 2.7.2 Partial F-Test

The **partial F-test** assesses whether a subset  $q$  of the  $p$  coefficients is zero.

| Concept                             | Description   |
|-------------------------------------|---|
| Null Hypothesis<br>( $H_0$ )        | $H_0 : \beta_{p-q+1} = \beta_{p-q+2} = \dots = \beta_p = 0$ <hr/> The subset of $q$ coefficients are zero. For convenience, the $q$ variables chosen for omission are at the end of the list. |
| Alternative Hypothesis<br>( $H_a$ ) | $H_a : \text{at least one } \beta_j \text{ is non-zero}$ <hr/> At least one of the $q$ coefficients is non-zero.  |
| F-Statistic                         | $F = \frac{(RSS_0 - RSS)/q}{RSS/(n - p - 1)}$ <hr/> $RSS_0$ uses all variables except the last $q$ .  |

## 2.8 ANOVA Table

| Sum of Squares (SS) | Formula                                | Degrees of Freedom (df) | Mean Square (MS)                                  |
|---------------------|--|-------------------------|---|
| Regression SS       | $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ | $p$                     | $\text{Reg MS} = \frac{\text{Reg SS}}{p}$         |
| Error SS (RSS)      | $\sum_{i=1}^n (y_i - \hat{y}_i)^2$     | $n - p - 1$             | $\text{MSE} = s^2 = \frac{\text{RSS}}{n - p - 1}$ |
| Total (TSS)         | $\sum_{i=1}^n (y_i - \bar{y})^2$       | $n - 1$                 |   |

## 2.9 Subset Selection

| Method                      | Algorithm   | Notes   |
|-----------------------------|---|---|
| Best Subset Selection       | <ol style="list-style-type: none"> <li>1. Start with the null model <math>\mathcal{M}_0</math> with no predictors.</li> <li>2. For each <math>k = 1, 2, \dots, p</math>: <ol style="list-style-type: none"> <li>a. Fit all <math>\binom{p}{k}</math> models with <math>k</math> predictors.</li> <li>b. Select the best one based on lowest RSS or highest <math>R^2</math>. Call it <math>\mathcal{M}_k</math>.</li> </ol> </li> <li>3. Choose the best overall model from <math>\mathcal{M}_0, \dots, \mathcal{M}_p</math> using selection criteria (e.g. <math>C_p</math>).</li> </ol>           | <ul style="list-style-type: none"> <li>- Builds a simple model with only key predictors.</li> <li>- Risk of overfitting.</li> <li>- Requires fitting <math>2^p</math> models – not feasible for large <math>p</math>.</li> <li>- Slower than alternatives; often avoided when <math>p</math> is large.</li> </ul>     |
| Forward Stepwise Selection  | <ol style="list-style-type: none"> <li>1. Start with the null model <math>\mathcal{M}_0</math>.</li> <li>2. For <math>k = 0, 1, \dots, p - 1</math>: <ol style="list-style-type: none"> <li>a. Evaluate all <math>p - k</math> models by adding one new predictor to <math>\mathcal{M}_k</math>.</li> <li>b. Choose the one with the lowest RSS or highest <math>R^2</math> as <math>\mathcal{M}_{k+1}</math>.</li> </ol> </li> <li>3. Choose the best overall model from <math>\mathcal{M}_0, \dots, \mathcal{M}_p</math> using selection criteria.</li> </ol>                                     | <ul style="list-style-type: none"> <li>- More computationally efficient than best subset selection.</li> <li>- With <math>p = 20</math>, it fits only 211 models vs. over 1 million for best subset selection.</li> <li>- Performs well in practice but doesn't guarantee finding the absolute best model.</li> </ul> |
| Backward Stepwise Selection | <ol style="list-style-type: none"> <li>1. Start with the full model <math>\mathcal{M}_p</math> containing all <math>p</math> predictors.</li> <li>2. For each <math>k = p, p - 1, \dots, 1</math>: <ol style="list-style-type: none"> <li>- a. Evaluate all models formed by removing one predictor from <math>\mathcal{M}_k</math>.</li> <li>- b. Choose the best <math>\mathcal{M}_{k-1}</math> based on lowest RSS or highest <math>R^2</math>.</li> </ol> </li> <li>3. Choose the best overall model from <math>\mathcal{M}_0, \dots, \mathcal{M}_p</math> using selection criteria.</li> </ol> | <p>Similar to forward stepwise selection.</p>   |



|                     |  |   |
|---------------------|--|---|
| Hybrid Approach     | Sequentially add variables, but also remove any that no longer improve the model.  | Combines the benefits of best subset selection with the computational efficiency of stepwise methods.   |
| Stepwise Regression | <ol style="list-style-type: none"> <li>1. Run all simple regressions with one variable. Choose the one with the largest t-ratio. If it's below a set threshold (e.g., 2), stop.</li> <li>2. Add variables one at a time based on the most significance. The t-ratio must be above the threshold.</li> <li>3. Remove variables one at a time based on the least significance. The t-ratio must be below the threshold.</li> <li>4. Alternate between adding and removing until no changes meet the criteria.</li> </ol> | <ul style="list-style-type: none"> <li>- Stepwise regression is fast but can overfit and miss the best model.</li> <li>- It ignores nonlinear effects, outliers, and joint variable interactions.</li> <li>- Relies solely on t-ratios and lacks expert input.</li> <li>- Modern tools allow for the practical use of more complex algorithms.</li> </ul> |

## 2.10 Choosing the Best Model from Subset Selection

| Model   | Formula  |  |
|---|--|--|
| Variable Definitions  | <ul style="list-style-type: none"> <li>- <math>d</math>: the total number of predictors in the model including the intercept.</li> <li>- <math>n</math>: the number of observations in the dataset.</li> <li>- <math>\hat{\sigma}^2</math> (MSE): an estimate of the variance of the model's error <math>\epsilon</math>.</li> </ul>   |  |
| Mallows's $C_p$<br><a href="#">Example</a>                      | $C_p = \frac{RSS_d}{\hat{\sigma}^2} - n + 2d$  | $C_p = \frac{1}{n} (RSS_d + 2d\hat{\sigma}^2)$ |
|   | Lower $C_p$ values indicate better fitting models. The two formulas will identify the same model even if they yield different values.  |  |
| Akaike Information Criterion (AIC)<br><a href="#">Example</a>   | $AIC = \frac{1}{n} (RSS_d + 2d\hat{\sigma}^2)$ <hr/> <p>A lower AIC indicates a better fit, with a penalty for the number of parameters to discourage overfitting. For linear models with Gaussian errors, AIC is proportionate to <math>C_p</math>.</p>   |  |
| Bayesian Information Criterion (BIC)<br><a href="#">Example</a> | $BIC = \frac{1}{n} (RSS_d + \log(n)d\hat{\sigma}^2)$ <hr/> <p>A lower BIC indicates a better fit, with a penalty for the number of parameters to discourage overfitting. BIC penalizes the number of predictors more heavily than AIC, especially as the sample size (<math>n</math>) increases.</p>   |  |
| Adjusted $R^2$<br><a href="#">Example</a>                       | $\text{Adjusted } R^2 = 1 - \frac{RSS/(n - d - 1)}{TSS/(n - 1)}$ <p>Here, <math>d</math> excludes the intercept term.</p> <hr/> <p>Higher adjusted <math>R^2</math> values indicate better fitting models.</p> <p><math>R^2</math> always increases or stays the same when more predictors are added to the model, regardless of whether those predictors are truly relevant. Adjusted <math>R^2</math> penalizes the addition of noise variables to account for the number of predictors, <math>d</math>, in the model.</p> |  |

## 2.11 Residual Analysis

### 2.11.1 Information

The purpose of residual analysis is to check the residuals for patterns or relationships with other variables. It also plays an important role in improving model formulation by identifying additional explanatory variables.

Discrepancies to look for include:

1. **Lack of Independence:** The deviations  $\{\epsilon_i\}$  are not independent.
2. **Heteroscedasticity:** Variability of observations is not constant.
3. **Relationships with Explanatory Variables:** If an explanatory variable can help explain  $\epsilon$ , then it can also help predict  $y$ .
4. **Nonnormal Distributions:** Significant deviation from normality nullifies usual inference procedures. Detected through QQ plots.
5. **Unusual Points:** Outliers or influential observations may disproportionately affect the regression model.

Three strategies for handling outliers include:

1. **Include and Comment**
2. **Delete the Observation**
3. **Create a Binary Variable**

Steps to follow after a preliminary model fit:

1. Display the distribution of residuals to identify outliers.
2. Assess the correlation between residuals and additional explanatory variables to identify linear relationships.
3. Plot residuals against additional explanatory variables to detect nonlinear relationships.

### 2.11.2 Standardized Residuals

| Name   | Formula                              | Notes   |
|--|--------------------------------------|---|
| Standardized Residual                            | $\frac{e_i}{s}$                      | Easy to compute and understand.<br>$s$ approximates the residual standard deviation.  |
| Standardized Residual<br><a href="#">Example</a> | $\frac{e_i}{s\sqrt{1-h_{ii}}}$       | More precise than the first definition.<br>Incorporates the leverage, $h_{ii}$ , and uses the standard error, $se(e_i) = s\sqrt{1-h_{ii}}$ .                                |
| Studentized Residual<br><a href="#">Example</a>  | $\frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}$ | Best for identifying outliers.<br>Excludes the $i$ th observation when estimating the standard error.<br>Follows a $t$ -distribution with $n - (p + 1)$ degrees of freedom. |

## 2.12 Influential Points

| Concept  | Description   |
|--|---|
| Influential Point                              | An <b>influential point</b> is a data point that has a disproportionate impact on the regression line.  |
| Outlier  | An <b>outlier</b> is an observation that is unusual in the vertical direction (unusual response value).   |
| Leverage                                       | $h_{ii} = \mathbf{x}_i^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{x}_i \text{ for } \frac{1}{n} \leq h_{ii} \leq 1$ <p>The <b>leverage</b> of an observation measures the influence that the observation's predictor values have on its fitted value.</p> <hr/> <p>Average leverage</p> $\bar{h} = \frac{1}{n} \sum_{i=1}^n h_{ii} = \frac{p+1}{n}$ <hr/> <p>Leverage in simple linear regression</p> $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^n (x_j - \bar{x})^2}$ |
| High Leverage Point<br><a href="#">Example</a> | <p>A <b>high leverage point</b> is an observation that is unusual in the horizontal directional (unusual predictor value). It can drag the regression line toward itself.</p> <hr/> $h_{ii} > \frac{3(p+1)}{n}$   |
| Addressing High Leverage Points                | <ol style="list-style-type: none"> <li>1. Include the observation with commentary</li> <li>2. Delete the observation</li> <li>3. Choose another variable</li> <li>4. Use a nonlinear transformation</li> </ol>  |
| Cook's Distance                                | <p>Cook's distance measures the change in predicted values when the <math>i</math>-th observation is removed from the model.</p> <div> <math display="block">D_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{(p+1)s^2}</math> <math display="block">D_i = \left( \frac{e_i}{se(e_i)} \right)^2 \frac{h_{ii}}{(p+1)(1-h_{ii})}</math> </div>  |

## 2.13 Collinearity

| Concept  | Description   |
|--|---|
| Collinearity   | <b>Collinearity (multicollinearity)</b> occurs when two or more predictor variables in a regression model are highly correlated, making it difficult to isolate their individual effects on the response variable.  |
| Variance Inflation Factor (VIF)<br><a href="#">Example</a> | $VIF_j = \frac{1}{1 - R_j^2}, \quad \text{for } j = 1, 2, \dots, p$ <p>The VIF quantifies how much the variance of a regression coefficient is inflated due to collinearity with other predictors.</p> <p><math>VIF_j &gt; 90\%</math> indicates severe collinearity.</p> |
| VIF Adjusted Standard Errors                               | $se(b_j) = s\sqrt{VIF_j} \frac{1}{s_{x_j}\sqrt{n-1}}$   |
| Addressing Collinearity                                    | <ol style="list-style-type: none"> <li>1. Center variables</li> <li>2. Acknowledge only</li> <li>3. Substitute with transformed variables</li> </ol>  |

## 2.14 Homoscedasticity and Heteroscedasticity

### 2.14.1 Definitions

| Concept                                       | Description   |
|---|---|
| Homoscedasticity                              | <p><b>Homoscedasticity</b> is the condition in which the variance of the errors is constant across all levels of the independent variable(s).</p> <hr/> $\text{Var}(\epsilon_i) = \sigma^2 \Rightarrow \text{Var}(y_i) = \sigma^2$  |
| Heteroscedasticity                            | <p><b>Heteroscedasticity</b> is the condition in which the variance of the errors is not constant across all levels of the independent variable(s).</p>   |
| Addressing Heteroscedasticity                 | <ol style="list-style-type: none"> <li>1. Heteroscedasticity-consistent standard errors</li> <li>2. Weighted least squares</li> <li>3. Transformation of variables</li> </ol>   |
| Heteroscedasticity-Consistent Standard Errors | $se_r(b_j) = \sqrt{(j+1)\text{st diagonal element of } \widehat{\text{Var}}(b)}$ <hr/> <p><b>Heteroscedasticity-consistent standard errors</b> (robust standard errors) adjust for heteroscedasticity without modifying the coefficients themselves.</p>                                      |
| Weighted Least Squares                        | $\hat{\beta}_{WLS} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}$ <hr/> <p><b>Weighted least squares</b> assigns different weights to observations based on the variability of their residuals, giving less weight to observations with higher variance.</p> |
| Transformation of Variables                   | <p><b>Transforming</b> the dependent variable <math>y</math> can effectively address heteroscedasticity by stabilizing the variance of the error terms.</p>   |

### 2.14.2 Breusch-Pagan Test

| Concept         | Description  |
|-----------------|--|
| Purpose         | The <b>Breusch-Pagan test</b> is used to detect heteroscedasticity in a regression model.  |
| Hypothesis Test | $H_0 : \gamma = 0$ <p>The variance of the residuals is constant (homoscedastic).</p>   |
|                 | $H_A : \text{Var}(y_i) = \sigma^2 + \mathbf{z}_i' \boldsymbol{\gamma}$ <p>The variance of the errors is not constant (heteroscedastic).</p>  |
| Steps           | <ol style="list-style-type: none"> <li>1. Run the regression model and obtain the residuals <math>e_i</math>.</li> <li>2. Compute <math>e_i^{*2} = e_i^2 / s^2</math>.</li> <li>3. Regress the standardized squared residuals <math>e_i^{*2}</math> on <math>\mathbf{z}_i</math>.</li> <li>4. Calculate the test statistic as <math display="block">\text{LM} = \frac{\text{Regress SS}_z}{2}</math>.</li> <li>5. Compare the test statistic to the chi-square distribution with degrees of freedom equal to the number of predictors <math>p</math> (excluding the intercept).</li> </ol> |



## 2.15 Ridge Regression

| Concept                             | Description   |
|-------------------------------------|---|
| Shrinkage Methods                   | <b>Shrinkage methods</b> build on ordinary least squares (OLS) by adding a penalty to the regression coefficients. The two most common shrinkage methods are ridge regression and the lasso.  |
| Ridge Regression                    | <b>Ridge regression</b> modifies the OLS approach by adding a shrinkage penalty to the size of the coefficients. Ridge regression shrinks the coefficients towards zero, but it does not set any of them to exactly zero.   |
| Ridge Regression Objective Function | Minimize:<br>$RSS + \lambda \sum_{j=1}^p \beta_j^2$   |
| Residual Sum of Squares (from OLS)  | $RSS = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2$   |
| L2 (Euclidean) Norm                 | $\ \beta\ _2 = \sqrt{\sum_{j=1}^p \beta_j^2}$ <p>The <b>L2 norm</b> represents the straight-line distance from the origin to the point defined by the vector in <math>p</math>-dimensional space. <a href="#">Geometrically</a>, it creates the shape of the penalty region (a circle in 2D, sphere in 3D, etc.).</p> |
| Tuning Parameter                    | The <b>tuning parameter</b> , $\lambda$ , controls the strength of the shrinkage penalty. Its purpose is to balance the trade-off between fitting the training data well and keeping the model coefficients small to avoid overfitting.   |
| Shrinkage Penalty                   | $\lambda \sum_{j=1}^p \beta_j^2 = \lambda (\ \beta\ _2)^2$ <p>As <math>\lambda</math> increases, the L2 norm decreases. Geometrically, the <b>shrinkage penalty</b> uses the L2 norm to push the coefficients inward.</p>   |

|                                |   |
|--------------------------------|---|
| Scale Equivariance             | <p>Least squares regression is <b>scale equivariant</b>, meaning that multiplying a predictor variable by a constant does not affect the relative contribution of that predictor to the model.</p> <p>Ridge regression is <b>not scale equivariant</b> because the shrinkage penalty depends on the magnitude of the coefficients. To address this, it is best to standardize the predictors before applying ridge regression:</p> $\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}$ |
| Advantages of Ridge Regression | <ol style="list-style-type: none"> <li>1. Lower test MSE can be achieved compared to OLS.</li> <li>2. Useful when dealing with datasets with a large number of predictors.</li> <li>3. The ability to adjust <math>\lambda</math> allows for optimization of overall model performance.</li> <li>4. Ridge regression is computationally more efficient than best subset selection, and can be computed almost as quickly as an OLS model.</li> </ol>  |

## 2.16 Lasso

### 2.16.1 Definitions

The concepts below extend from the [previous section](#).

| Concept   | Description  |
|---|--|
| Least Absolute Shrinkage and Selection Operator (LASSO) | <p><b>Lasso</b>, or the lasso, modifies the OLS approach by adding a shrinkage penalty to the size of the coefficients. Unlike ridge regression, lasso sets some coefficients exactly to zero.</p> <p>The ability of lasso to set some coefficients exactly to zero means it can apply <b>feature selection</b> by removing less important predictors.</p> |
| Ridge Regression Objective Function                     | <p>Minimize:</p> $RSS + \lambda \sum_{j=1}^p  \beta_j $  |
| Residual Sum of Squares (from OLS)                      | $RSS = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2$  |
| L1 Norm   | $\ \beta\ _1 = \sum_{j=1}^p  \beta_j $ <p><a href="#">Geometrically</a>, the <b>L1 norm</b> creates a diamond-shaped constraint region in parameter space.</p>   |

### 2.16.2 A Geometric Interpretation of Ridge Regression and Lasso

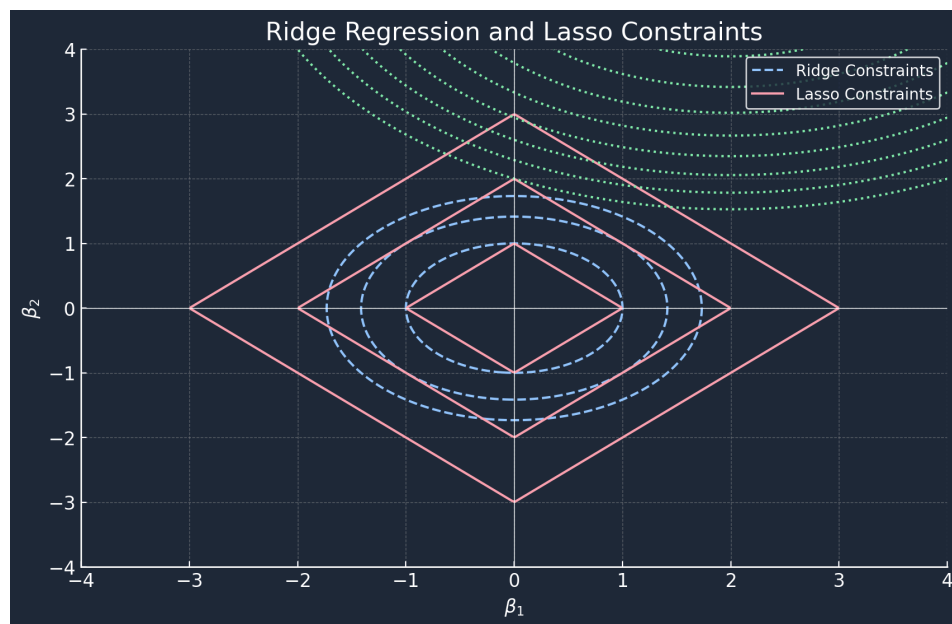
Ridge regression aims to find the coefficient vector that minimizes:

$$\min_{\beta} \{\text{RSS}\} \quad \text{subject to} \quad \sum_{j=1}^p \beta_j^2 \leq s$$

Lasso aims to find the coefficient vector that minimizes:

$$\min_{\beta} \{\text{RSS}\} \quad \text{subject to} \quad \sum_{j=1}^p |\beta_j| \leq s$$

Consider a simple example with two predictors ( $p = 2$ ).



The constraint  $\beta_1^2 + \beta_2^2 \leq s$  forms circular regions and the constraint  $|\beta_1| + |\beta_2| \leq s$  forms diamond-shaped regions for varying values of  $s$ . The solution to the optimization problem is the point where the RSS contour (dotted green curve) first touches a constraint boundary.

- For ridge regression, the smooth boundary of the circle means the solution is less likely to intersect at exactly zero ( $\beta_1 = 0$  or  $\beta_2 = 0$ ). Hence, ridge regression does not perform feature selection, since the coefficients are never exactly zero.
- For lasso, the diamond-shaped constraint has sharp corners at  $\beta_1 = 0$  and  $\beta_2 = 0$ . When the RSS contour intersects the diamond, it's likely to intersect at a corner where one of the coefficients is zero. This property makes the lasso perform feature selection.

## 2.17 Binary Dependent Variables

| Concept            | Description   |
|--------------------|---|
| Definition         | The <b>linear probability model</b> is a regression model that estimates the probability of a binary outcome as a linear function of the predictors.  |
| True Model Form    | $y_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i$ <hr/> $y_i$ is a binary variable that takes on a value of 0 or 1.   |
| Expected Value     | $\mathbb{E}(y_i) = \mathbf{x}_i' \boldsymbol{\beta} = \pi_i$ <p>where <math>\pi_i</math> is the probability that the outcome equals 1.</p>  |
| Variance           | $\text{Var}(y_i) = \mathbf{x}_i' \boldsymbol{\beta} (1 - \mathbf{x}_i' \boldsymbol{\beta})$   |
| Drawbacks          | <ol style="list-style-type: none"> <li>1. The expected value of the response is not inherently restricted to the <math>[0, 1]</math> interval, and may not be a valid probability.</li> <li>2. The variance of the error term in the linear probability model is not constant, leading to heteroscedasticity.</li> <li>3. The response variable <math>y_i</math> can only take values of 0 or 1, whereas the residuals are continuous.</li> </ol>   |
| Alternative Models | <p>Alternative models (e.g. <a href="#">logit</a> and <a href="#">probit</a> models) of the form</p> $\pi_i = \pi(\mathbf{x}_i' \boldsymbol{\beta}) = \Pr(y_i = 1   \mathbf{x}_i)$ <p>overcome the drawbacks of the linear probability model, where <math>\pi(\cdot)</math> is a predefined function.</p> <hr/> <p>Interpretation: The model form specifies that the conditional probability of <math>y_i = 1</math> given the predictors is a function of the linear combination of the predictors and the coefficients.</p> |

## 2.18 Logit and Probit Models

### 2.18.1 Logit Models

| Concept  | Description   |
|--|---|
| Logistic Regression                                  | <b>Logistic regression</b> (logit regression) is a statistical model that estimates the probability of a binary outcome by modeling the log-odds of the event as a linear combination of predictor variables.                                       |
| Logit Function                                       | $\text{logit}(p) = \ln \left( \frac{p}{1-p} \right)$  |
| Logistic Regression Model<br><a href="#">Example</a> | $\text{logit}(\pi_i) = \mathbf{x}_i' \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i1} + \dots \beta_p x_{ip}$  |
| Logistic Function                                    | $\pi(z) = \frac{1}{1 + \exp(-z)} = \frac{e^z}{1 + e^z}$ <hr/> <p>The logistic function maps any real-valued number to the (0, 1) interval, making it suitable for modeling probabilities.</p>   |
| Odds   | $\frac{p}{1-p}$   |
| Log-Odds   | $\ln \left( \frac{p}{1-p} \right)$  |
| Odds Ratio   | $e^{\beta_j} = \frac{\Pr(y_i = 1   x_{ij} = 1) / (1 - \Pr(y_i = 1   x_{ij} = 1))}{\Pr(y_i = 1   x_{ij} = 0) / (1 - \Pr(y_i = 1   x_{ij} = 0))}$ <hr/> <p>The <b>odds ratio</b> compares the odds of an event occurring in two different groups.</p> |

### 2.18.2 Probit Models

| Concept           | Description   |
|-------------------|---|
| Probit Regression | <p><b>Probit regression</b> is a statistical model that estimates the probability of a binary outcome by mapping a linear combination of predictor variables through the standard normal cumulative distribution function.</p> <hr/> $\pi(z) = \Phi(z)$ |

### 2.18.3 Threshold Interpretation

| Concept                                    | Description   |
|--|---|
| Underlying Linear Model                    | <p>The logit and probit models can be interpreted to have an underlying linear model:</p> $y_i^* = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i^*$ <p>where <math>y_i^*</math> is a continuous variable that captures the underlying continuous process, but is not directly observable from the data.</p> |
| Threshold Model<br><a href="#">Example</a> | $y_i = \begin{cases} 0 & \text{if } y_i^* \leq 0 \\ 1 & \text{if } y_i^* > 0 \end{cases}$   |

### 2.18.4 Parameter Estimation

| Concept                                       | Description  |                                 |
|---|--|---------------------------------|
| Log-Likelihood Function for a Binary Variable | $l(\beta) = \sum_{i=1}^n [y_i \ln \pi(\mathbf{x}'_i \beta) + (1 - y_i) \ln(1 - \pi(\mathbf{x}'_i \beta))]$ |                                 |
| Likelihood Ratio Test                         | $LRT = 2 \left( L(\hat{\beta}_{MLE}) - L_0 \right)$  |                                 |
| Max-Scaled $R^2$                              | $R_{ms}^2 = \frac{R^2}{R_{max}^2}$   |                                 |
|   | $R^2 = 1 - \frac{\exp(L_0/n)}{\exp(L(\hat{\beta}_{MLE})/n)}$   | $R_{max}^2 = 1 - \exp(L_0/n)^2$ |
| Pseudo- $R^2$                                 | $\text{Pseudo-}R^2 = \frac{L(\hat{\beta}_{MLE}) - L_0}{L_{max} - L_0}$                                     |                                 |



## 2.19 Nominal Dependent Variables

### 2.19.1 Generalized Logit Model

| Concept  | Description  |
|--|--|
| Nominal Dependent Variable                         | A <b>nominal dependent variable</b> is a categorical outcome whose values represent distinct, unordered groups with no inherent ranking.   |
| Creating Dummy Variables                           | A dummy variable can be added to a regression model when a qualitative predictor is binary. When a qualitative predictor has more than two levels, multiple dummy variables can be created.  |
| Generalized Logit Model                            | For the $j$ -th category ( $j = 1, 2, \dots, c$ ) and $i$ -th observation, define the linear predictor as:<br>$V_{i,j} = \mathbf{x}_i' \boldsymbol{\beta}_j$   |
| Baseline Category ( $c$ )                          | $P(y_i = c) = \frac{1}{\sum_{k=1}^c \exp(V_{i,k})}$<br>Category $c$ is the baseline category, which anchors all comparisons. Assume $\boldsymbol{\beta}_c = 0$ .   |
| Category Probabilities<br><a href="#">Example</a>  | $P(y_i = j) = \pi_{i,j} = \frac{\exp(V_{i,j})}{\sum_{k=1}^c \exp(V_{i,k})}$<br>Represents an estimate of the probability of outcome $j$ from linear predictors.  |
| Log-Odds Interpretation<br><a href="#">Example</a> | $\ln \frac{P(y_i = j)}{P(y_i = c)} = V_{i,j} - V_{i,c} = \mathbf{x}_i' \boldsymbol{\beta}_j$<br>Each $\boldsymbol{\beta}_j$ captures the change in the <b>log-odds</b> of category $j$ versus $c$ for a unit change in predictor $k$ . |
| Special Case                                       | When $c = 2$ and $\boldsymbol{\beta}_c = 0$ , the generalized logit model reduces to the <a href="#">logit model</a> .   |

### 2.19.2 Other Models

| Concept                 | Description   |
|-------------------------|---|
| Multinomial Logit Model | <p>The <b>multinomial logit model</b> assumes the same set of coefficients, <math>\beta</math>, for explanatory variables across all alternatives:</p> $V_{i,j} = \mathbf{x}'_{i,j}\beta$   |
|                         | <p>The generalized logit model is a special case of the multinomial logit model.</p>  |
|                         | $\ln \left( \frac{\Pr(y_i = h)}{\Pr(y_i = k)} \right) = (\mathbf{x}_{i,h} - \mathbf{x}_{i,k})'\beta$ <p>A feature of the multinomial logit model is the <b>independence of irrelevant alternatives</b>, which implies that the ratio of the probabilities of choosing any two alternatives <math>j</math> and <math>k</math> is independent of the presence or characteristics of other alternatives.</p> |
| Nested Logit Model      | <p>The <b>nested logit model</b> addresses the issue of independence of irrelevant alternatives by creating nested structures.</p> <p>It decomposes the probability of selecting a specific category into two components: the probability of selecting a particular nest and the conditional probability of choosing an option within that nest.</p>  |
| Mixed Logit Model       | <p>A <b>mixed logit model</b> is a multinomial logit model that relies on both <math>\beta</math> and <math>\beta_j</math>.</p> $V_i = \mathbf{x}'_{i,1,j}\beta + \mathbf{x}'_{i,2,j}\beta_j$   |

## 2.20 Ordinal Dependent Variables

| Concept                    | Description   |
|----------------------------|---|
| Ordinal Dependent Variable | An <b>ordinal dependent variable</b> is a categorical outcome whose values represent a natural ordering among the categories.   |
| Cumulative Probability     | <p><b>Cumulative probabilities</b> are used to model how explanatory variables influence the probability of being at or below each category level of an ordinal outcome.</p> <hr/> <p>For an ordinal variable <math>y</math> with <math>c</math> categories:<br/> <math>\Pr(y \leq j) = \pi_1 + \pi_2 + \cdots + \pi_j, \quad j = 1, 2, \dots, c</math></p> |
| Cumulative Logit           | $\text{logit}(\Pr(y \leq j)) = \ln \left( \frac{\Pr(y \leq j)}{1 - \Pr(y \leq j)} \right)$ $= \ln \left( \frac{\pi_1 + \pi_2 + \cdots + \pi_j}{\pi_{j+1} + \pi_{j+2} + \cdots + \pi_c} \right)$   |
| Cumulative Logit Model     | $\text{logit}(\Pr(y \leq j)) = \alpha_j$ $\alpha_1 \leq \alpha_2 \leq \cdots \leq \alpha_{c-1}$   |
| Proportional Odds Model    | $\text{logit}(\Pr(y \leq j)) = \alpha_j + \mathbf{x}_i' \boldsymbol{\beta}$   |
| Cumulative Probit Model    | $\Pr(y_i \leq j) = \Pr(y_i^* \leq \alpha_j) = \Phi(\alpha_j - \mathbf{x}_i' \boldsymbol{\beta})$  |

## 2.21 Poisson Regression

| Concept                   | Description   |   |
|---------------------------|---|---|
| Poisson Regression        | <b>Poisson regression</b> is a type of <a href="#">generalized linear model</a> used to model count data. It assumes that the response variable $y$ follows a Poisson distribution.   |   |
| Exposure                  | $E[y_i] = E_i \times \mu$ $E[y_i] = \mu_i = \exp(\mathbf{x}'_i \boldsymbol{\beta})$ <hr/> $E_i$ represents the <b>exposure</b> for the $i$ -th observation and $\mu$ is the mean from the Poisson distribution. This adjustment is made to account for different levels of exposure among observations. |   |
| Logarithmic Link Function | $\ln(\mu_i) = \ln(E_i) + \mathbf{x}'_i \boldsymbol{\beta}$  |   |
| Partial Derivative        | $\frac{\partial E[y_i]}{\partial x_{ij}} \times \frac{1}{E[y_i]} = \beta_j$ <hr/> $\beta_j$ represents the proportional change in the mean for a one-unit change in $x_{ij}$ .  |   |
| Goodness-of-Fit           | <b>General Form</b><br>$\frac{\sum (observation - estimate)^2}{estimate}$   | <b>Pearson's Chi-Square Statistic</b><br>$\sum_{j=1}^n \frac{(n_j - n\hat{p}_j)^2}{n\hat{p}_j}$ |
|                           | <b>Pearson Residual</b><br>$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$   | <b>Pearson Goodness-of-Fit Statistic</b><br>$\chi^2 = \sum_{i=1}^n r_i^2$                       |

## 2.22 Other Count Models

| Concept                         | Description  |
|---------------------------------|--|
| Drawbacks of Poisson Regression | <p>The Poisson regression model's simplicity can be too restrictive due to its assumption of <b>equidispersion</b> (mean equals variance).</p> <p>To address this, a common adjustment is to assume <math>\text{Var}(y_i) = \phi\mu_i</math>, where <math>\phi &gt; 0</math> accounts for dispersion.</p>  |
| Negative Binomial Distribution  | <p>The negative binomial distribution models the number of successes until an experience is stopped.</p> <hr/> <p>Advantages of using a dependent variable that follows the negative binomial distribution:</p> <ol style="list-style-type: none"> <li>1. It has more flexibility due to more parameters.</li> <li>2. The Poisson distribution is a special case of the negative binomial distribution.</li> <li>3. The negative binomial distribution can be derived as a mixture of Poisson distributions with a gamma-distributed rate parameter <math>\lambda</math>.</li> <li>4. The negative binomial distribution allows for straightforward estimation of features.</li> </ol> |
| Zero-Inflated Models            | <p><b>Zero-inflated models</b> are useful for understanding the probability of observing zero counts as a combination of genuine zeros from the count distribution and inflated zeros from non-reporting.</p>  |
| Hurdle Models                   | <p><b>Hurdle models</b> offer another approach to handle datasets with an excess number of zeros. These models are motivated by sequential decision-making processes. For example, in healthcare, an individual's decision to seek care (first hurdle) is distinct from the amount of care received (second hurdle).</p>   |
| Heterogeneity Models            | <p><b>Heterogeneity models</b> introduce one or more random parameters (e.g. <math>\alpha_i</math>, the heterogeneity component) to capture unobserved characteristics in the model.</p>   |
| Latent Class Models             | <p><b>Latent class models</b> aim to classify and homogenize observations by using a discrete random variable to modify basic count distributions. This means the classification is not directly observed but inferred through the model.</p>  |

## 2.23 Generalized Linear Models

### 2.23.1 Definitions

| Concept                                  | Description   |                                    |
|--|---|------------------------------------|
| Generalized Linear Model                 | <p>In ordinary linear regression, the response variable and residuals are assumed to be normally distributed with constant variance.</p> <p><b>Generalized linear models</b> (GLMs) extend this framework by allowing the response to follow any distribution in the linear exponential family.</p> |                                    |
| Linear Exponential Family                | <p>The <b>linear exponential family</b> is a class of probability distribution functions that can be expressed in the form:</p> $f(y; \theta, \phi) = \exp \left( \frac{y\theta - b(\theta)}{\phi} + S(y, \phi) \right)$  |                                    |
|  | $E[y] = \mu = b'(\theta)$   | $\text{Var}(y) = \phi b''(\theta)$ |
| Systematic Component                     | <p>For an observation <math>i</math> with <math>p</math> predictors, the <b>systematic component</b> is:</p> $\eta_i = \mathbf{x}'_i \boldsymbol{\beta} = x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{ip}\beta_p$   |                                    |
| Link Function<br><a href="#">Example</a> | <p>The <b>link function</b> is <math>g(\cdot)</math> where:</p> $\eta_i = g(\mu_i) = \mathbf{x}'_i \boldsymbol{\beta}$ <p>The purpose of the link function is to connect the mean of the response variable, <math>\mu_i</math>, to the linear predictor, <math>\eta_i</math>.</p>                   |                                    |
| Mean Function                            | <p>The inverse of the link function, <math>g^{-1}(\cdot)</math>, is the <b>mean function</b>:</p> $\mu_i = g^{-1}(\mathbf{x}'_i \boldsymbol{\beta})$ <p>It is used to obtain the <b>mean response</b>, <math>\mu_i = E[y_i]</math>, from the linear predictor.</p>                                  |                                    |

### 2.23.2 Canonical Link Function

The **canonical link function**  $g(\cdot)$  is a specific type of link function where:

$$g(\mu) = \theta = (b')^{-1}(\mu)$$

This relationship ensures that the linear predictor  $\eta$  is equal to the canonical parameter  $\theta$ , leading to more straightforward interpretations and simpler computations.

Below are mean functions  $b'(\theta)$  and canonical link functions  $g(\mu)$  for common distributions in the linear exponential family:

| Distribution     | Mean Function $b'(\theta)$      | Canonical Link $g(\mu)$               |
|------------------|---------------------------------|---------------------------------------|
| Normal           | $\theta$                        | $\mu$                                 |
| Bernoulli        | $\frac{e^\theta}{1 + e^\theta}$ | $\ln\left(\frac{\mu}{1 - \mu}\right)$ |
| Poisson          | $e^\theta$                      | $\ln(\mu)$                            |
| Gamma            | $-\frac{1}{\theta}$             | $-\frac{1}{\mu}$                      |
| Inverse Gaussian | $(-2\theta)^{-1/2}$             | $-\frac{1}{2\mu^2}$                   |

### 2.23.3 Linear Regression Sampling Assumptions on GLMs

| Linear Regression  | Generalized Linear Model  |
|--|---|
| $E[y_i] = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{ip}$ | Generalized through the link function.                                  |
| $\{x_{i1}, \dots, x_{ip}\}$ are nonstochastic variables.     | Also applies to GLMs.   |
| $Var(y_i) = \sigma^2$  | $Var(y_i) = \phi v(\mu_i)$<br>See the <a href="#">following table</a> . |
| $\{y_i\}$ are independent random variables.                  | Also applies to GLMs.   |
| $\{y_i\}$ are normally distributed.                          | Not required for GLMs.  |

#### 2.23.4 Variance Function

| Distribution     | Variance Function $v(\mu)$ |
|------------------|----------------------------|
| Normal           | 1                          |
| Bernoulli        | $\mu(1 - \mu)$             |
| Poisson          | $\mu$                      |
| Gamma            | $\mu^2$                    |
| Inverse Gaussian | $\mu^3$                    |

#### 2.23.5 The Tweedie Distribution

| Concept                  | Description  |
|--------------------------|--|
| The Tweedie Distribution | The <b>Tweedie distribution</b> is a flexible family of probability distributions that includes several special cases such as the normal, Poisson, gamma, and inverse Gaussian distributions.  |
| PDF                      | $f_S(y) = \exp \left( -\frac{1}{\phi} \left( \frac{\mu^{2-p}}{2-p} + y \cdot \frac{\mu^{p-1}}{p-1} \right) + S(y, \phi) \right)$   |
| Mean                     | $E[S_N] = \mu$   |
| Variance                 | $\text{Var}(S_N) = \phi \mu^p$ <p>The power parameter <math>p</math> determines the specific type of distribution where <math>1 &lt; p &lt; 2</math>.</p> <p>From the <a href="#">variance function table</a>, the Tweedie distribution can be viewed as a choice between the Poisson and gamma distributions.</p> |



## 2.24 Estimation in GLMs

### 2.24.1 Maximum Likelihood Estimation for Canonical Links

| Concept   | Description   |
|---|---|
| Linear Exponential Family (PDF)                   | $f(y; \theta, \phi) = \exp \left( \frac{y\theta - b(\theta)}{\phi} + S(y, \phi) \right)$  |
| Linear Exponential Family Log-Likelihood Function | $\ell(\beta) = \sum_{i=1}^n \log f(y_i; \theta_i, \phi_i)$  |
| Log-Likelihood Function for Canonical Links       | $\ell(\beta) = \sum_{i=1}^n \left( \frac{y_i(\mathbf{x}_i' \beta) - b(\mathbf{x}_i' \beta)}{\phi_i} + S(y_i, \phi_i) \right)$   |
| Information Matrix                                | $\mathbf{I}(\hat{\beta}) = -E \left[ \frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta'} \right] = \sum_{i=1}^n \left( \frac{b''(\mathbf{x}_i' \hat{\beta})}{\phi/w_i} \mathbf{x}_i \mathbf{x}_i' \right)$  |
| Overdispersion                                    | <p><b>Overdispersion</b> occurs when the observed variability in the response variable is greater than what is expected under the assumed distribution of the GLM.</p> <p>In the presence of overdispersion, the variance can be approximated by:</p> $\text{Var}(y_i) = \sigma^2 \phi b''(\theta_i) / w_i$ |

### 2.24.2 Goodness-of-Fit Statistics for GLMs

| Concept                                       | Description   |
|---|---|
| Sum of Squares<br><a href="#">Example</a>     | $\sum_i (y_i - \bar{y})^2 = \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2 \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$ <hr/> <p>In <a href="#">linear regression</a>, the cross-product term equals zero because the residuals <math>(y_i - \hat{y}_i)</math> are uncorrelated with the fitted values <math>(\hat{y}_i)</math>. For nonlinear models, the cross-product is rarely zero.</p> |
| Pearson Chi-Square Statistic                  | $\chi^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\phi v(\hat{\mu}_i)}$ <hr/> <p>In nonlinear models, <math>R^2</math> is not applicable, due to the cross-product term (<math>\text{TSS} \neq \text{RSS} + \text{Regression SS}</math>). <b>The Pearson chi-square statistic</b> is an alternative to <math>R^2</math>.</p>  |
| Deviance Statistic<br><a href="#">Example</a> | $D(\hat{\theta}) = 2\phi [l(\text{saturated model}) - l(\text{fitted model})]$ <hr/> <p>The deviance statistic measures the difference between the fitted model and the saturated model (the model with the best possible fit).</p>   |
| Deviance Statistic (Specific Cases)           | <p>Normal Distribution</p> $D(\hat{\mu}) = \sum_i (y_i - \hat{\mu}_i)^2$ <hr/> <p>Bernoulli Distribution</p> $D(\hat{\pi}) = \sum_i \left[ y_i \ln \left( \frac{y_i}{\hat{\pi}_i} \right) + (1 - y_i) \ln \left( \frac{1 - y_i}{1 - \hat{\pi}_i} \right) \right]$ <hr/> <p>Poisson Distribution</p> $D(\hat{\mu}) = \sum_i \left[ y_i \ln \left( \frac{y_i}{\hat{\mu}_i} \right) + (y_i - \hat{\mu}_i) \right]$ |

### 2.24.3 Residual Analysis for GLMs

| Concept             | Description   |
|---------------------|---|
| Raw Residuals       | In linear models, residuals are the difference between observed responses and fitted values. For GLMs, these are termed <b>raw residuals</b> and are denoted $y_i - \hat{\mu}_i$ . Raw residuals are not reliable for GLMs due to heteroscedasticity. |
| Cox-Snell Residuals | $e_i = R(y_i; \mathbf{x}_i, \hat{\boldsymbol{\theta}})$   |
| Pearson Residuals   | $R(y_i; \mathbf{x}_i, \boldsymbol{\theta}) = \frac{y_i - \mu_i}{\sqrt{\text{Var}(y_i)}}$  |
| Anscombe Residuals  | $R(y_i; \mathbf{x}_i, \boldsymbol{\theta}) = \frac{h(y_i) - E[h(y_i)]}{\sqrt{\text{Var}(h(y_i))}}$  |
| Deviance Residuals  | $R(y_i; \mathbf{x}_i, \boldsymbol{\theta}) = \text{sign}(y_i - \hat{\mu}_i) \sqrt{2 \left[ \ln f(y_i; \boldsymbol{\theta}_{i,\text{SAT}}) - \ln f(y_i; \hat{\boldsymbol{\theta}}_i) \right]}$   |

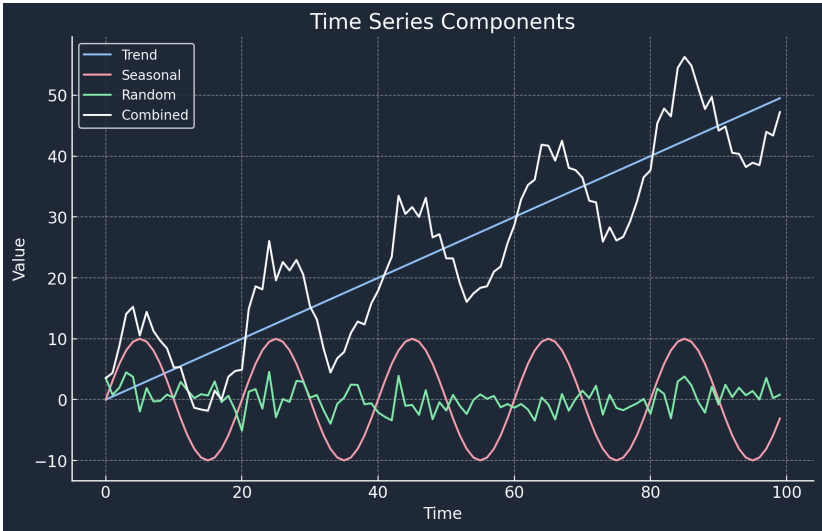
### **3. Time Series Models (Learning Objective 3)**

### 3.1 Introduction to Time Series

#### 3.1.1 Key Terms

| Term                 | Definition  | Example  |
|----------------------|---|--|
| Stochastic Process   | An ordered sequence of random variables indexed by time or space, used to model random evolution over time. | Evolution of a stock price over time.  |
| Longitudinal Data    | Data consisting of repeated measurements over time for the same subjects.                                   | Blood pressure recorded for patients over 10 years.                                    |
| Time Series          | A sequence of data points recorded at regular intervals, indexed by time $t$ .                              | Daily temperature recordings over a year.  |
| Cross-Sectional Data | Data collected at a single point in time across multiple subjects.  | Income, education, and occupation data from a survey of 1,000 individuals at one time. |
| Panel Data           | A combination of cross-sectional and longitudinal data. Tracks multiple entities over time.                 | Yearly income data for multiple households tracked over 5 years.                       |

3.1.2 Time Series Models

| Concept                | Description   |                                     |  |
|------------------------|---|-------------------------------------|--|
| Time Series Models     | <b>Time series models</b> capture and predict the behavior of variables over time to identify patterns, trends, and seasonal effects.   |                                     |  |
| Forecasting Components | <div>- Trends (<math>T_t</math>) reflect long-term movements.</div> <div>- <a href="#">Seasonal patterns</a> (<math>S_t</math>) capture periodic fluctuations.</div> <div>- Irregular or random patterns (<math>\epsilon_t</math>) represent unpredictable short-term changes.</div> <div></div>                                   |                                     |  |
| Model Forms            | $y_t = T_t + S_t + \epsilon_t$  | $y_t = T_t \times S_t + \epsilon_t$ | $y_t = T_t \times S_t \times \epsilon_t$ |
| Trend Examples         | <div>1. Linear trend: <math>y_t = \beta_0 + \beta_1 t + \epsilon_t</math></div> <div>2. Quadratic trend: <math>y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon_t</math>.</div> <div>3. Binary trend: <math>y_t = \beta_0 + \beta_1 z_t + \epsilon_t</math> where <math>z_t = \{0, 1\}</math>.</div> <div>4. Regime-switching trend: A trend that switches between two different regimes at a specified time.</div> |                                     |  |



## 3.2 Stationarity

### 3.2.1 Stationarity

| Concept                | Definition  |
|------------------------|---|
| Weak Stationarity      | <ol style="list-style-type: none"> <li>1. The mean <math>E(y_t)</math> is constant and does not depend on <math>t</math>.</li> <li>2. The covariance <math>Cov(y_s, y_t)</math> depends only on the lag <math> t - s </math>.</li> </ol> <hr/> <p>An implication of the second point is constant variance:<br/> <math>Cov(y_t, y_t) = Var(y_t) = Var(y_s) = \sigma^2</math> (homoscedasticity).</p> |
| Strong Stationarity    | The entire distribution, and not just the mean and variance, of $y_t$ is constant over time.  |
| Detecting Stationarity | Use a control chart with an upper control limit (e.g. $\bar{y} + 3s_y$ ) and a lower control limit (e.g. $\bar{y} - 3s_y$ ).  |

### 3.2.2 White Noise

| Concept                           | Definition   |
|-----------------------------------|--|
| Definition                        | <p><b>White noise:</b></p> <ol style="list-style-type: none"> <li>1. Resembles a stationary series, <math>\{y_t\}</math>, with no discernible pattern over time (i.i.d. random variables with zero mean and constant variance).</li> <li>2. Used as a benchmark for randomness.</li> <li>3. Once all systematic patterns, trends, and correlations have been filtered out from a time series, the remaining residuals ought to resemble white noise (verify with a plot).</li> </ol> |
| Forecast Interval for White Noise | $\bar{y} \pm t_{T-1, \alpha/2} s_Y \sqrt{1 + \frac{1}{T}}$   |

### 3.2.3 Random Walk

| Concept  | Definition  |
|--|---|
| Definition                                       | <p>A <b>random walk</b> is a time series, <math>\{y_t\}</math>, where each value is the previous value plus a random step (i.e. white noise, <math>\{c_t\}</math>).</p> <hr/> <p>The random walk is non-stationary because both the mean and variance depend on <math>t</math>:</p> $E[y_t] = y_0 + t\mu_c$ $Var(y_t) = Var(y_0 + c_1 + c_2 + \dots + c_t)$ |
| Forecast Estimate                                | $\hat{y}_{T+l} = y_T + l\bar{c}$ <hr/> <p>The point estimate for a forecast <math>l</math> lead-time units in the future,</p> $y_{T+l} = y_T + c_{T+1} + c_{T+2} + \dots + c_{T+l}.$  |
| 95% Forecast Interval<br><a href="#">Example</a> | $y_T + l\bar{c} \pm 2s_c\sqrt{l}$ <hr/> <p><math>y_{T+l}</math> is expected to be in this interval approximately 95% of the time.</p>   |
| Detecting a Random Walk                          | <ol style="list-style-type: none"> <li>1. Use a control chart.</li> <li>2. If the series follows a random walk model, the differenced series should follow a white noise process.</li> <li>3. A random walk model will exhibit a higher standard deviation in the original series than in the differenced series.</li> </ol>                                |



### 3.3 Forecast Evaluation

#### 3.3.1 Out-of-Sample Validation Process

| Step                 | Description  |
|----------------------|--|
| 1. Divide the Sample | Split the sample of size $T$ into two subsamples:<br>- Model Development Subsample: $t = 1, \dots, T_1$<br>- Model Validation Subsample: $t = T_1 + 1, \dots, T_1 + T_2$   |
| 2. Fit the Model     | Using the model development subsample ( $t = 1, \dots, T_1$ ), fit a candidate model to the dataset.   |
| 3. Forecast          | With the model from Step 2 and the dependent variables up to and including $t - 1$ , forecast the dependent variable $\hat{y}_t$ for $t = T_1 + 1, \dots, T_1 + T_2$ .   |
| 4. Compute Residuals | Use actual observations and the fitted values from Step 3 to compute one-step forecast residuals ( $e_t = y_t - \hat{y}_t$ ) for the model validation subsample. Summarize these residuals with comparison statistics. |
| 5. Choose a Model    | Repeat Steps 2 through 4 for each candidate model. Choose the model with the smallest set of comparison statistics.  |

### 3.3.2 Statistics for Comparing Forecasts

| Statistic  | Formula   |
|--|---|
| Mean Error (ME)<br><a href="#">Example</a>                       | $ME = \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} e_t$ <hr/> Measures recent trends not anticipated by the model.  |
| Mean Percentage Error (MPE)<br><a href="#">Example</a>           | $MPE = \frac{100}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \frac{e_t}{y_t}$ <hr/> Measures error relative to the actual value, indicating trends.  |
| Mean Square Error (MSE)<br><a href="#">Example</a>               | $MSE = \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2} e_t^2$ <hr/> Detects more patterns than ME.   |
| Mean Absolute Error (MAE)<br><a href="#">Example</a>             | $MAE = \frac{1}{T_2} \sum_{t=T_1+1}^{T_1+T_2}  e_t $ <hr/> Detects more patterns than ME, with units same as the dependent variable.  |
| Mean Absolute Percentage Error (MAPE)<br><a href="#">Example</a> | $MAPE = \frac{100}{T_2} \sum_{t=T_1+1}^{T_1+T_2} \left  \frac{e_t}{y_t} \right $ <hr/> Similar to MAE, MAPE detects more than trend patterns. Similar to MPE, MAPE examines error relative to the actual value. |

### 3.4 Autoregressive Models

#### 3.4.1 Autocorrelation

| Concept   | Definition   |
|---|--|
| Definition  | <b>Autocorrelation</b> is a measure of how much a time series is linearly correlated with a lagged version of itself.  |
| Correlation Statistic<br>( $r$ )                              | $r = \frac{1}{(T-1)s_x s_y} \sum_{t=1}^T (x_t - \bar{x})(y_t - \bar{y})$   |
| Lag-1 Autocorrelation<br>( $r_1$ )<br><a href="#">Example</a> | $r_1 = \frac{\sum_{t=2}^T (y_{t-1} - \bar{y})(y_t - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2}$   |
| Lag-k Autocorrelation<br>( $r_k$ )<br><a href="#">Example</a> | $r_k = \frac{\sum_{t=k+1}^T (y_{t-k} - \bar{y})(y_t - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2}$   |
| Interpretation of $r_k$                                       | <ul style="list-style-type: none"> <li>- Positive autocorrelation: a high value at time <math>t</math> implies high value at time <math>t + k</math></li> <li>- Negative autocorrelation: a high value at time <math>t</math> implies a low value at time <math>t + k</math>, indicating a mean-reverting behavior</li> <li>- Zero autocorrelation: implies no linear relationship between the values of the series at different times, suggesting that the series behaves like white noise</li> </ul> |

### 3.4.2 AR(1) Model

| Concept                                       | Definition  |
|---|---|
| Order 1 Autoregressive Model                  | <p>An <b>order 1 autoregressive model</b> (<math>AR(1)</math>) is a time series process that depends linearly on the immediately preceding value plus random noise.</p> <hr/> $y_t = \beta_0 + \beta_1 y_{t-1} + \epsilon_t, \quad t = 2, \dots, T$ |
| Stationarity                                  | $\beta_1$ must lie strictly between -1 and 1 ( $-1 < \beta_1 < 1$ ) to ensure that the $AR(1)$ model is stationary.   |
| Case: $\beta_1 = 1$                           | <p>The model simplifies to a random walk:</p> $y_t - y_{t-1} = \beta_0 + \epsilon_t$  |
| Case: $\beta_1 = 0$                           | <p>The model reduces to a white noise process:</p> $y_t = \beta_0 + \epsilon_t$   |
| Lag-k Autocorrelation Function ( $\rho_k$ )   | $\rho_k = \text{Corr}(y_t, y_{t-k}) = \frac{\text{Cov}(y_t, y_{t-k})}{\sqrt{\text{Var}(y_t)\text{Var}(y_{t-k})}} = \frac{\text{Cov}(y_t, y_{t-k})}{\sigma_y^2}$ <hr/> $\rho_k = 0 \text{ when } \beta_1 = 0 \text{ (white noise)}$                  |
| Fitting a Model                               | Match observed autocorrelations, $r_k$ , with theoretical expectations, $\rho_k$ , to determine if an autoregressive model is a good fit.   |
| Conditional Least Squares Estimates           | $\hat{\beta}_1 \approx r_1$ and $\hat{\beta}_0 \approx \bar{y}(1 - r_1)$  |
| Residuals                                     | $e_t = y_t - (\hat{\beta}_0 + \hat{\beta}_1 y_{t-1})$   |
| Variance                                      | $\sigma_y^2(1 - \beta_1^2) = \sigma^2$  |
| Mean Squared Error<br><a href="#">Example</a> | $s^2 = \frac{1}{T-3} \sum_{t=2}^T (e_t - \bar{e})^2$  |
| Smoothed Series                               | $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 y_{t-1}$   |

|  |  |
|--|--|
| <p>K-Step Ahead<br/>Forecast and Forecast<br/>Interval</p> | $\hat{y}_{T+k} = \hat{\beta}_0 + \hat{\beta}_1 \hat{y}_{T+k-1}$ <hr/> $\hat{y}_{T+k} \pm t_{\alpha/2} \cdot s \sqrt{1 + \hat{\beta}_1^2 + \hat{\beta}_1^4 + \dots + \hat{\beta}_1^{2(k-1)}}$ |
|--|--|



### 3.5 Smoothing

| Concept  | Definition   |  |
|--|--|--|
| Smoothing  | <b>Smoothing</b> techniques are used to reduce noise and better reveal the underlying trend in a time series.  |  |
| Moving Average<br>(Running Average)  | $\hat{s}_t = \frac{y_t + y_{t-1} + \cdots + y_{t-k+1}}{k} = \hat{s}_{t-1} + \frac{y_t - y_{t-k}}{k}$ <hr/> The adjustment term accounts for a new data point entering the average and an old data point leaving the average. |  |
| Double Smoothing<br><a href="#">Example</a>                                | Consider the Model<br>$y_t = \beta_0 + \beta_1 t + \epsilon$   |  |
|  | <p>Smoothed Series</p> $\hat{s}_t^{(1)} = \frac{y_t + \cdots + y_{t-k+1}}{k}$ <p>Doubly Smoothed Series</p> $\hat{s}_t^{(2)} = \frac{\hat{s}_t^{(1)} + \cdots + \hat{s}_{t-k+1}^{(1)}}{k}$                                   | <p>Trend Estimate</p> $\hat{\beta}_{1,T} = \frac{2(\hat{s}_T^{(1)} - \hat{s}_T^{(2)})}{k-1}$ <p>Forecast</p> $\hat{y}_{T+l} = \hat{s}_T + \hat{\beta}_{1,T} l$ |
| Weighted Least Squares (WLS)   | $WSS_T(b_0^*, \dots, b_k^*) = \sum_{t=1}^T w_t (y_t - (b_0^* + b_1^* x_{t1} + \cdots + b_k^* x_{tk}))^2$ <hr/> A generalization of ordinary least squares that accounts for variability in the observations.                 |  |
| One-Step<br>Prediction Error   | $y_t - \hat{s}_{t-1}$ <hr/> A measure of how well a forecasting model predicts the next observation.   |  |
| Sum of Squared<br>One-Step<br>Prediction Errors<br><a href="#">Example</a> | $SS(w) = \sum_{t=1}^T (y_t - \hat{s}_{t-1})^2$ <hr/> Used to select the optimal smoothing parameter, $w$ .   |  |

### 3.6 Exponential Smoothing

| Concept   | Definition   |   |
|---|--|---|
| Definition  | <b>Exponential smoothing</b> builds on simple moving averages by assuming that the most recent observations are more relevant.   |   |
| Formula   | $\hat{s}_t = \frac{y_t + wy_{t-1} + \dots + w^{t-1}y_1 + w^ty_0}{1/(1-w)}$ $\hat{s}_t = \hat{s}_{t-1} + (1-w)(y_t - \hat{s}_{t-1}) = (1-w)y_t + w\hat{s}_{t-1}$                        |   |
| Double Exponential Smoothing<br><a href="#">Example</a> | Consider a model with a <a href="#">trend component</a> :<br>$T_t = \beta_0 + \beta_1 t$   |   |
|   | <p><b>Smoothed Series</b></p> $\hat{s}_t^{(1)} = (1-w)y_t + w\hat{s}_{t-1}^{(1)}$ <p><b>Doubly Smoothed Series</b></p> $\hat{s}_t^{(2)} = (1-w)\hat{s}_t^{(1)} + w\hat{s}_{t-1}^{(2)}$ | <p><b>Intercept Estimate</b></p> $\hat{\beta}_{0,T} = 2\hat{s}_T^{(1)} - \hat{s}_T^{(2)}$ <p><b>Trend Estimate</b></p> $\hat{\beta}_{1,T} = \frac{1-w}{w}(\hat{s}_T^{(1)} - \hat{s}_T^{(2)})$ <p><b>Forecast</b></p> $\hat{y}_{T+l} = \hat{\beta}_{0,T} + \hat{\beta}_{1,T}l$ |

### 3.7 Seasonal Adjustments

| Concept                              | Definition  |
|--------------------------------------|---|
| Definition                           | <b>Seasonal</b> time series models capture patterns that repeat over fixed intervals of time (e.g. yearly cycles).  |
| Seasonal Base (SB)                   | The <b>seasonal base</b> is the period over which a seasonal pattern repeats itself (e.g. for monthly data exhibiting yearly seasonality, the seasonal base is 12 months).  |
| Seasonal Component                   | $S_t = \sum_{i=1}^m (a_i \sin(f_i t + b_i)) = \sum_{i=1}^m (\beta_{1i} \sin(f_i t) + \beta_{2i} \cos(f_i t))$ <hr/> $f_i = \frac{2\pi i}{SB}$   |
| Regression Formula                   | $y_t = \beta_0 + S_t + \epsilon_t = \beta_0 + \sum_{i=1}^m (\beta_{1i} \sin(f_i t) + \beta_{2i} \cos(f_i t)) + \epsilon_t$ <hr/> <p>Run multiple linear regression with <math>p = 2m</math> variables.</p>  |
| Seasonal Autoregressive Model        | <b>Seasonal autoregressive models</b> (SAR models) extend the autoregressive model by only incorporating lagged values at seasonal periods.   |
| $SAR(P)$                             | $y_t = \beta_0 + \beta_1 y_{t-SB} + \beta_2 y_{t-2SB} + \cdots + \beta_P y_{t-PSB} + \epsilon_t$ <hr/> <p>e.g. SAR(1), SB=12 (monthly data with yearly seasonality):</p> $y_t = \beta_0 + \beta_1 y_{t-12} + \epsilon_t$  |
| Holt-Winters Seasonal Additive Model | <p>The <b>Holt-Winters seasonal additive model</b> is an exponential smoothing model that accounts for level, trend, and seasonality, using weighted averages, <math>w_1, w_2, w_3</math>, that update over time.</p> <hr/> $y_t = \beta_0 + \beta_1 t + S_t + \epsilon_t$ $\hat{y}_{T+l} = b_{0,T} + b_{1,T} l + \hat{S}_T(l)$ |



### 3.8 Unit Root Test

| Concept   | Definition  |
|---|---|
| Definition  | A <b>unit root</b> is a characteristic of a time series that indicates that the series is non-stationary. A <b>unit root test</b> is used to determine the presence of a unit root.   |
| <b>Consider the Time Series Model</b><br>$y_t = \mu_0 + \phi(y_{t-1} - \mu_0) + \mu_1(\phi + (1 - \phi)t) + \epsilon_t$ |   |
| Parameters  | $\mu_0$ : intercept term<br>$\mu_1$ : time trend<br>$\phi$ : autoregressive parameter<br>$\epsilon_t$ : error term  |
| Special Cases   | Random Walk ( $\phi = 1$ )<br>$y_t = \mu_1 + y_{t-1} + \epsilon_t$<br><br>AR(1) ( $\phi < 1$ and $\mu_1 = 0$ )<br>$y_t = \beta_0 + \phi y_{t-1} + \epsilon_t$<br><br>Linear Trend ( $\phi = 0$ )<br>$y_t = \mu_0 + \mu_1 t + \epsilon_t$  |
| Dickey-Fuller Test  | <p>The <b>Dickey-Fuller (DF) test</b> is a t-test used to determine whether a time series has a unit root.</p> <hr/> <p>Running a regression model where <math>y_t</math> is potentially a random walk can be problematic due to non-stationarity.</p> <p>Instead, use least squares on the differenced model:</p> $y_t - y_{t-1} = \beta_0 + (\phi - 1)y_{t-1} + \beta_1 t + \epsilon_t$ <hr/> $H_0 : \phi = 1 \text{ (unit root / random walk)}$ $H_a : \phi < 1$ |
| Augmented Dickey-Fuller Test  | <p>The DF test assumes the errors are not autocorrelated. To address this, the <b>augmented Dickey-Fuller test</b> includes lagged differences on the right hand side of the equation to account for autocorrelation in the error terms:</p> $\sum_{j=1}^p \phi_j (y_{t-j} - y_{t-j-1})$  |

### 3.9 ARCH and GARCH Models

| Concept              | Definition   |
|----------------------|--|
| Definition           | <p><b>Volatility clustering</b> in time series occurs when periods of high volatility are followed by more high volatility, and periods of low volatility are followed by more low volatility.</p> <p>ARCH and GARCH models address changing variance over time.</p> |
| Conditional Variance | $\sigma_t^2 = \text{Var}_{t-1}(\epsilon_t) = \mathbb{E}[(\epsilon_t - \mathbb{E}(\epsilon_t \Omega_{t-1}))^2 \Omega_{t-1}]$  |
| $ARCH(p)$            | $\sigma_t^2 = w + \gamma_1\epsilon_{t-1}^2 + \gamma_2\epsilon_{t-2}^2 + \dots + \gamma_p\epsilon_{t-p}^2 = w + \gamma(B)\epsilon_t^2$  |
| $ARCH(1)$            | $\sigma_t^2 = w + \gamma_1\epsilon_{t-1}^2$  |
| $GARCH(p, q)$        | $\sigma_t^2 = w + \sum_{i=1}^p \gamma_i\epsilon_{t-i}^2 + \sum_{j=1}^q \delta_j\sigma_{t-j}^2$ <p>The GARCH model captures both short-term and long-term patterns in volatility.</p>   |

## **4. Decision Trees (Learning Objective 4)**

## 4.1 Introduction to Decision Trees

| Concept                          | Description  |
|----------------------------------|--|
| Decision Node<br>(Internal Node) | An internal node that splits into two or more branches.  |
| Leaf Node<br>(Terminal Node)     | The final output node that doesn't split further and contains the class label or the continuous value. |
| Branch<br>(Edge)                 | A connection between nodes representing the outcome of a test at a decision node.                      |

## 4.2 Regression Trees

| Concept                    | Description   |
|----------------------------|---|
| Definition                 | A <b>regression tree</b> is a type of decision tree used for predicting (continuous) numerical values.  |
| Building a Regression Tree | <ol style="list-style-type: none"> <li>1. Divide the predictor space of <math>X_1, X_2, \dots, X_p</math> into <math>J</math> distinct boxes (or regions) <math>R_1, R_2, \dots, R_J</math>.</li> <li>2. For every observation within region <math>R_j</math>, the regression tree makes the same prediction. This prediction is typically the mean of the response values, <math>\hat{y}_{R_j}</math>, for the training observations in <math>R_j</math>.</li> </ol> |
| Goal                       | <p>Select the regions to minimize the RSS:</p> $\text{RSS} = \sum_{j=1}^J \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$ <p>Minimizing the RSS finds the optimal splits that reduce the within-region variance.</p>  |

### 4.3 Recursive Binary Splitting

| Concept          | Description   |
|------------------|---|
| Definition       | <b>Recursive binary splitting</b> is a process used to create a regression tree, without having to evaluate every possible way to partition the feature space into $J$ boxes.   |
| Approach Summary | <ol style="list-style-type: none"> <li>1. Begin with the entire predictor space and progressively split it into smaller and smaller regions.</li> <li>2. At each step, make the best possible split based on a criterion, such as minimizing RSS.</li> <li>3. Apply the process recursively to each sub-region.</li> </ol>  |
| Approach Detail  | <ol style="list-style-type: none"> <li>1. For each <math>X_j</math>, evaluate different potential cutpoints <math>s</math>. For a given <math>j</math> and <math>s</math>, partition the data into two regions: <math>R_1(j, s)</math> and <math>R_2(j, s)</math>.</li> <li>2. Calculate the RSS for this split as: <math display="block">\sum_{i: x_i \in R_1(j, s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j, s)} (y_i - \hat{y}_{R_2})^2</math> <p>Identify the pair <math>(j, s)</math> that results in the smallest RSS and split the dataset into regions <math>R_1(j, s)</math> and <math>R_2(j, s)</math>.</p> </li> <li>3. Apply the same process recursively to each resulting region until a stopping criterion is met.</li> </ol> |
| Overfitting      | <p>Binary splitting can lead to overfitting, due to using potentially small nodes with few data points.</p> <p>Pruning can mitigate the risk of overfitting.</p>  |

## 4.4 Pruning

| Concept            | Description  |
|--------------------|--|
| Definition         | <p><b>Cost complexity pruning</b> (weakest link pruning) involves removing leaves of a decision tree to prevent overfitting to training data and improve generalization to test data.</p> <p>The primary goal of pruning is to select a subtree that minimizes the test error to strike a better balance between bias and variance.</p>  |
| Approach           | <ol style="list-style-type: none"> <li>1. Use recursive binary splitting to grow a full decision tree <math>T_0</math> using training data.</li> <li>2. For various tuning parameters, <math>\alpha</math>, create subtrees <math>T</math> such that: <math display="block">\sum_{m=1}^{ T } \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha  T </math> </li> <li>3. Select the optimal value of <math>\alpha</math> using K-fold cross-validation by minimizing the average MSE across the K folds.</li> <li>4. The subtree in step 2 with the optimal <math>\alpha</math> is the pruned tree that balances complexity and performance.</li> </ol>   |
| Selecting $\alpha$ | <p>The tuning parameter <math>\alpha</math> acts as a regularization term that balances the trade-off between the complexity of the tree and its fit to the training data.</p> <p>When <math>\alpha = 0</math>, the criterion reduces to the SSE. This often leads to overfitting because the tree will have many leaves, each potentially capturing noise in the data.</p> <ul style="list-style-type: none"> <li>- When <math>\alpha</math> is large, the term <math>\alpha  T </math> becomes significant. This encourages smaller, simpler trees. If <math>\alpha</math> is very large, the pruning process might remove many branches, potentially leading to underfitting.</li> </ul> <p>Find an <math>\alpha</math> that provides a good balance between the SSE and the complexity penalty by minimizing the average cross-validated MSE (step 3).</p> |

## 4.5 Classification Trees

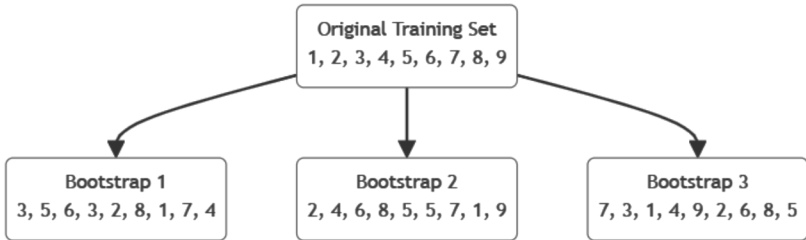
| Concept  | Description  |
|--|--|
| Building a Classification Tree                       | A <b>classification tree</b> is also constructed by recursively splitting the data into subsets based on feature values. Each split is chosen to maximize the most commonly occurring class.   |
| Node Purity  | Node purity measures how similar the response values are within a node.  |
| Classification Error Rate<br><a href="#">Example</a> | <p>A simple measure of node impurity, indicating the proportion of observations that do not belong to the most commonly occurring class.</p> $E = 1 - \max_k(\hat{p}_{mk})$ <p>where <math>\hat{p}_{mk}</math> is the proportion of observations in region <math>m</math> that belong to class <math>k</math>.</p> |
| Gini Index<br><a href="#">Example</a>                | $G = \sum_{k=1}^K \hat{p}_{mk}(1 - \hat{p}_{mk})$ <p>where <math>\hat{p}_{mk}</math> is defined above and <math>K</math> is the number of classes.</p>   |
| Entropy (Cross-Entropy)<br><a href="#">Example</a>   | $D = - \sum_{k=1}^K \hat{p}_{mk} \log(\hat{p}_{mk})$ <p>where <math>\hat{p}_{mk}</math> is defined above and <math>K</math> is the number of classes.</p>  |

## 4.6 Trees vs Linear Models

| Concept             | Regression Trees  | Linear Regression Models                      |
|---------------------|---|---|
| Model Form          | $f(X) = \sum_{m=1}^M c_m \cdot 1_{(X \in R_m)}$                     | $f(X) = \beta_0 + \sum_{j=1}^p X_j \beta_j$   |
| Explainability      | Easier to explain and understand, even by non-experts               | More difficult to explain and understand      |
| Decision-Making     | Mirrors human decision-making process                               | Does not mirror human decision-making process |
| Qualitative Data    | Does not require dummy variables                                    | Requires dummy variables                      |
| Predictive Accuracy | Lower   | Higher  |
| Changes in Data     | Small changes can lead to significant changes in the tree structure | More robust to changes in data                |



## 4.7 Bagging

| Concept              | Description   |
|----------------------|---|
| Definition           | <b>Bagging (bootstrap aggregation)</b> reduces the high variance of decision trees by averaging predictions from multiple models trained on bootstrapped datasets.  |
| Bootstrap Method     | <p>Involves sampling with replacement from a single dataset to create multiple bootstrapped training sets.</p> <hr/>    |
| Average Predictions  | <p>Using population data:</p> $\hat{f}_{\text{avg}}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^b(x)$ <hr/> <p>Using bootstrapped data:</p> $\hat{f}_{\text{bag}}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^{*b}(x)$  |
| Regression Trees     | <p>Build <math>B</math> unpruned regression trees from <math>B</math> bootstrapped training sets and average their predictions.</p> <p>Each tree has high variance, but low bias. Averaging the trees reduces the high variance while maintaining low bias.</p> |
| Classification Trees | Use majority vote (most common class) among $B$ predictions to determine the predicted class.   |
| Out-of-Bag (OOB)     | Fit the model with 2/3 of data; remaining 1/3 (OOB) used for testing. For large $B$ , the OOB error is comparable to the LOOCV error and more computationally efficient than CV.  |

## 4.8 Random Forests

Bagging grows multiple decision trees using bootstrap samples and considers all predictors,  $p$ , at each split. This often leads to highly correlated trees, especially when strong predictors are in the dataset. As a result, the variance reduction from averaging is limited.

**Random forests** solve this by adding randomness. At each split, only a random subset of predictors ( $m < p$ , typically  $m \approx \sqrt{p}$ ) is considered. This prevents strong predictors from dominating every tree, leading to more decorrelated trees. When  $m = p$ , random forests are equivalent to bagging.

The selection of  $m$  is crucial for balancing bias and variance in the model. A smaller  $m$  is useful when a large number of predictors are correlated. Random forests ( $m < p$ ) typically lead to a slight improvement over bagging ( $m = p$ ). Similar to bagging, a large number of trees  $B$  will not lead to overfitting.

## 4.9 Boosting

| Concept                           | Description   |
|-----------------------------------|---|
| Definition                        | <b>Boosting</b> builds models sequentially, leveraging the information from previously constructed models to correct the errors of its predecessors.  |
| Algorithm                         | <ol style="list-style-type: none"> <li>1. Set the initial prediction function <math>\hat{f}(x) = 0</math> and the residuals <math>r_i = y_i</math> for all training data points <math>i</math>.</li> <li>2. For <math>b = 1, 2, \dots, B</math>, repeat the following steps: <ul style="list-style-type: none"> <li>- Fit a regression tree <math>\hat{f}^b</math> with <math>d</math> splits (resulting in <math>d + 1</math> terminal nodes) to the training data <math>(X, r)</math>.</li> <li>- Update the prediction function by adding a shrunk version of the new tree:<br/> <math display="block">\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)</math> </li> <li>- Update the residuals: <math>r_i \leftarrow r_i - \lambda \hat{f}^b(x_i)</math></li> </ul> </li> <li>3. The final boosted model is given by:<br/> <math display="block">\hat{f}(x) = \sum_{b=1}^B \lambda \hat{f}^b(x)</math> </li> </ol> |
| Number of Trees ( $B$ )           | Unlike bagging and random forests, boosting can overfit if the number of trees $B$ is too large, though this overfitting occurs slowly. Use cross-validation to select the optimal value for $B$ .  |
| Shrinkage Parameter ( $\lambda$ ) | The shrinkage parameter (learning rate), $\lambda \in [0, 1]$ , ensures that the model learns slowly and avoids overfitting. Typical values are 0.01 or 0.001. A very small $\lambda$ can necessitate a larger $B$ to achieve good performance.   |
| Interaction Depth ( $d$ )         | The interaction depth controls the complexity of each tree in the boosted ensemble. Often, $d = 1$ works well, making each tree a stump with a single split.  |
| Performance                       | When evaluating the test error as a function of the total number of trees and the interaction depth $d$ , we observe the following: <ul style="list-style-type: none"> <li>- Stumps (<math>d = 1</math>) perform well if a sufficient number of them are used.</li> <li>- The model with stumps outperforms the model with trees of depth two (<math>d = 2</math>).</li> <li>- Both models outperform a random forest.</li> </ul>   |

## **5. Unsupervised Learning Techniques (Learning Objective 5)**

## 5.1 Introduction to Unsupervised Learning

| Concept                                      | Definition  |
|--|---|
| Unsupervised Learning                        | <p><b>Unsupervised learning</b> is a type of learning algorithm that focuses solely on the features <math>X_1, X_2, \dots, X_p</math> without considering the response variable, <math>Y</math>.</p> <p>The goal is to uncover patterns, groupings, or structure in the data without predefined labels.</p> |
| <a href="#">Principal Component Analysis</a> | <p><b>Principal component analysis</b> (PCA) is a dimensionality reduction method that transforms a large set of variables into a smaller one that still contains most of the original data's information.</p>  |
| <a href="#">K-Means Clustering</a>           | <p><b>K-means clustering</b> partitions the data into a predefined number of clusters by minimizing the variance within each cluster.</p>   |
| <a href="#">Hierarchical Clustering</a>      | <p><b>Hierarchical clustering</b> builds a tree-like structure of clusters by iteratively merging or splitting clusters based on their similarities.</p>  |

## 5.2 Principal Components Regression

### 5.2.1 Linear Combinations of Predictors

| Concept                | Description   |   |
|------------------------|---|---|
| Original Predictors    | $X_1, X_2, \dots, X_p$  |   |
| Linear Combination     | $Z_1, Z_2, \dots, Z_M$ where $M < p$  |   |
|                        | $Z_m = \sum_{j=1}^p \phi_{jm} X_j$  |   |
| Selecting $\phi_{jm}$  | <a href="#">Principal Component Analysis</a> or <a href="#">Partial Least Squares</a>                         |   |
| Regression Model       | $y_i = \theta_0 + \sum_{m=1}^M \theta_m z_{im} + \epsilon_i$  |   |
| Goal                   | Reduce $p + 1$ coefficients $(\beta_0, \dots, \beta_p)$ to $M + 1$ coefficients $(\theta_0, \dots, \theta_M)$ |   |
|                        | $\sum_{m=1}^M \theta_m z_{im} = \sum_{j=1}^p \beta_j x_{ij}$  | $\beta_j = \sum_{m=1}^M \theta_m \phi_{jm}$ |
| Bias-Variance Tradeoff | Reducing $M$ , where $M \ll p$ , introduces bias but significantly reduces variance.                          |   |

### 5.2.2 Principal Components Regression

| Concept                  | Description  |
|--------------------------|--|
| Definition               | <b>Principal components regression</b> combines principal component analysis with linear regression by using principal components as predictors.   |
| Steps                    | <ol style="list-style-type: none"><li>1. <a href="#">PCA</a> applied to the predictors <math>X_1, \dots, X_p</math> to reduce dimensionality (unsupervised)</li><li>2. Regression on selected principal components <math>Z_1, \dots, Z_M</math> (supervised)</li></ol> |
| Assumption               | Directions with greatest variance in predictors are most associated with the response $Y$ .  |
| Dimensionality Reduction | Uses only the first $M$ principal components where $M < p$ to reduce overfitting and improve generalization.   |
| Bias-Variance Tradeoff   | Bias decreases and variance increases as more principal components are included.   |
| Performance              | PCR performs well when few principal components are needed. Many components can lead to overfitting.   |

### 5.2.3 Partial Least Squares

In PCR, the response  $Y$  is not used to determine the principal component directions, so there is no guarantee that the directions explaining the predictors will also be effective for predicting the response. **Partial least squares (PLS)** is a supervised method that aims to overcome this limitation by incorporating the response variable in the identification of new feature directions. In practice, PLS does not perform much better than ridge regression or PCR. PLS can reduce bias at the cost of increased variance.

## 5.3 Principal Component Analysis

### 5.3.1 Definitions

| Concept   | Description  |
|---|--|
| Definition  | <b>Principal component analysis (PCA)</b> transforms the original variables into a new set of uncorrelated variables, called principal components, ordered by the amount of variance they explain in the data.                             |
| Original Features                                 | $X_1, X_2, \dots, X_p$   |
| The $m$ -th Principal Component (PC)<br>( $Z_m$ ) | $Z_m = \sum_{j=1}^p \phi_{jm} X_j = \phi_{1m} X_1 + \phi_{2m} X_2 + \dots + \phi_{pm} X_p$   |
| Loading Vector<br>( $\phi_m$ )                    | $\phi_m = (\phi_{1m}, \phi_{2m}, \dots, \phi_{pm})^T$ <p><math>\phi_{jm}</math> reflects the weight of the original feature <math>X_j</math> in forming the principal component <math>Z_m</math> where:</p> $\sum_{j=1}^p \phi_{jm}^2 = 1$ |
| Score<br>( $z_{im}$ )                             | $z_{im} = \sum_{j=1}^p \phi_{jm} x_{ij} = \phi_{1m} x_{i1} + \phi_{2m} x_{i2} + \dots + \phi_{pm} x_{ip}$ <p>The scores are the values of the principal components for each observation in the dataset.</p>                                |



### 5.3.2 Methodology

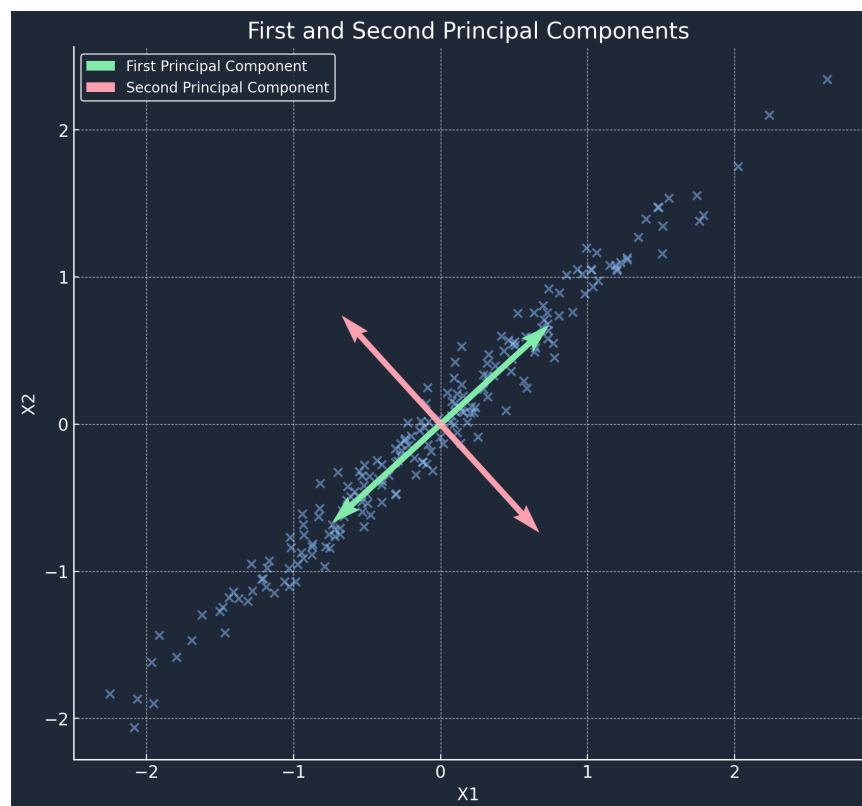
The goal of PCA is to choose the loading vectors  $\phi_m$  that maximize the variance of the principal components  $Z_m$ .

Consider the optimization problem for finding the first principal component ( $m = 1$ ):

$$\max_{\phi_{11}, \phi_{21}, \dots, \phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{j1} x_{ij} \right)^2 \right\} \text{ subject to } \sum_{j=1}^p \phi_{j1}^2 = 1$$

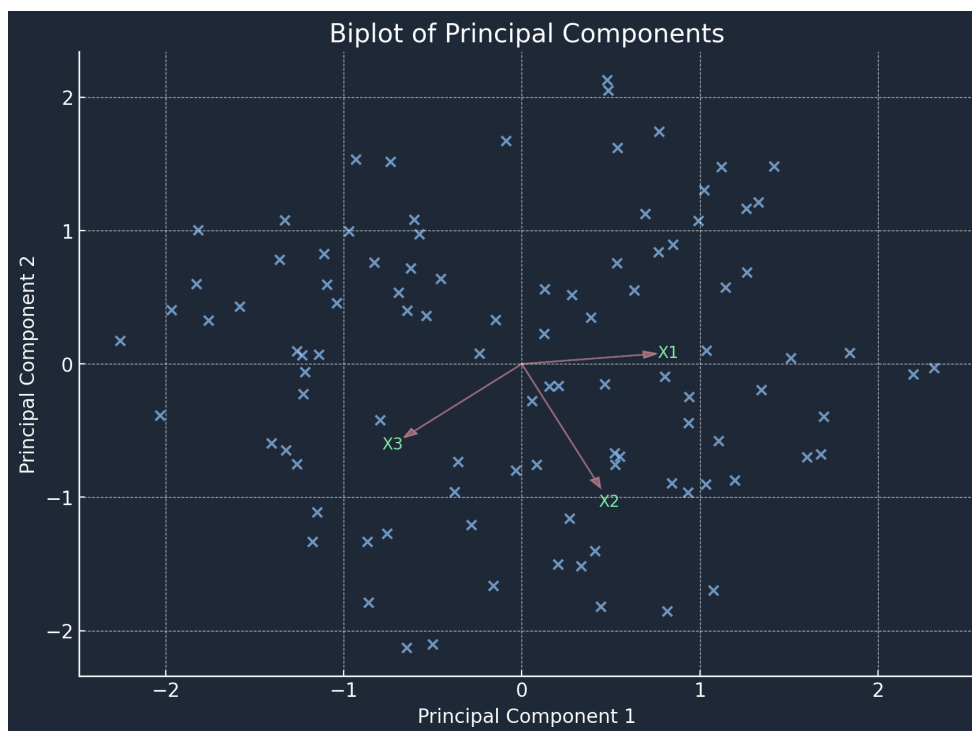
Once  $Z_1$  has been determined, the second principal component,  $Z_2$ , is the linear combination of  $X_1, \dots, X_p$  that has the maximum variance while being uncorrelated with  $Z_1$ .

Visually,  $Z_1$  is the vector that defines the line that is as close as possible to the data. In a two-dimensional example ( $p = 2$ ), once the loading vector  $\phi_1$  is found, there is only one possible direction (up to a sign flip) for  $\phi_2$ , which is orthogonal to  $\phi_1$ .



In higher-dimensional datasets ( $p > 2$ ), multiple principal components can be identified, each being orthogonal to all previously determined components.

Once the principal components are computed, they can be used to create a **biplot**, displaying both the scores of observations and the loadings of variables in a reduced-dimensional space.



### 5.3.3 Proportion of Variance Explained

| Concept                               | Description   |
|---------------------------------------|---|
| Definition                            | The <b>proportion of variance explained (PVE)</b> measures how much of the total data variance is captured by each principal component.                                     |
| Total Variance                        | $\sum_{j=1}^p \text{Var}(X_j) = \sum_{j=1}^p \frac{1}{n} \sum_{i=1}^n x_{ij}^2$ <p>For simplicity, assume the variables have mean zero.</p>                                 |
| Variance of the $m$ -th PC            | $\frac{1}{n} \sum_{i=1}^n z_{im}^2 = \frac{1}{n} \sum_{i=1}^n \left( \sum_{j=1}^p \phi_{jm} x_{ij} \right)^2$   |
| PVE of the $m$ -th PC                 | $\frac{\sum_{i=1}^n z_{im}^2}{\sum_{j=1}^p \sum_{i=1}^n x_{ij}^2} = \frac{\sum_{i=1}^n \left( \sum_{j=1}^p \phi_{jm} x_{ij} \right)^2}{\sum_{j=1}^p \sum_{i=1}^n x_{ij}^2}$ |
| PVE as $R^2$<br>Approximation for $X$ | $1 - \frac{\sum_{j=1}^p \sum_{i=1}^n \left( x_{ij} - \sum_{m=1}^M z_{im} \phi_{jm} \right)^2}{\sum_{j=1}^p \sum_{i=1}^n x_{ij}^2} = 1 - \frac{\text{RSS}}{\text{TSS}}$      |
| Scree Plot<br><a href="#">Example</a> | A <b>scree plot</b> visualizes the PVE for each PC. Look for an elbow point where the explained variance drops significantly to determine the number of PCs for the model.  |

## 5.4 K-Means Clustering

### 5.4.1 Definitions

| Concept   | Description   |
|---|---|
| Goal  | Given a set of observations $\{x_1, x_2, \dots, x_n\}$ where each observation is a $p$ -dimensional vector, K-means clustering aims to partition the $n$ observations into $K$ pre-determined clusters, $C_1, C_2, \dots, C_K$ .                                  |
| Properties  | <p>Every observation belongs to one and only one cluster.<br/> <math>C_1 \cup C_2 \cup \dots \cup C_K = \{1, 2, \dots, n\}</math></p> <hr/> <p>The clusters are non-overlapping.<br/> <math>C_k \cap C_{k'} = \emptyset</math> for all <math>k \neq k'</math></p> |
| Euclidean Distance                                  | <p>Euclidean distance is the straight-line distance between two points, calculated as:</p> $d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$   |
| Within-Cluster Variation<br><a href="#">Example</a> | $W(C_k) = \frac{1}{ C_k } \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$ <hr/> <p>The most common choice for this measure is the squared Euclidean distance, <math>d^2</math>.</p>   |
| Optimization Goal                                   | $\min_{C_1, \dots, C_K} \left\{ \sum_{k=1}^K W(C_k) \right\} = \min_{C_1, \dots, C_K} \left\{ \sum_{k=1}^K \frac{1}{ C_k } \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right\}$   |
| Local Minimum                                       | <p>The final solution may be a local minimum rather than a global minimum.</p> <p>The algorithm is often run multiple times, and the solution with the lowest within-cluster variation is selected.</p>   |
| Handling Outliers                                   | Clustering algorithms may not be suitable when a small subset of observations are significantly different from the rest. Mixture models offer a more flexible approach for handling outliers.   |

5.4.2 K-Means Clustering Algorithm

| Step              | Description  |
|-------------------|--|
| 1. Initialization | Randomly assign a number from 1 to $K$ to each observation, indicating the initial cluster assignment.   |
| 2. Iteration      | Repeat the following steps until the cluster assignments stop changing:<br>a. For each of the $K$ clusters, compute the cluster centroid. <a href="#">Example</a> .<br>b. Assign each observation to the cluster whose centroid is closest, based on Euclidean distance. <a href="#">Example</a> . |

5.4.3 Algorithm Visual Example



## 5.5 Hierarchical Clustering

### 5.5.1 Definitions

| Concept                  | Description   |
|--------------------------|---|
| Hierarchical Clustering  | <b>Hierarchical clustering</b> produces a dendrogram to show cluster arrangements and merging order. Unlike k-means clustering, hierarchical clustering does not require the number of clusters to be specified in advance. |
| Dendrogram               | A <b>dendrogram</b> is a tree-like diagram that shows the sequences of merges. Each leaf node represents a single observation, and each internal node represents a cluster formed by merging two clusters.                  |
| Node Height              | The height of the nodes in the dendrogram indicates the level of <b>dissimilarity</b> (or distance) at which clusters are merged.<br><br>The most common dissimilarity measure is Euclidean distance.                       |
| Agglomerative Clustering | <b>Agglomerative (bottom-up) clustering</b> is the most common type of hierarchical clustering, starting with each observation as its own cluster and merging (fusing) the most similar clusters step-by-step.              |

### 5.5.2 Hierarchical Clustering Algorithm

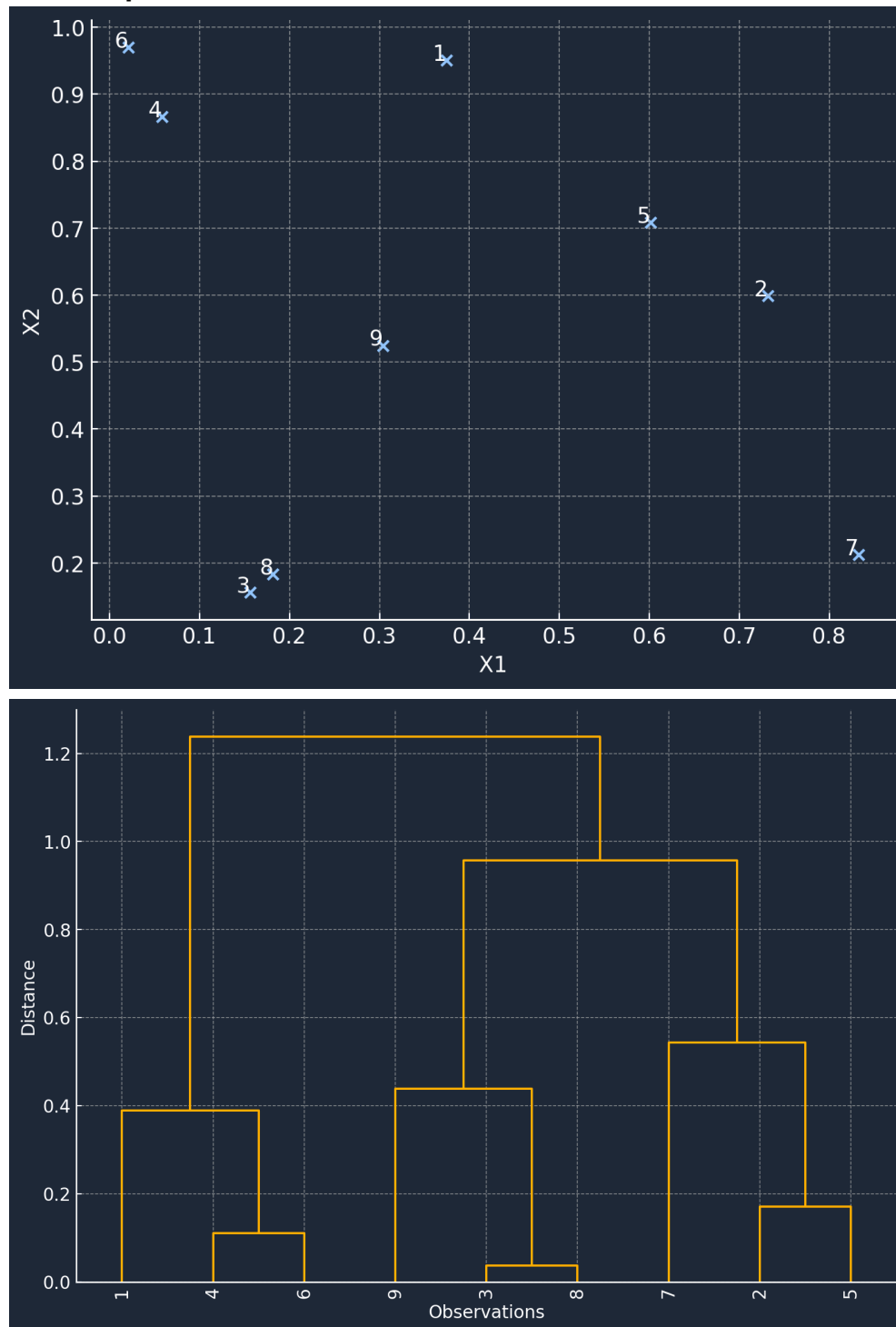
| Step              | Description  |
|-------------------|--|
| 1. Initialization | Start with $n$ observations, each as its own cluster. Compute all pairwise dissimilarities between the observations.   |
| 2. Iteration      | For $i = n, n - 1, \dots, 2$<br>a. Identify closest clusters: Examine pairwise intercluster dissimilarities among the $i$ clusters and identify the two clusters with the smallest dissimilarity. Fuse them.<br>b. Update dissimilarities: Update the pairwise dissimilarities among the remaining $i - 1$ clusters. |

### 5.5.3 Calculating Dissimilarity - Linkage Methods

| Concept          | Linkage Method  | Characteristics   |
|------------------|---|---|
| Complete Linkage | Maximum intercluster dissimilarity. <a href="#">Example.</a>                  | Produces compact clusters; sensitive to outliers.   |
| Single Linkage   | Minimal intercluster dissimilarity. <a href="#">Example.</a>                  | Can produce elongated or chained clusters; less compact.  |
| Average Linkage  | Mean intercluster dissimilarity. <a href="#">Example.</a>                     | Balanced clusters; less sensitive to outliers.  |
| Centroid Linkage | Dissimilarity between the centroids of the clusters. <a href="#">Example.</a> | Can produce inversions, where clusters are fused below either of the individual clusters in the dendrogram. |

Intercluster dissimilarity measures how different two clusters are, using the maximum, minimum, or average pairwise distance between observations depending on the linkage method (complete, single, or average, respectively).

### 5.5.4 Visual Example



The second image shows the resulting dendrogram from a hierarchical clustering algorithm on the data set from the first image.