Cheatsheet on "Visualization in Computational Data Analytics," "Model Selection with One Predictor in Computational Data Analytics," "Linear Regression," "Assumptions of Linear Regression," and "Diagnostics Analytics":

Visualization in Computational Data Analytics:

1. Data Exploration:
   - Initial understanding of data's structure, distribution, and patterns.
   - Essential for data exploration.

2. Pattern Recognition:
   - Reveals hidden patterns or trends.
   - Identifies anomalies or outliers.

3. Communication:
   - Powerful means of conveying complex information.
   - Aids in decision-making processes.

4. Comparisons:
   - Easy comparison of data across categories, time periods, or groups.
   - Simplifies drawing insights and conclusions.

5. Dimensionality Reduction:
   - Techniques like PCA and t-SNE for visualizing high-dimensional data in lower dimensions.

6. Interactive Visualizations:
   - Created using tools like D3.js and Tableau.
   - Enables dynamic data exploration.

7. Data Cleaning:
   - Highlights data quality issues.
   - Aids in preprocessing.

Model Selection with One Predictor in Computational Data Analytics:

1. Univariate Analysis:
   - Examines the relationship between a single predictor and the target variable.
   - Descriptive statistics and basic visualizations.

2. Correlation Analysis:
   - Assesses the correlation between the single predictor and the target variable.

3. Hypothesis Testing:
   - Utilizes tests like t-tests or ANOVA to determine the predictor's significance.

4. Feature Selection:
   - Evaluates the importance of the single predictor.
   - May exclude weak predictors.

5. Model Building:
   - If significant, build a simple regression model (e.g., linear regression).

6. Model Evaluation:
   - Assess model performance with metrics (e.g., R-squared, MAE).

7. Model Validation:
   - Use techniques like cross-validation to ensure model generalizability.

8. Model Comparison:
   - Compare the model with other potential predictors or models.

Linear Regression:

- Introduction:
  - Supervised machine learning model.
  - Establishes a linear relationship between input (independent) and output (dependent) variables.

- Equations:
  - Simple Linear Regression: `y = b0 + b1x`
  - Multiple Linear Regression: `y = b0 + b1x1 + b2x2 + ... + bnxn`

- Objective:
  - Find the best-fit line, minimizing error.

- Assumptions of Linear Regression:
  1. Linearity
  2. Normality
  3. Homoscedasticity
  4. Independence
  5. Error Terms Distribution
  6. No Autocorrelation

- Techniques:
  - Ordinary Least Squares (OLS), Gradient Descent, Regularization.

- Applications:
  - Marketing, finance, insurance, etc.

Assumptions of Linear Regression:

1. Linearity:
   - Dependent variable linearly related to independent variables.

2. Normality:
   - Both variables should be normally distributed.

3. Homoscedasticity:
   - Variance of error terms should be constant.

4. Independence/No Multicollinearity:
   - Independent variables uncorrelated, no multicollinearity.

5. Error Terms Distribution:
   - Error terms should be normally distributed.

6. No Autocorrelation:
   - Error terms independent of each other.

Diagnostics Analytics:

- What Is Diagnostic Analytics?:
  - Explains "Why did this happen?"
  - Identifies causative factors in data.

- Importance:
  - Gain insights into factors affecting events.
  - Improve decision-making.

- Types of Analytics:
  - Descriptive, Predictive, Prescriptive, Diagnostic.

- How Does Diagnostic Analytics Work?:
  - Data drilling, data mining, correlation analysis.
  - Identify anomalies, gather data, establish causal connections.

- Process:
  1. Identify Anomalies
  2. Discovery
  3. Establish Causal Connections

- Use Cases:
  - Healthcare, retail, manufacturing, human resources.

- Benefits:
  - Understand reasons behind past events.
  - Informed decision-making.

- Drawbacks:
  - Focus on historical data.
  - Complement with predictive and prescriptive analytics.

Likelihood Frequentist:

- Introduction to Likelihood:
  - Likelihood vs. Probability.

- Maximum Likelihood Estimation (MLE):

  - Estimating model parameters from data.

- Models:
  - Formal representation of events or processes.

- Introduction to Maximum Likelihood Estimation for Machine Learning:
  - Solving density estimation problems.
  - MLE and likelihood function.

- Problem of Probability Density Estimation:
  - Estimating joint probability distribution for a dataset.
  - MLE and MAP approaches.

- Maximum Likelihood Estimation:
  - Optimization to maximize likelihood.
  - Using log-likelihood function.

- Relationship to Machine Learning:
  - Application in supervised and unsupervised learning.

- Fitting a Line using Likelihood:
  - Linear regression as an MLE problem.
  - Derivation and goal of MLE equation.

These cheatsheets cover key concepts in data analytics, visualization, model selection, linear regression, assumptions, diagnostics analytics, and likelihood frequentist.

Visualization in Computational Data Analytics:

1. Data Exploration: Visualization is a critical step in data exploration. It helps analysts gain an initial understanding of the data's structure, distribution, and patterns.

2. Pattern Recognition: Visualizing data can reveal hidden patterns or trends that might not be apparent from raw data. Graphs, charts, and plots make it easier to identify anomalies or outliers.

3. Communication: Visualizations serve as a powerful means of communication. They make it easier to convey complex information to non-technical stakeholders, helping in decision-making processes.

4. Comparisons: Visual representations allow for the easy comparison of data across different categories, time periods, or groups, making it simpler to draw insights and conclusions.

5. Dimensionality Reduction: Techniques like PCA (Principal Component Analysis) and t-SNE (t-distributed Stochastic Neighbor Embedding) are used for visualizing high-dimensional data in lower dimensions to uncover structures or clusters.

6. Interactive Visualizations: Interactive visualizations, created using tools like D3.js or Tableau, enable users to explore data dynamically, promoting deeper insights and understanding.

7. Data Cleaning: Visualizations can highlight data quality issues, such as missing values or inconsistencies, making it easier to address these problems during the preprocessing stage.

Model Selection with One Predictor in Computational Data Analytics:

1. Univariate Analysis: In cases where you have one predictor variable, univariate analysis involves examining the relationship between the single predictor and the target variable. This could be done using descriptive statistics and basic visualizations.

2. Correlation Analysis: Assess the correlation between the single predictor and the target variable to understand the strength and direction of the relationship.

3. Hypothesis Testing: Perform hypothesis tests like t-tests or ANOVA to determine if the single predictor significantly affects the target variable.

4. Feature Selection: Evaluate the importance of the single predictor in the context of your modeling goals. If it's a weak predictor, you might consider excluding it from your model.

5. Model Building: If the single predictor is deemed significant and relevant, you can build a simple regression model, like a linear regression, to predict the target variable using this predictor.

6. Model Evaluation: Assess the model's performance using appropriate metrics (e.g., R-squared, Mean Absolute Error, etc.) to determine how well the single predictor explains the variability in the target variable.

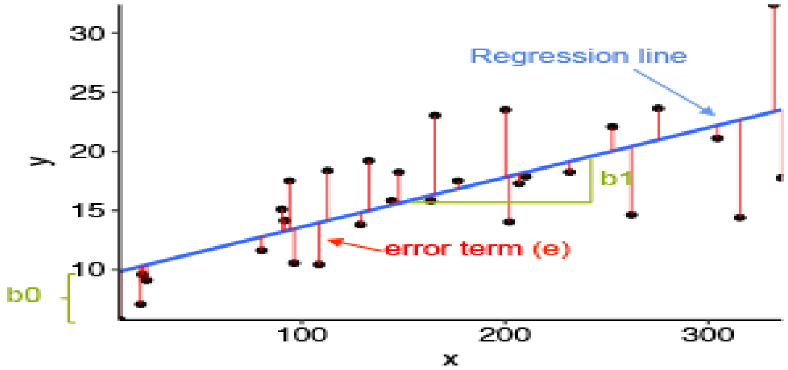
7. Model Validation: Use techniques like cross-validation to ensure the model's generalizability and robustness.

8. Model Comparison: If you have other potential predictors or models, compare the model built with this single predictor to models with additional predictors to determine which one performs better in terms of predictive accuracy and generalization.

These topics are related to Computational Data Analytics as they involve the process of analyzing and extracting insights from data, whether through visualization techniques for data exploration or model selection for predictive modeling, which is a key aspect of data analytics and machine learning.

Linear Regression PPT Notes
Certainly, here are detailed notes on Linear Regression:

Introduction
- Linear Regression is a supervised Machine Learning model that finds the best-fit linear line between independent and dependent variables, establishing a linear relationship.
- It assumes a linear relationship between input variables (independent variables 'x') and the output variable (dependent variable 'y').
- Simple Linear Regression is used for a single input variable, while Multiple Linear Regression is employed when there are multiple input variables.



Equations

- Simple Linear Regression Equation: `y = b0 + b1x`, where `b0` is the intercept, `b1` is the coefficient or slope, `x` is the independent variable, and `y` is the dependent variable.

$$y = b_o + b_1 x$$

- Multiple Linear Regression Equation: `y = b0 + b1x1 + b2x2 + b3x3 + ... + bnxn`, where `b0` is the intercept, `b1`, `b2`, `b3`, ..., `bn` are coefficients or slopes of independent variables `x1`, `x2`, `x3`, ..., `xn`, and `y` is the dependent variable.

$$y = b_o + b_1 x_1 + b_2 x_2 + b_3 x_3 \ldots . + b_n x_n$$

Objective

- The primary goal of a Linear Regression model is to find the best-fit linear line with optimal values for the intercept and coefficients that minimize the error (the difference between actual and predicted values).

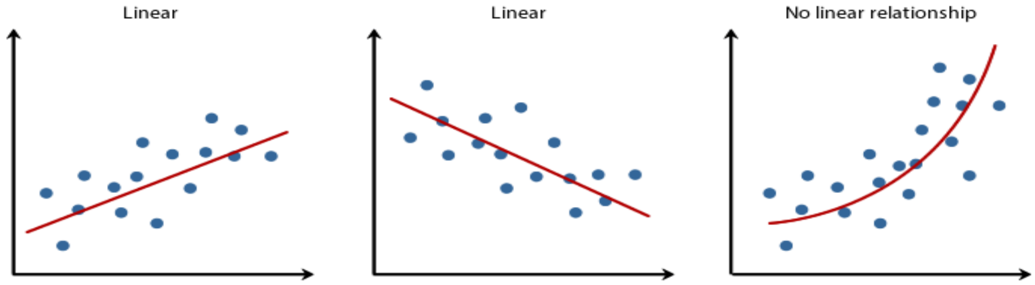
Mathematical Approach

- Residual/Error: Residual/Error represents the difference between actual values and predicted values.
- Sum of Residuals/Errors: The sum of the differences between actual and predicted values.
- Square of Sum of Residuals/Errors: The square of the sum of the differences between actual and predicted values.

$$\sum {e_i}^2 = \sum (Y_i - \hat{Y}_i)^2$$

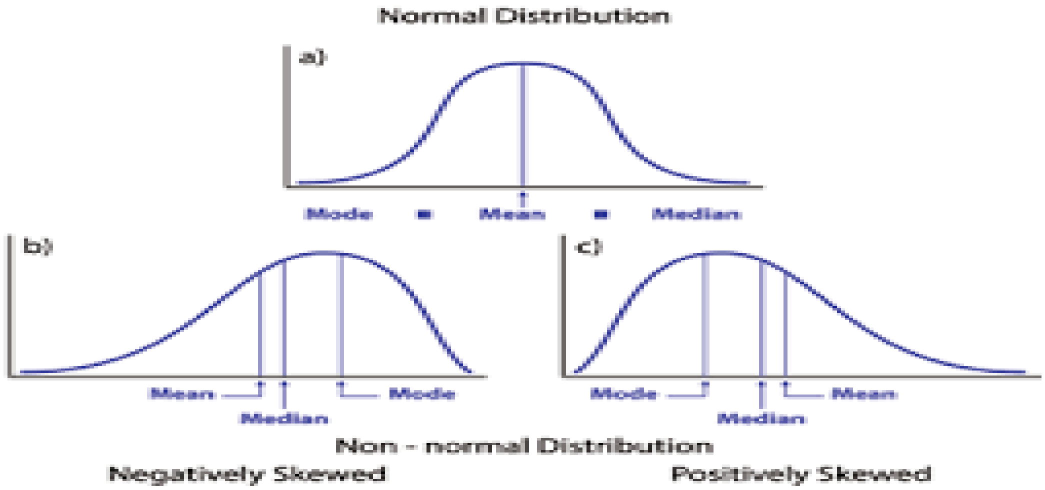
Assumptions of Linear Regression

1. Linearity: The dependent variable should be linearly related to independent variables. This can be verified through scatter plots.

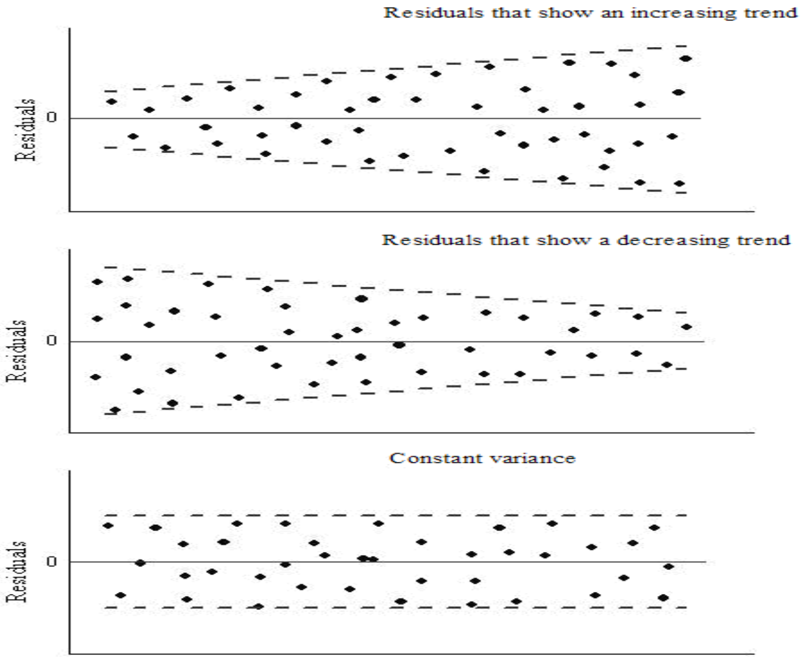


Copyright 2014. Laerd Statistics.

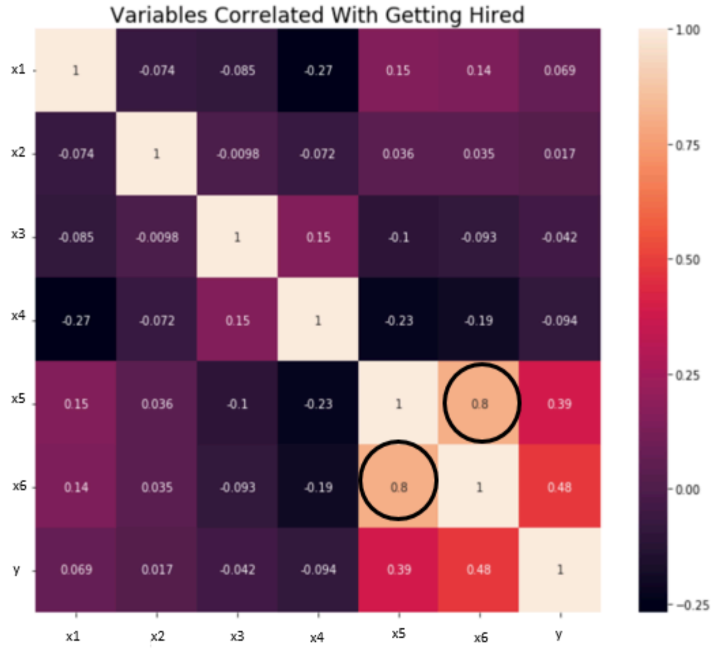
2. Normality: Both the independent and dependent variables should be normally distributed.



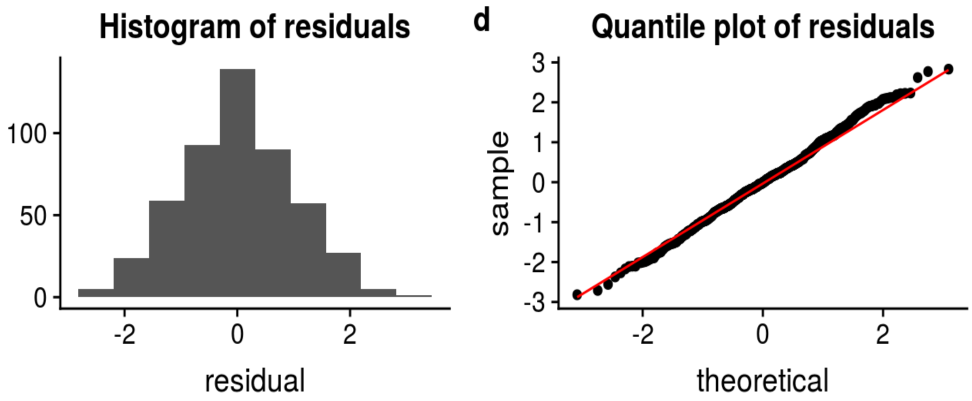
3. Homoscedasticity: The variance of error terms should be constant, meaning the spread of residuals should be consistent.



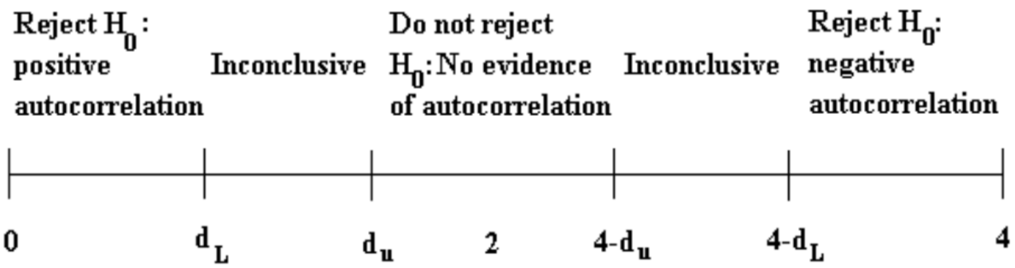
4. Independence/No Multicollinearity: Independent variables should be uncorrelated, with no significant multicollinearity.



5. Error Terms Distribution: Error terms should be normally distributed.



6. No Autocorrelation: Error terms should be independent of each other.


Reject H_0: positive autocorrelation
Inconclusive
Do not reject H_0: No evidence of autocorrelation
Inconclusive
Reject H_0: negative autocorrelation
0
d_L
d_u
2
$4\text{-}d_u$
$4\text{-}d_L$
4

Dealing with Assumption Violations

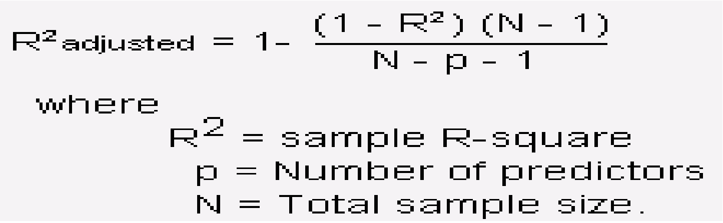
- Violations of assumptions can lead to decreased model accuracy. Techniques for handling these violations include data transformations, feature selection, or regularization methods.

Evaluation Metrics for Regression Analysis

1. R-squared (Coefficient of Determination): Measures the proportion of the variance in the dependent variable explained by the independent variables.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

2. Adjusted R-squared: Adjusts R-squared for the number of independent variables to provide a more accurate representation of model performance.



$$R^2_{\text{adjusted}} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

p = Number of predictors

N = Total sample size.

3. Mean Squared Error (MSE): The mean of the squared differences between actual and predicted values.

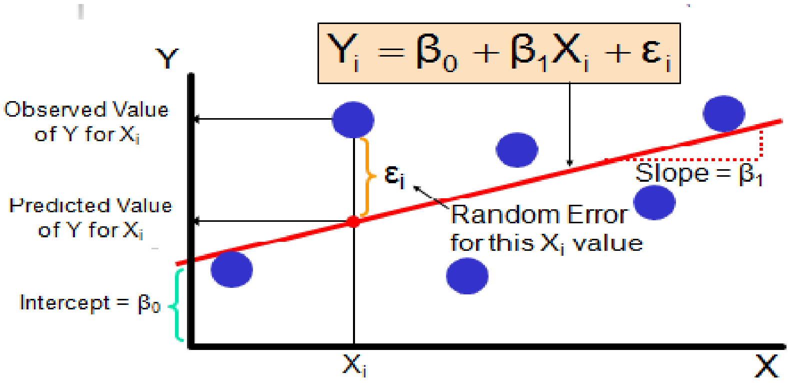
$$MSE = \frac{1}{n} \Sigma \underbrace{\left(y - \widehat{y} \right)^2}_{\text{The square of the difference between actual and predicted}}$$

4. Root Mean Squared Error (RMSE): The square root of the MSE, penalizing larger errors.

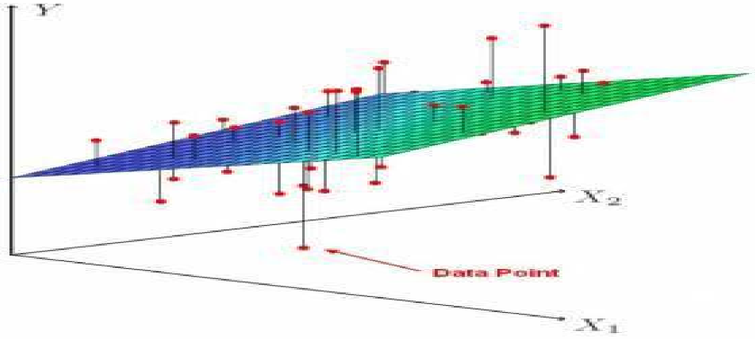
$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Model Representation

- In Simple Linear Regression with one input variable and one output variable, the model takes the form: `Y(pred) = b0 + b1x`.


$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
Y
Observed Value of Y for X_i
Predicted Value of Y for X_i
Intercept = β_0
ε_i
Random Error for this X_i value
Slope = β_1
X_i
X

- In higher dimensions (with multiple input variables), the line becomes a plane or hyperplane.



Violations of Assumptions

- Violations of linearity, independence, homoscedasticity, and normality assumptions can lead to inaccurate model results and predictions.

Assumptions of Linear Regression:

Linear regression is a widely used statistical technique for modeling relationships between dependent and independent variables. To justify the use of linear regression for inference and prediction, several key assumptions must be satisfied:

1. Linearity and Additivity:

- (i) The relationship between the dependent variable (Y) and each independent variable (X) is both linear and additive.
- (a) The expected value of Y is a straight-line function of each independent variable while holding all other variables constant.
- (b) The slope of this line remains constant and does not depend on the values of other independent variables.
- (c) The effects of different independent variables on the expected value of Y are additive, meaning they do not interact in a multiplicative or nonlinear manner.

2. Statistical Independence of Errors:

- (ii) The errors (residuals) are statistically independent of each other.
- Specifically, there should be no correlation between consecutive errors in the case of time series data. Each observation's error is not influenced by the error of the previous observation.

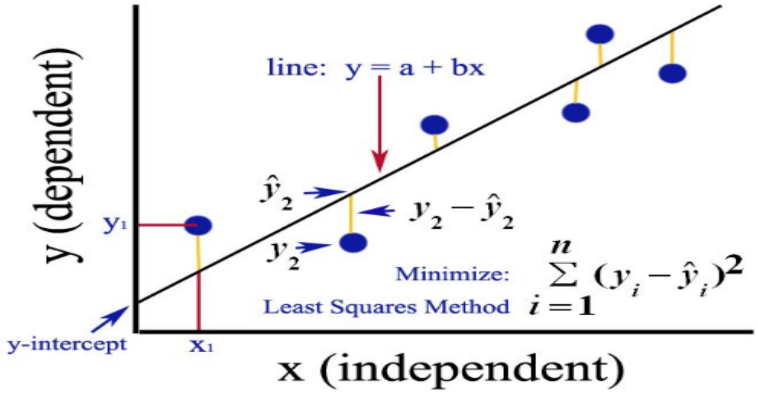
3. Homoscedasticity (Constant Variance) of Errors:
- (a) For time series data, the variance of errors should be constant over time. There should be no systematic change in the spread of errors as time progresses.
- (b) For predictions, the variance of errors should be constant for all values of the independent variables. The spread of errors should not change as you make predictions.
- (c) The variance of errors should not vary systematically with any of the independent variables. In other words, the errors should exhibit constant variance across the entire range of independent variable values.

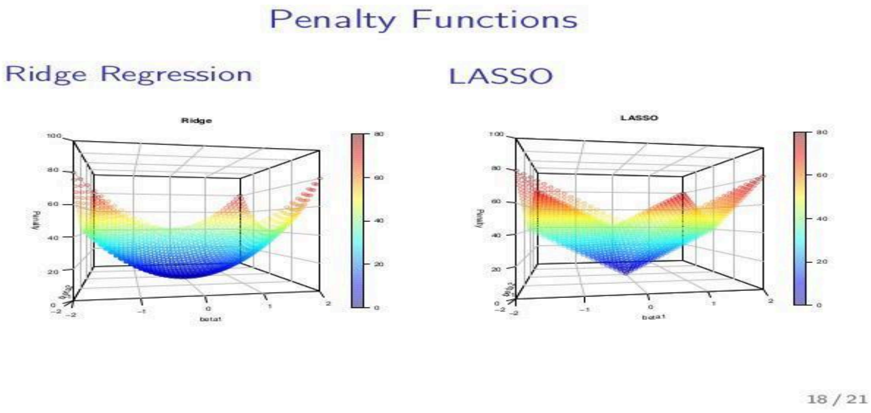
4. Normality of the Error Distribution:
- (iv) The errors (residuals) should follow a normal distribution.
- This assumption implies that the errors are symmetrically distributed around zero and follow a bell-shaped curve.

Impact of Violating Assumptions:
If any of these assumptions are violated, meaning that there are nonlinear relationships between dependent and independent variables or the errors exhibit correlation, heteroscedasticity (varying error variance), or non-normality, the following consequences may occur:
- Forecasts: Predictions made by the regression model may be inaccurate.
- Confidence Intervals: Confidence intervals for parameter estimates may be unreliable.
- Scientific Insights: Conclusions drawn from the analysis may be inefficient, biased, or misleading.

Therefore, it is essential to check these assumptions when applying linear regression models and consider alternative methods or adjustments when they are not met. Violations of these assumptions can lead to unreliable results and hinder the interpretability of the model.

Techniques for Building a Linear Regression Model
1. Ordinary Least Squares (OLS): Minimizes the sum of squared differences between observed and predicted values to fit a multiple linear regression model.





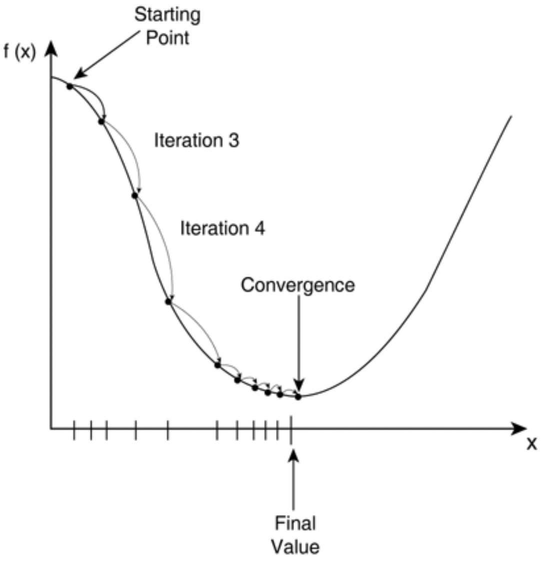
Penalty Functions

Ridge Regression

LASSO

18 / 21

2. Gradient Descent: Iteratively minimizes the error in the model by updating coefficients using a learning rate.



3. Regularization: Extensions of linear regression that reduce complexity and account for collinearity, including Lasso and Ridge Regression.

Applications of Linear Regression
- Linear regression can be applied in various fields such as marketing, finance, and insurance to evaluate trends, make forecasts, analyze marketing effectiveness, assess risk, and optimize decision-making processes.

Real-time Example
- Linear regression can be used to predict student grades based on the number of hours studied, with the objective of minimizing the prediction error.

These detailed notes cover the fundamental concepts, assumptions, techniques, evaluation metrics, and applications of Linear Regression.

Diagnostics PPT Notes

Diagnostics Analytics: Detailed Notes

What Is Diagnostic Analytics?

- Definition: Diagnostic analytics is a branch of advanced analytics that focuses on answering the question, "Why did this happen?" It involves examining data or content to understand the root causes of observed patterns and events.

- Techniques: Diagnostic analytics utilizes various techniques, including data drilling, data mining, and correlation analysis, to delve into data and identify causative factors.

- Additional Data: In some cases, to investigate the root causes of trends, companies may need to incorporate external data sources alongside internal data.

- Purpose: The primary purpose of diagnostic analytics is to help organizations gain insights into the factors driving their past events and make informed decisions to remedy issues and improve future outcomes.

Importance of Diagnostic Analytics

- Diagnostic analytics is vital for companies in gaining a comprehensive understanding of their business performance.

- It aids in discerning the influence of both internal and external factors on outcomes, helping companies make better-informed decisions.

- It is particularly valuable for identifying the reasons behind trends and events and enables companies to replicate success and rectify problems.

- For example, if a specific marketing campaign led to increased product sales, diagnostic analytics can uncover this and guide the allocation of more resources to similar campaigns.

Types of Analytics

- Diagnostic analytics is one of the four primary types of business analytics, alongside descriptive, predictive, and prescriptive analytics.

- Descriptive analytics focuses on summarizing and highlighting historical data trends to answer "What happened?"

- Predictive analytics looks into how future trends might unfold and their potential impact.

- Prescriptive analytics suggests actions to respond to future trends and improve business outcomes.

How Does Diagnostic Analytics Work?

- Diagnostic analytics employs techniques like data drilling, data mining, and correlation analysis to identify the causes of trends.

- Data drilling involves a deeper examination of specific aspects of the data to discover what is driving observed trends.

- Data mining searches for patterns and associations within data, revealing the most common factors linked to specific events.

- Correlation analysis assesses the strength of relationships between different variables in the data.

Process of Diagnostic Analytics

- The diagnostic analytics process typically comprises three stages:

  1. Identify Anomalies: Recognize trends or anomalies that require explanation, sometimes using statistical analysis to confirm their significance.

  2. Discovery: Gather data that can explain the anomalies, which may include external data sources alongside internal data.

  3. Establish Causal Connections: Use techniques like probability theory, regression analysis, filtering, and time-series data analytics to determine causal relationships among variables and uncover the root causes of anomalies.

Three Diagnostic Analytics Categories

- The diagnostic analytics process can be categorized into three stages: identifying anomalies, conducting data discovery, and establishing causal relationships.

Use Cases of Diagnostic Analytics

- Diagnostic analytics is applicable in various industries, including healthcare, retail, manufacturing, and human resources.

- It can be used to investigate the causes of trends such as revenue fluctuations, product popularity, employee turnover, and production bottlenecks.

Benefits of Diagnostic Analytics

- Diagnostic analytics helps companies understand the reasons behind past events, facilitating informed decision-making and a data-driven culture.

- It allows businesses to identify contributing factors that may not be immediately apparent, enabling more effective solutions and improvements.

Drawbacks of Diagnostic Analytics

- A limitation of diagnostic analytics is its focus on historical data; it does not provide insights into future events.

- It may require further investigation to establish definitive cause-and-effect relationships between variables.

- To address future trends, businesses should complement diagnostic analytics with predictive and prescriptive analytics.

Likelihood Frequentist PPT Notes

Likelihood Frequentist: Detailed Notes

Introduction to Likelihood

- Likelihood describes how to find the best distribution of the data for some feature or situation in the data given a certain value of some feature or situation.
- Probability describes how to find the chance of something given a sample distribution of data.

Maximum Likelihood Estimation (MLE)

- MLE is a frequentist approach for estimating the parameters of a model given some observed data.
- General approach:
  1. Observe some data.
  2. Write down a model for how the data was generated.
  3. Set the model parameters to values that maximize the likelihood of the parameters given the data.

Models

- A model is a formal representation of beliefs, assumptions, and simplifications surrounding an event or process.
- Example: Coin Flip
  - Factors to consider: the coin's properties, initial position, force exerted, angle of force, center of mass, gravity.
- Simplified models are often used for practicality even when the real world is complex.

Introduction to Maximum Likelihood Estimation for Machine Learning

- Density estimation is about estimating the probability distribution for a sample of observations.
- Maximum Likelihood Estimation is a common framework for solving density estimation problems.
- It involves defining a likelihood function to calculate the conditional probability of observing data given a probability distribution and distribution parameters.

Problem of Probability Density Estimation

- Probability density estimation involves estimating the joint probability distribution for a dataset.
- It's challenging when the sample is small and noisy.
- Two common approaches: Maximum a Posteriori (MAP) and Maximum Likelihood Estimation (MLE).

Maximum Likelihood Estimation

- MLE is an optimization problem to find parameters that maximize the likelihood function.
- Likelihood function is the conditional probability of observing the data given the model parameters.
- It's common to use the log-likelihood function to avoid numerical instability when multiplying many small probabilities.
- Minimizing the negative log-likelihood (NLL) is often preferred in optimization problems.

Relationship to Machine Learning

- MLE can be applied to supervised and unsupervised machine learning.
- MLE can be used to estimate conditional probabilities, e.g., predicting the output (y) given the input (X).
- Linear regression and logistic regression can be framed as MLE problems.
- MLE provides a consistent way to approach predictive modeling as an optimization problem.

Fitting a Line using Likelihood

- Linear regression can be framed as a maximum likelihood problem.
- The likelihood function is derived from assuming Gaussian noise.
- The goal is to maximize the product of the probabilities for each data point, which is equivalent to the NLL minimization.

- Linear regression provides a prediction that is the mean of a Gaussian distribution, which can be used to calculate prediction intervals.

Derivation of MLE Equation
- A brief derivation of the MLE equation for linear regression is provided.
- The likelihood function is derived using the assumption of normally distributed errors.
- The MLE equation for linear regression is presented.

These detailed notes cover the concepts of Likelihood Frequentist, Maximum Likelihood Estimation, the role of likelihood in machine learning, and its application in linear regression. The notes also include equations and explanations to help understand the concepts better.

TIP: READ MAM PPT BEFORE EXAM