20 Exam Questions Part A

Q:1 What is the main difference between linear regression and logistic regression?

A) Linear regression can only model binary outcomes, while logistic regression can model continuous outcomes.

B) Linear regression assumes that the relationship between the predictor variables and the response variable is linear, while logistic regression assumes that the relationship is non-linear.

C) Linear regression models the mean of the response variable as a linear function of the predictor variables, while logistic regression models the log-odds of the response variable as a linear function of the predictor variables.

D) Linear regression can only model continuous outcomes, while logistic regression can model binary outcomes.

Answer: C) Linear regression models the mean of the response variable as a linear function of the predictor variables, while logistic regression models the log-odds of the response variable as a linear function of the predictor variables.

Explanation: Linear regression is used to model the relationship between a continuous response variable and one or more continuous predictor variables, while logistic regression is used to model the probability of a binary outcome as a function of one or more predictor variables. In logistic regression, the log-odds of the response variable is modeled as a linear function of the predictor variables.

Q:2 What is the purpose of the Akaike Information Criterion (AIC) in model selection?

A) To minimize the sum of squared residuals in the model.

B) To maximize the likelihood of the data under the model.

C) To balance model fit and complexity by penalizing models with more parameters.

D) To test the significance of individual predictor variables in the model.

Answer: C) To balance model fit and complexity by penalizing models with more parameters.

Explanation: The AIC is a metric used in model selection that takes into account both the goodness-of-fit of the model and the number of parameters used in the model. It aims to balance model fit and complexity, by penalizing models with more parameters.

Q:3 What is the purpose of the F-test in linear regression?

A) To test the significance of the intercept term in the model.

B) To test the significance of the overall model fit.

C) To test the significance of individual predictor variables in the model.

D) To test the normality of the residuals in the model.

Answer: B) To test the significance of the overall model fit.

Explanation: The F-test in linear regression is used to test the overall significance of the model fit, by comparing the variance explained by the model to the variance not explained by the model.

Q:4 Which of the following is a common assumption of linear regression?

A) Homoscedasticity of the residuals.

B) Independence of the predictor variables.

C) Multicollinearity between the predictor variables.

D) Non-normality of the residuals.

Answer: A) Homoscedasticity of the residuals.

Explanation: Homoscedasticity refers to the assumption that the variance of the residuals is constant across all levels of the predictor variables. This assumption is important because if the

variance of the residuals varies with the predictor variables, then the standard errors of the coefficients will be biased and the hypothesis tests will be invalid.

Q:5 What is the purpose of the deviance in logistic regression?

A) To measure the goodness-of-fit of the model.

B) To test the normality of the residuals in the model.

C) To test the significance of individual predictor variables in the model.

D) To estimate the probability of the response variable being in a certain category.

Answer: A) To measure the goodness-of-fit of the model.

Explanation: The deviance in logistic regression is a measure of the goodness-of-fit of the model, similar to the sum of squared residuals in linear regression. It compares the fit of the model to the fit of a null model, where all predictor variables are removed.

Q:6 What is the purpose of the coefficient of determination (R-squared) in linear regression?

A) To measure the goodness-of-fit of the model.

B) To test the significance of individual predictor variables in the model.

C) To estimate the probability of the response variable being in a certain category.

D) To measure the correlation between the predictor and response variables.

Answer: A) To measure the goodness-of-fit of the model.

Explanation: The coefficient of determination (R-squared) in linear regression is a measure of the proportion of variance in the response variable that is explained by the predictor variables in the model. It ranges from 0 to 1, where higher values indicate better fit.

Q:7 Which of the following is a common assumption of logistic regression?

A) Homoscedasticity of the residuals.

B) Independence of the predictor variables.

C) Linearity of the predictor variables.

D) Normality of the residuals.

Answer: B) Independence of the predictor variables.

Explanation: Independence of the predictor variables is a common assumption of logistic regression, as correlated predictor variables can lead to biased coefficient estimates and invalid hypothesis tests.

Q:8 What is the purpose of the likelihood function in maximum likelihood estimation?

A) To estimate the probability of the response variable being in a certain category.

B) To measure the goodness-of-fit of the model.

C) To test the significance of individual predictor variables in the model.

D) To maximize the probability of observing the data given the model parameters.

Answer: D) To maximize the probability of observing the data given the model parameters.

Explanation: Maximum likelihood estimation is a method used to estimate the parameters of a statistical model by finding the parameter values that maximize the likelihood function, which is the probability of observing the data given the model parameters.

Q:9 Which of the following is a common assumption of time series models?

A) Stationarity of the data.

B) Normality of the residuals.

C) Independence of the observations.

D) Linearity of the predictor variables.

Answer: A) Stationarity of the data.

Explanation: Stationarity of the data is a common assumption of time series models, as non-stationary data can lead to biased parameter estimates and invalid hypothesis tests.

Q:10 What is the purpose of the autocorrelation function (ACF) in time series analysis?

A) To estimate the probability of the response variable being in a certain category.

B) To measure the correlation between the predictor and response variables.

C) To test the significance of individual predictor variables in the model.

D) To identify the presence of autocorrelation in the data.

Answer: D) To identify the presence of autocorrelation in the data.

Explanation: The autocorrelation function (ACF) is a measure of the correlation between a time series and a lagged version of itself. It is used to identify the presence of autocorrelation in the data, which can affect the accuracy of parameter estimates and hypothesis tests.

Q:11 Which of the following is a common assumption of ARIMA models?

A) Homoscedasticity of the residuals.

B) Independence of the predictor variables.

C) Stationarity of the data.

D) Linearity of the predictor variables.

Answer: C) Stationarity of the data.

Explanation: Stationarity of the data is a common assumption of ARIMA models, as non-stationary data can lead to biased parameter estimates and invalid hypothesis tests.

Q:12 What is the purpose of the Bayesian Information Criterion (BIC) in model selection?

A) To select the model with the highest likelihood function.

B) To select the model with the lowest sum of squared residuals.

C) To select the model with the lowest Akaike Information Criterion (AIC).

D) To balance model fit and complexity by penalizing the number of parameters in the model.

Answer: D) To balance model fit and complexity by penalizing the number of parameters in the model.

Explanation: The Bayesian Information Criterion (BIC) is a criterion for model selection that balances the fit of the model to the data with the complexity of the model, by penalizing the number of parameters in the model. The model with the lowest BIC is preferred, as it provides the best trade-off between model fit and complexity.

Q:13 Which of the following is a common assumption of ANOVA?

A) Normality of the residuals.

B) Independence of the predictor variables.

C) Linearity of the predictor variables.

D) Homoscedasticity of the residuals.

Answer: A) Normality of the residuals.

Explanation: Normality of the residuals is a common assumption of ANOVA, as non-normal residuals can lead to biased estimates of variance and invalid hypothesis tests.

Q:14 What is the purpose of the F-test in ANOVA?

A) To test the significance of individual predictor variables in the model.

B) To test the significance of the overall model.

C) To measure the goodness-of-fit of the model.

D) To identify the presence of multicollinearity in the predictor variables.

Answer: B) To test the significance of the overall model.

Explanation: The F-test in ANOVA is used to test the significance of the overall model, by comparing the variance explained by the model to the variance not explained by the model.

Q:15 What is the purpose of the random effects in mixed-effects models?

A) To account for the variability between observations within each group.

B) To account for the variability between groups.

C) To estimate the fixed effects in the model.

D) To measure the correlation between the predictor and response variables.

Answer: B) To account for the variability between groups.

Explanation: Random effects in mixed-effects models are used to account for the variability between groups, by allowing the intercept and/or slope coefficients to vary randomly across groups.

Q:16 Which of the following is true about convolutional neural networks (CNNs)?

A) CNNs are only useful for image classification

B) CNNs use fully connected layers for feature extraction

C) CNNs can learn spatial features in an image

D) CNNs cannot handle variable input sizes

Answer: C) CNNs can learn spatial features in an image

Explanation: CNNs are a type of neural network that are particularly well-suited for image processing and analysis. They use convolutional layers to learn spatial features in an image, which makes them more effective than traditional neural networks for tasks like image classification.

Q:17 Which of the following is a limitation of supervised learning?

A) It requires labeled data for training

B) It can only be used for classification tasks

C) It cannot handle high-dimensional data

D) It is not scalable to large datasets

Answer: A) It requires labeled data for training

Explanation: Supervised learning is a machine learning technique that involves training a model on labeled data. This means that each data point is associated with a label or target value. However, obtaining labeled data can be time-consuming and expensive, which is a major limitation of supervised learning.

Q:18 Which of the following is true about the cost function used in supervised learning?

A) The cost function is used to measure the accuracy of the model

B) The cost function is used to minimize the error between the predicted and actual output

C) The cost function is used to measure the complexity of the model

D) The cost function is not used in supervised learning

Answer: B) The cost function is used to minimize the error between the predicted and actual output

Explanation: The cost function is a measure of how well the model is performing on the training data. In supervised learning, the goal is to minimize the difference between the predicted output of the model and the actual output, which is achieved by minimizing the cost function.

Q:19 Which of the following is an example of a supervised learning algorithm?

A) K-means clustering

B) Decision tree

C) Support vector machine

D) Principal component analysis

Answer: C) Support vector machine and B) Decision tree

Explanation: Support vector machine (SVM) is a popular supervised learning algorithm that is commonly used for classification and regression tasks. It works by finding the hyperplane that best separates the classes in the data. Decision tree is a type of supervised learning algorithm used for classification and regression tasks. It works by recursively splitting the data based on the features, and building a tree-like model to predict the target variable.

Q:20 Which of the following is true about overfitting in supervised learning?

A) Overfitting occurs when the model performs well on the training data but poorly on the test data

B) Overfitting occurs when the model is too simple and underperforms on the training data

C) Overfitting occurs when the model is too complex and performs poorly on both the training and test data

D) Overfitting occurs when the model is too sensitive to outliers in the data

Answer: A) Overfitting occurs when the model performs well on the training data but poorly on the test data

Explanation: Overfitting is a common problem in supervised learning where the model becomes too complex and starts to fit noise in the training data. This causes the model to perform well on the training data but poorly on the test data, which indicates that the model has not generalized well.