30/08/23

M T W T F S S
Page No.:
Date:
YOUVA

Postlab 1

9237
Ryan V

1. The fundamental presumptions behind linear regression are :-

i) There always exists a linear relationship between the independent and dependent variable.
ii) Homoscedasticity which is the assumption of constant variance. This means the variance of residuals should remain constant across all levels of independent variables.
iii) No or little multicollinearity exists. That is two or more independent variables are highly correlated.

2 Heteroscedasticity is the opposite of homoscedasticity. Here the variance of residuals varies across all levels of independent variables.
In other words, the spread or dispersion of the residuals is not constant across all levels of independent variables.
This can be problematic as it violates the assumption of homoscedasticity.

3 R-Squared (Coefficient of Determination) and Adjusted R-squared are model evaluation metrics.

R-Squared measures the proportion of variance in the dependent variable.
Its range is between 0 and 1. The value

closer to 1 indicates that the model is good and can explain large proportion of variance.

Adjusted R-squared takes into account the no. of independent variables in the model.
It penalizes the addition of irrelevant variables that don't contribute significantly to the model's explanatory power.
Adjusted R-squared is always lower than or equal to the R-squared value.

30/08/23

M T W T F S S
Page No.:
Date:
YOUVA

Postlab 2

9237
Ryan V

1. Linear regression (SLR) models a linear relation between a dependent variable and one or more independent variable.
Multivariate linear regression is a broader concept which uses multiple independent variables to model a linear relationship with the target or dependent variable.
Here various types of independent variables are used to predict the dependent variable.

2. Multivariate regression involves predicting a dependent variable using two or more independent variables
Example :-

Predicting the value of a house using the location, no. of rooms and sq.ft area is an example of multivariate regression.

3. We use multivariable regression models when we can't model a linear relationship using a single independent variable.
Multiple independent variables are used to capture and understand the complex relationship.
They help us to understand the influence of different factors on the target variable.
This provides a more realistic representation of the model.

4. 1) Linearity :- Relationships between the dependent variable and the independent variables are assumed to be linear.
2) Independence :- The multiple observations are assumed to be independent of each other. The residuals should not exhibit any temporal or spatial dependence.
3) Normality :- Residuals should follow a normal distribution.
4) Homoscedasticity :- Residuals should have constant variance across all levels of independent variables.

5. Multivariate normality refers to the assumption that residuals in a regression model follow a multivariate normal distribution. Linear Regression models the relationship between dependent & independent variables using LE.

6. Univariate regression involves predicting a single dependent variable using one independent variable whereas multivariate regression uses two or more independent variables. Multivariate regression captures more complex relationships by considering the joint influence of several variables.

7. Multivariate analysis helps to uncover relationships among multiple variables simultaneously. It provides insights into the influence of variables on each other and the outcome of interest.

8 It is primarily quantitative as it involves the statistical examination of numerical data from multiple variables.
It aims to quantify relationships, patterns and dependencies among these variables.
Qualitative information can also be incorporated.

30/08/23

M T W T F S S
Page No.:
Date
YOUVA

Postlab 3

9237
Ryan V

1. The logistic function is an S shaped curve that models the relationship between an input variable and probability of binary outcome. $(z = 1/1+e^{-z})$
The range of values is between 0 and 1.

2. Logistic Regression is well liked for its versatility in modelling binary outcomes and possibilities.
It handles both categorical and continuous predictor variables effectively, providing interpretable coefficients that indicate the direction and magnitude of variable effects.
It converts logits value ranging from $-\infty$ to ∞ to range of 0 and 1.

3 A logistic regression model's probability can be stated as a conditional probability using logistic function.

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1 + \cdots \beta_p X_p)}}$$

where Y is the outcome,
$X_1, X_2, \cdots X_p$ are predictor variables,
$\beta_0, \beta_1 \cdots \beta_p$ are coefficients.

4 The model outputs the logits ie log odds and logistic function outputs the probability.

Logistic Model $= \alpha + 1X_1 + 2X_2 + \cdots kX_k$

Logistic function $= f(z) = 1/1 + e^{-(\alpha + 1X1 + 2X_2 + \cdots)}$

3/5/23

M T W T F S S
Page No.:
Date:
YOUVA

Postlab 4

9237
Ryan V.

1 Decision tree helps to capture non-linear relationships between features and target variables.
It can handle categorical as well as numerical data.
It is used for regression and classification problems.
It is also robust to outliers.
They are most suitable for tabular data.

2 Inductive bias is a preference for simpler shorter trees over loger ones.
It favours attributes that offer more discriminatory power earlier in the tree building process.

3 Decision trees handles missing values by either using most common value of that attribute or by making decisions based on available data and choosing paths that best align with average values.

4 For continuous attributes, decision tree selects a threshold value that best splits the data into subsets.
Subsets should be such that the entropy is low or information gain is high.
The tree then branches accordingly.

5 Information Gain is measure that quantifies the reduction in entropy by splitting using an attribute.
The disadvantages includes bias towards attributes with many values, preference towards attributes with large no. of distinct values.
To overcome these disadvantages, information gain ratio is used.