Worksheet Set 1

PYTHON

1. C
2. B
3. C
4. A
5. D
6. C
7. A
8. A
9. A and C
10. B and C

11, 12, 13, 14 and 15 are in the python sheet

MACHINE LEARNING

1. A
2. B
3. B
4. C
5. C
6. B
7. D
8. D
9. A
10. B
11. B
12. B and C
13. Regularization is a set of methods for reducing overfitting in machine learning models. Regularization trades a marginal decrease in training accuracy for an increase in generalizability. Regularization provides a range of techniques to correct for overfitting in machine learning models. As such, regularization is a method for increasing a model's generalizability (that is, it is ability to produce accurate predictions on new datasets) Regularization provides this increased generalizability at the sake of increased training error. In other words, regularization methods typically lead to less accurate predictions on training data but more accurate predictions on test data.
14. There are different Regularization algorithms?

- Ridge Regression
- LASSO (Least Absolute Shrinkage and Selection Operator) Regression
- Elastic-Net Regression

Ridge Regression

Ridge regression is a method for analyzing data that suffer from multi-collinearity.

LASSO Regression

LASSO is a regression analysis method that performs both feature selection and regularization in order to enhance the prediction accuracy of the model.

Elastic-Net Regression

Elastic-Net is a regularized regression method that linearly combines the L1 and L2 penalties of the LASSO and Ridge methods respectively.

15 The error term represents the difference between the observed values and the predicted values in a regression model. It captures all the factors influencing the dependent variable that the model does not account for, due to limitations in data, variable selection, or other hidden factors. Consider an economist attempting to model the consumption spending of households based on their income level. Even with a well-specified model that includes income as an independent variable, the actual consumption patterns will not align perfectly with the model's predictions. This discrepancy is captured by the error term. For instance, if the model predicts a household with an income of $50,000 will spend $30,000 annually, but they spend $32,000, the error term for this household reflects a $2,000 unexplained amount. This discrepancy could result from factors not included in the model, such as the households' saving habits, preferences, or access to credit.

The error term is pivotal in regression analysis for several reasons:
– **Model Accuracy:** It helps in assessing the fit of the model. A smaller error term on average indicates that the model explains a large portion of the variation in the dependent variable.
– **Inference:** It is crucial for performing statistical tests on the estimated parameters. The error term's properties, such as its distribution, are essential in determining the efficiency and unbiasedness of parameter estimators.
– **Specification:** A systematic pattern in the error terms can indicate model misspecification, such as omitted variables, incorrect functional form, or heteroscedasticity.

The error term is interpreted as the component of the dependent variable that is not explained by the independent variables in the model. It reflects the impact of all factors affecting the dependent variable other than those explicitly included in the mode

The error term represents the combined effect of all omitted and unmeasurable factors affecting the dependent variable. Since it encompasses factors that are either not known or not measurable, the error term itself cannot be observed directly. Instead, its presence is inferred through discrepancies between observed and predicted values.

In theory, the error term for an individual observation could be zero if the model perfectly predicts the value of the dependent variable for that observation. However, for the model as a whole, it's highly unlikely all error terms would be zero due to the multitude of unobserved factors that could influence the dependent variable. In practice, the aim is to minimize the error term's magnitude on average, acknowledging that some level of error is inevitable due to model simplification and the inherent randomness in data.

In summary, the error term is a fundamental concept in econometrics and statistical modeling, encapsulating all the influences on the dependent variable not captured by the model. Understanding its role and characteristics is crucial for building accurate models and making reliable inferences from data.

STATISTICS

1. A
2. A
3. B
4. D
5. C
6. B
7. B
8. A
9. C

10. Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, it shows that data near the mean are more frequently occurring than data far from the mean. The normal distribution curve is bell shaped.

    In a normal distribution, the mean is zero and the standard deviation is 1. It has zero skew and a kurtosis of 3.

    The empirical rule (also known as 3-sigma rule) for all normal distributions is that 68.2% of the observations will appear within plus or minus one standard deviation of the mean; 95.4% will fall within +/- two standard deviations; and 99.7% within +/- three standard deviations.

11. The following approaches are used to deal with missing data: drop rows with missing data, drop columns with missing data, mean imputation, median imputation, mode imputation, random sample imputation, and multiple imputations. Each of these methods has pros and cons.

12. A/B testing, or split testing, is a quantitative user research method. In A/B testing, researchers show different users two versions of the same design to identify which one performs better. The A refers to the original design, while the B refers to the variation of the A design.

13. Mean imputation is an acceptable practise of dealing with missing data, especially, when the missing data is a numerical data which the mean can be gotten or calculated from the other data available. Although, there is an issue if the missing data is a categorical issue. Also, the assumption of 'Missing Completely at Random also diminishes the use of Mean imputation. Summarily, the choice of which methods to use has to do the peculiarities of the missing data so that biases are not introduced which can greatly affect the prediction of the data.

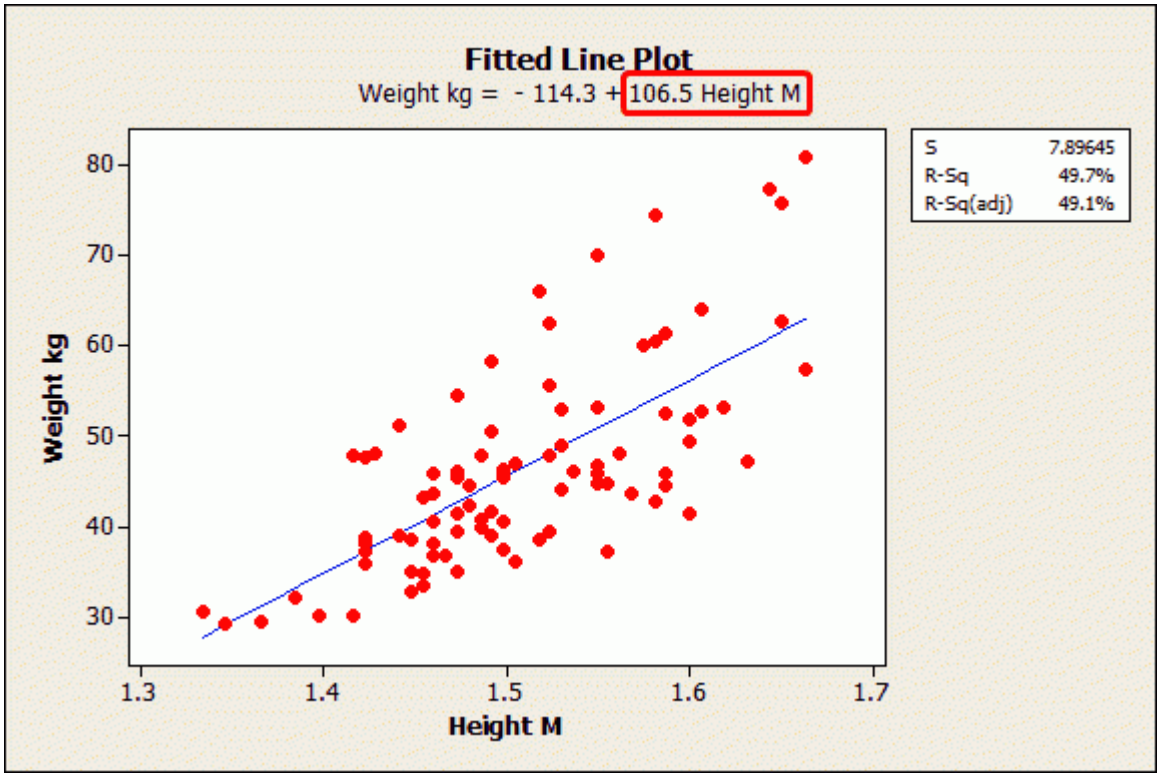
14. Linear regression explains the relationships between at least one explanatory variable and an outcome variable. Linear regression can fit curvature and interaction effects.

The explanatory variables in linear regression is referred to as independent variables (IV) and the outcome as dependent variables (DV). When a linear model has one IV, the procedure is known as simple linear regression. When there are more than one IV, statisticians refer to it as multiple regression. These models assume that the average value of the dependent variable depends on a linear function of the independent variables.

Linear regression serves two primary purposes—understanding the relationships between variables and prediction.

1. The coefficients represent the estimated magnitude and direction (positive/negative) of the relationship between each independent variable and the dependent variable.
2. The equation allows you to predict the mean value of the dependent variable given the values of the independent variables that you specify.

Linear regression finds the constant and coefficient values for the IVs for a line that best fit your sample data. The graph below shows the best linear fit for the height and weight data points, revealing the mathematical relationship between them. Additionally, you can use the line's equation to predict future values of the weight given a person's height.



Linear regression was one of the earliest types of regression analysis to be rigorously studied and widely applied in real-world scenarios. This popularity stems from the relative ease of fitting linear models to data and the straightforward nature of analyzing the statistical properties of these models.

15. The branches of statistics are

Descriptive statistics: It is the first part of statistics that deals with the collection of data. Descriptive statistics are used to do various kinds of analysis on different studies.

Example of Descriptive Statistics

- Central tendency measures
- Variability measures

The central tendency measures and variability measures use tables, general discussions, and charts.

Measures of Central Tendency

Central tendency measures specifically help statisticians evaluate the distribution center of values. These tendency measures are:

Mean

Mean is a conventional method used to describe the central tendency. Typically, calculate the average of values, count all values, and then divide them with the number of available values.

Formula of Mean

m= Sum of the terms/numbers of terms

For Example: Calculate the mean of the following data:10, 10, 40, 50, 20
Solution: m= Sum of the terms/numbers of terms10+10+40+50+20/5130/5 = 26Thus, mean=26

Median

It is the result that is in the middle of a set of values. An easy way to calculate the median is to edit the results in numerical journals and locate the result that is in the center of the distributed sample.

Formula of Median

To solve the median, there are two formulas;

- **When n is odd,**

(n+1 / 2)th observation

- **When n is even,**

median=(n/2)th + (n/2+ 1)th observation / 2

For Example 1: Find the median of 4, 1, and 7.
Solution: As the given are odd numbers of observations, then we apply the formula,median= (n+1)/2Thus, the median is3 + 1 / 2 = 4 / 2 = 2Median = 2

For Example 2: Find the median of the data 2, 4, 6, 8, 10, 12.
Solution: As the given numbers are even of observation, then we use the formula (n is even),
We will pick the middle numbers, 6 and 8.Thus, the median will be6+8/214/2 median= 7

Mode

The mode is the frequently occurring value in the given data set.

For Example: Find the mode of the given data,4, 2, 4, 3, 2, 2
Solution: Arrange the numbers2, 2, 2, 3, 4, 4Now you can see that the number 2 is found 3 times. Thus, the mode of the given data is 2.

Measures of Variability

The variability measure helps statisticians to analyze the distribution that is spreading from a specific data set. Some of the variables of variability include quartiles, ranges, variances, and standard deviation.

Inferential Statistics

Inference statistics are techniques that enable statisticians to use the information collected from the sample to conclude, bring decisions, or predict a defined population.

Inference statistics often speak in terms of probability by using descriptive statistics. Besides, a statistician uses these techniques for data analysis, drafting, and making conclusions from limited information. That is obtained by taking samples and testing how reliable they are.

Most future predictions and generalizations on a smaller specimen population study are in the inference statistics' scope. Besides, most social sciences experiments deal with studying a small sample population that helps determine the community's behaviour.

Designing a real experiment, the researcher can bring conclusions relevant to his study. When making conclusions, it should be cautious not to draw wrongly or biased.

Example of Inferential Statistics

Suppose you want to get an idea about the percentage of the people who love shopping at FILA. We take the sample of the population and find the proportion of individuals who love the FILA brand. With the assistance of probability, this sample proportion allows us to make a few assumptions about the population proportion. This study belongs to inferential statistics.

Different types of inferential statistics include:

- **Regression analysis:** It is a set of statistical methods used to estimate relationships between a dependent variable and one or more independent variables. It includes several variations, like linear, multiple linear, and nonlinear. The most well-known models are simple linear and multiple linear.
- **Analysis of variance (ANOVA):** ANOVA is a statistical method that distributes observed variance data into various components. A one-way ANOVA is applied for three or more data groups to gain information about the relationship between the dependent and independent variables.
- **Analysis of covariance (ANCOVA):** It is used to test categorical variables' main and interaction effects on constant dependent variables and keep control for the impact of selected other constant variables. The control variables are known as covariates.
- **Statistical significance (t-test):** It is used to determine a significant difference between the means of two groups related to particular features. A t-test studies the t-statistic, the t-distribution values, and the degree of freedom to learn the statistical significance.
- **Correlation analysis:** It is a statistical method that is used to find the relationship between two variables or datasets and discover how strong the relationship may be.