TASK:

In what situations might you choose to use ANOVA over linear regression or vice versa? Provide a real-life example dataset and explain how you would approach the data analysis using either ANOVA or linear regression. Justify your choice of one of the two approaches. How can Python be used to develop a useful tool in this context?

Introduction:

ANOVA (Analysis of Variance) and linear regression are both statistical techniques used to analyze relationships between variables. While they share similarities, they are applied in different contexts based on the nature of the data and the research question at hand. Understanding when to use ANOVA over linear regression, or vice versa, is crucial for effective data analysis.

What Is Analysis of Variance (ANOVA)?

Analysis of variance (ANOVA) is an analysis tool used in statistics that splits an observed aggregate variability found inside a data set into two parts: systematic factors and random factors. The systematic factors have a statistical influence on the given data set, while the random factors do not. Analysts use the ANOVA test to determine the influence that independent variables have on the dependent variable in a regression study.

Example of ANOVA Model Preferred

Suppose a biologist wants to understand whether or not four different fertilizers lead to the same average plant growth (in inches) during a one-month period. To test this, she applies each fertilizer to 20 plants and records the growth of each plant after one month. In this scenario, the biologist should use a one-way ANOVA model to analyze the differences between the fertilizers because there is one predictor variable and it is categorical.

What is linear regression?

Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.

Example of Regression Model Preferred

Suppose a real estate agent wants to understand the relationship between square footage and house price. To analyze this relationship, he collects data on square footage and house price

for 200 houses in a particular city. In this scenario, the real estate agent should use a simple linear regression model to analyze the relationship between these two variables because the predictor variable (square footage) is continuous.

ANOVA vs. Linear Regression:

ANOVA is typically used when the main objective is to compare means across multiple groups. It is well-suited for situations where the independent variable is categorical and the dependent variable is continuous. On the other hand, linear regression is employed when the focus is on predicting the value of a continuous dependent variable based on one or more continuous or categorical independent variables.

Dataset: https://www.kaggle.com/datasets/drmaryeslander/education-ocuupation-salary

References:

Frost, J. (2023, October 26). How F-tests work in analysis of variance (ANOVA). Statistics By Jim. https://statisticsbyjim.com/anova/f-tests-anova/

GfG. (2022, February 21). How to perform Welch's Anova in python. GeeksforGeeks. https://www.geeksforgeeks.org/how-to-perform-welchs-anova-in-python/

Investopedia. *Analysis of Variance (ANOVA) Explanation, Formula, and Applications. https://www.investopedia.com/terms/a/anova.asp*

IBM. *What is linear regression?. https://www.ibm.com/topics/linear-regression*

Tim. (2022, February 23). Anderson-Darling Test & Statistic: Definition, examples. Statistics How To. https://www.statisticshowto.com/anderson-darling-test/