TASK:

What are some potential limitations or assumptions of ANOVA and linear models, and how might these affect the validity of the results? How can these limitations be addressed or mitigated? How can Python be used to develop a useful tool in this context?

Definition:

- ANOVA is a statistical technique that assesses potential differences in a scale-level dependent variable by a nominal-level variable having 2 or more categories. For example, an ANOVA can examine potential differences in IQ scores by Country (US vs. Canada vs. Italy vs. Spain). Developed by Ronald Fisher in 1918, this test extends the t and the z test which have the problem of only allowing the nominal level variable to have two categories. This test is also called the Fisher analysis of variance.
- We can define linear regression as being used to learn the linear relationship between the target and one or more forecasters, and it is probably one of the most popular and effective deductive algorithms in the field of statistics. Linear regression attempts to demonstrate the link between two variables by fitting a linear equation to the observed information. One of the variables is considered an explanatory variable and the other a dependent variable.

Limitations and Assumptions of ANOVA and Linear Models

Assumptions of Linear Regression

To conduct a simple linear regression, one has to make certain assumptions about the data:

- Homogeneity of variance (homoscedasticity)- One of the main predictions in a simple linear regression method is that the size of the error stays constant. This simply means that in the value of the independent variable, the error size never changes significantly.
- Independence of observations- All the relationships between the observations are transparent, which means that nothing is hidden, and only valid sampling methods are used during the collection of data.
- Normality- There is a normal rate of flow in the data.

Assumptions of ANOVA There are three assumptions that should be met when computing an ANOVA:

- normal population distribution. The distribution of values within each group should be normally distributed. If a transformation is applied to the data, it should be applied to all batches
- Homogeneity of variance: The variance between the batches (homogeneity of variance) should be similar.

- The data are independent.

Addressing or mitigating the limitations of ANOVA and linear models involves a combination of careful data preprocessing, model diagnostics, and potentially using alternative statistical methods. Here are some strategies:

1. Check for Normality and Homoscedasticity: Conduct normality tests on the residuals (e.g., Shapiro-Wilk test) and homogeneity of variance tests (e.g., Levene's test). If these assumptions are violated, consider transforming the data (e.g., using logarithmic or square root transformations) or using robust regression techniques that are less sensitive to these assumptions.
2. Address Nonlinearity: If the relationship between the independent and dependent variables is nonlinear, consider using polynomial regression, spline regression, or generalized additive models (GAMs) that can capture nonlinear relationships more flexibly.
3. Account for Non-independence: In cases of non-independent data (e.g., repeated measures or clustered data), use mixed-effects models or generalized estimating equations (GEE) that can appropriately account for the correlation structure in the data.

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*

Penn State University. *ANOVA Assumptions. https://online.stat.psu.edu/stat500/lesson/10/10.2/10.2.1*

statistics solutions. *ANOVA (Analysis of Variance). https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/anova/*

analytics steps. *Simple Linear Regression: Applications, Limitations & Examples. https://www.analyticssteps.com/blogs/simple-linear-regression-applications-limitations-examples*