TASK:

Linear regression is a powerful tool for modeling relationships between variables in a dataset. However, there are limitations to linear regression that may result in biased or inaccurate predictions. Discuss at least two common assumptions of linear regression models and the potential consequences of violating these assumptions. How can Python be used to diagnose and address violations of these assumptions in a linear regression model? Provide at least one example of a linear regression model that violates one of these assumptions and how it can be improved.

Definition:

Linear regression is a statistical method used to model the relationship between one or more independent variables (predictors) and a dependent variable (response).

The general idea of regression

All models define the outcome (Y) as a function of one or more parameters and an independent variable (X).

Why linear regression is important:

Linear-regression models are relatively simple and provide an easy-to-interpret mathematical formula that can generate predictions. Linear regression can be applied to various areas in business and academic study.

You'll find that linear regression is used in everything from biological, behavioral, environmental and social sciences to business. Linear-regression models have become a proven way to scientifically and reliably predict the future. Because linear regression is a long-established statistical procedure, the properties of linear-regression models are well understood and can be trained very quickly.

The goals of regression

Scientists use regression with one of three distinct goals:

- To fit a model to your data in order to obtain best-fit values of the parameters, or to compare the fits of alternative models.
- To fit a smooth curve in order to interpolate values from the curve, or perhaps to draw a graph with a smooth curve.
- To make predictions.

Discuss at least two common assumptions of linear regression models and the potential consequences of violating these assumptions.

To answer the question regarding the assumptions and potential consequences of linear regression violations, we will begin by explaining 3 common assumptions:

1. Linear and Additive

If you fit a linear model to a non-linear, non-additive data set, the regression algorithm would fail to capture the trend mathematically, thus resulting in an inefficient model. Also, this will result in erroneous predictions on an unseen data set.

1. Multicollinearity

It occurs when the independent variables show moderate to high correlation. In a model with correlated variables, it becomes a tough task to figure out the true relationship of a predictors with response variable. In other words, it becomes difficult to find out which variable is actually contributing to predict the response variable.

1. Homoscedasticity:

This assumption states that the variance of the residuals is constant across all levels of the independent variables. If this assumption is violated (i.e., if there is heteroscedasticity), it can lead to inefficient coefficient estimates and incorrect standard errors, affecting the hypothesis tests for the coefficients.

data set:

https://www.kaggle.com/datasets/levyedgar44/income-and-happiness-correction

References:

Analyticsvidhya. 6 Assumptions of Linear Regression :Plots and Solutions. https://www.analyticsvidhya.com/blog/2016/07/deeper-regression-analysis-assumptions-plots-solutions/

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

Graphpad. The goal of regression. https://www.graphpad.com/guides/prism/latest/curve-fitting/reg_the-goal-of-regression.htm