Name:Aman Kumar

Roll No: 2020492

**Omitted variable:**

Problem -

When a significant variable is omitted from a regression analysis, the estimates of the coefficients of the variables that are included are skewed. In order to account for their impacts on the outcome variable, it is best to include all pertinent variables in the regression analysis.

Evidences:

For instance, the researchers who conducted this study on the association between coffee intake and mortality risk discovered that the estimates of the effect of coffee on mortality risk were biased as a result of the variable of smoking status being left out of the analysis.

DOI:<https://doi.org/10.1017/S0007114513003814>

Optimal Solution:

The optimal strategy is to adjust for their impacts on the outcome variable by including all pertinent factors in the regression analysis and also the best way to correct this error is to analyse all pertinent variables. In order to avoid confounding the relationship of interest, researchers should carefully assess what factors might be left out of the study and include them if practical.

**Overfitting:**

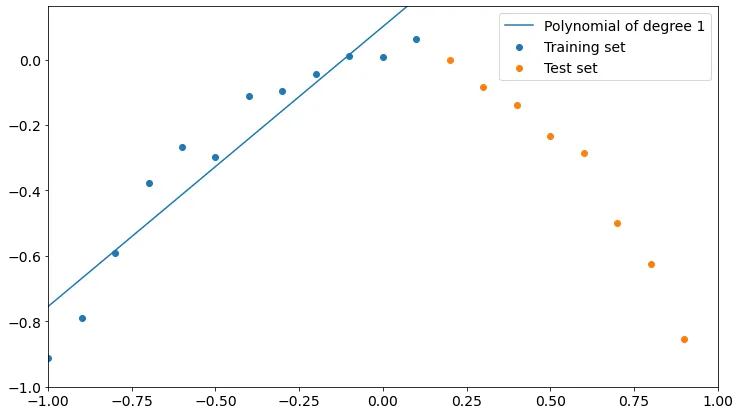
Problem:

When a model is overfitted, the performance of the model on fresh data is overestimated because the model is too sophisticated for the quantity of data at hand. The best course of action is to utilise a less complex model or regularisation techniques to make the model less complex.

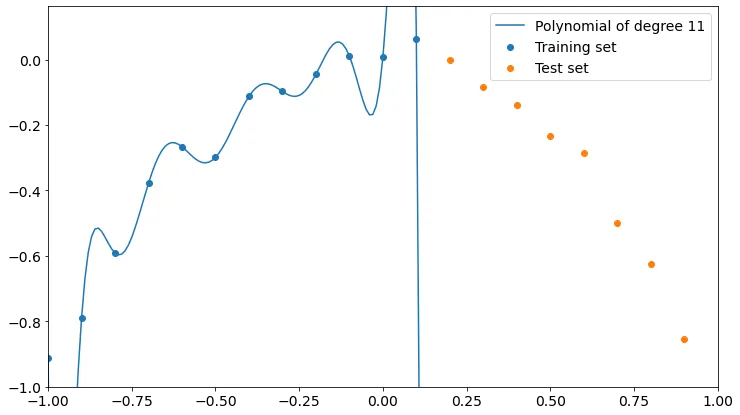
Evidences:

The researchers who conducted this research on the prediction of breast cancer using gene expression data discovered that overfitting and poor performance on fresh data were caused by employing a model that was too complex.

DOI:<https://doi.org/10.1073/pnas.201162998>



Although it may fit better, a polynomial of degree one fits training data better than test data. We could claim that the model is performing poorly because it is not properly learning from training.



The polynomial currently matches the training points better, but it is entirely incorrect for the test points. More degrees appear to bring us closer to overfitting training data and poor test data accuracy. A high degree polynomial is a more complex model than a low-degree one because, as you may recall, the higher the degree of a polynomial, the more parameters are involved in the learning process.

Optimal Solution:

The best course of action is to utilise a less complex model or regularisation techniques to make the model less complex and also the most effective approach for avoiding overfitting is to utilise a simpler model or regularisation approaches. To evaluate how well their models work on fresh data, researchers should utilise cross-validation or other techniques, and then modify the complexity of their models as necessary. Overfitting can also be avoided by using regularisation techniques like L1 or L2 regularisation.

**Non-representative samples:**

Problem-

When the sample population does not fairly reflect the larger population of interest, non-representative samples are produced. As a result, estimations of population parameters may be skewed. Using a sampling technique that guarantees the sample is representative of the larger population is the best course of action.

Evidence-

The researchers who conducted this research on the connection between vitamin D and mortality risk discovered that employing an unrepresentative sample of older people resulted in inaccurate assessments of the impact of vitamin D on mortality risk.

DOI:<https://dx.doi.org/10.7326/M21-3324>

Optimal Solution-

Using a sample size that guarantees the sample is representative of the larger population is the optimal approach and also, using a representative sample of the target population is the best solution for this error. By using random sampling or by choosing a sample that has been shown to be representative of the population under investigation, researchers should make every effort to ensure that their sample is accurate.

**Measurement Error:**

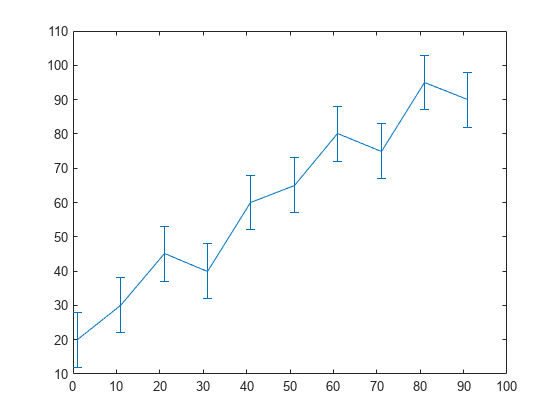
Problem-

When the measurement technique employed to gather the data is imprecise or inaccurate or the instruments used in experiment or research were broken and not giving the true result then, biased estimations of the actual population parameter result . Using an exact and precise measurement technique is the best course of action.

Evidence-

The researchers who conducted this research on the connection between air pollution and mortality risk discovered that utilising an insufficient measurement technique resulted in inaccurate estimations of the impact of air pollution on mortality risk.

DOI:<https://doi.org/10.3390/atmos11010009>



Line plot with error bars-A, each data point on the line plot of the y is marked with a vertical error bar. The sizes of each error bar above and below the data points are determined by the values in err.

**errorbar(y,err)**

Optimal Solution-

Using an exact and precise measurement technique and always insure that your instruments are not broken and calibrated before use, and using more precise measures of the variables of interest is the ideal solution for this error. Instead than depending solely on self-reported data, researchers should employ validated metrics, direct observation, or medical records to get data.

**Non-normality of data:**

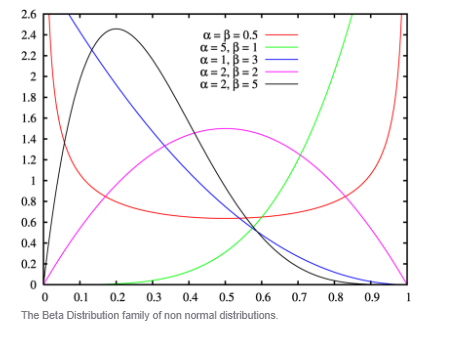
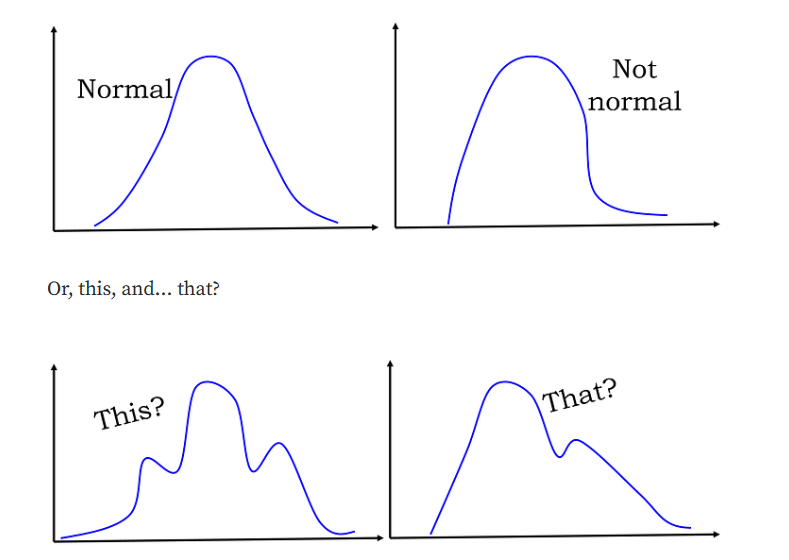
Problem-

When the distribution of the data is not normal, it is said that the data are non-normal, and this causes estimations of the population parameters to be skewed. Using non-parametric techniques or transforming the data to obtain normalcy is the best course of action.

Evidence-

The authors of this research on the connection between job stress and blood pressure discovered that performing a parametric analysis on non-normal data resulted in inaccurate estimations of the influence of job stress on blood pressure.

DOI:<https://doi.org/10.2105/ajph.2012.301153>



A non-normal model naturally fits many data sets. For instance, the frequency of accidents typically follows a Poisson distribution, and product lifetimes typically follow a Weibull distribution. Unfortunately, there may be instances where your data does not fit the normal distribution even though it should. Reasons could be outliers, multiple distributions may be combined in your data, insufficient data and data may be inappropriately graphed.

Optimal Solution-

The optimal approach is to alter the data to achieve normality or to employ non-parametric techniques and using non-parametric techniques or transforming the data to make it more normal before performing a parametric analysis is the best course of action for this error. Prior to performing parametric analyses, researchers should determine whether their data are normally distributed and then select the best procedures depending on those results.