Heteroscedasticity and Data Smoothing

Atanu Maity, Abhijit Ghosh, Sangita Roy, Swati Barui, Arpita Santra, Soumen Pal, Sandhya Pattanayak

Narula Institute Of Technology, ECE Department, Agarpara, Kolkata

Abstract

Data smoothing refers to a statistical method of getting rid of outliers from datasets to shape the patterns greater noticeable. It is accomplished by the use of algorithms to dispose of statistical noise from datasets. The utilization of information smoothing can assist forecast patterns, like share market, economics, computer vision, data mining, etc.

Keywords: Heteroscedasticity, data smoothing, outlier detection

Introduction

It's a process of studying available data and drawing valuable insights/ information from it with any software and algorithms. It's being used every day and everywhere to enable the business/lifestyle/information of nature to take smart and accurate decisions.Exploratory Data Analysis is to find out the patterns and relationships among the data variables and also gives us the summary of the data set hence the modeling becomes easy.There are 4 levels of data analysis: i) Descriptive, ii) Diagnostic, iii) Predictive, iv) Prescriptive. **Graphical and Analytical representation of data:**It's the simplest technique of representing data with its insight trends and pattern across the variable. Some major types of graphs are: i) Line, ii) Bar, iii) Histogram, iv) pie, v) Scatter, vi) Box and Whisker, VII) time series.**Descriptive Statistics of data set:**It's useful for calculating numerical values/variables. It's building a better understanding of data by standard deviation which tells the spread from mean/ average/expected values. If the SD is low then most of the number is close to the average value and a high SD means the number is spread out.**Data exploration:**Data exploration needs to understand the data and make sure it's ready to use for modeling. Skimming through the values of the variables of interest and seeing whether you can notice any pattern/anomaly by just looking at the data is called 'Eyeballing'[1,2]. **Outlier:**It's a data point that distances from other points in a Box and whisker plot whose values lie outside the usual range of the data. Outlier does not mean the higher values also mean the lower values also. There are 3 parts of an outlier:

i) Inter quartile range (IQR) = (q3-q1), ii) Lower limit = {q1 – (1.5 * IQR)}, iii) Upper Limit = {q3 + (1.5*IQR)}

So outliers occur due to various reasons: i) genuine variability in the data, ii) recording error of a sound, iii) Sampling error, iv) Experimental error etc.Many times some important values are not present in the data in a particular observation which is called 'missing values'. The missing values can reduce the information of a data. So missing values should be treated. There are different ways to treat outliers/missing values: i) Deletion, ii) imputing, iii) data Transformation, iv) Binning. **Correlation and ANOVAs:** Correlation is a measure of dependent/association between two variables i.e. how one variable changes with others. Most often it means how close both are. Correlation between two variables is: I) Positive II) Negative III) Zero. Those variables should take in modeling whose correlation having a high (positive/negative) above a certain cut-off value with targeted value.**ANOVA** stands for analyzing variance. It checks if the mean i.e. average value of the targeted variable across different levels/unique values of a categorical variable are equal/not. It accesses the importance of one/more levels by comparing the means of the target variable at different levels of categorical variables. Two values that are obtained from ANOVA: i) F-value = a large value, ii) p-value < 0.05. **Train and Test Data set:**The data on which the model is built is called 'trained data' which is used by the model to learn.Once the model is trained, it's examined as to how well it has learned using another subset of the original data which is called 'test data'. The model predicts the target variable value for the test data and then the predicted values are compared with actual values and checked as to how many of them were correctly predicted.**Examples of some kind of data:**i) Image is one type of data that we process with some technique. Image processing is a method to perform some operation on an image, to get an enhanced image, or to extract some useful information from it. It's one type of signal processing in which input is an image and output may be image/characteristic/features. ii) Noisy information is one type of data that can significantly impact prediction on meaningful data. Therefore we have to reduce the noise to get the better performance of thEqual variance assumed data. Among many noise handling techniques, polishing techniques generally improve classification accuracy than filtering and robust technique.iii) Sound is another important data set for today's lifestyle. Sound processing is also trending to businesses with ML, AI, and UI, etc.**Feature Scaling:**This is all about scaling the feature variable into the same range. The variables are scaled to have a similar magnitude and range so that model is not biased towards a particular variable.It's just for those algorithms where some measure of distance between data points is involved like logistic regression linear, linear regression, K nearest neighbor, principal component analysis etc.The most popular technique for feature scaling are:

i) Standardization: It rescales the feature values so that they have the properties of a normal standard distribution with mean as zero and standard deviation of one.

$$x_1 = (x - \mu)/\sigma$$

ii) Min-Max Scaling: The value range for transformed variables lies between [0, 1].

$$x_1 = (x - (x)\,)/((x)\, - (x)\,),$$

iii) Normalization: Range is fixed from -1 to 1.

$$x_1 = (x - mean(x))/((x)\, - (x)\,)$$

Residual Plot:

A residual value is a measure of how much a regression line vertically misses a data point. Regression lines are the best fit of a set of data. A residual plot has the residual values on the vertical axis and a horizontal axis displays the independent variable.So, it's a scatter plot of the difference between prediction and actual[3,4]. A residual plot is typically used to find the regression. Some data sets are not good candidates for regression. Like: Heteroscedastic data, non-linearly associated data, outlier data, etc.

Residual value = (Prediction – Actual)

Homoscedasticity:

It's a 'Greek' word that means data with the same (homo) and scatters (skedasis). Simply we can say that the data has the same scatter. So basically, if we put the data in a graph with a reference line at the middle then we can find that the point must be about the same distance from the line which is called a regression line.The word is to be noted that I have said 'distance' not variance. When viewing a graph, it's easier to look at the distance from the point to the line to determine if a set of data shows Homoscedasticity. Technically it's the variance that counts and that what you have to calculate with this formula assumes

$$\sigma = \sum(X - ?\,\mu\mu\mu\mu u)/N?$$

As variance is just standard deviation squared, as a condition where the standard deviation is equal to all points. So if the ratio of the largest variance to the smallest is 1.5 or below the data is called 'Homoscedasticity'.The assumption of equal variance is that the different samples have the same variance, even if they come from different data. The assumption of equal variance is also used in the linear regression model, which predicts that the data is homoscedastic. So it's possible that if your data is widely spread out then regression will not work properly.In linear regression assumption, the spread of the residual is constant across the plot. Anytime that you want to disturb the assumption, there is a chance that you can't trust the statistical result. There are two reasons:i) While coefficient estimate's bias will not affect, it does make them less precious. Lower precision increases the correct population/data value.ii) In Heteroskedastic, the variance of the coefficient estimates increases so it produces a p-value that is smaller than the usual. This value will not be observed by the Ordinary Least Squares Regression, so among that variance, OLSR calculates the t-value and F- value which will lead you to believe that the model is statically significant but it's not.

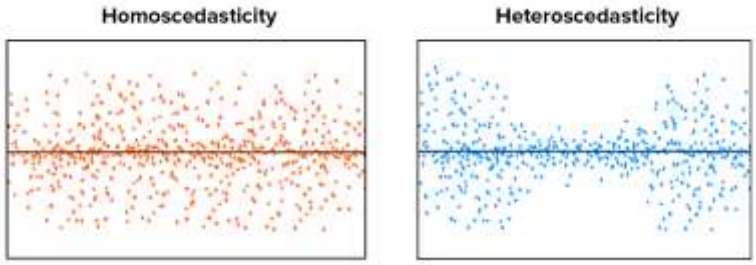

Figure 1. A plot of Homoscedasticity and Heteroscedasticity

Heteroscedasticity:

So here it means different dispersion. Technically, it refers to data with unequal variability (scatter) across the set of second, predictor variables. For the theoretical purpose, we all concentrate on those data which is going to be distributed equally but outside of our classroom this situation rarely happens in our nature. Most of the data are Heteroskedasis. **Heteroskedasticity in Regression Analysis:**In regression analysis, we talk about hetero-scatter in the context of residual or error term. Especially, it's a systematic change in the spread of the residuals over the range of measure values. This is a problem as OLS regression assumes that all the residuals are drawn from the data that has a context variance. To satisfy the regression assumption and be able to trust the result, the residual must have a constant variance. **Plotting a graph with an example:**Let's take an example: the weight and height of the students in a classroom must be unequal. This creates a cone-shaped graph for variability.

Figure 2. Plotting an example graph

Plotting variation of student's height/weight would result in a funnel that starts off small and spreads out as you move to the right of the graph. However, the cone can be in either direction (left-right or right-left).When the cone spreads out to the right, a small value of X gives a small scatter while larger values of X give a larger scatter to Y. When the cone spreads out to the left, a small value of X gives a larger scatter while larger values of X give a small scatter concerning Y. Heteroscedasticity can also be found in daily observation of the financial market, predicting sports over a season, and many other volatile situations that produce high-frequency data plotted over time.**Cause of Heteroskedasticity:**In more cases, the variance increases proportionally with this factor but remains constant as a percentage. For example, a 5% change in a normal market is much smaller than a 5% change in the shear market. So we have to care about the wide range of values. Because a large number of values are associated with these problems.Let's take another example, a cross-sectional study that involves that a house can have very large values for Mumbai but a small value for West Bengal. And also cross-sectional studies of incomes can have a range that extends from poverty to billionaires.**Categories:**There are two general types of Heteroscedasticity:i) Pure Heteroscedasticity refers to a cause where you correctly specify the model, and that causes the non-constant variance in the residual plot.ii) Impure Heteroscedasticity refers to cases where you incorrectly specify the model, and that causes the non-constant variance. When you leave an important variable out of a model, the omitted effect is absorbed into the error term. If the effect of the omitted variable varies throughout the observed range of data, it can produce the telltale sign of Hetero-scatter in the residual plot[5].

Experiment:

Regression Analysis:

In statistics, it's hard to stare at a set of random numbers in a table and try to make any sense of it. To understand the Regression let's take an example:

Table 1	
Year	Annual
1991	728
1991	728
1992	645
1993	675
1994	665
1995	818
1996	971
1997	867
1998	881
1999	901
2000	701
2001	678
2002	652
2003	791
2004	834
2005	764
2006	619
2007	839
2008	767
2009	816
2010	823
2011	1048
2012	597
2013	636
2014	537

Table 2		
Year	Annual	Rank
2011	1048	100%
1996	971	96%
1999	901	92%
1998	881	88%
1997	867	83%
2007	839	79%
2004	834	75%
2010	823	71%
1995	818	67%
2009	816	63%
2003	791	58%
2008	767	54%
2005	764	50%
1991	728	42%
1991	728	42%
2000	701	38%
2001	678	33%
1993	675	29%
1994	665	25%
2002	652	21%
1992	645	17%
2013	636	13%
2006	619	8%
2012	597	4%
2014	537	0%

If you have global records on average 'rainfall' in every state of your country and you are asked to predict how much rainfall will happen this year in your city. So, looking at the table you may predict that 10-20 cm3, that may be a good prediction but you can do this better by 'Regression'. Essentially, regression is the 'better guesses at using a set of data to make some kind of prediction. It's fitting a set of points to a graph. There's a whole host of tools that can do the regression for you. This also gives you an R-Squared value. This is 0.0721. This number tells you how good your model is! The range of the value is 0 – 1, zero means terrible and one means perfect. The output would include a summary for regression, which includes: i) Multiple correlation coefficients(R), ii) Coefficient of determinations(R^2), iii) Adjusted R^2., iii) The standard error of the estimate.

Conclusion

In statistical methods, data smoothing is an undoubtedly important topic for refinement of information. It was seen that unwanted data are random in nature and their dispersions are not even or uniform. From that point of view, these dispersions are heteroscedastic. Above discussed work shows that data smoothing in statistical methods is the obvious choice to remove outliers from data sets shaping the patterns from mixed data sets. State-of-the Art algorithms are there to remove noise from mixed data. The application areas are to forecast patterns, like share market, economics, computer vision, data mining, etc.

References

1. Baltagi , B. H. , Jung , B. C. , Song , S. H. (2009). Testing for heteroskedasticity and serial correlation in a random effects panel data model . *J. Econometrics* 154 : 122 – 124 . [Crossref], [Web of Science ®], [Google Scholar]
2. Box , G. E. P. , Cox , D. R. (1964). An analysis of transformations . *J. Roy. Statist. Soc. B* 26 : 211 – 252 . [Google Scholar]
3. Candelon , B. , Gil-Alana , L. A. (2004). Fractional integration and business cycle features . *Empir. Econ.* 60 : 343 – 359 . [Crossref], [Google Scholar]
4. Cavaliere , G. (2004). Unit root tests under time-varying variances . *Econometric Rev.* 23 : 259 – 292 . [Taylor & Francis Online], [Google Scholar]
5. E Klann, R Ramlau,, Regularization by fractional filter methods and data smoothing,29 February 2008, IOP Publishing Ltd.