Project 1 : Data quality issues

Suraj Eswaran

University of the Cumberlands

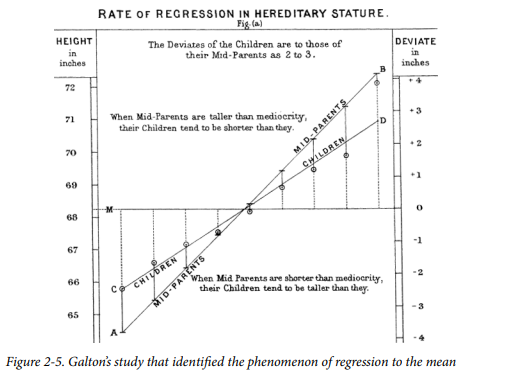
2024 Spring – Statistics For Data Science (MSDS – 531 – A01) – First Bi-Term

Professor Dr. Danny T. Barnes

01/14/2024

# This project is mostly based on issues with data quality: see more carefully chapter 2 Data and Sampling Distributions from ‘Practical Statistics for Data Scientists’. Cover in the project the following:

## Explain figures 2.5 and 2.6 on page 38.



From Figure 2-5, the graph explains the phenomenon of regression to the mean by Franscis Galton(Bruce et al., 2020) by plotting average heights of parents with respect to heights of children. In this research, he had understood that parents who are shorter tend to have taller children and vice-versa. Comparing the line of mid-parents with respect to children line in the graph, we could observe the slope of between those line less than or equal to 45 degrees. At the point zero of deviation, the median at which both parents and children are of same height is some where between 68 to 69 inches.

From the graph, we can infer an information in such a way that taller children with short parents tend to project negative standard deviation whereas shorter children with taller parents resulting in positive standard deviation.

## 

From Figure 2-6, we have a histogram of annual incomes of 1000 loan applicants with different sample sizes. At the first graph, data of annual income was taken when n=1(that is they have taken the sample only once), where the second graph is projected with collection of 5 different samples of 1000 loan applicants and third graph shows the collection of 20 different samples of 1000 loan applicants. As we increase the number of sample size, the graph tends to be thinner and taller.

Taking sample means from the population infers a change in the shape of graph but the mean which is calculated from sample means will resulting the same but there will slight decrease in standard deviation of sample means as the number of observation is inversely proportional to standard deviation. From the figure, the third graph will have lesser standard deviation of sample means while the first graph will have a higher standard deviation.

# Discuss the following:

## How can you tell if the data is an outlier or if it is something important?

Outliers are such kind of data which does not follow into the pattern or line of the dataset. Outliers can be extremely high value or extremely low value, as it depends on the trend of the data. It is necessary to detect outliers as it can significantly impact the quality of the data. Let’s take an example where we can decide whether the data is an outlier or not. If we were given a list of numbers [4,8,300,3,62,17,81,32] , where our simple approach would be ordering them in an ascending order, resulting in [3,4,8,17,32,62,81,300]. Looking at this result, we will be considering 300 is an outlier as it does not seem to follow a pattern and they are extremely higher value.

An arithmetic method name interquartile range where we calculate it from box plot using third quartile and first quartile difference. If any data is greater than or equal to sum of third quartile with 1.5 times of Interquartile range, that is considered as higher outlier value whereas any data which is lesser than difference of first quartile with 1.5 times of Interquartile range, is considered as lower outlier value. There are few other ways at which data can be decided whether the data is an outlier or not. Data Analyst while exploring those kinds of data in a dataset, they convert those value to either “NULL” or “NA” so that those data will be omitted.

**Which data is the noise and how is the noise different from outliers?**

Noisy data is a wrong kind of data which are irrelevant to the dataset that we have. For example, a set of weights that we have, and if there are negative value on it, then that is considered as a noisy data. Sometimes, noisy data can be obtained either due to wrong spelling, numerical mistakes or typing mistakes. Outliers points are a kind of genuine data points, but noisy points are wrong data points. Noisy data can also affect while training those to machine learning model as it might affect the efficiency.

Noisy data can be removed by manually evaluation of dataset where we can sort those data in ascending order or perform some regression where we can see those points on being detected.

# When there are missing values, explain the pros and cons of the following strategies:

**Elimination of Data Objects**

There can be dataset where there will be missing values due to several reasons like lack of information or not stored properly. So, in order to avoid those, there are several ways to deal with those data. One such is eliminating those data objects where we will remove those missing information from the dataset. The main advantage of eliminating those data objects is that having those data can spoil the result of our observations. For example, if we have a dataset where we have empty data in one of the columns and perform regression with invalid data can result in wrong accuracy value of it. For these kinds of analysis, it is better to remove those objects, so that we will get correct analysis. The second advantage is that it reduces the data size where removing those data, can help us save time in data analysis.

But eliminating those data objects can sometimes result in missing some important information as they can be deciding factor in analyzing those data. Although the data size reduces, there are chances that it can affect the precision of the model as more the sample size will result in finding the proper precision value.

**Elimination of Missing Values:**

Another way of handling missing value is by eliminating the missing value rather than whole data object. With this method, we will be able to keep that data object with us as we are not eliminating whole data object, we will be able to retain those and use the rest of the information. Also, analysis can be done if we use information of other objects with respect to missing values where we can obtain more accuracy than eliminating the whole object.

But there is a disadvantage where we need time to understand that missing data before using that with other data objects, as sometimes it can redirect to the wrong direction. This resultant value can not be the actual resulted value that we have obtained.

# What are the limitations of analyzing real data with missing values and why is it impossible to really know such data?

While analyzing real data with missing values, we face lots of issue where it takes time to deal with those data. Having missing data in real data can result in reduction of sample size. For example, if we are conducting a survey where we are collecting information and opinion from the customers/users, where they might be questions which is not required to answer. Tabulating those data, we will get to see few missing response for that particular question, thus end up in missing value scenario. In these scenarios, we will be having reduced sample size, which will affect the accuracy of the study.

If we use those missing value’s object for our study, there will be chance of uncertainty in our model analysis. With this result, there won’t be any guarantee that accuracy level will be high as missing values could not be calculated.

Some missing values are filled by analyzing with other information objects and utilize those values onto that objects. For example, if we have data of housing prices where some data objects won’t have Garage Space details, but using other information object and filling it won’t be given us the accuracy it might not be correct data. Thus, handling those missing data has been a challenge while working with real time data, although we can train other data and test it with objects for predicting it using machine learning.

# References

Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical statistics for data scientists: 50+ essential concepts using R and Python. O'Reilly Media.