Learning demand with missing data

Narayanan U. Edakunni and Viveka Kulharia

October 27, 2015

1 Introduction

Public transport is a crucial element of a city/town in any part of the world. It is safer, cheaper and a sustainable mode of transportation. There are different forms of ticketing solutions available for these transit networks. In the past, many of the transit agencies have been interested in recording and tracking the revenue earned in a day. The revenue is tracked by recording the number of tickets sold and the consequent money earned. However, with the advance of technology like smart cards based ticketing, it is possible to obtain detailed data to perform advanced analysis of the data. Data from smart card based ticketing would include the boarding bus stop, alighting bus stop, the time of boarding and restricted identity of the ticket holder. This data allows different types of analysis to be performed on them. The analysis can then be used to improve the operations of the transit agency. However, due to various reasons, some of the data points could be missing with only some summary statistics of the data being available. In this invention, *we propose a system and method to fit a function to a data when exact values of some data points are missing. We make use of the summary statistics of the missing values to come up with better fit for the function.*

In this disclosure, we specifically look at the case when there is missing ticketing data. This could happen in different scenarios:

- Transit agencies have traditionally been relying on paper based tickets and very often record only the revenue(alternatively number of tickets sold) collected at the end of the day. When these transit agencies transition to electronic ticketing based technology, for a considerable period of time electronic ticketing co-exist with paper based ticketing. In these circumstances, electronic ticketing provides granular data about demand whereas the manual ticketing gives only a summary information like the total tickets sold. Hence, we can treat the granular data points lost due to manual ticketing as missing values. We compliment the rest of the data with the summary statistics of the missing values to improve the analysis over the data.
- There can be failures in the ticket collection due to which the system might fall back to manual ticketing. This again can be modelled as data missing

from the system with some summary statistic available for the missing data.

- When transit agencies share the data, there might be issue of privacy regarding the commuters. This might require the agencies to share only partial information with some data points deliberately excluded. In this case we can use the summary statistics of the hidden data to come up with better fits to the data.

2 Prior art

There have been instances and applications where the data involved in the analysis contains some missing values. However, the invention proposed is the first instance where we look at missing values in the demand data of a transit network. There have been lots of research looking at the best way to impute the values of the missing data[1, 3]. There have also been research looking at how to use any additional information about the missing values in learning more about the data[2]. However, the methods proposed depend on the specific application and the analysis that is being performed. In the current invention, we look at the problem of predicting demand as a function of time when some of the historical data is missing. We then develop a method that uses the summary statistics available for the missing data to build a more accurate model of the whole data.

3 Description of the invention

In this invention we propose a novel system and method to fit a function mapping time of the day to demand when some of the data points used to fit the function are missing. We first build a model of the mapping from time of day to the demand at that time using the historical data collected from the transit agency. Alternatively, the model could also map the demand in the past to future demand, usually referred to as autoregressive model. The historical data will consist of number of tickets sold/swiped at a particular time of the day over a number of days. We use this data to fit a function that would express demand as a function of the time of day. In our analysis we will assume that the relation is linear though we can extend it easily to the case when the mapping is non-linear. We build our analysis for a generic case where the independent variable can be multi-dimensional. The design matrix would be represented by X_N which is an $N \times m$ matrix with m being the the number of features and N being the number of data points observed. Hence, X_N would represent the historical data. The corresponding demand would be represented by Y_N. We assume that out of the N observations n are missing. The corresponding variables would be X_n and Y_n respectively. We then assume that the Y is a linear function of X corrupted by additive Gaussian noise ϵ and can be represented

as:

$$[Y_n^T Y^T]^T = [X_n^T X^T]^T \beta + \epsilon \tag{1}$$

where β is the regression coefficient. In order to estimate β we minimise the squared error of the prediction with respect to the observed response and is given by -

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\| Y_N - X_N \beta \right\|. \tag{2}$$

The corresponding solution for $\hat{\beta}$ is given by,

$$\hat{\beta} = (X_N^T X_N)^{-1} X_N^T Y_N \tag{3}$$

When we have side information like the sum of the missing responses we can include them in the regression as a constraint leading to -

$$\hat{\beta} = \operatorname{argmin}_\beta \left\| Y - X\beta \right\| \quad \text{subject to} \tag{4}$$

$$Y_n^T \mathbf{1} = c \tag{5}$$

where $\mathbf{1}$ is a column vector of ones with length n and c is the sum of the missing response values in Y_n. When solved we get,

$$\hat{\beta} = \left(X^T X + \frac{X_n^T \mathbf{1}\mathbf{1}^T X_n}{n} \right)^{-1} \left(X^T Y + \frac{X_n^T \mathbf{1} c}{n} \right) \tag{6}$$

which is quite different from the normal regression estimate given in eq.(3).

4 Evaluation

In this section, we demonstrate the use of the method developed in this invention to build a predictive model when there are missing values in the observed data. We assume that the independent variable X_N is of dimension 10. We first uniformly sample $N = 100$ data points from the 10 dimensional space to stand for the input observations. We then used a known and fixed value of the regression coefficients β to compute the response variable Y_N. For a given value of n we randomly delete n values from the rows of X_N and Y_N to obtain X and Y. We further compute the sum of Y_n to obtain the sum constant c. The matrix X, the matrix X_n of input variables with missing responses and the response values Y denote the type of data we would work with. In the particular case of ticketing data, the missing values in Y would correspond to the part where manual ticketing was used and the sum of tickets sold would correspond to c. We now use this data to estimate in two different ways-

- Use X and Y to estimate the regression coefficient using eq.(3). Note that in this case we are not using the side information.
- In the second case we use X,X_n, Y and c to compute the regression coefficient using eq.(6).

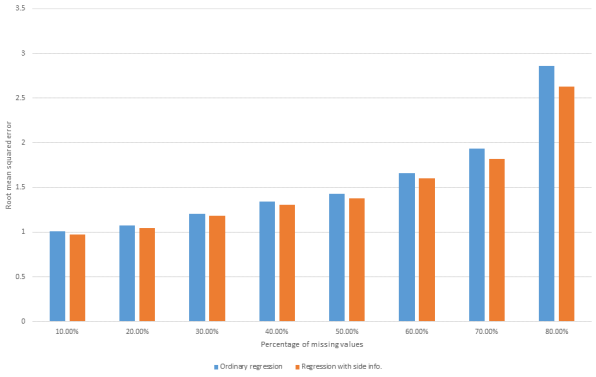

Figure 1: Comparison of regression approaches to handle missing value

We then use a separate hold out set of the data to estimate the responses and compute the root mean squared error(RMSE) of the predictions from the actual responses. We repeat the experiment over 10 different realisations of the data and the RMSE is averaged over these repeats. Furthermore, we vary the percentage of missing values in the training data to study the effect of missing value on the quality of the regression. The result of the evaluation is shown in fig.(1). We can see from the figure that as the number of missing values increases, the regression method using side information outperforms the conventional regression estimate. As the number of missing values increase, the usable data decreases and the conventional regression is not able to determine the parameters of the model accurately. The evaluation demonstrates the efficacy of the method proposed in this disclosure.

5 Detectability

Any invention/solution that uses auxiliary statistics of missing data to estimate properties of demand from the ticketing data.

6 Claims

- A novel method and system to deal with missing data.
- A method to use auxiliary data statistics of missing data to produce more accurate model when the actual data is missing.
- A novel application of the method in constructing a model of demand on public transport based on both finely granular and coarse data of ticketing.

References

[1] Andrew Gelman and Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[2] O. Ozan Koyluoglu and Naveen Ramakrishnan. Learning with missing data. Technical report, Ohio State University, 2010.

[3] R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. New York: Wiley and Sons, 2002.