Proceedings of the 2011 International Conference on Industrial Engineering and Operations Management
Kuala Lumpur, Malaysia, January 22 – 24, 2011

A Non-Parametric Predictive Model for Missing Data: A Case of Philippine Public Hospitals

Victor John M. Cantor, Richard C. Li, Martha Lauren L. Tan, Rachelle Joy S. Yu
Department of Industrial Engineering
De La Salle University, Manila 1004, Philippines

Abstract

Organizations have an abundance of data but actually have incomplete information. Incomplete information happens when there is missing or unreliable data which can lead to wrong decisions. A predictive model was developed using Hurwicz criterion by means of linear programming (LP). This predictive model estimates the organization's missing or unreliable data using external data from other similar organizations. The proposed predictive model was tested using data from Philippine public hospitals under the Department of Health. The model was able to provide a range of data that can test the validity of incomplete information encompassing the optimistic and pessimistic decisions made by the organization.

Keywords
Hurwicz criterion, linear programming, predictive model, performance measurement

1. Introduction
Missing data pertains to performance-related data that are unavailable due to several reasons that include: administrative fault in which the staff failed to collect the data, malfunctioning of equipment that resulted to data corruption and refusal of respondent to answer the questions from a survey among others. Performance measurement is data intensive thus in any case wherein data is missing, results may appear skewed and the reliability of the measurement is questionable [1].

Missing data exists and inevitable in all companies in different sectors thus should be accounted for. Exclusion or elimination of the cases or categories that contains missing data produces a distorted result of efficiency as some inefficient systems may be considered efficient in the absence of considering a significant input or output in evaluation. Moreover, the impact of missing data is detrimental not only through its potential hidden biases of the results but also in its practical impact on the sample size available for analysis. Therefore, estimation of missing data is left for selection. In choosing so, several estimation methodologies exist and must be further analyzed with regards to their assumptions and applicability with the study.

There are two kinds of estimation methodologies, the parametric which assumes data come from a type of probability distribution and makes inferences about the parameters of the distribution and the non-parametric which often referred to as distribution free methods as they do not rely on assumptions that the data are drawn from a given probability distribution. For the purpose of this study, the non-parametric technique was opted to not go into several assumptions about the data to be gathered.

2. Review of Related Literature

2.1 Missing Data

Since efficiency measurements for most organizations produce a wide range of outputs and use numerous inputs, data for these inputs and outputs must be well measured and collected in order to guarantee the accuracy of the measurement. However, more often than not it is not easy to get track of the records and information that are needed as much as sometimes there is no systematized recording or documentation system that would allow storage of records more accurately and timely.

In this premise, the problem of missing data arises frequently in practice. Missing data are performance – related data that are unavailable during the collection of data to be used for measuring performance [2]. For example, with the study of [3], consider a large survey of families conducted in 1967 with many socioeconomic variables recorded, and a follow-up survey of the same families in 1970. Not only is it likely that there will be a few missing values scattered throughout the data set, but also it is likely that there will be a large block of missing values in the 1970 data because many families studied in 1967 could not be located in 1970.

Another with [4], considering the 160 studies that have been identified as having missing data, only 54 (33.75%) explicitly acknowledged the problem. In contrast, 106 (66.25%) of the studies that have been identified as having missing data did not mention the problem, and missing values were inferred from degrees of freedom values that were inconsistent with the stated sample size and design characteristics.

Missing data are a common problem in quantitative research studies. Standard statistical procedures were developed for data sets, so missing values represent a considerable nuisance to the analyst. Traditionally, missing data were dealt by means of various ad hoc methods that attempted to "fix" that data before analysis. The blanket removal of cases with missing data (i.e., list-wise deletion) is one such strategy. Another method involves substituting missing values with the variable mean. Unfortunately, these ad hoc traditional methods can seriously bias sample statistics and have been criticized in the methodological literature [5].

2.2 Parametric Estimation Methodology

Regression analysis is used to predict the missing values of a variable based on its relationship to other variables in the data set. Although it has the appeal of using relationships already existing in the sample as the basis of prediction, this method also has several disadvantages. First, it reinforces the relationships already in the data. As the use of this method increases, the resulting data become more characteristic of the sample and less generalizable. Second, unless stochastic terms are added to the estimated values, the variance of the distribution is understood. Third, this method assumes that the variable with missing data has substantial correlations with other variables. If these correlations are not sufficient to produce a meaningful estimate, then other methods, such as mean substitution, are preferable. Finally, the regression procedure is not constrained in the estimates it makes. Thus, the predicted values may not fall in the valid ranges for variables, thereby requiring some form of additional adjustment. Even with all of these potential problems, the regression method of imputation holds promise in those instances for which moderate levels of widely scattered missing data are present and for which the relationships between variables are sufficiently established so that the researcher is confident that using this method will not impact the generalizability of the results. However, for most cases the data are being collected for analysis does not follow normal distribution thus for the purpose of this study a non-parametric estimation method is more preferred.

2.3 LP Predictive Estimation

A non-parametric technique that finds the best weights to be used in the linear equation based on the data set and its objective is to minimize the prediction values errors, that is, both negative and positive residuals. Mathematically, its objective function is to minimize the standard error of estimated using the residuals as the variables.

2.4 Hurwicz Criterion

The Hurwicz criterion attempts to find a middle ground between the extremes posed by the optimist and pessimist criteria. Instead of assuming total optimism and pessimism, Hurwicz criterion incorporates a measure of both by assigning a certain percentage weight to optimism and the balance to pessimism. A weighted average can be computed for every action alternative with an alpha-weight, α, called the coefficient of realism. Realism here means that the unbridled optimism of Maximax is replaced by an attenuated optimism as denoted by the α. Note that $0 \leq \alpha \leq 1$. Thus, a better name for the coefficient of realism is coefficient of optimism. An $\alpha = 1$ denotes absolute optimism (Maximax) while an $\alpha = 0$ indicates absolute pessimism (Maximin). The α is selected subjectively by the decision maker. Hurwicz criterion will be incorporated in the estimation of missing to consider the differing level of optimism of different decision makers.

3. Methodology

The study follows a systematic methodology in developing an LP predictive estimation model up until the validation of the estimation logic. Initially, the development of the LP predictive estimation model with the inclusion of Hurwicz criterion was done. The entire model was encoded in Microsoft Excel for data validation. Data were gathered from literature and from different database sources about healthcare performance in the Philippines wherein complete information was available. The intention of collecting a complete set of data was to purposely have the freedom to delete some data which will represent the missing data; and the deleted information will then become the basis of the desirability of the estimated value of the LP predictive estimation model. After the collection of data was completed, the LP predictive estimation model was run using the gathered data and predictive results were obtained. Several runs were done for each data set that were collected using different estimation techniques (i.e., regression analysis, mean substitution, and case substitution) in order to compare the predictive results of these estimation techniques with that of the LP predictive estimation model. Finally, after the comparison of results with the different estimation techniques, analysis of model logic and extraction of insights from the model results were conducted to ensure that the LP predictive estimation model gave a sound and logical results.

4. LP Predictive Estimation Methodology

4.1 LP Predictive Estimation Model

To adjust data sets where there are missing data occurrences, an LP predictive estimation model is suggested to estimate the missing data. The LP predictive estimation model is non-parametric in nature; as such there is no assumption of normality of data. Whether the data is non-normal or normal the LP predictive estimation model can be used to estimate the data. The model is as follows:

- Indices
  i = data set
- Parameters
  C_i = actual value of parameter to be estimated in data set i
  X_i = predictor parameter for data set i
- Variables
  A_1 = positive slope coefficient
  A_2 = negative slope coefficient
  B_1 = positive y slope coefficient
  B_2 = negative y slope coefficient
  δ_i = estimate error for data set i
- Objective Function
  The objective of the model is to find the best combination of coefficients to minimize the sum of estimate error from what is estimated by the model, and the actual value of the parameter to be estimated (this is denoted by the variable δ_i).

$$Min \sum_{i=1}^{n} \delta_i \tag{1}$$

- Constraints
  Estimate errors are obtained and calculated through error constraints limits the error should be the difference of the predicted value is greater that the actual value (negative error), and the other type would be that the predicted value is less than the actual value (positive error). To differentiate positive and negative errors two constraints are introduced for the two possible error types.
  - Negative Error Constraint
    The inequality constrains that if an estimated value is less than the actual value, the estimate error should be greater than zero.

$$C_i - \left[(A_1 - A_2) * X_i + B_1 - B_2 - \delta_i \geq 0, \forall i \right] \tag{2}$$

  - Positive Error Constraint

This constraint limits that if an estimated value is less than the actual value, the estimate error should be greater than zero.

$$C_i - \left[\left(A_1 - A_2\right) * X_i + B_1 - B_2 + \delta_i \geq 0, \forall i\right] \quad (3)$$

4.2 The Prediction Methodology

- Using the model, pair wise comparison of Standard Error is done to select the most appropriate predictor for the data to be predicted. Predictor with least standard error is selected.

- Use function to estimate data range

  The LP estimation model generates a single point estimate of the missing data, however since the values is just an estimate there is no guarantee that the calculated value will represent the 'real' value of the missing data. An interval estimate would better suit to represent the range of values in which the missing value can be expected. From each of the pair wise combination comparisons the standard error of the linear function was calculated using the following formula:

$$S_e = \sqrt{\frac{SSE}{n-2}} \quad (4)$$

  It is usually true that approximately 68% of the estimated values of y will be within Se, and approximately 95% of estimated y will be within 2Se [6]. In the three data set used it was found that approximately 83% are within Se. For the purpose of the estimate intervals, these are approximated by $\pm$ Se.

- Depending on the degree of optimism/pessimism set Hurwicz criterion

  Since missing data are expressed in terms of intervals, the concept of Hurwicz criterion is suggested such that the degree of optimism or pessimism is incorporated in obtaining a performance measure. The Hurwicz criterion attempts to find a middle ground between the extremes posed by the optimist and pessimist criteria. Instead of assuming total optimism or pessimism, Hurwicz incorporates a measure of both by assigning a certain percentage weight to optimism and the balance to pessimism.

- Using data range and Hurwicz criterion, predict missing data

5. Results and Analysis

5.1 Data

For the purpose of numerical validation, a case study is done on the urban health system of Local Government Units (LGU) in Metro Manila. There are 17 LGUs to be considered. The data set and information used for this case study is gathered from the official website and resources of the respective cities and from the materials of the study conducted by [7]. Data used are population, health establishments, health personnel, treatments serviced and cases examined. Health personnel refer to all the personnel involved in assisting the needs of the people in terms of their health problems. Some of the personnel involved in aiding the people are medical doctor, registered nurses, nutritionists or dietician, dentists, midwives, other technical and non-technical health worker. The value of the health personnel is solved using weighted average method. The health personnel is equal to the sum product of the number of personnel per manpower type and weight per manpower type.

Health establishments are the total number of health facilities in the LGU such as hospitals, health centers and clinics. Treatments serviced refer to the total number of service provided to the constituents of each city. It concerns treatments and vaccines that each LGU provided to the citizens of the city. This includes the following treatments and vaccines: HEPA B immunization, tetanus toxoid, vitamin A, deworming, dental, tuberculosis, rabies, and leprosy treatment. In addition, cases examined refers to the total number of patients examined by the health

personnel in each LGU namely tuberculosis cases, STD cases, measles cases, dengue fever cases, and pneumonia cases. As can be seen in Table 1, there are cases of missing data for health establishments and health personnel.

Table 1: LGUs Health Data Set Used for Case Study

City/Municipality	Population	Health Establishments	Health Personnel	Treatments Serviced	Cases Examined
Malabon	324,356.00	49.00	460.00	99,735.00	6,658.00
Navotas	231,717.00	24.00	249.00	84,038.00	3,110.00
Valenzuela	555,272.00	77.00	532.00	202,084.00	16,479.00
Caloocan City	1,445,209.00	80.00	496.00	409,966.00	9,828.00
Marikina	445,510.00	-	262.00	123,555.00	4,388.00
Pasig	557,297.00	83.00	-	178,332.00	15,995.00
Pateros	60,378.00	-	66.00	17,136.00	1,402.00
Taguig	630,161.00	60.00	774.00	185,806.00	4,589.00
Quezon City	2,468,417.00	-	905.00	1,008,935.00	58,466.00
Makati	389,478.00	73.00	410.00	129,037.00	20,365.00
Mandaluyong	264,825.00	58.00	388.00	89,485.00	9,674.00
San Juan	107,899.00	24.00	212.00	31,923.00	2,348.00
Manila	1,468,023.00	148.00	-	441,860.00	44,131.00
Las Piñas	575,359.00	-	464.00	234,935.00	25,557.00
Muntinlupa	348,443.00	46.00	898.00	97,746.00	5,695.00
Parañaque	550,831.00	49.00	463.00	181,893.00	2,602.00
Pasay City	285,649.00	46.00	383.00	113,106.00	14,181.00

To validate results of the LP estimation technique, standard error of estimates of the resulting function was used to compare the resulting data estimations. Linear function coefficients were obtained through regression analysis as well as the LP estimation methodology.

5.2 Pair-wise Regression Results

Using regression analysis for estimation for number of health establishments, the best predictor would be the data for treatments serviced with a standard error of 13.76. As for health personnel estimates, the best predictor would be the health establishment data set with standard error of 198.92 (See Table 2).

Table 2: Pair-wise Regression Results in Predicting Establishment and Personnel Variables

Y	X	SSE	Se	Coefficients	
				X	Intercept
Establishment	Population	3590.181818	14.5664856	3.59985E-05	36.471768
Establishment	Personnel	3590.181818	18.552989	0.034534731	36.743149
Establishment	Treatments	3590.181818	13.7602885	0.000136283	33.142326
Establishment	Cases	3590.181818	14.5334453	0.002143271	34.659588
Personnel	Population	412746.5455	204.93803	0.000163696	402.23726
Personnel	Establishment	412746.5455	198.928675	3.970297782	267.12777
Personnel	Treatments	412746.5455	206.716466	0.000526608	400.85072
Personnel	Cases	412746.5455	214.068607	-0.000930268	486.71523

5.3 Pair-wise LP Estimation Results

Using LP estimation methodology for the estimation of establishments, the best predictor would be treatments with a standard error of 10.66. As for health personnel estimates the best predictor would be the establishments with standard error of 157.95 (See Table 3).

For both regression and LP estimation, the predictors chosen are the same, but the LP estimation methodology resulted in lower standard error.

Table 3: Pair-wise Regression Results in Predicting Establishment and Personnel Variables

Y	X	SSE	Se	Coefficients	
				X	Intercept
Establishment	Population	2009.874556	11.2079061	2.93215E-05	37.624352
Establishment	Personnel	3250.404038	14.2530787	0.035369775	32.623794
Establishment	Treatments	1817.170965	10.6570721	0.000108898	35.355698
Establishment	Cases	1963.368569	11.0774788	0.00175093	37.342307
Personnel	Population	411489.9354	160.368703	0.00017649	341.26105
Personnel	Establishment	399166.8033	157.949122	4.806451613	161.90323
Personnel	Treatments	421997.3396	162.403306	0.00055623	338.22578
Personnel	Cases	426534.4262	163.274008	-0.00298373	470.76367

5.4 Pair-wise LP Estimation Results

Resulting estimates are logical as the resulting estimates are quite proportional to other inputs. Example for the city of Marikina, its data for population, personnel, treatments and cases are small to mid size in magnitude, it is only logical that the estimate for health establishment is also small to mid size. This behavior is observed with other cities as well where estimates of data are obtained.

Table 4 and Table 5 show various options of results that can be obtained using the LP estimation methodology. These varying types of results can be used depending on the needs and use of the estimated data. Point estimate simply gives the resulting estimate of the predicting function; while a range of values (minimum and maximum) can be used to have the confidence that the real value or data falls between the predicted range. Hurwicz criterion on the other hand, gives the flexibility for the user on how optimistic he or she is on the LGU or city in terms of input usage or output production. For example, in estimating for the health establishment data for Marikina (the user is a bit pessimistic in the efficiency of Marikina Health Care System); since there is a degree of pessimism, a lower alpha criterion can be used (0.4) to predict a higher value but still within the max and min range. Hurwicz criterion can give the user estimation flexibility, depending on the degree in which they know the operations of the LGU or city in which a data is being estimated.

Table 4: Estimate Results for Health Establishments

Estimates for Health Establishments					
City	Point estimate	Min value	Max Value	With Hurwicz	
				alpha = 0.4	alpha = 0.6
Marikina	48.81	38.15	59.47	50.94	46.68
Pateros	37.22	26.56	47.88	39.35	35.09
Quezon City	145.23	134.57	155.88	147.36	143.09
Las Piñas	60.94	50.28	71.60	63.07	58.81

Table 5: Estimate Results for Health Personnel

Estimates for Health Personnel					
City	Point estimate	Min value	Max Value	With Hurwicz	
				alpha = 0.4	alpha = 0.6
Pasig	560.84	402.89	718.79	592.43	529.25
Manila	873.26	715.31	1031.21	904.85	841.67

5.5 Conclusion

The LP estimation methodology had similar results with regression analysis in terms of the best predictor for a linear estimation function. However, using the LP estimation methodology resulted in lower standard error of estimate. Advantages of using the LP estimation methodology include: lower standard error of estimate, and that it is not limited to the assumptions of data. LP estimation can be used even when data sets are small (less than 30) or when data exhibits non-normality. Also, resulting estimates were logical in terms of the proportionality with other input data set across city/LGU. The estimation methodology also allows for varying types of data estimates, such as ranges and using of Hurwicz criterion to allow for user flexibility for optimistic and pessimistic estimates.

References

1. Kao, C., and Liu, S.T., 2000, "Data Envelopment Analysis with missing data: An application to University Libraries in Taiwan," The Journal of the Operational Research Society, 51(8), 897-905.
2. Little, R.J.A. and Rubin, D.B., 2002, Statistical Analysis with Missing Data, 2nd edition, New York: John Wiley.
3. Rubin, D.B., 1976, "Inference and missing data," Biometrika, 63, 581-592.
4. Roth, P.L., 1994, "Missing data: A conceptual review for applied psychologists," Personnel Psychology, 47(3), 537-560.
5. Enders, C., and Peugh, J., 2004, "Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement," Review of Educational Research, 74(4), 525-556.
6. Winston, W.L., 2004, Operations research: Applications and algorithms, 4th edition, California: Thomson.
7. Manalang, A., 2009, "A baseline comparative of the urban health systems in the national capital region," APIEMS conference proceedings 2009.