Using Predictive Models to Predict Outpatient Volume in Thailand using R Programming

Karn Yongsiriwit

College of Digital Innovation and Communication Technology, Rangsit University, Pathum Thani, Thailand
Corresponding author, E-mail: Karn.y@rsu.ac.th

Abstract

With the rapid adoption of information technology in healthcare over the last decade, a massive amount of data is accumulated and turned into Big data in healthcare. Therefore, exploring and utilizing such data are certainly necessary to gain valuable insight into the strategic planning of healthcare services in our nation. A relatively recent field of study and research, namely data science, plays a critical role in studying the prediction of data by constructing predictive models. However, not many pieces of research are conducted to explore the possibility of data prediction in healthcare, especially in Thailand. In this work, we use R programming to demonstrate the construction of two predictive models, Linear regression and Seasonal Naive, based on the actual number of outpatients in Thailand between 2014 and 2017. These two predictive models predict the daily average outpatient volume for each month in Thailand in 2018. Thereafter, we evaluate these two predictive models with the actual data in 2018 and compare the evaluation results. The results of our work can be a guideline for constructing a predictive model to predict patient volume, which is useful for the strategic planning and allocation of healthcare resources to handle upcoming patient volumes in the future.

Keywords**:** *Big Data in healthcare, Data Science, Data Analytics, Predictive Models, Linear Regression, Seasonal Naive, Outpatient Volume in Thailand, R Programming*

1. Introduction

During the last decade, the use of information systems and electronic health record (HER) for health services in hospitals to manage patient data has generated a massive amount of data, widely known as Big data (Sam, 2012). Such data is Terabyte-scale and will increase to Petabyte-scale in the near future. Performing data analytics on big data in healthcare is essential to obtain valuable knowledge which can be a guideline for planning policy to improve the quality of life in particular areas. Existing researches related to big data analytics, however, mainly focus on business and economic data and nearly neglect the data in healthcare. Therefore, this article presents and demonstrates the process of building two predictive models, namely Linear regression and Seasonal Naive. Such models are built from the actual data of outpatient volumes in Thailand between 2014 and 2017. We evaluate these two predictive models with the actual patient volumes in 2018. Finally, we compare the evaluation results of the two models. We believe that this article can be a guideline for using R programming to construct a predictive model to predict patient volume. Thus, hospitals can efficiently allocate healthcare resources to handle upcoming patient volumes.

Nowadays, data science is an increasingly demanding field of research to perform big data analytics. Many data science tools have been developed to facilitate the process of big data analytics such as KNIME, RapidMiner, Pentaho, R, Python, Julia, and Java. R is a programming language and open-source software environment for statistical computing and graphics (R Core Team, 2016). During the last few years, R has continuously increased the number of users, surveyed by Revolutions Daily news (2015). In addition, various data analytics methods have been developed as packages in the R language to promote the ease and efficiency of performing data analytics. In this work, we construct data predictive models to predict upcoming patient volumes in Thailand. Those who are interested in the R program can study our work as a guideline for further similar use.

The predictive model can be used in a variety of research related to health informatics. For example, Varun, Sreenivas, and Jack (2013) analyzed the health insurance data in the United States. Concretely, they constructed a predictive model to detect anomaly submission of health insurance claims using various data mining methods such as social network analysis and text mining. Batra et al. (2014) defined a standard for storing health information in medical electronic devices which can support planning

for healthcare policies. Nuaimi (2014) constructed predictive models in predicting the demand for healthcare services in Abu Dhabi. Costas, Sakib, Haik, and Majid (2016) presented an automated approach to classify and code diagnosis, symptoms, and procedures using machine learning and data mining techniques. Shu-Kay and Geoffrey (2016) have developed a mixture of model-based statistical pattern recognition for medical applications with hierarchically structured data. Rafael et al. (2016) have analyzed the number of patient in the emergency department of a hospital using various statistical principles such as linear regression and artificial neural networks. Parkpoom (2018) presented an approach to discover groups of the diseases frequently found in Thai preschool children, or the children of ages between 0-3 years, using an association rules mining, namely the apriori algorithm. All the work mentioned above, however, did not precisely demonstrate the process of constructing the predictive models for healthcare data. In this work, we select the R programming language as a tool to show the basic of data analytics process for constructing the predictive models to predict upcoming patient volumes in the future.

2. Objectives

This work aims to demonstrate the process of constructing predictive models to predict upcoming patient volumes in the future with R programming.

3. Materials and Methods

In this work, we use the R programming language as a tool to demonstrate the process of construction and evaluation of two predictive models, namely Linear regression and Seasonal Naïve, on the actual number of outpatients in Thailand. Concretely, we define the process as depicted in Figure 1. The process has 5 steps: (1) Data collection and preparation, (2) Data partitioning, (3) Predictive model construction, and (4) Data Prediction.

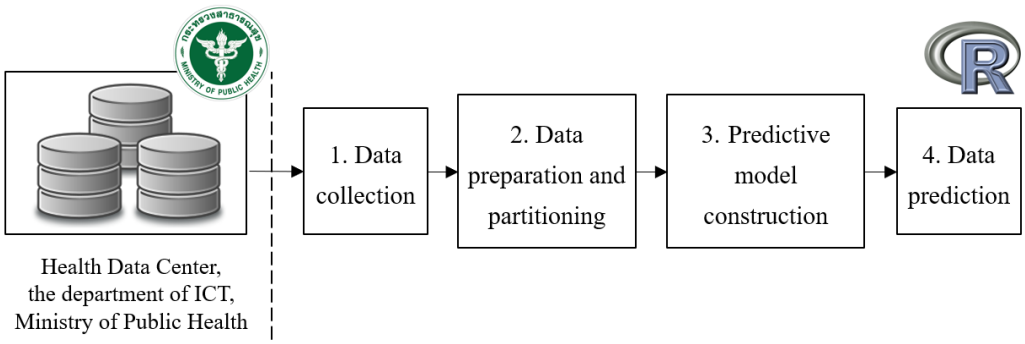

Figure 1 The process of constructing predictive models using R programming

3.1 Data collection

We appreciated the support of the department of Information and Communication Technology (ICT), Ministry of Public Health, for providing us the access to Health Data Center which has the data scale of more than 9 Terabyte structured based on Thai healthcare data model standard (Ministry of Public Health, 2017). Such data represents all the aspects related to providing healthcare services for registered medical facilities in Thailand. Table 1 depicts an excerpt of the data, that are, outpatient and hospital data. Outpatient data is on the right and hospital data is on the left. At outpatient data, each record details an outpatient (*pid*) in a hospital (*hospcode*) at a particular date (*date_serve*) diagnosed with a disease (*diagcode*). For the hospital data, each record represents a hospital (*hospcode*) in a province (*provcode*) with a hospital name (*hosname*).

Table 1 Excerpt of outpatient and hospital data from Health Data Center

hospcode	pid	date_serv	diagcode
11106	000003	2016-10-04	E100
11106	000003	2016-10-04	J100
11137	000055	2016-10-05	I110
11137	000057	2016-10-06	A000
13834	000123	2016-10-07	E100

hospcode	provcode	hosname
11106	11	Ratrin Hospital
11137	11	Sivawej Hospital
11219	12	Anan Phatthana Hospital
13834	30	Kum Phaya Health Promoting Hospital

We compute and collect an actual daily number of outpatients for medical facilities between 2014-2018 in Thailand by querying on the Health Data Center. The results shows in Table 2. The column *count* represents the total number of outpatients in a hospital (*hospcode*) during the date (*date_serv*). We record such data in a CSV format, that is, a delimited text file that uses a comma to separate values.

Table 2 Excerpt of daily outpatient number in Thailand

date_serv	hospcode	count
2014-01-01	00934	12
2015-01-01	00935	17
2016-01-02	00933	141
2017-06-09	99921	393
2018-12-31	99932	109

3.2 Data preparation and partitioning

Table 3 shows the process of data preparing and partitioning using R programming. At line 1, the data of outpatient volumes between 2014 and 2018 is loaded into the program as *df*. Following, we add new columns *year* and *month* into the *df* based on the column *date_serv*. Therefore, we select a hospital that we want to construct predictive models by specifying a hospital code in *hoscode* (at line 5). We select the Khon Kaen Hospital (*hoscode = 10670*) because this hospital is one of the top ten hospitals who have the highest average yearly volume of patients based on our dataset. Thus, we compute the average patient volume per day in every month for the selected hospital at line 6-7 with the help of *dplyr* library.

Before constructing the predictive models, we partition our outpatient volume data into training and test data. Training data (*train*) is the daily average outpatient volume for each month between 2014 and 2017, while test data (*test*) is only the data in 2018 as shown in Table 3 (at line 8-9). Hereafter, we will construct the models from the training data and evaluate such models with the test data.

Table 3 Data preparation and partitioning using R programming

```
1  df <- read.csv("data2014-2018.csv")
2  df$date_serv <- as.Date(df$date_serv, format="%Y-%m-%d")
3  df$year <- format(df$date_serv,"%Y")
4  df$month <- format(df$date_serv,'%m')
5  hoscode = 13754
6  library(dplyr)
7  df.monthly <- df %>%
     filter(hos == hoscode) %>%
     group_by(year, month) %>%
     summarise(total = mean(total))
8  train <- filter(df.monthly, year != 2018)
9  test <- filter(df.monthly, year == 2018)
```

3.3 Predictive Model Construction

Linear regression is a basic and commonly used method for predictive analysis. It allows studying of relationships between two continuous (quantitative) variables. On the other hand, Seasonal naïve is one of the forecasting approaches accounting for seasonality by setting each prediction to be equal to the last observed value of the same season. Such an approach is suitable for our dataset due to a high level of seasonality, that is, the specific months tend to have a greater volume than other months on every year as shown in Figure 2. Therefore, in this work, we decided to demonstrate the process of constructing these two predictive models.

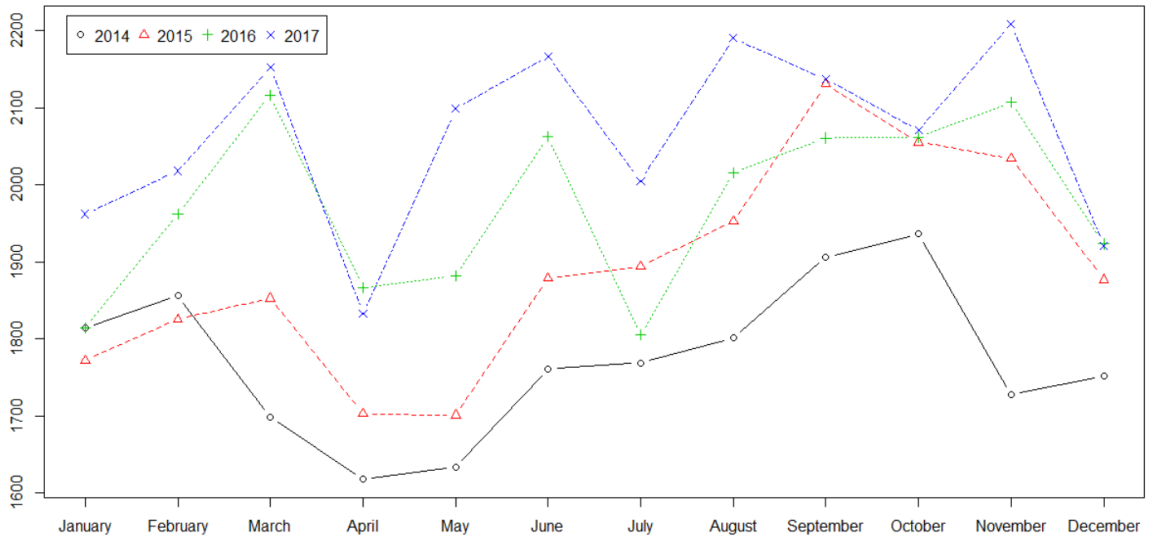

Figure 2 The average patient volume for each month between 2014 and 2017

Table 4 shows the constructing of predictive models using R programming. From the training data, we create a time series with a frequency equal to 12 (monthly) at line 1. Hereafter, we construct two predictive models: (1) Linear Regression (using a *tslm* function at line 2) considering the trend and the season of the data and (2) Seasonal naïve (using a *snaive* function at line 3). Our models are constructed and fitted based on the training data which is the monthly average outpatient volume between 2014 and 2017 of the Khon Kaen Hospital. Note that we simply construct such models with no tuning of their parameters for the sake of simplicity. Thus, readers can easily understand and follow the steps. Following, we will use these two predictive models to predict the monthly average outpatient volume in 2018.

Table 4 Predictive model construction using R programming

1	ts <- ts(train$total, frequency=12)
2	fit.lm <- tslm(ts ~ trend + season)
3	fit.snaive <- snaive(ts)

3.4 Data Prediction

Table 5 shows the usage of the R programming to predict data from the predictive models (Linear regression at line 2 and Seasonal naive at line 3) with the use of *forecast* function from the *forecast* library (at line 1). The data prediction for Khon Kaen Hospital using Linear regression and Seasonal naive models is shown in Figure 2 and 3, respectively. The black line represents the average monthly number of outpatients between 2014 and 2017, while the blue line presents the predicted monthly number of patients in 2018.

Table 4 Data predictiong using R programming

```
1  library(forecast)
2  fc.lm <- forecast(fit.lm, h=12)
3  fc.snaive <- forecast(fit.snaive, h=12)
4  plot(fc.lm)
5  plot(fc.snaive)
```

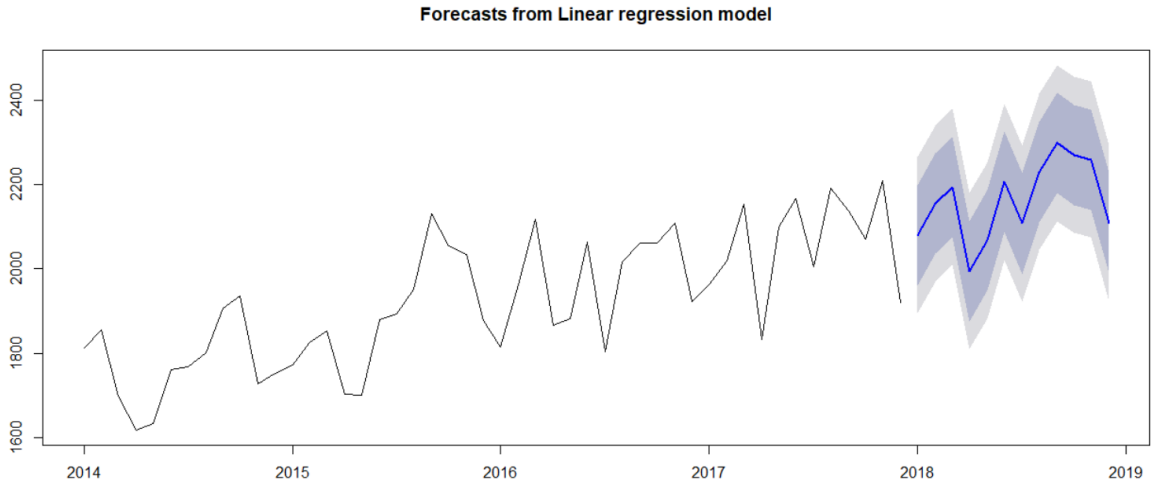

Figure 2 The data prediction from the Linear regression model

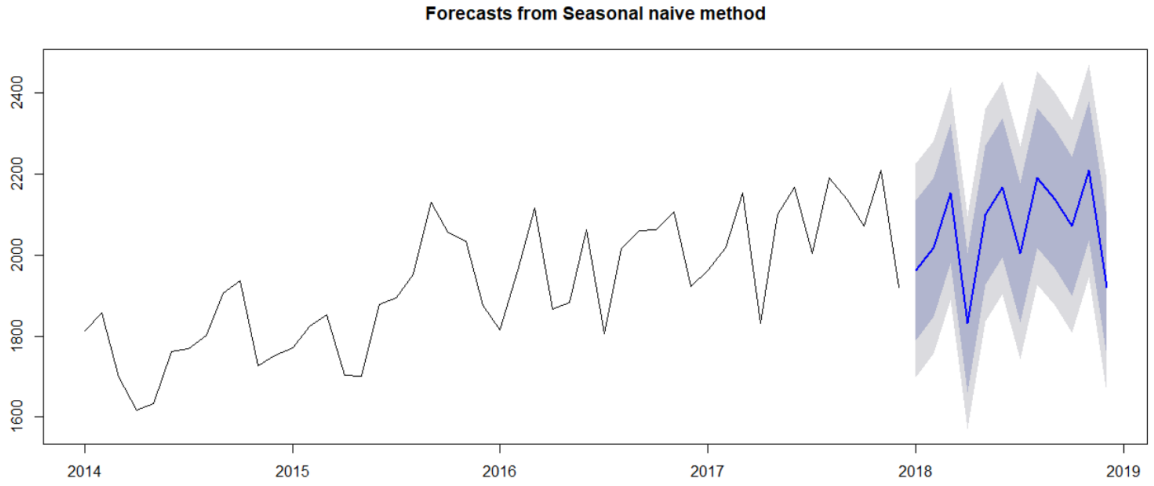

Figure 3 The data prediction from the Seasonal naive

4. Results and Discussion

We evaluate our two models, Linear regression and Seasonal naïve, by comparing the prediction results with the actual average monthly number of outpatients for Khon Kaen Hospital in 2018. The result is shown in Table 5. By considering the values in the table, it is still difficult to say whether Linear regression or Seasonal naïve is a more suitable model for predicting patient volume. For example, in January, March, and May, Seasonal naïve can predict value closer to the actual data than Linear regression. However, in February and April, Linear regression can predict value closer to the actual data than Seasonal naïve.

Table 5 Comparison of the prediction results with actual data

Period	Actual data	Linear regression	Seasonal naive
January-2018	2076.258	2079.837	2059.200
February-2018	2123.893	2154.824	2115.439
March-2018	2063.355	2194.216	2249.684
April-2018	1921.000	1994.288	1929.741
May-2018	2145.839	2067.983	2195.813
June-2018	2079.433	2206.513	2263.508
July-2018	2058.387	2107.337	2102.039
August-2018	2299.839	2229.305	2287.491
September-2018	2182.533	2298.096	2234.641
October-2018	2214.742	2270.249	2167.781
November-2018	2189.833	2258.563	2305.508
December-2018	1932.419	2107.523	2018.200

We use root-mean-square error (RMSE) to measure the accuracy of our models which is the difference between the predicted values from our models and the actual values. We calculate RMSE using R programming as shown in Table 6. We use library *Metrics* in line 1. Then, we call function *rmse* to compute RMSE for the predicted values of Linear regression and Seasonal naive models (at line 2 and 3). The RMSE are 255.45 and 632.74, respectively. Linear Regression has lower RMSE value than Seasonal naïve. From the result from Queen Savang Vadhana Memorial Hospital, Linear Regression is more suitable for predicting outpatient volumes than Seasonal naïve.

Table 6 Measuring RMSE using R programming

```
1  library(Metrics)
2  rmse(test$total, fc.lm$mean)
   255.45
3  rmse(test$total, fc.snaive$mean)
   632.7513
```

We also construct two predictive models for other hospitals that belong to the top 10 highest average number of patients per year (2014-2018). The result shows in Table 7. We can see that it is not all the case for Linear regression to have a lower value than Seasonal naïve. Seasonal naïve outperforms Linear regression for Chiangrai Prachanukroh Hospital, Khon Kaen Hospital, and Hatyai Hospital. We can see that the accuracy of predictive models depends on the data. Therefore, building different models and evaluating them is crucial to select the most suitable model for the data.

Table 7 Measuring RMSE for predictive models in different hospitals

Hospital	root-mean-square error (RMSE)	
	Linear regression	Seasonal naive
Queen Savang Vadhana Memorial Hospital	255.450	632.751
Samut Sakhon Hospital	549.931	577.314
Chiangrai Prachanukroh Hospital	324.906	277.926
Maharaj Nakorn Chiang Mai Hospital	222.063	229.403
Khon Kaen Hospital	93.483	85.481
Hatyai Hospital	176.589	151.148
Rajburi Hospital	125.249	295.634
Buddhachinaraj Hospital	103.961	129.471
Sunpasitthiprasong Hospital	144.509	147.338
Udon Thani Hospital	176.357	314.443

5. Conclusion

In this paper, we use R programming to demonstrate the process of constructing predictive models, namely Linear regression and Seasonal Naïve, using an actual number of patients in Thailand. With the use of several libraries in R programming, the tasks for constructing such models, predicting data, and evaluating the models are not difficult. We measure the differences between the predicted values and the actual values using root-mean-square error (RMSE). The results show that RMSE values for different hospitals are varied due to the data differences. Therefore, building different models and evaluating them is crucial to select the most suitable model. We believe that our work can be useful for those who want to learn the basic of data science and data analytics for Big data in healthcare. Lastly, the results of this work can help to efficiently allocate healthcare resources to handle upcoming patient volumes in the future.

For future work, we aim to explore other predictive models for predicting patient volumes. We also aim at looking for other factors that have an effect on a daily patient volume such as weather, temperature, different seasons, holidays, etc.

6. Acknowledgements

We appreciated the support of the department of Information and Communication Technology (ICT), Ministry of Public Health, for providing us the access to Health Data Center to obtain the dataset of daily patient volume for every registered medical facilities in Thailand between 2014 and 2018.

7. References

Sam, M. (2012). From Databases to Big Data. IEEE Internet Computing. *International Journal of IEEE Internet Computing, 16* (3), 4 – 6.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from http://www.R-project.org/.

Revolutions Daily news. (2015). New surveys show continued popularity of R. Retrieved from http://blog.revolutionanalytics.com/2015/11/new-surveys-show-continued-popularity-of-r.html, access on 27/03/2017.

Varun, C., Sreenivas, R., Sukumar, J. C. (2013), Knowledge discovery from massive healthcare claims data, paper presented in the nineteenth International Conference on Knowledge Discovery and Data Mining, Montreal, Chicago, IL, USA.

Batra, S., Sachdeva, P., Mehndiratta, H. Jyotsana, P. (2014), Mining Standardized Semantic Interoperable Electronic HealthcareRecords, *Biomedical Informatics and Technology,* 404, of the series Communications in Computer and Information Science 2014, pp 179-193

Nuaimi, N. A. (2014). Data mining approaches for predicting demand for healthcare services in Abu Dhabi, Proceedings of the 10th International Conference on Innovations in Information Technology, Al A in 2014, pp. 42-47.

Costas, S., Sakib, S., Haik, K., Majid, S. (2016). A Big-Data platform for Medical Knowledge Extraction from Electronic Health Records: Automatic Assignment of ICD-9 Codes, paper presented in the ninth ACM International Conference on Pervasive Technologies Related to Assistive Environments, Corfu Island, Greece.

Shu-Kay, N., Geoffrey, J. M. (2016). Finding group structures in "Big Data" in healthcare research using mixture models, paper presented in IEEE International Conference on Bioinformatics and Biomedicine, Shenyang, China.

Rafael, C., Flavio, S. F., Filipe, R. L., Jeruza N., Ricardo, S. K., Beatriz, D. S. (2016). Forecasting Daily Volume and Acuity of Patients in the Emergency Department. *Computational and Mathematical Methods in Medicine. 2016*, 3863268:1-3863268:8.

Parkpoom, C. (2018). Association Rules Mining for Diseases of Preschool Children in Thailand Using R Programming, RSU International Research Conference 2018, 187-195.

Ministry of Public Health. (2017). Thai healthcare data model standard references, 2017. Retrieved from https://www.moph.go.th/