ISSN- 2394-5125 VOL 07, ISSUE 01, 2020

Analysis of Medical Data using Linear Regression Technique: Numerical and Graphical Application

Dr. Manish Pandey[1], Dr. Parul Sharma[2], Rakesh Kumar[3]

[1]Director, International Institute for Technical Teachers, Dehradun

[2] Govt Post Graduate College for Women, Sec-14, Panchkula

[3] Assistant Professor, Tula's Institute, Dehradun

1. Abstract

We begin by exploring ideas in pattern extraction from biological datasets. Association rules have been studied extensively in the Knowledge Discovery in Databases (KDD) field for pattern extraction, and there exists many efficient algorithms to perform such task. The support and confidence thresholds are usually used to guide the search for interesting patterns. From our literature survey, we observed that most of the pattern mining methods are exhaustive; some practical difficulties arise when the number of items in each record is very large. Data mining has been defined as "the nontrivial extraction of implicit, previously unknown, and potentially useful information from data" [19] and "the science of extracting useful information from large data sets or databases" [20]. It is the core principle of the knowledge discovery process, which also includes data selection, preprocessing and cleaning, transformation and reduction, evaluation, and visualization.

Descriptive mining automatically extracts new or useful information from large databases and presents the discovered information in intuitively understandable terms for human analysis. Association rule mining is the most well-studied *descriptive mining* method in the Knowledge Discovery in Databases (KDD) field [21-24]. Their primary strength lies in their significant expressive power and their being relatively simple to comprehend, thus making them suitable for incorporation into decision-making processes.

Health care generates mountains of administrative data about patients, hospitals, bed costs, claims, etc. Clinical trials, electronic patient records and computer supported disease management will increasingly produce mountains of clinical data. This data is a strategic resource for health care institutions. Database Management Systems gave access to the data stored but this was only a small part of what could be gained from the data. Traditional On-Line Transaction Processing Systems (OLTPs), are good at putting data into databases quickly, safely and efficiently but are not good at delivering meaningful analysis in return.

Data storage became easier as the availability of large amounts of computing power at low cost became available i.e. the cost of processing power and storage has fallen and has made data cheap. There was also the introduction of new machine learning methods for knowledge representation based on logic programming in addition to traditional statistical analysis of data. The new methods tend to be computationally intensive hence a demand for more processing power. Improved data and information handling capabilities have contributed to the rapid development of new opportunities for knowledge discovery.

2. Introduction

In modern healthcare, evidence-based medicine is a new direction. Its task is to diagnose, medicate and prevent diseases using medical evidence. Traditionally medical data about a large patient population is analyzed to perform the healthcare management and medical research. In order to obtain the best evidence for given disease, external clinical expertise as well as internal clinical experience are pre-requisite for healthcare practitioners at right time and in the right manner. External evidence-based knowledge can not be applied directly to the patient without adjusting it to the patient's health condition. Through integration of data warehousing, On Line Analytical Processing (OLAP) and data mining techniques are useful in the healthcare area for care givers and clinical managers, as these provide an easy decision support platform [10]. This research proposes use of knowledge Discovery in Medical Databases, where medical experts can express their concerns and preferences to guide knowledge exploration from the data sets. When applying the derived knowledge patterns in medical work, the domain experts can further justify the decision support information and then refine the scope of the knowledge. With the following key features

- Centralized clinical data repository to securely manage data while providing distributed web based access for collaborative research.
- Organization and management of clinical research along multiple hierarchical studies, each with its own set of authorized users, protocols, patient cohorts, and saved queries.
- The system supports sharing resources across projects in a secure and transparent manner.
- Dynamic generation of web-based case report forms (CRFS) for clinical assessment instruments from user-defined clinical parameters, protocols and validation logic.
- Double Data Entry capabilities for clinical data to ensure accuracy when transcribing data from paper forms.
- Intuitive workflows and user interfaces integrated within the context of a familiar web-based environment; this improves access and minimizes the nee for dedicated training.
- Extensive interfaces for data query and retrieval across projects, individuals and phenotypic parameters, with export of datasets in standardized formats for statistical analysis.

3. Significance of the Study

For those in Academia and Research, the study would among others

- Provide good grounding in data mining and its related concepts, like data warehousing, data marts and so on.
- Enable research by organisations involved in Marketing, Telecoms, Insurance and Banks into using data mining techniques to increase Returns on Investment (ROI).

In the case of practitioners, the results of the study would

- Enable medical practitioners discharge their duties efficiently and effectively,
- Assist the hospital authorities to better manage their facilities and Staff
- Assist the hospital authorities to contribute effectively to the efficient and smooth running of the various insurance schemes that are being implemented
- Provide an easy to use interface with the capability of quickly drilling down to increasingly detailed patient information
- Improve quality of patient care through the use of data, trend and comparative analysis of clinical data
- Provide detailed clinical information to any Health Research Data warehouse, which will be used to analyze large amounts of data in long-term research projects
- Provide health services research through data management and analysis for better Policy Support.

The result will be useful to the government health managers, nongovernmental healthcare providers and their managers, and policy makers in health and related sectors of the economy. Finally, the Ministry of Health would benefit immensely since as part of its management information and performance monitoring, it hopes to improve upon the application of appropriate technology for data collection, storage, analysis and dissemination of health information.

Executive support systems (ESS) support the strategic level by providing a generalized computing and communications environment to assist senior management’s decision making. Laudon and Laudon (2000:37) indicated that the various types of systems in the organisation exchange data with one another. Transaction processing systems are a major source of data for other systems, especially MIS and DSS. The information needs of the various functional areas and organisational levels are too specialized to be served by a single system. With the increasing amount of Data from these systems, there is the need to organize it in such a way that Information and Knowledge can readily be made available for decision making purposes.

4. Time Series Analysis for Forecasting

A time series is a sequence of data points, measured typically at successive times spaced at uniform time intervals. Examples of time series are the daily closing value of the Dow Jones index or the annual flow volume of the Nile River at Aswan. Time series *analysis* comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series *forecasting* is the use of a model to forecast future events based on known past events: to predict data points before they are measured. An example of time series forecasting in econometrics is predicting the opening price of a stock based on its past performance. Time series are very frequently plotted via line charts.

Time series data have a natural temporal ordering. This makes time series analysis distinct from other common data analysis problems, in which there is no natural ordering of the observations (e.g. explaining people's wages by reference to their education level, where the individuals' data could be entered in any order). Time series analysis is also distinct from spatial data analysis where the observations typically relate to geographical locations (e.g. accounting for house prices by the location as well as the intrinsic characteristics of the houses). A time series model will generally reflect the fact that observations close together in time will be more closely related than observations further apart. In addition, time series models will often make use of the natural one-way ordering of time so that values for a given period will be expressed as deriving in some way from past values, rather than from future values.

5. A Summary of Forecasting Methods

A time series is a set of ordered observations on a quantitative characteristic of a phenomenon at equally spaced time points. One of the main goals of time series analysis is to forecast future values of the series. A trend is a regular, slowly evolving change in the series level. Changes that can be modeled by low-order polynomials. The use of intuitive methods usually precludes any quantitative measure of confidence in the resulting forecast. The statistical analysis of the individual relationships that make up a model, and of the model as a whole, makes it possible to attach a measure of confidence to the model’s forecasts.

Once a model has been constructed and fitted to data, a sensitivity analysis can be used to study many of its properties. In particular, the effects of small changes in individual variables in the model can be evaluated. For example, in the case of a model that describes and predicts interest rates, one could measure the effect on a particular interest rate of a change in the rate of inflation. This type of sensitivity study can be performed only if the model is an explicit one.

6. Linear Regression Equations

If we expect a set of data to have a linear correlation, **it is not necessary for us to plot the data** in order to determine the constants *m* (slope) and *b* (y-intercept) of the equation y = mx + c. Instead, we can apply a statistical treatment known as linear regression to the data and determine these constants.

Given a set of data (x_i, y_i) with *n* data points, the slope, y-intercept and correlation coefficient, *r*, can be determined using the following:

$$m = \frac{n\sum(xy) - \sum x \sum y}{n\sum\left(x^2\right) - \left(\sum x\right)^2}$$

$$b = \frac{\sum y - m\sum x}{n}$$

$$r = \frac{n\sum(xy) - \sum x \sum y}{\sqrt{\left[n\sum\left(x^2\right) - \left(\sum x\right)^2\right]\left[n\sum\left(y^2\right) - \left(\sum y\right)^2\right]}}$$

(Note that the limits of the summation, which are *i* to *n*, and the summation indices on *x* and *y* have been omitted.), implicitly applying regression to the sample data.

It may appear that the above equations are quite complicated, however upon inspection, we see that their components are nothing more than simple algebraic manipulations of the raw data. We can expand our spread sheet to include these components.

1. First, add three columns that will be used to determine the quantities **xy**, **x^2** and **y^2**, for each data point.
2. Next, use Excel to evaluate the following: **∑x**, **∑y**, ∑ **(xy)**, ∑ **(x^2)**, ∑ **(y^2)**, **(∑x)^2**, **(∑y)^2**. Recall that the symbol, **S**, means "summation". Additionally, the term **xy** is the product of **x** and **y**, that is: **x * y**. Also, the term ∑ **(x^2)** is very different than the term **(Sx)^2**. Be careful with your order of operations!
3. Now use Excel to count the number of data points, **n**. in this example is:
4. Finally, use the above components and the linear regression equations given in the previous section to calculate the **slope (m)**, **y-intercept (b)** and **correlation coefficient (r)** of the data. If you are careful, your spread sheet should look like ours. Note that our equations for the slope, y-intercept and correlation coefficient are highlighted in yellow.

6.1 Application of linear regression technique for the calculation of line equation for the month 2007:

Table 1 shows the details for the calculation of slope and constant for a straight line, where

x	: number of months in a year
y1	: number of patient in the year

2007

xy1 : multiplication of month and number of patients

xx : square of month number

y1y1 : square of number of patient in a month

After applying the following linear regression formula of slope and constant for the line equation, we will get the value of m and b:

Where n = 12 (for 12 months in a year), after applying the formula for the slope of the line and the constant b for the line equation y = mx + b, we get:
m = 7.566433566
b = 506.3181818
Now for the plot of the graph for x = month of the year, y = variable from line equation after applying the value of slope (m) and constant (b), and y1 = number of patients in the month. We get:

6.2 Application of linear regression technique for the calculation of line equation for the month 2008:

Table 4 shows the details for the calculation of slope and constant for a straight line, where

x : number of months in a year

y2 : number of patient in the year 2008

xy2 : multiplication of month and number of patients

xx : square of month number

y2y2 : square of number of patient in a month

After applying the following linear regression formula of slope and constant for the line equation, we will get the value of m and b:

Where n = 12 (for 12 months in a year)
m = 32.1
b = 464

Now for the plot of the graph for x = month of the year, y = variable from line equation after applying the value of slope (m) and constant (b), and y2 = number of patients in the month. We get:

6.3 Application of linear regression technique for the calculation of line equation for the month 2009:

Table 7 shows the details for the calculation of slope and constant for a straight line, where

x : number of months in a year
y3 : number of patient in the year 2009
xy3 : multiplication of month and number of patients
xx : square of month number
y3y3 : square of number of patient in a month

After applying the following linear regression formula of slope and constant for the line equation, we will get the value of m and b:

Where n = 12 (for 12 months in a year)
m = 98.35
b = 323

Now for the plot of the graph for x = month of the year, y = variable from line equation after applying the value of slope (m) and constant (b), and y3 = number of patients in the month. We get:

6.4 Application of linear regression technique for the calculation of line equation for the month 2010 (predicted value after indexing):

Table 10 shows the details for the calculation of slope and constant for a straight line, where

x : number of months in a year
y4 : number of patient in the year 2010
xy4 : multiplication of month and number of patients
xx : square of month number
y4y4 : square of number of patient in a month

After applying the following linear regression formula of slope and constant for the line equation, we will get the value of m and b:

Where n = 12 (for 12 months in a year):
m = 47.980918
b = 370.23427

Now for the plot of the graph for x = month of the year, y = variable from line equation after applying the value of slope (m) and constant (b), and y1 = number of patients in the month. We get:

6.5 Comparison of slopes for four year:

We get the different slope for different years

m (2007) = 7.566
m (2008) = 32.1
m (2009) = 98.35
m (2010) = 47.98

Figure 5 shows the comparison of slops of different four years: On comparing the four different slopes of four years we got that the predicted slope of the fourth year is about the average of the three year slopes hence we can say that the fourth year prediction is the good quality prediction for the patient data. The statistical reports in the following pages shows the various reports and analysis of patient data.

7. Conclusion

A time series is a sequence of observations which are ordered in time. Inherent in the collection of data taken over time is some form of random variation. There exist methods for reducing of canceling the effect due to random variation. Widely used techniques are "smoothing". These techniques, when properly applied, reveals more clearly the underlying trends. In other words, smoothing techniques are used to reduce irregularities (random fluctuations) in time series data. They provide a clearer view of the true underlying behavior of the series. The volume of patient medical data at the various hospitals has been increasing over the years. However, the data is not properly managed. As a result of this, majority of out-patients do not have full medical record. With this situation, the physician’s time is wasted since they have to collect this information again and in addition, it becomes very difficult for them to keep track of the patients. This reduces the ability to carry out high quality clinical research in the hospitals, and compromises the continuity of healthcare as well as the quality of healthcare delivery in the hospital. A Data Mart can be designed to collect, store, organize and retrieve the medical information of patients. The Data Mart can be built upon and deployed in for easy detection of false claims and disease management. A simple way of detecting trend in seasonal data is to take averages over a certain period. If these averages change with time we can say that there is evidence of a trend in the series.

References

[1] Abbott, P. A., Goodwin, L., Cullen, P. and Delaney, C. (1999). ‘The ABC’s of Data Mining: A primer for Health care Professionals” AMIA 1999 annual Seminar. Workshop/ Session W11. http://medicine.ucsd.edu/f99/index.html 18th June 2004**.**

[2] Bamgboye, E.A and Familusi (1990). “Morbidity trends at the children’s emergency room, University College Hospital, Ibadan, Nigeria”. Afr. J. Med. Sci. Vol 19. Pp 49-56

[3] Boashash, B. (ed.), (2003) *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference,* Elsevier Science, Oxford, 2003 ISBN ISBN 0080443354

[4] Berry, M. and Linoff, G. (2000). Data mining Techniques. John Wesley and Sons, Inc. USA: In Bicen, Pelin and Oktay Firat, S.U. (2003). “Knowledge Discovery in databases (KDD) and Data Mining: An application of customer segmentation analysis in banking sector”. Bulletin of the International Statistical Institute, 54th Session, Volume LX, Invited Papers, August 2003, Berlin, , Germany. Pp.136

[5] Biritwum, R. B, Gulaid, J and Amaning, A. O.(2000) “Pattern of diseases or conditions leading to hospitalization at Korle Bu Teaching Hospital, Ghana in 1996”. Ghana Medical Journal , Vol. 34, Number 4, Dec 2000, pp 197 -205

[6] Bloomfield, P. (1976). *Fourier analysis of time series: An introduction*. New York: Wiley.

[7] Connolly, T. and Begg, C. (2005*)* Database systems. *3rd Ed.* Addison-Wesley. Pp. 1239-1240

[8] Carbone, P(2000): Data mart Logo Courtesy (August 2000, Volume 4 No. 2) retrieved on 26th June 2004 www.mitre.org/news/the_edge/august_00/carbone.html

[9] Edelstein, Herbert (1996). "Technology How To: Mining Data Warehouses." Information Week (January 8, 1996): In Laudon, Kenneth C and Laudon, Jane Price (2000): Management Information systems: Organisation and Technology in Networked Enterprise, 6th ed. Prentice-Hall, Upper Saddle River, New Jersey, pp 469-470.

[10] Fong, A.C.M, Hui, S.C., and Jha, G., Data Mining for Decision Support, IEEE IT Professional, 4(2), 9-17, March/April, 2002.

[11] George, C., OLAP, Relational and Multidimensional Database Systems, Acm Sigmod Record, 25(30), 64-69, Sept. 1996.

[12] Inmon, W. H., Welch, J. D., Glassey, L. Katherine (1997). Managing the Data Warehouse: Practical Techniques for monitoring Operations and Performances, Administering Data and tools and managing Change and Growth. John Wiley & sons, Inc. USA. Pp 65-74.

[13] Jayanthi Ranjan, "Applications of Data Mining Techniques in Pharmaceutical Industry", Journal of Theoretical and Applied Information Technology, 2007 JATIT. All rights reserved, www.jatit.org.

[14] Kantor, J (2001): "A Case Study for ROI: With IntelR Processor-Based Data Mining ", In Company newsletter of Clalit Health Services, Israel's second largest health service provider (HSP).

[15] Krippendorf, M and Song, I (2002): The Translation of Star Schema into Entity-Relationship diagrams". College of Information Science and Technology Journal, Drexel University.

[16] Palaniappan, Sellappan, Chua Sook Ling Clinical Decision Support Using OLAP With Data Mining, IJCSNS International Journal of Computer Science and Network Security, 290 VOL.8 No.9, September 2008.

[17] Panos, V., and Timos, S., A Survey on Logical Models for OLAP Databases. ACM Sigmod Record, 28(4), 64-69, Dec.1999.

[18] Ralph, K. and Margy, R., The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling (2nd ed.), Canada: John Wiley & Sons, Inc, 2002.

[19] Robert, S.C., Joseph, A.V. and David, B., Microsoft Data Warehousing: Building Distributed Decision Support Systems, London: Idea Group Publishing, 1999.

[20] Sarwagi, S., Explaining Differences in Multidimensional Aggregate, Proceedings of the 25th International Conference on Very Large Data Bases, Scotland, United Kingdom, 42-53, September, 1999.

[21] Surajit, C., Umeshwar, D., and Ganti, V., Database Technology for Decision Support Systems, IEEE Computer, 34(12), 48-55, Dec. 2001.

[22] Surajit, C. and Umeshwar, D., An Overview of Data Warehousing and OLAP Technology, ACM Sigmod Record,26(1), 65-74, 1997.

[23] Theus, Martin (2003). "A day in the life of a Data Miner". Bulletin of the International Statistical Institute, 54th Session, Volume LX, Invited Papers, August 2003, Berlin, , Germany. Pp.298-301

[24] Usama, M. F., Data Mining and Knowledge Discovery: Making Sense Out of Data, IEEE Expert, 20-25, 1996, October.

[25] Usama F., Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases. Proceedings of the 9th International Conference on Scientific and Statistical Database Management (SSDBM '97), Olympia, WA., 2-11, 1997.

[26] Zaiane, R. O. (1999). CMPUT690 Principles of Knowledge Discovery in Databases lecture Notes. Department of Computing Science, University of Alberta.

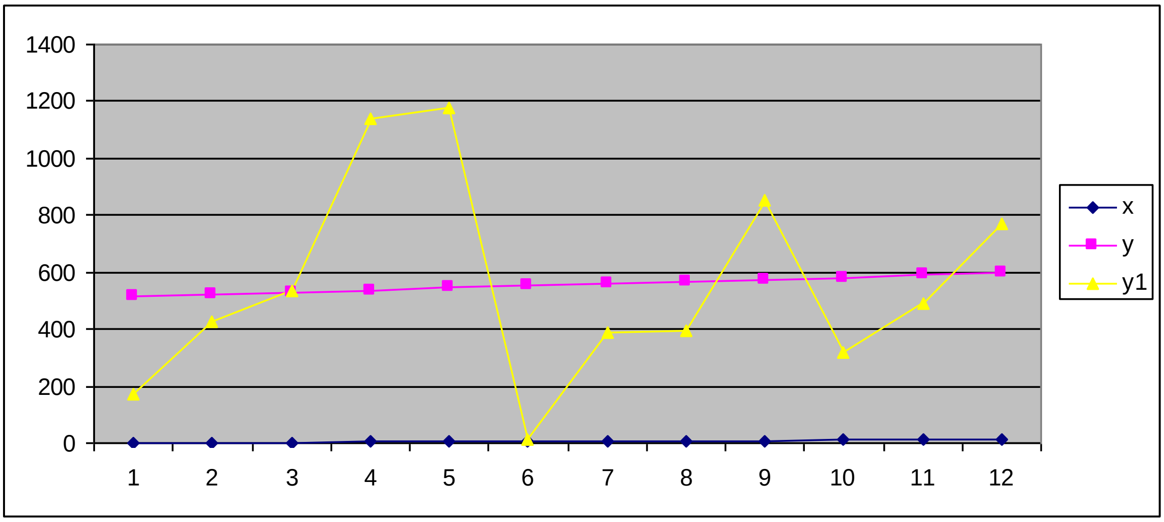

Figure 1: Graph for the number of patient in a year (y1), slope of the line after applying linear regression formula

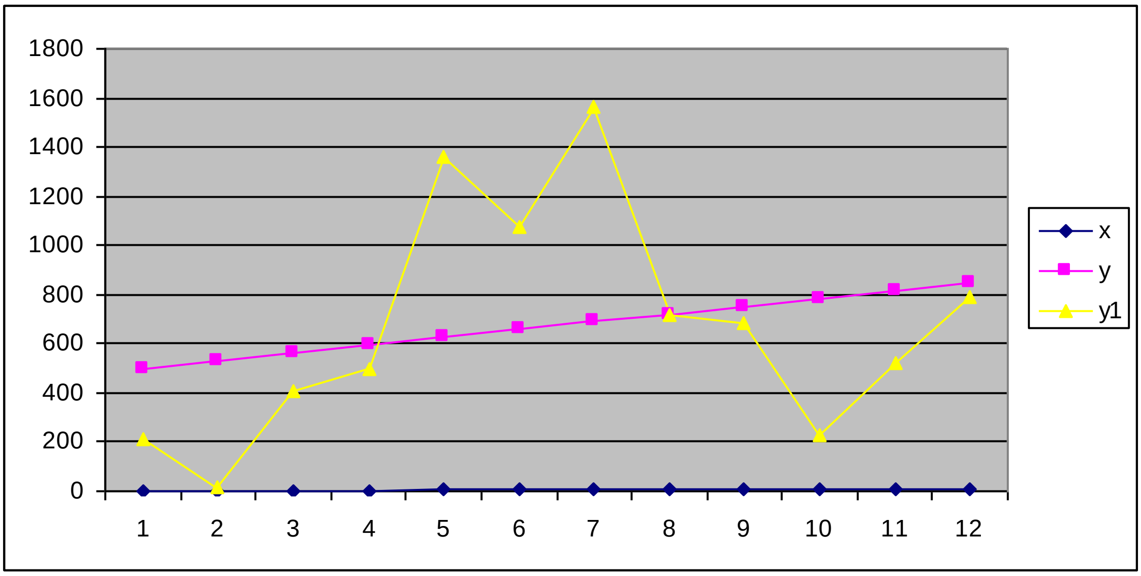

Figure 2: Graph for the number of patient in a year (y2), slope of the line after applying linear regression formula

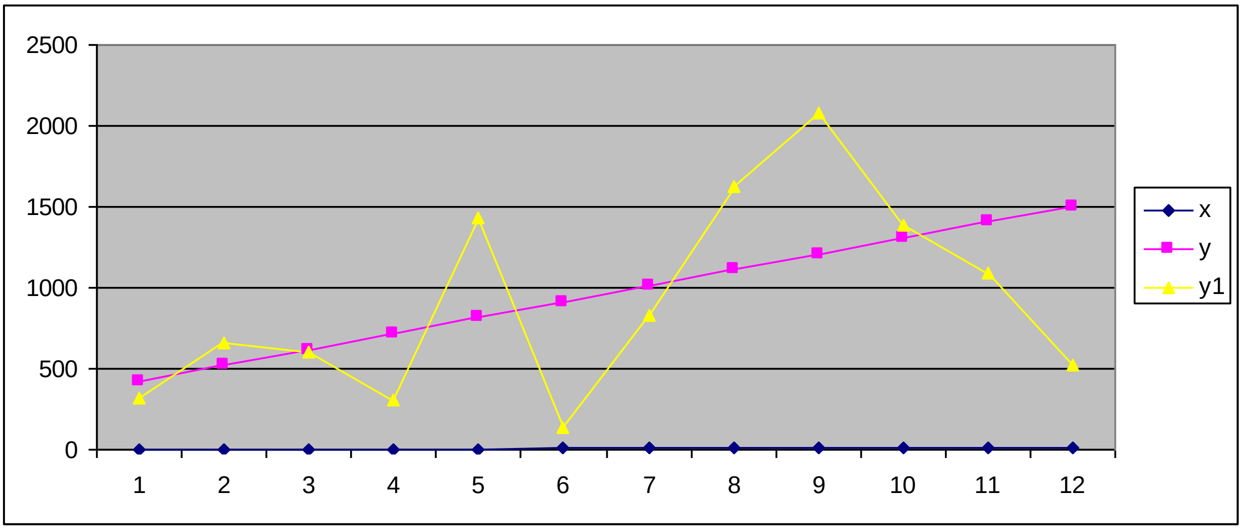

Figure 3: Graph for the number of patient in a year (y3), slope of the line after applying linear regression formula

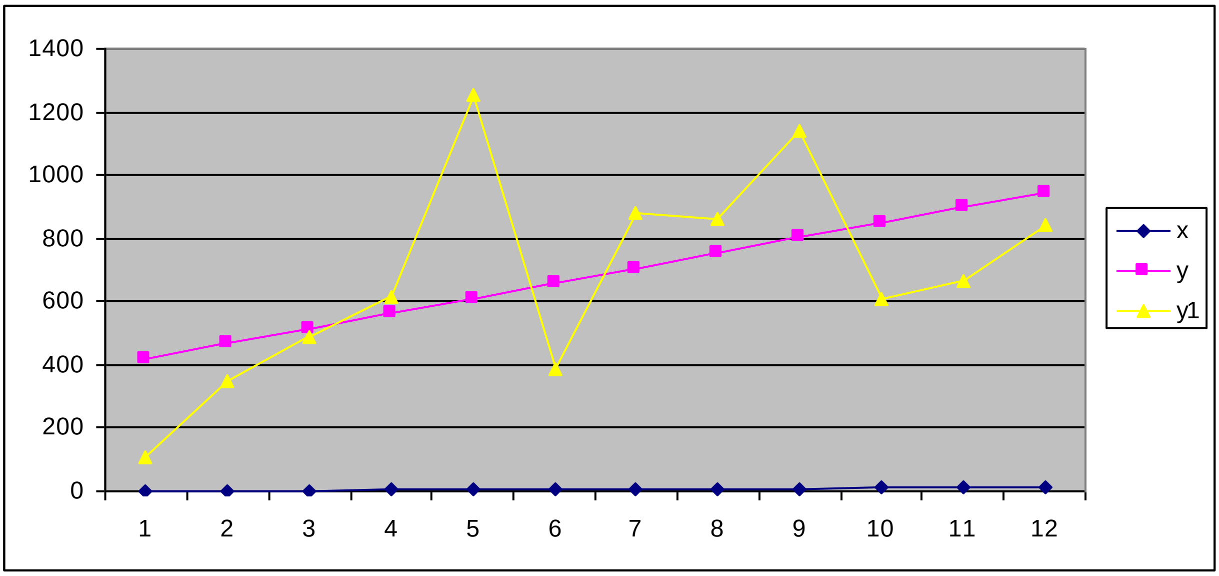

Figure 4: Graph for the number of patient in a year (y4), slope of the line after applying linear regression formula

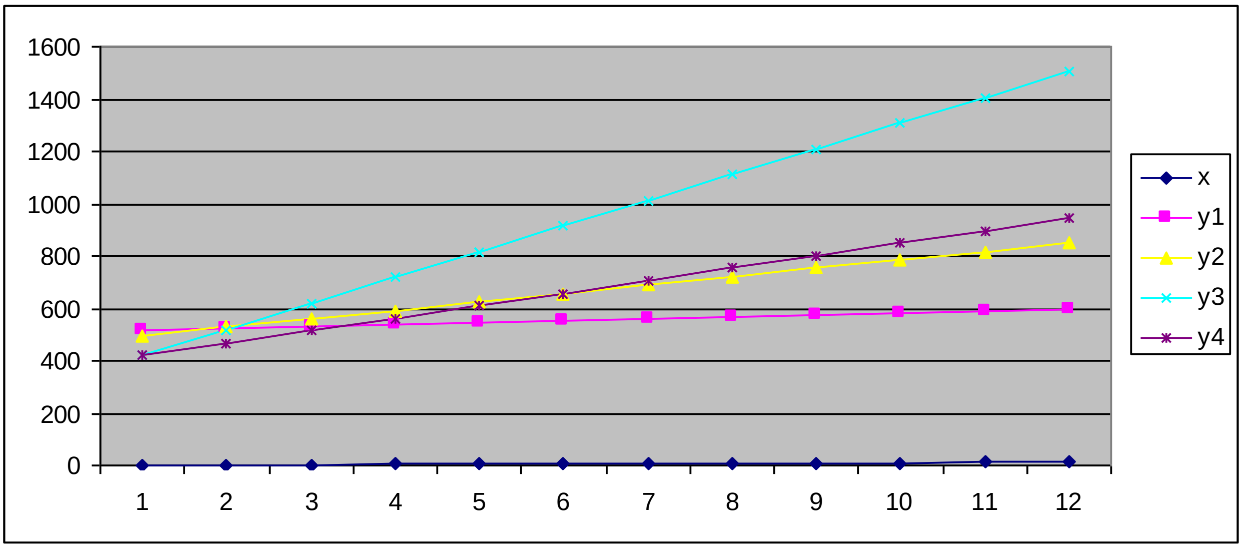

Figure 5: Comparison of slops for four years