Applying Data Mining Technology in Modeling and Predicting Number of Students in Bedia Center

Ola Rayyan
AL-Quds Open University, Palestinian
orayyan@qou.edu

Abstract

In this document we review the concept of Data Mining, and we will show how this technology help in taking decisions base on historical stored data. Data Mining uses several kinds of modeling techniques, and we will focus on Regression Model technique which is used to predict the value of a response (dependent) variable from one or more predictor (independent) variables where the variables are numeric. The forms of regression are linear, multiple, weighted, polynomial, nonparametric and robust.

In this case study we will predict and estimate the number of students will enroll in Bedia Educational Center which is branch of Al-Quds Open University. The estimation is based on the main resource of data which is the historical data for those students were enrolled in Salfeet Educational Region; it is also another branch for Al-Quds Open University. This prediction and estimation based on simple linear regression modeling technique. The reason of using this technique is the steady increasing in the number of students at the university in general and at Salfeet Educational Region in specific. This case study will give the decision makers at Al-Quds Open University view about the number of students in the future, which help them to take the right decision for the situation of Bedia Educational Center. Also they can manage its arrangements with more precise estimations such as revenues, expenses and employment stuff.

***Keywords*:** Data Mining, Regression, Simple Linear Regression

1. Introduction

AL-Quds Open University is distributed university through the Palestine, because it is university for open education. This type of universities required to be spread at all districts of Palestine. The requirements to open new educational branch depends on the number of students will enroll in this new branch. For this reason we need to estimate the number of students based on the previous historical data for another educational region like Salfeet Educational Region using regression model.

The main problems is learning the regression model, spend more time to learn how to use the modeling tool like SPSS. Also exporting the data from the database and transforming it to suitable my work and aggregate it to do the mining analysis for the data.

Regression models are used to predict one variable from one or more other variables. Regression models provide the scientist with a powerful tool, allowing predictions about past, present, or future events to be made with information about past or present events. The scientist employs these models either because it is less expensive in terms of time and/or money to collect the information to make the predictions than to collect the information about the event itself, or, more likely, because the event to be predicted will occur in some future time.

In order to construct a regression model, both the information which is going to be used to make the prediction and the information which is to be predicted must be obtained from a sample of objects or individuals. The relationship between the two pieces of information is then modeled with a linear transformation. Then in the future, only the first information is necessary, and the regression model is used to transform this information into the predicted. In other words, it is necessary to have information on both variables before the model can be constructed.

For example, the personnel officer of the widget manufacturing company might give all applicants a test and predict the number of widgets made per hour on the basis of the test score. In order to create a regression model, the personnel officer would first have to give the test to a sample of applicants and hire all of them. Later, when the number of widgets made per hour had stabilized, the personnel officer could create a prediction model to predict the widget production of future applicants. All future applicants would be given the test and hiring decisions would be based on test performance.

A notational scheme is now necessary to describe the procedure:

Xi is the variable used to predict, and is sometimes called the independent variable. In the case of the widget manufacturing example, it would be the test score.

Yi is the observed value of the predicted variable, and is sometimes called the dependent variable. In the example, it would be the number of widgets produced per hour by that individual.

Y'i is the predicted value of the dependent variable. In the example it would be the predicted number of widgets per hour by that individual.

The goal in the regression procedure is to create a model where the predicted and observed values of the variable to be predicted are as similar as possible. For example, in the widget manufacturing situation, it is desired that the predicted number of widgets made per hour be

as similar to observed values as possible. The more similar these two values, the better the model. The next section presents a method of measuring the similarity of the predicted and observed values of the predicted variable.

In This report we will talk about the background of data mining, and we will also talk about the case study and tools used in our work. We will discuss the result and evaluate these results.

2. Background

2.1 What Motivated Data Mining? Why Is It Important?

The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and science exploration.
Data mining can be viewed as a result of the natural evolution of information technology. An evolutionary path has been witnessed in the database industry in the development of the following functionalities: data collection and database creation, data management (including data storage and retrieval, and database transaction processing), and data analysis and understanding (involving data warehousing and data mining)[1].

2.2 What Is Data Mining?

Simply stated, data mining refers to extracting or "mining" knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. "Knowledge mining," a shorter term may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Thus, such a misnomer that carries both "data" and "mining" became a popular choice. There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging[1]. Many people treat data mining as a synonym for another popularly used term, \Knowledge Discovery in Databases", or KDD. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases. Knowledge discovery as a process is depicted in Figure 1.2, and consists of an iterative sequence of the following steps:

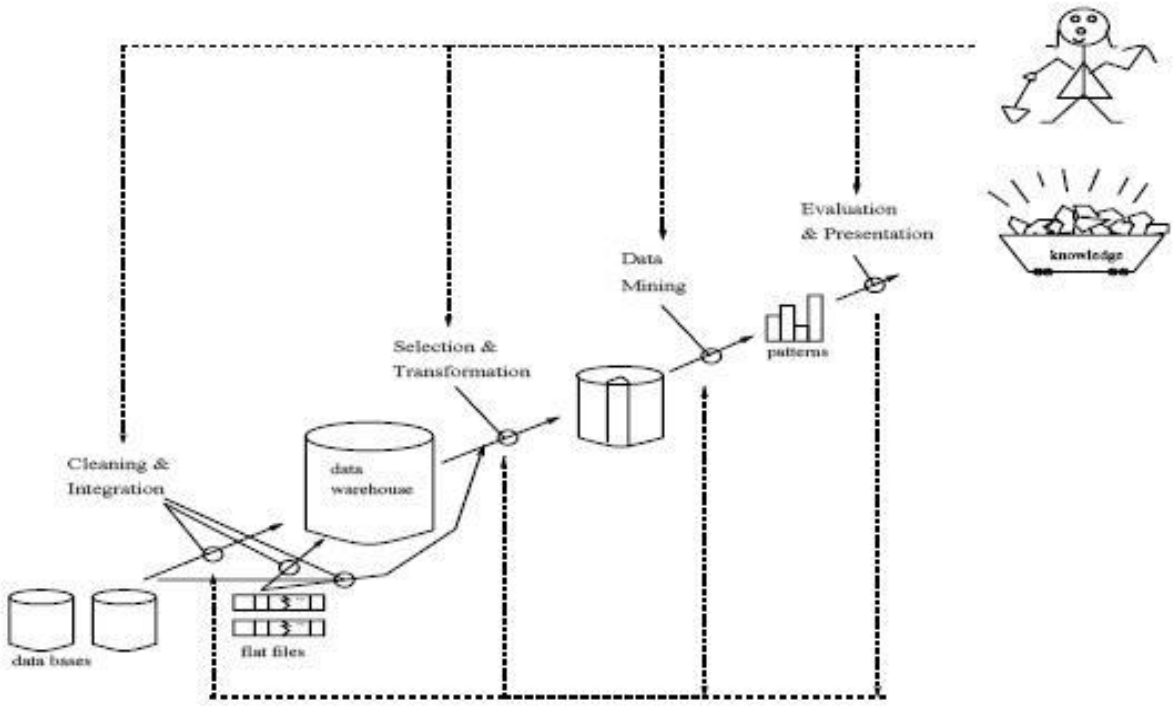


Figure 1.2: Data mining as a process of knowledge discovery.

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources maybe combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)
6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some **interestingness measures.**
7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

2.3 Data Mining-On What Kind of Data?

In this section, we examine a number of different data stores on which mining can be performed. In principle, data mining should be applicable to any kind of information repository. This includes relational databases, data warehouses, transactional databases, advanced database systems, flat files, and the World Wide Web. Advanced database systems include object-oriented and object-relational databases, and specific application- oriented databases, such as spatial databases, time-series databases, text databases, and multimedia databases. The challenges and techniques of mining may differ for each of the repository systems.

2.3.1 Relational Databases

A database system, also called a database management system (DBMS), consists of a collection of interrelated data, known as a database, and a set of software programs to manage and access the data. The software programs involve mechanisms for the definition of database structures; for data storage; for concurrent, shared, or distributed data access; and for ensuring the consistency and security of the information stored, despite system crashes or attempts at unauthorized access[2].

2.3.2 Data Warehouses

Data warehouse is a repository of information collected from multiple sources, stored under a unified schema, and which usually resides at a single site. Data warehouses are constructed via a process of data cleaning, data transformation, data integration, data loading, and periodic data refreshing.

2.3.3 Transactional Databases

In general, a transactional database consists of a file where each record represents a transaction. A transaction typically includes a unique transaction identity number (Trans ID), and a list of the items making up the transaction (such as items purchased in a store). The transactional database may have additional tables associated with it, which contain other information regarding the sale, such as the date of the transaction, the customer ID number, the ID number of the sales person and of the branch at which the sale occurred, and so on.

2.3.4 Advanced Database Systems and Advanced Database Applications

Relational database systems have been widely used in business applications. With the advances of database technology, various kinds of advanced database systems have emerged and are undergoing development to address the requirements of new database applications.
The new database applications include handling spatial data (such as maps), engineering design data (such as the design of buildings, system components, or integrated circuits), hypertext and multimedia data (including text, image, video, and audio data), time-related data (such as historical records or stock exchange data), and the World Wide Web (a huge, widely distributed information repository made available by the Internet). These applications require efficient data structures and scalable methods for handling complex object structures, variable-length records, semi structured or unstructured data, text and multimedia data, and database schemas with complex structures and dynamic changes[3].

2.4 Data Mining Functionalities-What Kinds of Patterns Can Be Mined?

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and

predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions[4].

2.4.1 Concept/Class Description: Characterization and Discrimination

Data can be associated with classes or concepts. These descriptions can be derived via (1) data characterization, by summarizing the data of the class under study (often called the target class) in general terms, or (2) data discrimination, by comparison of the target class with one or a set of comparative classes (often called the contrasting classes), or (3) both data characterization and discrimination.

Data characterization is a summarization of the general characteristics or features of a target class of data. The data corresponding to the user-specified class are typically collected by a database query.

Data discrimination is a comparison of the general features of target class data objects with the general features of objects from one or a set of contrasting classes. The target and contrasting classes can be specified by the user, and the corresponding data objects retrieved through database queries.

2.4.2 Association Analysis

"What is association analysis?" Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis.

2.4.3 Classification and Prediction

Classification is the process of finding a set of models (or functions) that describe and distinguish data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. The derived model is based on the analysis of a set of training data (i.e., data objects whose class label is known).
"How is the derived model presented?" The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. Decision trees can be easily converted to classification rules. A neural network, when used for classification, is typically a collection of neuron-like processing units with weighted connections between the units.

2.4.4 Cluster Analysis

"What is cluster analysis?" Unlike classification and prediction, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not

known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the interclass similarity and minimizing the interclass similarity. That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. Clustering can also facilitate taxonomy formation, that is, the organization of observations into a hierarchy of classes that group similar events together.

2.4.5 Outlier Analysis

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. However, in some applications such as fraud detection, the rare events can be more interesting than the more regularly occurring ones. The analysis of outlier data is referred to as outlier mining.

2.4.6 Evolution Analysis

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct features of such an analysis include time-series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

3. Case Study

3.1 Prediction number of student's case study

In this case study we will try to find a model to estimate the number of students for Bedia Educational Center based on the number of students on at Salfeet Educational Region those from Bedia region. We will use the simple linear regression in this case study.

3.2 Linear Regression and Prediction

Linear regression uses the relationship between distributions of scores in making predictions. If there is a relationship between two distributions, it is possible to predict a person's score in one distribution on the basis of their score in the other distribution (e.g., using a score on an aptitude test to predict actual job performance). Simple regression refers to the situation where there are only two distributions of scores, X and Y. By convention, X is the predictor variable, and Y the criterion (or predicted) variable.

3.2 Definitions

a) A **scatterplot** is a graph of paired X and Y values
b) A **linear relationship** is one in which the relationship between X and Y can best be represented by a straight line.

c) A **curvilinear relationship** is one in which the relationship between X and Y can best be represented by a curved line.
d) A **perfect relationship** exists when all of the points in the scatter plot fall exactly on the line (or curve). An **imperfect relationship** is one in which there is a relationship, but not all points fall on the line (or curve).
e) A **positive relationship** exists when Y increases as X increases (i.e., when the slope is positive).
f) A **negative relationship** exists when Y decreases as X increases (i.e., when the slope is negative).

3.3 Equation for a straight line

The equation for a straight line is usually written as:

$$Y=bX+a$$

where b = slope of the line
$= (Y2 - Y1) / (X2 - X1)$
= "the rise" divided by "the run"

and

a = the Y-intercept
= the value of Y when $X = 0$

Perhaps an example will help to clarify what this means. Imagine that you have decided to start working out at a gym. The annual membership fee is £25, and in addition to that, you must pay £2 every time you go to the gym.1 If we let X = the number of times you go to the gym, and Y = the total cost, we would find that:

$$Y=2X+25$$

The Y-intercept is 25. That is, if you never go to the gym (X = 0), your total cost is £25. And the slope of the line (b) is 2: Every time you go to the gym, it costs you another £2. Putting this another way, every time there is an increase of 1 on the X-axis, there is an increase of 2 on the Y axis.

3.4 Miscellaneous points about linear regression

Linear regression is used to predict a Y score from a score on X. Bear in mind the following:
1) The relationship between X and Y must be **linear**. If the relationship is not linear, prediction will not be very accurate.

2) Normally, we are not interested in predicting Y scores that are already known. We derive our regression equation with sample data that consists of paired X and Y scores, but use the equation to predict Y scores when only X values are given. Because we use data collected from a sample to make these predictions, it is vital to have a **representative** sample when deriving a regression equation.

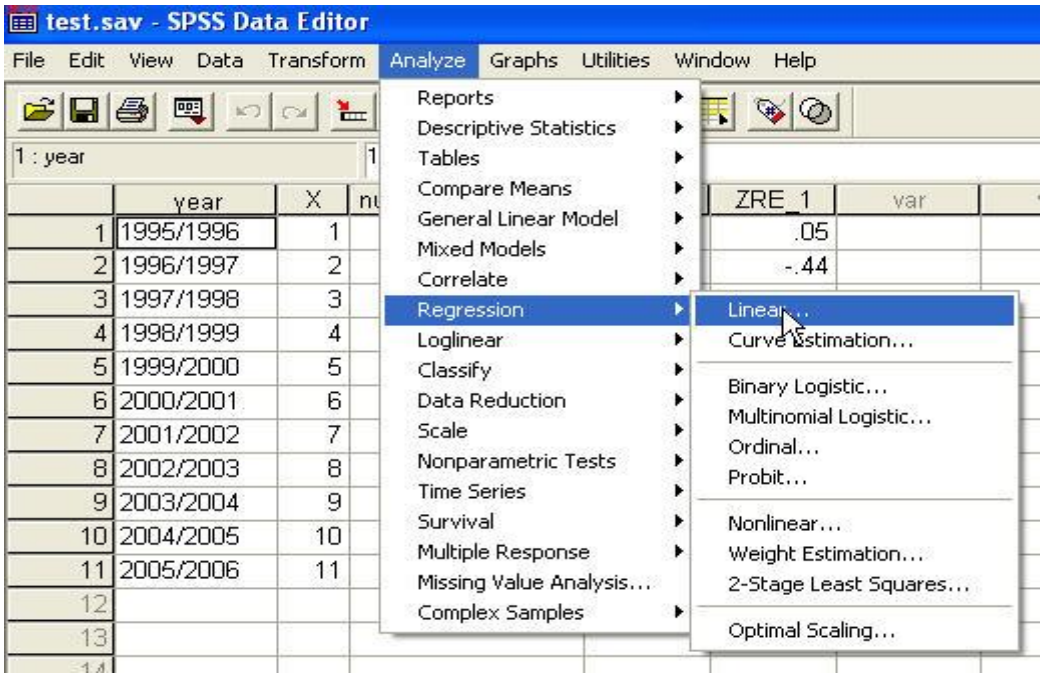
3) A regression equation is properly used only for the range of the variables on which it was based. We do not know whether the relationship between X and Y continues to be linear beyond the range of sample values.

4) Prediction is most accurate if the data have the property of **homoscedasticity** i.e., if the variability of the Y scores is constant at all points along the regression line.

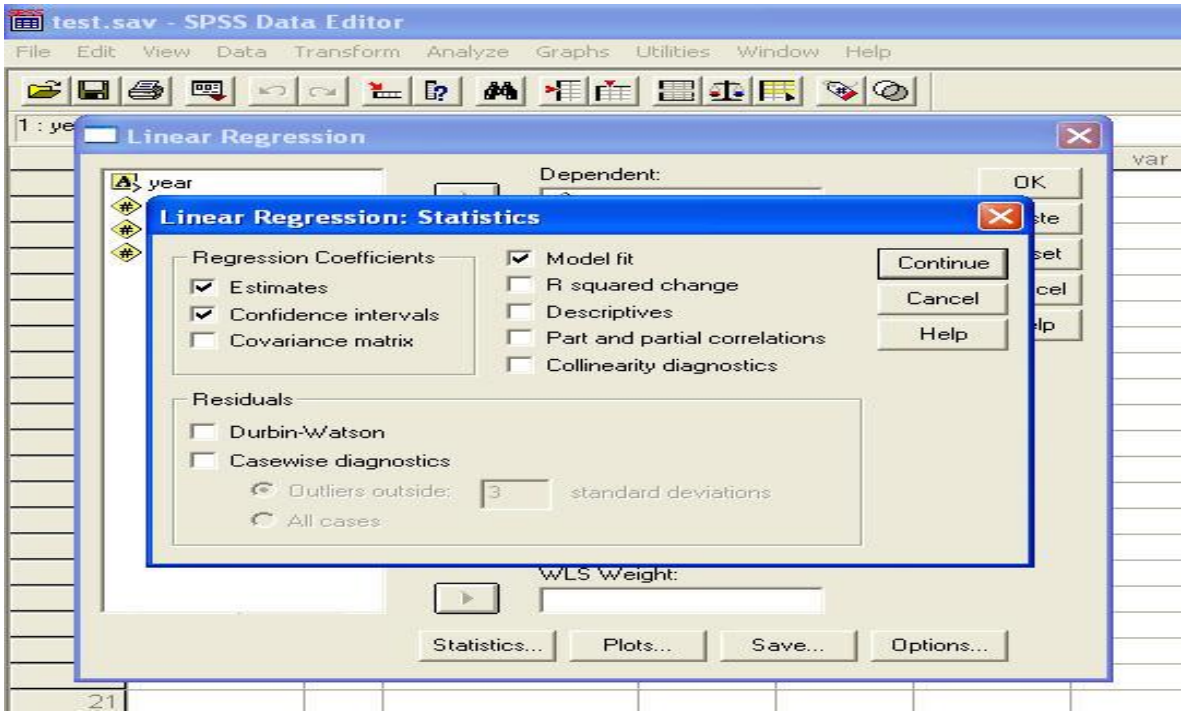
5) When X and Y are both normally distributed and the number of paired scores is large, the data in a bivariate frequency distribution often produce a so-called **bivariate normal** distribution. When you have such a distribution, the **standard error of estimate** can be used in the same way we used the standard deviation of a normal distribution. That is, we could say that about 68% of the scores in the scatterplot fall within 1 standard error of the regression line; and about 95% of the scores fall within 2 standard errors of the regression line.

3.5 Regression Analysis Using SPSS

The REGRESSION command is called in SPSS as follows:



Selecting the following options will command the program to do a simple linear regression and create two new variables in the data editor: one with the predicted values of Y and the other with the residuals.



The output from the preceding includes the correlation coefficient and standard error of estimate.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.984[a]	.969	.965	41.639

a. Predictors: (Constant), X

b. Dependent Variable: numberOfStudents

The regression coefficients are also given in the output.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	383.714	26.927		14.250	.000	322.801	444.627
	X	66.286	3.970	.984	16.696	.000	57.305	75.267

a. Dependent Variable: numberOfStudents

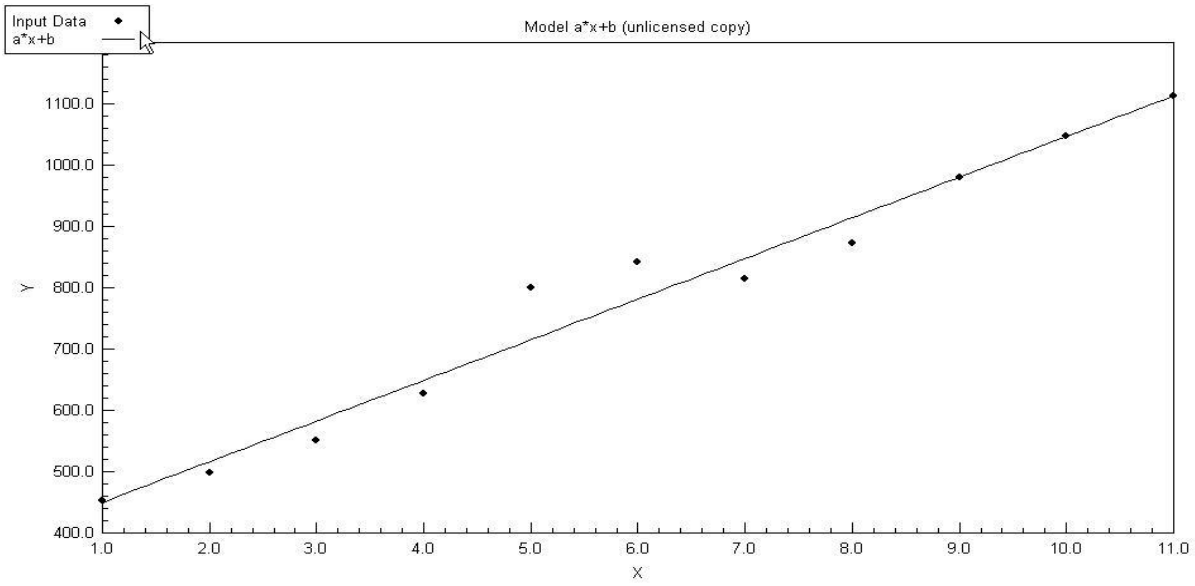
The optional save command generates two new variables in the data file.

ile Edit View Data Transform Analyze Graphs Utilities Window Help

: numberOfStudents | 800

	year	X	numberOfStudents	PRE_1	ZRE_1	var
1	1995/1996	1	452	450.00	.05	
2	1996/1997	2	498	516.29	-.44	
3	1997/1998	3	550	582.57	-.78	
4	1998/1999	4	628	648.86	-.50	
5	1999/2000	5	800	715.14	2.04	
6	2000/2001	6	842	781.43	1.45	
7	2001/2002	7	814	847.71	-.81	
8	2002/2003	8	872	914.00	-1.01	
9	2003/2004	9	980	980.29	.00	
10	2004/2005	10	1047	1046.57	.00	
11	2005/2006	11	1113	1112.86	.00	
12						
13						
14						
15						

Also this is the graph to best fit errors



4. Software Tools

4.1 SPSS

SPSS for Windows provides a powerful statistical analysis and data management system in a graphical environment, using descriptive menus and simple dialog boxes to do most of the work for you. Most tasks can be accomplished simply by pointing and clicking the mouse. In addition to the simple point-and-click interface for statistical analysis, SPSS for Windows provides:

Data Editor: A versatile spreadsheet-like system for defining, entering, editing, and displaying data.

Viewer: The Viewer makes it easy to browse your results, selectively show and hide output, change the display order results, and move presentation-quality tables and charts between SPSS and other applications.

Multidimensional pivot tables: Your results come alive with multidimensional pivot tables. Explore your tables by rearranging rows, columns, and layers. Uncover important findings that can get lost in standard reports. Compare groups easily by splitting your table so that only one group is displayed at a time.

High-resolution graphics: High-resolution, full-color pie charts, bar charts, histograms, scatterplots, 3-D graphics, and more are included as standard features in SPSS.
Database access: Retrieve information from databases by using the Database Wizard instead of complicated SQL queries.

Data transformations: Transformation features help get your data ready for analysis. You can easily subset data, combine categories, add, aggregate, merge, split, and transpose files, and more.

Electronic distribution: Send e-mail reports to others with the click of a button, or export tables and charts in HTML format for Internet and intranet distribution.

Online Help: Detailed tutorials provide a comprehensive overview; context-sensitive Help topics in dialog boxes guide you through specific tasks; pop-up definitions in pivot table results explain statistical terms; the Statistics Coach helps you find the procedures that you need; and Case Studies provide hands-on examples of how to use statistical procedures and interpret the results.

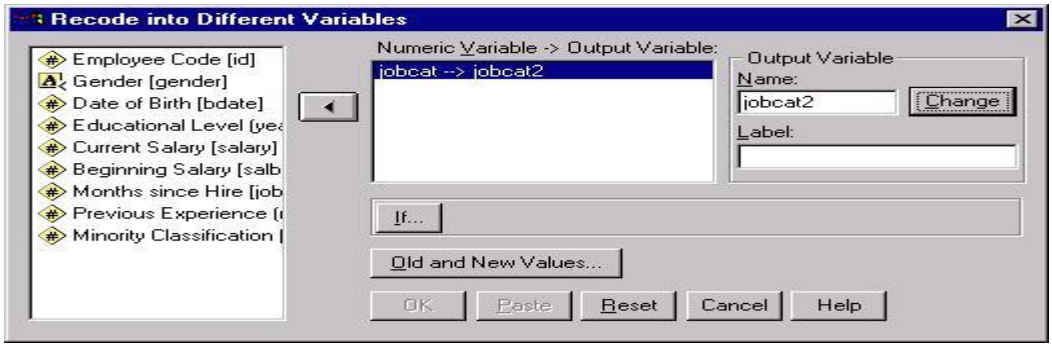
Command language: Although most tasks can be accomplished with simple point-and-click gestures, SPSS also provides a powerful command language that allows you to save and automate many common tasks. The command language also provides some functionality not found in the menus and dialog boxes.

4.2 Recoding Variables

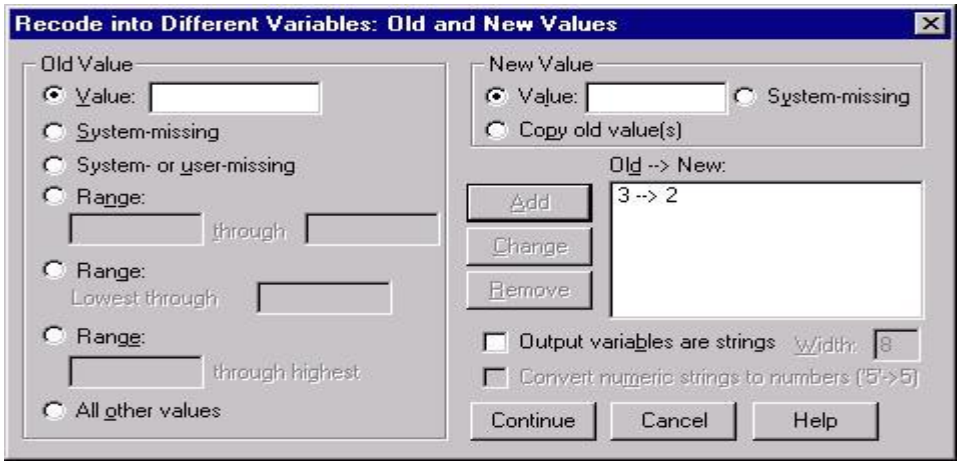
You can also modify the values of existing variables in your dataset. For example, if a dataset contains a variable that classifies an employee's status in three categories, but for a particular analysis you want to combine two of these classifications into a single category, then two of the values would need to be recoded into a single value so that there are two total groups.
The Recode option (or Alt+T+R) is available from the menu in the Data Editor:

Transform
Recode

Additionally, there are two options for recoding variables in the Recode submenu. The Into Same Variables (Alt+T+R+S) option changes the values of the existing variables, whereas the Into Different Variables (Alt+T+R+D) option is used to create a new variable with the recoded values. Both options are essentially the same, except that recoding into a different variable requires you to supply a new variable name. You should use the Into Different Variables option, because you may change your mind about your recoding scheme at a later date. Thus, if your do change you mind, you still have the original values.
The following example illustrates the use of the Recode option to recode values into a new variable. When that option is selected from the menu, the following dialog box will appear:



First, a variable from the existing dataset should be selected by clicking on that variable, then clicking the arrow button in the middle of the dialog box. This will result in the selected variable being displayed in the box labeled, Numeric Variable -> Output Variable. Next, you must supply the name of the new variable, and optionally you can supply a label for the new variable. After a new variable name has been supplied, click on the button labeled Old and new Values. This will result in the following dialog box:



The above dialog box is the same regardless of whether you are recoding values into the same variable or creating a new variable. The original value of the variable being recoded is entered in the box labeled Old Value, and the new value is entered in the box labeled New Value. After values are entered in these boxes, click on the button labeled Add to complete the recode process.
Continuing with the above example, a variable with three values, such as jobcat, could be recoded into a variable with two values by recoding one of the values. In the example dataset, jobcat has three values: 1, 2, and 3. If the goal were to combine cases with the values 2 and 3, this could be accomplished by recoding cases with the value 3 into 2's. For example, by

entering 3 in the box labeled Old Value and entering 2 in the box labeled New Value then clicking Add, all of the cases labeled 3 would take on the value 2. This can be repeated for as many of the values as necessary.
Values can also be recoded conditionally. The process for recoding values on the basis of a condition is essentially identical to the process for conditionally computing new variables discussed in the previous section: when you click on the If button in the main Recode dialog box, the same dialog box that was obtained from clicking If in the the Compute dialog box will appear with the same options.

4.3 Sorting Cases

Sorting cases allows you to organize rows of data in ascending or descending order on the basis of one or more variable. For example, the data could be sorted by job category so that all of the cases coded as job category 1 appear first in the dataset, followed by all of the cases that are labeled 2 and 3 respectively. The data could also be sorted by more than one variable. For example, within job category, cases could be listed in order of their salary. The Sort Cases (or Alt+ D+O) option is available under the Data menu item in the Data Editor:

Data → Sort Cases...
The dialog box that results from selecting Sort Cases presents only a few options:

To choose whether the data are sorted in ascending or descending order, select the appropriate button. You must also specify on which variables the data are to be sorted. The hierarchy of such a sorting is determined by the order in which variables are entered in the Sort by box. Variables are sorted by the first variable entered, then the next variable is sorted within that first variable. For example, if jobcat was the first variable entered, followed by salary, the data would first be sorted by jobcat, then, within each of the job categories, data would be sorted by salary.

5. Conclusion

Regression models are powerful tools for predicting a score based on some other score. They involve a linear transformation of the predictor variable into the predicted variable. The parameters of the linear transformation are selected such that the least squares criterion is met, resulting in an "optimal" model. The model can then be used in the future to predict either exact scores, called point estimates, or intervals of scores, called interval estimates.

In this case study we find the optimal model in this formula:

$$Y^{\wedge} = 383.714 + 66.286 * (X)$$

Where Y^ is the predicted number of students
And X is the year we want to predict.

This model has only on dependent variable which is number of students in Salfeet Educational Region. Also we can develop this model to take multiple dependent variables like number of student in tawjihi in Salfeet Region and the population increasing percentage, but this required additional historical data to be built.

6. References

1. J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

2. T. Imielinski and H. Mannila. A database perspective on knowledge discovery. Communications of ACM, 39:58-64, 1996.

3. G. Piatetsky-Shapiro, U. Fayyad, and P. Smith. From data mining to knowledge discovery: An overview. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.

4. G. Piatetsky-Shapiro and W. J. Frawley. Knowledge Discovery in Databases. AAAI/MIT Press, 1991.