SIMPLE LINEAR REGRESSION

by (Student’s Name)

Course Name

Professor’s Name

University Name

City, State

Date

Simple Linear Regression

Introduction

The use of linear regression is to predict a trend in data, or predict the value of a variable (dependent) from the value of another variable (independent), by fitting a straight line through the data. Linear regression represents a connecting link between the independent (carrier) variable and dependent (response) variable, which if graphed on X and Y-coordinates, results in a straight line. Linear regression shows the straight line which thoroughly represents, or predicts, the value of the response variable, given the noted value of the carrier variable (Frey 2006). This essay aims at reviewing the article introduction to simple linear regression by Dallal (2000).

Problem Statement

Dallal (2000) assumed a relationship between body mass (independent or carrier variable) and muscle strength (dependent or response variable), the more body mass the more muscle strength. However, this relationship is not without exceptions, which is reflected on the scatter plot of a regression model. Therefore, the author posed the question of how to illustrate the straight line, which accurately portrays the data, or predicts the value of the response variable.

Research Purpose Statement

In the given example, most cases would show a perfect regression. However, standardization of the procedure of putting in a straight line is necessary to provide better communication and common grounds for analysts working on the same data. Further, in the example regression equation given (Strength = -13.971 + 3.016 LBM [Lean Body mass]), one can draw two conclusions. First, a predicted muscle strength equals LBM multiplied by 3.016 minus 13.971. Second, the difference between muscle strength of two individuals is presumably 3.016 multiplied by the difference in their LBM.

Research Questions

Research question 1: Why do we need to fit a regression equation into a set of data?

It is clear from the previous example that there are reasons for fitting a regression equation into a set of data. These are: to describe the data and to predict an independent (response) variable from a dependent (carrier) one.

Research question 2: What is the underlying principle of calculating a straight line?

If the points signaling data in a scatter plot are close to a line, it means the line represents, matches, or gives a good fit of data. If not, then the line with most of the points closer to it than any other is the one that gives good fit of data.

Research question 3: How is linear regression (least squares) equation used to illustrate the best fitting line?

The standard used, as the name implies, is the sum of squared residuals (observed – predicted values) is minimal for the best fitting line. This applies to a line fitted to a set of sample data to promote generalization to a population from which this sample was taken. Yet for a population, there is a slightly different linear regression equation. The equation illustrates that an output (dependent) variable on the Y-axis can be predicted from an input (independent) variables on the X-axis after adding a random error.

Research question 4: Is the sample regression equation an accurate estimate of the population regression equation?

There is a reservation for accreditation of this statement, which is directed at the confidence bands in relation to the regression line. They are understood as the standard error of the mean (the standard deviation of the mean of the sampling distribution), yet with one exception that is the sampling mean of the dependent variables amplifies as it adds distance from the mean.

Sources of Data

Dallal (2000) stated that in the second part of his article (linked to the main article) cross- sectional data are used. This type of data has the advantages of being used if sampling method is not weighted and-or un-stratified. This method can also be used if the researcher is concerned only with minor or small probabilities. The longitudinal data results in more statistical power, however, in repeated cross-sectional analysis, new subjects added per analysis compensate for the inherent decreased statistical power (Yee and Niemeier 1996).

Data Collection Strategies and Methods

A good data collection strategy should have two objectives. The first is having motivated respondents affected by time consuming, trust in statistics, difficulty of questionnaire, and benefit included. The second objective should be having high quality data, which tailored to sample individuals, sampling method, and good instruments of data collection (Statistics Norway 2007).

There are many methods of data collection, and selection of a particular method depends on the available resources, reliability, resources of analysis and reporting, besides the skills and knowledge of the analyst. Some of these methods are case studies, behavior observation check lists, attitude and opinion surveys, questionnaires distributed by mail, e-mails, or phone calls. Other methods of data collection include time series (evaluating one variable over a period of time as a week) and individual or group interviews (The Ohio State University 2005).

Conclusion

Dallal (2000) inferred that simple linear regression means that we can predict a dependent variable from an independent one, thus, whenever we need to know from the beginning each time we add information. The regression line is important as it makes the estimation of a dependent variable more accurate and allows the estimation of a response variable for individuals with values of the carrier variable not included in the data. The author also inferred that there are two methods of predicting a variable either from within the range of values of independent variable of the sample given (interpolation) or outside this range (extrapolation). The author recommended the first method as it has the advantage of being safe, yet with concerns as regards the way to demonstrate the linearity of relationship between the two variables.

Bibliography

Dallal, G. (2000). Introduction to simple linear regression. [Online] Available from: <http://www.tufts.edu/~gdallal/slr.htm>. [Accessed: 14th January 2008].

Frey, B. (2006). *Statistics Hacks*. Sebastopol, CA: O'Reilly Media Inc.

Statistics Norway. (2007). *Strategy for Data Collection.* [Online] Available from: <http://www.ssb.no/vis/english/about_ssb/strategy/strategy_data_collection.pdf>. [Accessed: 7th February 2008].

The Ohio State University. (2005). *Bulletin Extension - Step Four: Methods of Data Collection.* [Online] Available from: <http://www.ohioline.ag.ohio-state.edu>. [Accessed: 7th February 2008].

Yee J. L. & Niemeier D. (1996). Advantages and Disadvantages: Longitudinal vs. Repeated Cross-Section Survey - A Discussion Paper. *Project Battelle*. 94. p.16-22.