Chapter 06.02
Introduction of Regression Analysis

After reading this chapter, you should be able to:

1. *know what regression analysis is,*
2. *know the effective use of regression, and*
3. *enumerate uses and abuses of regression.*

What is regression analysis?

Regression analysis gives information on the relationship between a response (dependent) variable and one or more (predictor) independent variables to the extent that information is contained in the data. The goal of regression analysis is to express the response variable as a function of the predictor variables. The duality of fit and the accuracy of conclusion depend on the data used. Hence non-representative or improperly compiled data result in poor fits and conclusions. Thus, for effective use of regression analysis one must

1. investigate the data collection process,
2. discover any limitations in data collected, and
3. restrict conclusions accordingly.

Once a regression analysis relationship is obtained, it can be used to predict values of the response variable, identify variables that most affect the response, or verify hypothesized causal models of the response. The value of each predictor variable can be assessed through statistical tests on the estimated coefficients (multipliers) of the predictor variables.

An example of a regression model is the linear regression model which is a linear relationship between response variable, y and the predictor variable, $x_i, i = 1,2...,n$ of the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon \qquad (1)$$

where

$\beta_0, \beta_1.......\beta_n$ are regression coefficients (unknown model parameters), and

ε is the error due to variability in the observed responses.

Example 1

In the transformation of raw or uncooked potato to cooked potato, heat is applied for some specific tune. One might postulate that the amount of untransformed portion of the

starch (y) inside the potato is a linear function of time (t) and temperature (θ) of cooking. This is represented as

$$y = \beta_0 + \beta_1 t + \beta_2 \theta + \varepsilon \tag{2}$$

Linear as used in linear regression refers to the form of occurrence of the unknown parameters, β_1 and β_2 as ,simple linear multipliers of the predictor variable. Thus, the two equations below are also both linear.

$$y = \beta_0 + \beta_1 t + \beta_2 t\theta + \beta_3 \theta + \varepsilon \tag{3}$$

$$y = \beta_0 + \beta_1 t\theta + \beta_2 \theta + \varepsilon \tag{4}$$

Comparison of Regression and Correlation

Unlike regression, correlation analysis assesses the simultaneous variability of a collection of variables. The relationship is not directional and interest is not on how some variables respond to others but on how they are mutually associated. Thus, simultaneous variability of a collection of variables is referred to as correlation analysis.

Uses of Regression Analysis

Three uses for regression analysis are for
1. prediction
2. model specification and
3. parameter estimation.
   Regression analysis equations are designed only to make predictions. Good predictions will not be possible if the model is not correctly specified and accuracy of the parameter not ensured. However, accurate prediction and model specification require that all relevant variables be accounted for in the data and the prediction equation be defined in the correct functional form for all predictor variables.
   Parameter estimation is the most difficult to perform because not only is the model required to be correctly specified, the prediction must also be accurate and the data should allow for good estimation. For example, multicolinearity creates a problem and requires that some estimators may not be used. Thus, limitations of data and inability to measure all predictor variables relevant in a study restrict the use of prediction equations.

Abuses of Regression Analysis

Let us examine three common abuses of regression analysis.
1. Extrapolation
2. Generalization
3. Causation

Extrapolation

If you were dealing in the stock market or even interested in it, then you might remember the stock market crash of March 2000. During 1997-1999, investors thought they would double their money every year. They started buying fancy cars and houses on credit,

and living the high life. Little did they know that the whole market was hyped on speculation and little economic sense. The Enron and MCI financial fiascos soon followed.

Let us look if we could have safely extrapolated the NASDAQ index[1] from past years. Below is the table of NASDAQ index, S, as a function of end of year number, t (Year 1 is the end of year 1994, and Year 6 is the end of year 1999).

Table 1 NASDAQ index as a function of year number.

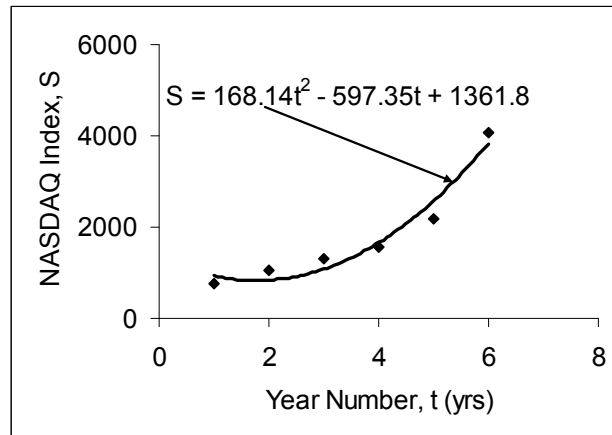| Year Number (t)	NASDAQ Index (S)
1 (1994)	752
2 (1995)	1052
3 (1996)	1291
4 (1997)	1570
5 (1998)	2193
6 (1999)	4069

Figure 1 The regression line of NASDAQ Index as a function of year number.

A relationship $S = a_0 + a_1 t + a_2 t^2$ between the NASDAQ index, S, and the year number, t, is developed using least square regression and is found to be $S = 168.14t^2 - 597.37t + 1361.8$. The data and the regression line are shown in Figure 1. The data is given only for Years 1 through 6 and it is desired to calculate the value for $t > 6$. This is extrapolation outside the model data. The error inherent in this model is shown in Table 2 and Figure 2. Look at the

---

[1] NASDAQ (National Association of Securities Dealers Automated Quotations) index is a composite index based on the stock market value of 3,000 companies. The NASDAQ index began on February 5, 1971 with a base value of 100. Twenty years later in 1995, NASDAQ index crossed the 1000 mark. It rose as high as 5132 on March 10, 2000 and currently is at a value of 2282 (February 19, 2006).

Year 7 and 8 that was not included in the data – the error between the predicted and actual values is 119% and 277%, respectively.

Table 2 NASDAQ index as a function of year number.

Year Number (t)	NASDAQ Index (S)	Predicted Index	Absolute Relative True Error (%)
1 (1994)	752	933	24
2 (1995)	1052	840	20
3 (1996)	1291	1083	16
4 (1997)	1570	1663	6
5 (1998)	2193	2579	18
6 (1999)	4069	3831	6
7 (2000)	2471	5419	119
8 (2001)	1951	7344	276

This illustration is not exaggerated and it is important that a careful use of any given model equations is always employed. At all times, it is imperative to infer the domain of independent variables for which a given equation is valid.
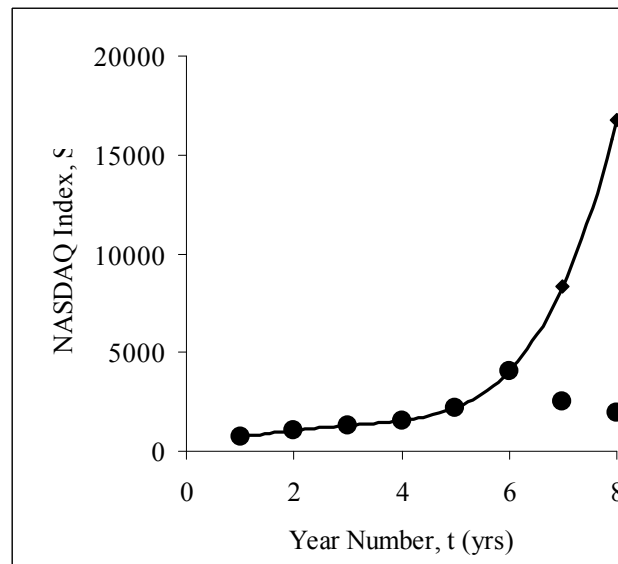

Figure 2 Extrapolated curve and actual data for Years 7 and 8.

Generalization

Generalization could arise when unsupported or over exaggerated claims are made. It is not often possible to measure all predictor variables relevant in a study. For example, a study carried out about the behavior of men might have inadvertently restricted the survey to Caucasian men only. Shall we then generalize the result as the attributes of all men irrespective of race? Such use of regression equation is an abuse since the limitations imposed by the data restrict the use of the prediction equations to Caucasian men.

Misidentification

Finally, misidentification of causation is a classic abuse of regression analysis equations. Regression analysis can only aid in the confirmation or refutation of a causal model - the model must however have a theoretical basis. In a chemical reacting system in which two species react to form a product, the amount of product formed or amount of reacting species vary with time. Although a regression equation of species concentration and time can be obtained, one cannot attribute time as the causal agent for the varying species concentration. Regression analysis cannot prove causality, rather it can only substantiate or contradict causal assumptions. Anything outside this is an abuse of regression analysis method.

Least Squares Methods

This is the most popular method of parameter estimation for coefficients of regression models. It has well known probability distributions and gives unbiased estimators of regression parameters with the smallest variance.

We wish to predict the response to n data points $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ by a regression model given by

$$y = f(x) \tag{6}$$

where, the function $f(x)$ has regression constants that need to be estimated.

For example

$f(x) = a_0 + a_1 x$ is a straight-line regression model with constants a_0 and a_1

$f(x) = a_0 e^{a_1 x}$ is an exponential model with constants a_0 and a_1

$f(x) = a_0 + a_1 x + a_2 x^2$ is a quadratic model with constants a_0, a_1 and a_2

A measure of goodness of fit, that is how the regression model $f(x)$ predicts the response variable y is the magnitude of the residual, E_i at each of the n data points.

$$E_i = y_i - f(x_i), \ i = 1, 2, ... n \tag{7}$$

Ideally, if all the residuals E_i are zero, one may have found an equation in which

all the points lie on a model. Thus, minimization of the residual is an objective of obtaining regression coefficients. In the least squares method, estimates of the constants of the models are chosen such that minimization of the sum of the squared residuals is achieved, that is

minimize $\sum_{i=1}^{n} E_i^2$.

Why minimize the sum of the square of the residuals?

Why not for instance minimize the sum of the residual errors, $\sum_{i=1}^{n} E_i$, or the sum of

the absolute values of the residuals, $\sum_{i=1}^{n} |E_i|$? Alternatively, constants of the model can be

chosen such that the average residual is zero without making individual residuals small. Will any of these criteria yield unbiased parameters with the smallest variance? All of these

questions will be answered when we discuss linear regression in the next chapter (Chapter 06.03).

Regression	
Topic	Introduction to Regression
Summary	Textbook notes for the introduction to regression
Major	All engineering majors
Authors	Egwu Kalu, Autar Kaw
Date	October 11, 2008
Web Site	http://numericalmethods.eng.usf.edu