Introduction to the Linear Regression Analysis

A. Historical Origin of the term Regression

The term ***regression*** was introduced by *Francis Galton*. In his landmark paper Regression Toward Mediocrity in Hereditary Stature, he compared the heights of parents and their children. He was particularly interested in the idea that the children of tall parents tended to be tall also, but a little shorter than their parents. Children of short parents tended to be short, but not quite as short as their parents. He referred to this as "regression to mediocrity" (or regression to the mean). In quantifying regression to the mean, he invented what we would call regression.

It is perhaps surprising that Galton's specific work on height is still relevant today. Some interesting questions from Galton's data come to mind. How would one fit a model that relates parent and child heights? How would one predict a child's height based on their parents? How would we quantify regression to the mean? In this subject, we will answer all of these questions.

B. Basic Notation and Terminology

Modeling refers to the development of mathematical expressions that describe in some sense the behavior of a random variable of interest. This variable may be the price of wheat in the world market, the number of deaths from lung cancer, the rate of growth of a particular type of tumor, or the tensile strength of metal wire. In all cases, this variable is called the ***dependent variable*** and denoted with Y. A subscript on Y identifies the particular unit from which the observation was taken, the time at which the price was recorded, the city in which the deaths were recorded, the experimental unit on which the tumor growth was recorded, and so forth.

We write $X_1, X_2, \ldots, X_n$ to describe n data points. For example, consider the data set: $Age = \{12{,}20{,}23\}$. Then we have

$$X_1 = 12, X_2 = 20, X_3 = 23 \text{ and } n = 3.$$

Most commonly the modeling is aimed at describing how the mean of the dependent variable (Y) changes with changing conditions; the variance of the dependent variable is assumed to be unaffected by the changing conditions. Other variables which are thought to provide information on the behavior of the dependent variable are incorporated into the model as ***predictor*** or ***explanatory variables***. These variables are called the ***independent variables*** and are denoted by X with subscripts as needed to identify different independent variables. Additional subscripts denote the observational unit from which the data were taken. The X_i are assumed to be known constants.

In the literature the terms dependent variable and explanatory variable are described variously. A representative list is:

Dependent Variable	**Independent Variable**
Explained Variable	Explanatory Variable
Predictand Variable	Predictor Variable
Regressand Variable	Regressor Variable
Response Variable	Stimulus Variable
Endogenous Variable	Exogenous Variable
Outcome Variable	Covariate Variable

Although it is a matter of personal taste and tradition, in this subject we will use the dependent variable/ explanatory variable or the more neutral, dependent variable and independent variable terminology.

In addition to the X_i, all models involve unknown constants, called ***parameters***, which control the behavior of the model. These parameters are denoted by Greek letters (β_i) and are to be estimated from the data.

The term ***random*** is a synonym for the term ***stochastic***. A random or stochastic variable is a variable that can take on any set of values, positive or negative, with a given probability.

Functional Versus Statistical Relation Between Two Variables

A ***Functional/Deterministic Relationship*** involves an exact relationship between two variables. Suppose, for simplicity, that Y and X are two real-valued variables. The functional relationship between *Y* and *X* are represented as

$$Y = f(x)$$

for some function *f*.

Take, for instance, the conversion relationship between temperature in degrees Celsius *(C)* and temperature in degrees Fahrenheit *(F)*. We know the relationship is:

$$F = \frac{9}{5}(15) + 32 = 59$$

There is exact (linear) relationship between degrees Celsius and degrees Fahrenheit.

A ***Statistical Relationship*** is not an exact relationship. It is instead a relationship in which *"trend"* exists between the predictor *X* and the response *Y*, but there is also some *"scatter"*. A notional description of a statistical relationship is therefore of the form

$$Y = f(x) + \varepsilon$$

where $f(X)$ describes a functional relationship and ε is a statistical error, e.g., a random variable. The function *f* describes the trend, i.e., the tendency of the changes in *Y* as *X* changes while describes the deviation from that trend.

In the real world, most variables are not simply the function of one other factor. They may be influenced by many things, including random events. or stochastic. The dependence of crop yield on temperature, rainfall, sunshine, and fertilizer, for example, is statistical in nature in the sense that the explanatory variables, although certainly important, will not enable the agronomist to predict crop yield exactly because of errors involved in measuring these variables as well as a host of other factors (variables) that collectively affect the yield but may be difficult to identify individually. Thus, there is bound to be some intrinsic or random variability in the dependent variable crop yield that cannot be fully explained no matter how many explanatory variables we consider. In regression analysis we are not concerned with such functional/deterministic relationships because it might be inadequate or cumbersome to represent relationships between the variables using deterministic relationships only.

Regression Versus Correlation

Closely related to but conceptually very much different from regression analysis is *correlation analysis*, where the primary objective is to measure the ***strength* or *degree of linear association between two variables***. For example, we may be interested in finding the correlation (coefficient) between smoking and lung cancer, between scores on statistics and mathematics examinations, between high school grades and college grades, and so on. In *regression analysis*, we are not primarily interested in such a measure. Instead, we try to ***estimate*** or ***predict the average value of one variable on the basis of the fixed values of other variables***. Thus, we may want to know whether we can predict the average score on a statistics examination by knowing a students score on a mathematics examination.

Regression Versus Causation

Regression deals with dependence amongst variables within a model. But it cannot always imply causation. For example, we stated above that rainfall affects crop yield and there is data that support this. However, this is a one-way relationship: crop yield cannot affect rainfall. It means there is no cause and effect reaction on regression if there is no causation. In short, we conclude that a *statistical relationship does not imply causation.*

Can regression be used for causation? Regression **can be used to determine a causal relationship between X and Y in a controlled environment**. However, to determine the certainty of the cause you may need to pay attention to the mechanism (the process through which the cause occurs).

C. Three Most Common Tasks for Regression Models:

1. Relationship
2. Modeling
3. Prediction

D. Regression Applications

Regression analysis of data is a very powerful statistical tool. It provides a technique for building a statistical predictor of a response and enables you to place a bound (an approximate upper limit) on your error of prediction. For example, suppose you manage a construction company and you would like to predict the profit y per construction job as a function of a set of independent variables $x_1, x_2, \ldots, x_n$. If you could find the right combination of independent variables and could postulate a reasonable mathematical equation to relate y to these variables, you could possibly deduce which of the independent variables were causally related to profit per job and then control these variables to achieve a higher company profit. In addition, you could use the forecasts in corporate planning. The following examples illustrate a few of the many successful applications of regression analysis to real-world problems.

Example 1: *Psychology*
"Moonlighters" are workers who hold two jobs at the same time. What are the factors that impact the likelihood of a moonlighting worker becoming aggressive toward his or her supervisor? This was the research question of interest in the *Journal of Applied Psychology (July 2005)*. Based on data collected from a large sample of moonlighters, the researchers fit several multiple regression models for supervisor-directed aggression score (y). The most important predictors of y were determined to be age, gender, level of self-esteem, history of aggression, level of supervisor abuse, and perception of injustice at either job.

Example 2: *Geography*
Can the population of an urban area be estimated without taking a census? In *Geographical Analysis (January 2007)* geography professors at the University of WisconsinMilwaukee and Ohio State University demonstrated the use of satellite image maps for estimating urban population. A portion of Columbus, Ohio, was partitioned into $n = 25$ census block groups and satellite imagery was obtained. Multiple regression was used to successfully model population density (y) as a function of proportion of block with low-density residential areas (x_1) and proportion of block with high density residential areas (x_2).

Example 3: *Accounting*
A study of Machiavellian traits (e.g., negative character traits such as manipulation, cunning, duplicity, deception, and bad faith) in accountants was published in *Behavioral Research in Accounting (January 2008)*. Multiple regression was employed to model the Machiavellian ("Mach") rating score of an accountant as a function of the independent variables age, gender, education, and income. Of these, only income was found to have a significant impact on Mach rating score.

Example 4: *Engineering*
During periods of high electricity demand, especially during the hot summer months, the power output from a gas turbine engine can drop dramatically. One way to counter this drop in power is by cooling the inlet air to the gas turbine. An increasingly popular cooling method uses high pressure inlet fogging. The performance of a gas turbine augmented with high pressure inlet fogging was investigated in the *Journal of Engineering for Gas Turbines and Power (January 2005)*. One key performance variable, the heat rate (kilojoules

per kilowatt per hour), was discovered to be related to several independent variables, including cycle speed (revolutions per minute), inlet temperature (C) exhaust gas temperature (C), cycle pressure ratio, and air mass ow rate (kilograms per second).

Example 5: *Management*

Do chief executive officers (CEOs) and their top managers always agree on the goals of the company? Goal importance congruence between CEOs and vice presidents (VPs) was studied in the *Academy of Management Journal (February 2008)*. The researchers used regression to model a VPs attitude toward the goal of improving efficiency (y) as a function of the two independent variables, level of CEO leadership (x_1) and level of congruence between the CEO and the VP (x_2). They discovered that the impact of CEO leadership on a VPs attitude toward improving efficiency depended on level of congruence.

Example 6: *Law*

For over 20 years, courts have accepted evidence of "battered woman syndrome" as a defense in homicide cases. An article published In the *Duke Journal of Gender Law and Policy (Summer 2003)* examined the impact of expert testimony on the outcome of homicide trials that involve battered woman syndrome. Based on data collected on individual juror votes from past trials, the article reported that "when expert testimony was present, women jurors were more likely than men to change a verdict from not guilty to guilty after deliberations." This result was obtained from a multiple regression model for likelihood of changing a verdict from not guilty to guilty after deliberations, y, as a function of juror gender (male or female) and expert testimony (yes or no).

Example 7: *Education*

The Standardized Admission Test (SAT) scores of 3,492 high school and college students, some of whom paid a private tutor in an effort to obtain a higher score, were analyzed in *Chance (Winter 2001)*. Multiple regression was used to successfully estimate the effect of coaching on the SAT-Mathematics score, y. The independent variables included in the model were scores on PSAT, whether the student was coached, student ethnicity, socioeconomic status, overall high school GPA, number of mathematics courses taken in high school, and overall GPA for the math courses.

Prepared by:

KATRINA D. ELIZON
Department of Mathematics and Statistics
PUP College of Science