CHAPTER – 2
REGRESSION
MODEL
<table>
<thead>
<tr>
<th>Sr. No.</th>
<th>Name of Topic</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>60</td>
</tr>
<tr>
<td>2.2</td>
<td>Basic Concept of Statistics</td>
<td>60</td>
</tr>
<tr>
<td>2.3</td>
<td>Difference between Causality and Correlation</td>
<td>71</td>
</tr>
<tr>
<td>2.4</td>
<td>Statistical Regression Analysis</td>
<td>77</td>
</tr>
<tr>
<td>2.5</td>
<td>Types of Regression Analysis</td>
<td>78</td>
</tr>
<tr>
<td>2.6</td>
<td>General Linear Model</td>
<td>86</td>
</tr>
<tr>
<td>2.7</td>
<td>Types of Analysis</td>
<td>94</td>
</tr>
<tr>
<td>2.8</td>
<td>Conclusion</td>
<td>102</td>
</tr>
<tr>
<td></td>
<td>Reference</td>
<td>103</td>
</tr>
</tbody>
</table>
2.1 INTRODUCTION

Regression analysis is the art and science of fitting straight lines to patterns of data. Regression analysis is a statistical tool for the investigation of relationships between variables. Usually, the investigator seeks to ascertain the causal effect of one variable upon another—the effect of a price increase upon demand, for example, or the effect of changes in the money supply upon the inflation rate. To explore such issues, the investigator assembles data on the underlying variables of interest and employs regression to estimate the quantitative effect of the causal variables upon the variable that they influence. The investigator also typically assesses the —statistical significance of the estimated relationships, that is, the degree of confidence that the true relationship is close to the estimated relationship. Regression techniques have long been central to the field of economic statistics (—econometrics!). Increasingly, they have become important to lawyers and legal policy makers in U.S.A. as well. Regression has been offered as evidence of liability under Title VII of the Civil Rights Act of 1964 ((Bazemore Vs Friday, 478 U.S. 385. 400{1986}), as evidence of racial bias in death penalty litigation , (McClesky Vs Kemp, 481 U.S. 279 {1987}), as evidence of damages in contract actions (Cotton Brothers Baking Co. v. Industrial Risk Insurers, 941 F.2d 380 (5th Cir. 1991). as evidence of violations under the Voting Rights Act, (Thornburgh v. Gingles, 478 U.S. 30 (1986). and as evidence of damages in antitrust litigation, Sprayrite Service Corp. v. Monsanto Co., 684 F.2d 1226 (7th Cir. 1982). among other things.

For the present study researcher is focusing on the overview of the all the regression models- how they work, what they assume, and highlights when key assumption do not hold.

2.2 BASIC CONCEPT OF STATISTICS

(I) Variables

Variables are things that researcher measure, control, or manipulate in research. They differ in many respects, most notably in the role they are given in any research and in the type of measures that can be applied to them.

(II) Correlational vs. Experimental Research

Most empirical research belongs clearly to one of these two general categories. In correlational research, researcher do not (or at least try not to) influence any variables but only measure them and look for relations (correlations) between some set of variables, such as blood pressure and cholesterol level. In experimental research, researchers manipulate some variables and then measure the effects of this manipulation on other variables. For example, a researcher might artificially increase
blood pressure and then record cholesterol level. Data analysis in experimental research also comes down to calculating "correlations" between variables, specifically, those manipulated and those affected by the manipulation. However, experimental data may potentially provide qualitatively better information: only experimental data can conclusively demonstrate causal relations between variables. For example, if researcher found that whenever researcher change variable A then variable B changes, then researcher can conclude that "A influences B." Data from correlational research can only be "interpreted" in causal terms based on some theories that researcher has, but correlational data cannot conclusively prove causality.

(III) Dependent vs. Independent Variables

Independent variables are those that are manipulated whereas dependent variables are only measured or registered. This distinction appears terminologically confusing to many because, "all variables depend on something." However, once researcher get used to this distinction, it becomes indispensable. The terms dependent and independent variable apply mostly to experimental research where some variables are manipulated, and in this sense they are "independent" from the initial reaction patterns, features, intentions, etc. of the subjects. Some other variables are expected to be "dependent" on the manipulation or experimental conditions. That is to say, they depend on "what the subject will do" in response. Somewhat contrary to the nature of this distinction, these terms are also used in studies where researcher do not literally manipulate independent variables, but only assign subjects to "experimental groups" based on some pre-existing properties of the subjects. For example, if in an experiment, males are compared to females regarding their white cell count (WCC), Gender could be called the independent variable and WCC the dependent variable.

(IV) Measurement Scales

Variables differ in how well they can be measured, i.e., in how much measurable information their measurement scale can provide. There is obviously some measurement error involved in every measurement, which determines the amount of information that researcher can obtain. Another factor that determines the amount of information that can be provided by a variable is its type of measurement scale. Specifically, variables are classified as (a) nominal, (b) ordinal, (c) interval, or (d) ratio.

(a) Nominal Variables

Nominal variables allow for only qualitative classification. That is, they can be measured only in terms of whether the individual items belong to some distinctively different categories, but researcher cannot quantify or even rank
order those categories. For example, all researcher can say is that two individuals are different in terms of variable A (e.g., they are of different race), but researcher cannot say which one "has more" of the quality represented by the variable. Typical examples of nominal variables are gender, race, color, city, etc.

(b) Ordinal Variables

Ordinal variables allow researcher to rank order the items researcher measure in terms of which has less and which has more of the quality represented by the variable, but still they do not allow researcher to say "how much more." A typical example of an ordinal variable is the socioeconomic status of families. For example, researcher knows that upper-middle is higher than middle but researcher cannot say that it is, for example, 18% higher. Also, this very distinction between nominal, ordinal, and interval scales itself represents a good example of an ordinal variable. For example, researcher can say that nominal measurement provides less information than ordinal measurement, but researcher cannot say "how much less" or how this difference compares to the difference between ordinal and interval scales.

(c) Interval Variables

Interval variables allow researcher not only to rank order the items that are measured, but also to quantify and compare the sizes of differences between them. For example, temperature, as measured in degrees Fahrenheit or Celsius, constitutes an interval scale. Researcher can say that a temperature of 40 degrees is higher than a temperature of 30 degrees, and that an increase from 20 to 40 degrees is twice as much as an increase from 30 to 40 degrees.

(d) Ratio Variables

Ratio variables are very similar to interval variables; in addition to all the properties of interval variables, they feature an identifiable absolute zero point, thus, they allow for statements such as x is two times more than y. Typical examples of ratio scales are measures of time or space. For example, as the Kelvin temperature scale is a ratio scale, not only can researcher say that a temperature of 200 degrees is higher than one of 100 degrees; researcher can correctly state that it is twice as high. Interval scales do not have the ratio property. Most statistical data analysis procedures do not distinguish between the interval and ratio properties of the measurement scales.
(V) Relations between variables

Regardless of their type, two or more variables are related if, in a sample of observations, the values of those variables are distributed in a consistent manner. In other words, variables are related if their values systematically correspond to each other for these observations. For example, Gender and WCC would be considered to be related if most males had high WCC and most females low WCC, or vice versa; Height is related to Weight because, typically, tall individuals are heavier than short ones; IQ is related to the Number of Errors in a test if people with higher IQ's make fewer errors.

(VI) Importance of relations between variables

Generally speaking, the ultimate goal of every research or scientific analysis is to find relations between variables. The philosophy of science teaches us that there is no other way of representing "meaning" except in terms of relations between some quantities or qualities; either way involves relations between variables. Thus, the advancement of science must always involve finding new relations between variables. Correlational research involves measuring such relations in the most straightforward manner. However, experimental research is not any different in this respect. For example, the above mentioned experiment comparing WCC in males and females can be described as looking for a correlation between two variables: Gender and WCC. Statistics does nothing else but help us evaluate relations between variables. Actually, all of the hundreds of procedures that are described in this online textbook can be interpreted in terms of evaluating various kinds of inter-variable relations.

(VII) Two Basic Features of Every Relation between Variables

The two most elementary formal properties of every relation between variables are the relation's (a) magnitude (or "size") and (b) its reliability (or "truthfulness").

(a) Magnitude

Magnitude (or "size"). The magnitude is much easier to understand and measure than the reliability. For example, if every male in sample was found to have a higher WCC than any female in the sample, it could say that the magnitude of the relation between the two variables (Gender and WCC) is very high in the sample. In other words, it could predict one based on the other.
(b) Reliability

Reliability (or "truthfulness"). The reliability of a relation is a much less intuitive concept, but still extremely important. It pertains to the "representativeness" of the result found in the specific sample for the entire population. In other words, it says how probable it is that a similar relation would be found if the experiment was replicated with other samples drawn from the same population. Remember that researcher is almost never "ultimately" interested only in what is going on in their sample; researcher is interested in the sample only to the extent it can provide information about the population. If any study meets some specific criteria then the reliability of a relation between variables observed in that sample can be quantitatively estimated and represented using a standard measure which is technically called p-value or statistical significance level.

(VIII) Statistical Significance or “P – Value”

The statistical significance of a result is the probability that the observed relationship (e.g., between variables) or a difference (e.g., between means) in a sample occurred by pure chance ("luck of the draw"), and that in the population from which the sample was drawn, no such relationship or differences exist. Using less technical terms, it could say that the statistical significance of a result tells something about the degree to which the result is "true" (in the sense of being "representative of the population").

More technically, the value of the p-value represents a decreasing index of the reliability of a result defined by Brownlee in 1960. The higher the p-value, the less it can believe that the observed relation between variables in the sample is a reliable indicator of the relation between the respective variables in the population. Specifically, the p-value represents the probability of error that is involved in accepting observed result as valid, that is, as "representative of the population." For example, a p-value of .05 (i.e., 1/20) indicates that there is a 5% probability that the relation between the variables found in sample is a "fluke." In other words, assuming that in the population there was no relation between those variables whatsoever, and researcher were repeating experiments such as one after another, it could expect that approximately in every 20 replications of the experiment there would be one in which the relation between the variables in question would be equal or stronger than in respective research. (Note that this is not the same as saying that, given that there is a relationship between the variables, it can expect to replicate the results 5% of the time or 95% of the time; when there is a relationship between the variables in the population, the probability of replicating the study and finding that relationship is related to the statistical power of the design.). In many areas of research, the p-value of .05 is customarily treated as a "border-line acceptable" error level.
(VIII) How to determine that a Result is "Really" Significant

There is no way to avoid arbitrariness in the final decision as to what level of significance will be treated as really "significant." That is, the selection of some level of significance, up to which the results will be rejected as invalid, is arbitrary. In practice, the final decision usually depends on whether the outcome was predicted a priori or only found post hoc in the course of many analyses and comparisons performed on the data set, on the total amount of consistent supportive evidence in the entire data set, and on "traditions" existing in the particular area of research. Typically, in many sciences, results that yield $p < .05$ are considered borderline statistically significant, but this level of significance still involves a pretty high probability of error (5%). Results that are significant at the $p < .01$ level are commonly considered statistically significant, and $p < .005$ or $p < .001$ levels are often called "highly" significant. But these classifications represent nothing else but arbitrary conventions that are only informally based on general research experience.

(IX) Statistical Significance and the Number of Analyses Performed

Needless to say, the more analyses researcher perform on a data set, the more results will meet "by chance" the conventional significance level. For example, if researcher calculate correlations between ten variables (i.e., 45 different correlation coefficients), then researcher should expect to find by chance that about two (i.e., one in every 20) correlation coefficients are significant at the $p < .05$ level, even if the values of the variables were totally random and those variables do not correlate in the population. Some statistical methods that involve many comparisons and, thus, a good chance for such errors include some "correction" or adjustment for the total number of comparisons. However, many statistical methods especially simple exploratory data analyses do not offer any straightforward remedies to this problem. Therefore, it is up to the researcher to carefully evaluate the reliability of unexpected findings.

(X) Strength vs. Reliability of a Relation between Variables

It is said before that strength and reliability are two different features of relationships between variables. However, they are not totally independent. In general, for in a sample of a particular size, the larger the magnitude of the relation between variables, the more reliable the relation

(XI) Why Stronger Relations between Variables are More Significant

Assuming that there is no relation between the respective variables in the population, the most likely outcome would be also finding no relation between these variables in the research sample. Thus, the stronger the relation found in the sample, the less likely it is that there is no corresponding relation in the population. As it is seen, the
magnitude and significance of a relation appear to be closely related, and researcher could calculate the significance from the magnitude and vice-versa; however, this is true only if the sample size is kept constant, because the relation of a given strength could be either highly significant or not significant at all, depending on the sample size.

(XII) Why Significance of a Relation between Variables Depends on the Size of the Sample

If there are very few observations, then there are also respectively few possible combinations of the values of the variables and, thus, the probability of obtaining by chance a combination of those values indicative of a strong relation is relatively high.

Consider the following illustration. If researcher is interested in two variables (Gender: male/female and WCC: high/low), and there are only four subjects in propose sample (two males and two females), then the probability that researcher will find, purely by chance, a 100% relation between the two variables can be as high as one-eighth. Specifically, there is a one-in-eight chance that both males will have a high WCC and both females a low WCC, or vice versa.

Now consider the probability of obtaining such a perfect match by chance if in sample consisted of 100 subjects; the probability of obtaining such an outcome by chance would be practically zero.

Let's look at a more general example. Imagine a theoretical population in which the average value of WCC in males and females is exactly the same. Needless to say, if it start replicating a simple experiment by drawing pairs of samples (of males and females) of a particular size from this population and calculating the difference between the average WCC in each pair of samples, most of the experiments will yield results close to zero. However, from time to time, a pair of samples will be drawn where the difference between males and females will be quite different from zero.

How often will it happen? The smaller the sample size in each experiment, the more likely it is that we will obtain such erroneous results, which in this case would be results indicative of the existence of a relation between Gender and WCC obtained from a population in which such a relation does not exist.
(XIII) Why Small Relations can be Proven Significant Only in Large Samples

The examples in the previous paragraphs indicate that if a relationship between variables in question is "objectively" (i.e., in the population) small, then there is no way to identify such a relation in a study unless the research sample is correspondingly large. Even if in sample is in fact "perfectly representative," the effect will not be statistically significant if the sample is small. Analogously, if a relation in question is "objectively" very large, then it can be found to be highly significant even in a study based on a very small sample.

Consider this additional illustration. If a coin is slightly asymmetrical and, when tossed, is somewhat more likely to produce heads than tails (e.g., 60% vs. 40%), then ten tosses would not be sufficient to convince anyone that the coin is asymmetrical even if the outcome obtained (six heads and four tails) was perfectly representative of the bias of the coin. However, is it so that 10 tosses is not enough to prove anything? No; if the effect in question were large enough, then ten tosses could be quite enough. For instance, imagine now that the coin is so asymmetrical that no matter how you toss it, the outcome will be heads. If it tossed such a coin ten times and each toss produced heads, most people would consider it sufficient evidence that something is wrong with the coin. In other words, it would be considered convincing evidence that in the theoretical population of an infinite number of tosses of this coin, there would be more heads than tails. Thus, if a relation is large, then it can be found to be significant even in a small sample.

(XIV) Can "No Relation" be a Significant Result?

The smaller the relation between variables, the larger the sample size that is necessary to prove it significant For example, imagine how many tosses would be necessary to prove that a coin is asymmetrical if its bias were only .000001%! Thus, the necessary minimum sample size increases as the magnitude of the effect to be demonstrated decreases. When the magnitude of the effect approaches zero, the necessary sample size to conclusively prove it approaches infinity. That is to say, if there is almost no relation between two variables, then the sample size must be almost equal to the population size, which is assumed to be infinitely large. Statistical significance represents the probability that a similar outcome would be obtained if we tested the entire population. Thus, everything that would be found after testing the entire population would be, by definition, significant at the highest possible level, and this also includes all "no relation" results.
(XV) How to Measure the Magnitude (Strength) of Relations between Variables

There are very many measures of the magnitude of relationships between variables that have been developed by statisticians; the choice of a specific measure in given circumstances depends on the number of variables involved, measurement scales used, nature of the relations, etc. Almost all of them, however, follow one general principle: they attempt to somehow evaluate the observed relation by comparing it to the "maximum imaginable relation" between those specific variables.

Technically speaking, a common way to perform such evaluations is to look at how differentiated the values are of the variables, and then calculate what part of this "overall available differentiation" is accounted for by instances when that differentiation is "common" in the two (or more) variables in question. Speaking less technically, we compare "what is common in those variables" to "what potentially could have been common if the variables were perfectly related."

Let's consider a simple illustration. Let's say that in our sample, the average index of WCC is 100 in males and 102 in females. Thus, we could say that on average, the deviation of each individual score from the grand mean (101) contains a component due to the gender of the subject; the size of this component is 1. That value, in a sense, represents some measure of relation between Gender and WCC. However, this value is a very poor measure because it does not tell us how relatively large this component is given the "overall differentiation" of WCC scores. Consider two extreme possibilities:

i. If all WCC scores of males were equal exactly to 100 and those of females equal to 102, then all deviations from the grand mean in our sample would be entirely accounted for by gender. We would say that in our sample, Gender is perfectly correlated with WCC, that is, 100% of the observed differences between subjects regarding their WCC is accounted for by their gender.

ii. If WCC scores were in the range of 0-1000, the same difference (of 2) between the average WCC of males and females found in the study would account for such a small part of the overall differentiation of scores that most likely it would be considered negligible. For example, one more subject taken into account could change, or even reverse the direction of the difference. Therefore, every good measure of relations between variables must take into account the overall differentiation of individual scores in the sample and evaluate the relation in terms of (relatively) how much of this differentiation is accounted for by the relation in question.
(XVI) Common "General Format" of Most Statistical Tests

Because the ultimate goal of most statistical tests is to evaluate relations between variables, most statistical tests follow the general format that was explained in the previous paragraph. Technically speaking, they represent a ratio of some measure of the differentiation common in the variables in question to the overall differentiation of those variables. For example, they represent a ratio of the part of the overall differentiation of the WCC scores that can be accounted for by gender to the overall differentiation of the WCC scores. This ratio is usually called a ratio of explained variation to total variation. In statistics, the term explained variation does not necessarily imply that we "conceptually understand" it. It is used only to denote the common variation in the variables in question, that is, the part of variation in one variable that is "explained" by the specific values of the other variable, and vice versa.

(XVII) How the "Level of Statistical Significance" is calculated

Specifically, the significance depends mostly on the sample size. As explained before, in very large samples, even very small relations between variables will be significant, whereas in very small samples even very large relations cannot be considered reliable (significant). Thus, in order to determine the level of statistical significance, we need a function that represents the relationship between "magnitude" and "significance" of relations between two variables, depending on the sample size. The function we need would tell us exactly "how likely it is to obtain a relation of a given magnitude (or larger) from a sample of a given size, assuming that there is no such relation between those variables in the population." In other words, that function would give us the significance (p) level, and it would tell us the probability of error involved in rejecting the idea that the relation in question does not exist in the population. This "alternative" hypothesis (that there is no relation in the population) is usually called the null hypothesis. It would be ideal if the probability function was linear and, for example, only had different slopes for different sample sizes. Unfortunately, the function is more complex and is not always exactly the same; however, in most cases we know its shape and can use it to determine the significance levels for our findings in samples of a particular size. Most of these functions are related to a general type of function, which is called normal.

(XVIII) Why the "Normal Distribution" is Important

The distribution of many test statistics is normal or follows some form that can be derived from the normal distribution. In this sense, philosophically speaking, the normal distribution represents one of the empirically verified elementary "truths about the general nature of reality," and its status can be compared to the one of fundamental laws of natural sciences. The exact shape of the normal distribution (the
characteristic "bell curve") is defined by a function that has only two parameters: mean and standard deviation.

A characteristic property of the normal distribution is that 68% of all of its observations fall within a range of ±1 standard deviation from the mean, and a range of ±2 standard deviations includes 95% of the scores. In other words, in a normal distribution, observations that have a standardized value of less than -2 or more than +2 have a relative frequency of 5% or less. Standardized value means that a value is expressed in terms of its difference from the mean, divided by the standard deviation.

If STATISTICA is accessed, one can explore the exact values of probability associated with different values in the normal distribution using the interactive Probability Calculator tool; for example, if it enter the Z value (i.e., standardized value) of 4, the associated probability computed by STATISTICA will be less than 0.0001, because in the normal distribution almost all observations (i.e., more than 99.99%) fall within the range of ±4 standard deviations.

(XIX) Are All Test Statistics Normally Distributed?

Not all, but most of them are either based on the normal distribution directly or on distributions that are related to and can be derived from normal, such as t, F, or Chi-square. Typically, these tests require that the variables analyzed are themselves normally distributed in the population, that is, they meet the so-called "normality assumption." Many observed variables actually are normally distributed, which is another reason why the normal distribution represents a "general feature" of empirical reality. The problem may occur when it try to use a normal distribution-based test to analyze data from variables that are themselves not normally distributed. In such cases, researcher has two general choices. First, researcher can use some alternative "nonparametric" test (or so-called "distribution-free test"); but this is often inconvenient because such tests are typically less powerful and less flexible in terms of types of conclusions that they can provide. Alternatively, in many cases researcher can still use the normal distribution-based test if researcher only makes sure that the size of their samples is large enough. The latter option is based on an extremely important principle that is largely responsible for the popularity of tests that are based on the normal function. Namely, as the sample size increases, the shape of the sampling distribution (i.e., distribution of a statistic from the sample; this term was first used by Fisher, 1928a) approaches normal shape, even if the distribution of the variable in question is not normal.
(XX) Consequences of Violating the Normality Assumption

Although many of the statements made in the preceding paragraphs can be proven mathematically, some of them do not have theoretical proof and can be demonstrated only empirically, via so-called Monte-Carlo experiments. Monte-Carlo studies were used extensively with normal distribution-based tests to determine how sensitive they are to violations of the assumption of normal distribution of the analyzed variables in the population. The general conclusion from these studies is that the consequences of such violations are less severe than previously thought. Although these conclusions should not entirely discourage anyone from being concerned about the normality assumption, they have increased the overall popularity of the distribution-dependent statistical tests in all areas of research.

2.3 DIFFERENCE BETWEEN CAUSALITY AND CORRELATION

Causation and Correlation are loosely used words in analytics. People tend to use these words interchangeably without knowing the fundamental logic behind them. Apparently, people get trapped in the phonetics of these words and end up using them at incorrect places. But, let me warn you that apart from the similar sounding names, there isn’t a lot common in the two phenomena. Their fundamental implications are very different.

Let’s understand the difference between Causation and Correlation using a few examples below.

(A) EXAMPLES

Example 1: X – Tier of B-school College a student gets offer for => Y – Salary after the graduation

Hypothesis – Students going to premium B-schools get higher salaries on an average.

Is this B-school a cause of getting better jobs?
Causal relation does not exist. For instance, only ambitious and intelligent people are selected from elite B-schools who further get much higher salary than the average. Hence, even if these students did not study in Tier 1 B-School, he/she still might get more than the average salaries. Hence, we have alternate reasoning issue in this case.

**Example 2: X – Smoking Cigarettes => Y – Level of Mental Stress**

**Hypothesis** – People who smoke are found to have higher level of stress. 
Is smoking the reason of stress?

Causal relation does not exist. We can reject hypothesis based on inverse causality. For instance, higher mental stress can actually influence a person to smoke.

**Example 3: X – Having Kids => Y – Maturity level**

**Hypothesis** – People get more matured after having kids?  
Are having kids a cause of attaining higher maturity levels?

Causal relation does not exist. Once again, we can reject hypothesis based on inverse causality. For instance, only mature people are likely to be prepared to have kids. We can also apply alternate reasoning with underlying cause as the age. Higher age leads to both, having kids and higher maturity levels.
Example 4: X – Altitude => Y – Temperature

**Hypothesis** – We witness lower temperature at high altitudes. Which means, the higher you go, the colder it would become. Is higher altitude a cause of lower temperature?

Causal relation does exist. We definitely know that inverse causality is not possible. Also alternate reasoning or mutual independence can be rejected.

**KEY POINTS IN ESTABLISHING CAUSATION**

Here are the key point (X => Y) pairs used in establishing causation:

1. **Alternate Reasoning:** If there is an alternate reason (Z) which indeed can influence both X and Y (Z => X & Z => Y are true), we can reject the hypothesis of X => Y.

2. **Inverse Causality:** If instead of X influencing Y, we have Y influencing X, we can reject X => Y hypothesis based on inverse causality.

3. **Mutual independence:** Sometimes X and Y might just be correlated and nothing else. In such cases we reject hypothesis based on mutual independence.

**CONCLUSIVELY DERIVE CAUSE-EFFECT RELATIONSHIP?**

In fields like pharmacy, it is very important to establish cause-effect pairs. And that’s why, there are enough researches done to find cause-effect pairs. Let’s understand the following definitions before we get down to mathematics:

1. **Randomized Experimental data:** An experiment is often defined as random assignment of observational units to different conditions, and conditions differ by the treatment of observational units. Treatment is a generic term, which translates most
CHAPTER – 2

REGRESSION MODELS

easily in medical applications (e.g. patients are treated differently under different conditions), but it applies to other areas as well.

2. Observational data: If we do not have the luxury to do a randomized experiment, we are forced to work on existing data sources. These events have already happened without any control. Hence, the selection is not random.

Deriving out causality from Observational data is very difficult and non-conclusive. For a conclusive result on causality, we need to do randomized experiments.

(d) WHY ARE OBSERVATIONAL DATA NOT CONCLUSIVE?

i. Observational Data not conclusive because the selection in observational data is not randomized. We can never conclude individual cause-effect pair.

ii. For instance, if the number of students graduating from Tier 1 B-school gets higher salary; this will not conclude causality relationship because the selection was based on initial performance.

iii. However, if we randomly select students for Tier 1 B-schools, this analysis will become more conclusive to establish causality.

(e) WHY DON’T WE DO RANDOM EXPERIMENT EVERY TIME TO ESTABLISH CAUSALITY?

There are multiple reasons you might be asked to work on observational data instead of experiment data to establish causality.

First is, the cost involved to do these experiments. For instance, if your hypothesis is giving free I-phone to customers, this activity will have an incremental gain on sales of Mac. Doing this experiment without knowing anything on causality can be an expensive proposal.

Second is, not all experiments are allowed ethically. For instance, if you want to know whether smoking contributes to stress, you need to make normal people smoke, which is ethically not possible.

(f) IN THAT CASE, HOW DO WE ESTABLISH CAUSALITY USING OBSERVATIONAL DATA?

There has been good amount of research done on this particular issue. The entire objective of these methodologies is to eliminate the effect of any unobserved variable. In this section, I will introduce you to some of these well-known techniques:
1. **Panel Model (Ordinary regression):** This method comes in very handy if the unobserved dimension is invariant along at least one dimension. For instance, if the unobserved dimension is invariant over time, we can try building a panel model which can segregate out the bias coming from unobserved dimension.

2. **Simulated Control:** Biggest concern with observation data is that we do not get treatment and non-treatment data for the same data point. For instance, referring to the smoking example above, a person can’t be a smoker and non-smoker at the same time.

   But, what if, we can find out a look alike for all treated in the non-treated group. And then compare the response of this treatment among look alikes. This is the most common method implemented currently in the industry.

   The look alike can be found using nearest neighbor algorithm, k-d tree or any other algorithm. Let’s take an instance; we have two people with same age group, gender, income etc. One of them starts smoking and another does not. Now the stress level can be compared over a period of time given no other condition changes among them.

   While this might sound a very exciting approach theoretically, it is usually difficult to carve out pure simulated / virtual control and this can some time lead to conclusions, which may not be correct. This is actually a topic for a different article in future.

3. **Instrumental Variable (IV):** This is probably the hardest one which I find to implement. Following are the steps to implement this technique:

   i. Find the cause – effect pair.
   
   ii. Find an attribute which is related to cause but is independent of the error which we get by regressing cause-effect pair. This variable is known as Instrumental Variable.
   
   iii. Now estimate the cause variables using IV.
   
   iv. Try regressing estimated cause – effect to find the actual coefficient of causality.

   In observational data, any regression technique applied between cause-effect pair gives a biased coefficient. Using this methodology, we come out with an unbiased estimation. For example (in cigarette – mental stress pair), we might think it could be influenced by reverse causality.

   Now, if we can find any information which is connected to cigarette consumption but not mental stress, we might be able to find the actual relationship. Generally IV is regulatory based variables. For example, here we found a tax regulation which increased only cigarette price and lead to lesser consumption of cigarettes across
board. We can now try finding out the mental stress using the 4-step method discussed above.

4. **Regression discontinuity design:** This is amongst one of my favorite choices. It this makes the observational data really close to experimental design.

In the graph shown below, we are finding a dimension which has a spike on treatment and non-treatment population ratio. Suppose, we want to test the effect of scholarship in college on the grades by the end of course for students. Note that, scholarship is offered to students scoring more than 80% in their pre-college examinations. Here’s the twist. Because these students are already bright, they might continue being on top in future as well. Hence, this is a very difficult cause-effect relation to crack!

![Chart 2.1 Regression Discontinuity Design](image)

But, what if, we compare students who scored just below 80% (say 79.9%) with those who scored just above 80% (say 80.1%) on grades by end of the college. The assumption being that 79.9% student won’t be much different from 80.1% student. And the only thing which can change is the effect of scholarship. This is known as Quasi Randomized Selection.

Hence, the results are very close to perfect conclusions on causality. The only challenge with this methodology is that getting such a dimension is very difficult which can give a pure break up between treated and non-treated population.
2.4 STATISTICAL REGRESSION ANALYSIS

Regression is a generic term for all methods attempting to fit a model to observed data in order to quantify the relationship between two groups of variables. The fitted model may then be used either to merely describe the relationship between the two groups of variables, or to predict new values.

(a) Regression Analysis

Regression analysis is a form of predictive modeling technique which investigates the relationship between a dependent (target) and independent variable(s) (predictor). This technique is used for forecasting, time series modeling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

Regression analysis is an important tool for modeling and analyzing data. Here, we fit a curve / line to the data points, in such a manner that the differences between the distances of data points from the curve or line is minimized. It’ll explain this in more details as follows.

Chart 2.1 Regression Analysis

(b) Reason for using Regression Analysis

As mentioned above, regression analysis estimates the relationship between two or more variables. Let’s understand this with an easy example:

Let’s say, you want to estimate growth in sales of a company based on current economic conditions. You have the recent company data which indicates that the growth in sales is around two and a half times the growth in the economy. Using this insight, we can predict future sales of the company based on current & past information.
There are multiple benefits of using regression analysis. They are as follows:

i. It indicates the significant relationships between dependent variable and independent variable.

ii. It indicates the strength of impact of multiple independent variables on a dependent variable.

Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.

2.5 TYPES OF REGRESSION ANALYSIS

There are various kinds of regression techniques available to make predictions. These techniques are mostly driven by three metrics.

For the creative ones, you can even cook up new regressions, if you feel the need to use a combination of the parameters above, which people haven’t used before. But before you start that, let us understand the most commonly used regressions:

1. Linear Regression

It is one of the most widely known modeling techniques. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.

Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).
It is represented by an equation $Y = a + bX + e$, where $a$ is intercept, $b$ is slope of the line and $e$ is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).

**Chart 2.3 Linear Regression Model**

The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable. Now, the question is —How do we obtain best fit line?

**How to obtain best fit line (Value of a and b)?**

This task can be easily accomplished by Least Square Method. It is the most common method used for fitting a regression line. It calculates the best-fit line for the observed data by minimizing the sum of the squares of the vertical deviations from each data point to the line. Because the deviations are first squared, when added, there is no cancelling out between positive and negative values.

**Chart 2.4 Linear Regression best fitted model**
We can evaluate the model performance using the metric R-square.

Important Points:

i. There must be linear relationship between independent and dependent variables

ii. Multiple regressions suffer from multicollinearity, autocorrelation, heteroskedasticity.

iii. Linear Regression is very sensitive to Outliers. It can terribly affect the regression line and eventually the forecasted values.

iv. Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable

v. In case of multiple independent variables, we can go with forward selection, backward elimination and step wise approach for selection of most significant independent variables.

2. Logistic Regression

Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the value of Y ranges from 0 to 1 and it can represent by following equation.

\[
\text{Odds} = \frac{p}{(1-p)} = \text{probability of event occurrence} / \text{probability of not event occurrence} \\
\ln(\text{odds}) = \ln\left(\frac{p}{(1-p)}\right) \\
\text{Logit}(p) = \ln\left(\frac{p}{(1-p)}\right) = b0+b1X1+b2X2+b3X3+...+bkXk
\]

Above, p is the probability of presence of the characteristic of interest. A question that you should ask here is —why have we used log in the equation?!

Since we are working here with a binomial distribution (dependent variable), we need to choose a link function which is best suited for this distribution. And, it is logit function. In the equation above, the parameters are chosen to maximize the likelihood of observing the sample values rather than minimizing the sum of squared errors (like in ordinary regression).
Important Points:

i. It is widely used for **classification problems**

ii. Logistic regression doesn’t require linear relationship between dependent and independent variables. It can handle various types of relationships because it applies a non-linear log transformation to the predicted odds ratio

iii. To avoid over fitting and under fitting, we should include all significant variables. A good approach to ensure this practice is to use a step wise method to estimate the logistic regression

iv. It requires **large sample sizes** because maximum likelihood estimates are less powerful at low sample sizes than ordinary least square

v. The independent variables should not be correlated with each other i.e. **no multicollinearity**. However, we have the options to include interaction effects of categorical variables in the analysis and in the model.

vi. If the values of dependent variable is ordinal, then it is called as **Ordinal logistic regression**

vii. If dependent variable is multi class then it is known as **Multinomial Logistic regression**.
3. Polynomial Regression

A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation: \( y = a + b x^2 \)

In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.

**Chart 2.6 Polynomial Regression Model**

![Polynomial Regression Model](chart)

**Important Points:**

i. While there might be a temptation to fit a higher degree polynomial to get lower error, this can result in over-fitting. Always plot the relationships to see the fit and focus on making sure that the curve fits the nature of the problem. Here is an example of how plotting can help:

![Plot Examples](plot)

ii. Especially look out for curve towards the ends and see whether those shapes and trends make sense. Higher polynomials can end up producing weird results on extrapolation.
4. Stepwise Regression

This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves no human intervention.

This feat is achieved by observing statistical values like R-square, t-stats and AIC metric to discern significant variables. Stepwise regression basically fits the regression model by adding/dropping co-variates one at a time based on a specified criterion.

Some of the most commonly used Stepwise regression methods are listed below:

i. Standard stepwise regression does two things. It adds and removes predictors as needed for each step.
ii. Forward selection starts with most significant predictor in the model and adds variable for each step.
iii. Backward elimination starts with all predictors in the model and removes the least significant variable for each step.

The aim of this modeling technique is to maximize the prediction power with minimum number of predictor variables. It is one of the method to handle higher dimensionality of data set.

5. Ridge Regression

Ridge Regression is a technique used when the data suffers from multicollinearity (Independent variables are highly correlated). In multicollinearity, even though the least squares estimate (OLS) are unbiased; their variances are large which deviates the observed value far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors.

Above, we saw the equation for linear regression.

\[ y = a + b \times x \]

This equation also has an error term. The complete equation becomes:

\[ y = a + b \times x + e \text{ (error term)} \]

[Error term is the value needed to correct for a prediction error between the observed and predicted value]


\[
y = a + b_1 x_1 + b_2 x_2 + \ldots + e,
\]

for multiple independent variables.

In a linear equation, prediction errors can be decomposed into two sub components. First is due to the biased and second is due to the variance. Prediction error can occur due to any one of these two or both components. Here, we’ll discuss about the error caused due to variance.

Ridge regression solves the multicollinearity problem through shrinkage parameter \( \lambda \) (lambda). Look at the equation below.

\[
\beta^* = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \| y - X \beta \|_2^2 + \lambda \| \beta \|_2^2 \right\}
\]

In this equation, we have two components. First one is least square term and other one is lambda of the summation of \( \beta_2 \) (beta squared) where \( \beta \) is the coefficient. This is added to least square term in order to shrink the parameter to have a very low variance.

**Important Points:**

i. The assumptions of this regression is same as least squared regression except normality is not to be assumed

ii. It shrinks the value of coefficients but doesn’t reaches zero, which suggests no feature selection feature

iii. This is a regularization method and uses

6. Lasso Regression

Similar to Ridge Regression, Lasso (Least Absolute Shrinkage and Selection Operator) also penalizes the absolute size of the regression coefficients. In addition, it is capable of reducing the variability and improving the accuracy of linear regression models. Look at the equation below:

\[
\beta^* = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \| y - X \beta \|_2^2 + \lambda \| \beta \|_1 \right\}
\]

Lasso regression differs from ridge regression in a way that it uses absolute values in the penalty function, instead of squares. This leads to penalizing (or equivalently constraining the sum of the absolute values of the estimates) values which causes some of the parameter estimates to turn out exactly zero. Larger the penalty applied, further the estimates get shrunk towards absolute zero. This results to variable selection out of given n variables.
Important Points:

i. The assumptions of this regression is same as least squared regression except normality is not to be assumed

ii. It shrinks coefficients to zero (exactly zero), which certainly helps in feature selection

iii. This is a regularization method and uses various regularization

iv. If group of predictors are highly correlated, lasso picks only one of them and shrinks the others to zero

7. Elastic Net Regression

Elastic Net is hybrid of Lasso and Ridge Regression techniques. It is trained with L1 and L2 prior as regularize. Elastic-net is useful when there are multiple features which are correlated. Lasso is likely to pick one of these at random, while elastic-net is likely to pick both.

\[
\hat{\beta} = \arg \min_{\beta} (\|y - X\beta\|^2 + \lambda_2 \|eta\|^2 + \lambda_1 \|eta\|_1).
\]

A practical advantage of trading-off between Lasso and Ridge is that, it allows Elastic-Net to inherit some of Ridge’s stability under rotation.

Important Points:

i. It encourages group effect in case of highly correlated variables

ii. There are no limitations on the number of selected variables

iii. It can suffer with double shrinkage

Beyond these 7 most commonly used regression techniques, one can also look at other models like Bayesian, Ecological and Robust regression.

How to select the right regression model?

Life is usually simple, when you know only one or two techniques. One of the training institutes I know of tells their students – if the outcome is continuous – apply linear regression. If it is binary – use logistic regression! However, higher the number of options available at our disposal, more difficult it becomes to choose the right one. A similar case happens with regression models.

Within multiple types of regression models, it is important to choose the best suited technique based on type of independent and dependent variables, dimensionality in
the data and other essential characteristics of the data. Below are the key factors that you should practice to select the right regression model:

i. Data exploration is an inevitable part of building predictive model. It should be your first step before selecting the right model like identify the relationship and impact of variables.

ii. To compare the goodness of fit for different models, we can analyses different metrics like statistical significance of parameters, R-square, Adjusted r-square, AIC, BIC and error term. Another one is the Mallow’s Cp criterion. This essentially checks for possible bias in your model, by comparing the model with all possible sub models (or a careful selection of them).

iii. Cross-validation is the best way to evaluate models used for prediction. Here you divide your data set into two group (train and validate). A simple mean squared difference between the observed and predicted values give you a measure for the prediction accuracy.

iv. If your data set has multiple confounding variables, you should not choose automatic model selection method because you do not want to put these in a model at the same time.

v. It’ll also depend on your objective. It can occur that a less powerful model is easy to implement as compared to a highly statistically significant model.

vi. Regression regularization methods (Lasso, Ridge and Elastic Net) works well in case of high dimensionality and multicollinearity among the variables in the data set.

### 2.6 GENERAL LINEAR MODEL

The roots of the general linear model surely go back to the origins of mathematical thought, but it is the emergence of the theory of algebraic invariants in the 1800's that made the general linear model, as we know it today, possible. The theory of algebraic invariants developed from the groundbreaking work of 19th century mathematicians such as Gauss, Boole, Cayley, and Sylvester. The theory seeks to identify those quantities in systems of equations which remain unchanged under linear transformations of the variables in the system. Stated more imaginatively (but in a way in which the originators of the theory would not consider an overstatement), the theory of algebraic invariants searches for the eternal and unchanging amongst the chaos of the transitory and the illusory. That is no small goal for any theory, mathematical or otherwise.

The wonder of it all is the theory of algebraic invariants was successful far beyond the hopes of its originators. Eigenvalues, eigenvectors, determinants, matrix decomposition methods; all derive from the theory of algebraic invariants. The contributions of the theory of algebraic invariants to the development of statistical theory and methods are numerous, but a simple example familiar to even the most
casual student of statistics is illustrative. The correlation between two variables is unchanged by linear transformations of either or both variables. We probably take this property of correlation coefficients for granted, but what would data analysis be like if we did not have statistics that are invariant to the scaling of the variables involved? Some thought on this question should convince you that without the theory of algebraic invariants, the development of useful statistical techniques would be nigh impossible.

The development of the linear regression model in the late 19th century, and the development of correlational methods shortly thereafter, are clearly direct outgrowths of the theory of algebraic invariants. Regression and correlational methods, in turn, serve as the basis for the general linear model. Indeed, the general linear model can be seen as an extension of linear multiple regression for a single dependent variable. Understanding the multiple regression models is fundamental to understanding the general linear model, so we will look at the purpose of multiple regressions, the computational algorithms used to solve regression problems, and how the regression model is extended in the case of the general linear model. A basic introduction to multiple regression methods and the analytic problems to which they are applied is provided in the Multiple Regression.

(a) The purpose of Multiple Regression

The general linear model can be seen as an extension of linear multiple regression for a single dependent variable, and understanding the multiple regression model is fundamental to understanding the general linear model. The general purpose of multiple regressions (the term was first used by Pearson, 1908) is to quantify the relationship between several independent or predictor variables and a dependent or criterion variable. For a detailed introduction to multiple regressions, also refer to the Multiple Regression section. For example, a real estate agent might record for each listing the size of the house (in square feet), the number of bedrooms, the average income in the respective neighborhood according to census data, and a subjective rating of appeal of the house. Once this information has been compiled for various houses it would be interesting to see whether and how these measures relate to the price for which a house is sold. For example, we might learn that the number of bedrooms is a better predictor of the price for which a house sells in a particular neighborhood than how "pretty" the house is (subjective rating). We may also detect "outliers," for example, houses that should really sell for more, given their location and characteristics.

Personnel professionals customarily use multiple regression procedures to determine equitable compensation. We can determine a number of factors or dimensions such as "amount of responsibility" (Resp) or "number of people to supervise" (No_Super) that we believe to contribute to the value of a job. The personnel analyst then usually
conducts a salary survey among comparable companies in the market, recording the salaries and respective characteristics (i.e., values on dimensions) for different positions. This information can be used in a multiple regression analysis to build a regression equation of the form:

\[
\text{Salary} = 0.5 \times \text{Resp} + 0.8 \times \text{No\_Super}
\]

Once this so-called regression equation has been determined, the analyst can now easily construct a graph of the expected (predicted) salaries and the actual salaries of job incumbents in his or her company. Thus, the analyst is able to determine which position is underpaid (below the regression line) or overpaid (above the regression line), or paid equitably.

In the social and natural sciences multiple regression procedures are very widely used in research. In general, multiple regression allows the researcher to ask (and hopefully answer) the general question "what is the best predictor of ..." For example, educational researchers might want to learn what are the best predictors of success in high-school. Psychologists may want to determine which personality variable best predicts social adjustment. Sociologists may want to find out which of the multiple social indicators best predict whether or not a new immigrant group will adapt and be absorbed into society.

(b) Computations for solving the Multiple Regression Equation

A one-dimensional surface in a two-dimensional or two-variable space is a line defined by the equation \( Y = b_0 + b_1X \). According to this equation, the \( Y \) variable can be expressed in terms of or as a function of a constant (\( b_0 \)) and a slope (\( b_1 \)) times the \( X \) variable. The constant is also referred to as the intercept, and the slope as the regression coefficient. For example, GPA may best be predicted as \( 1+.02 \times \text{IQ} \). Thus, knowing that a student has an IQ of 130 would lead us to predict that her GPA would be 3.6 (since, \( 1+.02 \times 130=3.6 \)). In the multiple regression case, when there are multiple predictor variables, the regression surface usually cannot be visualized in a two-dimensional space, but the computations are a straightforward extension of the computations in the single predictor case. For example, if in addition to IQ we had additional predictors of achievement (e.g., Motivation, Self-discipline) we could construct a linear equation containing all those variables.

In general then, multiple regression procedures will estimate a linear equation of the form:

\[
Y = b_0 + b_1X_1 + b_2X_2 + ... + b_kX_k
\]

where \( k \) is the number of predictors. Note that in this equation, the regression coefficients (or \( b_1 \) ... \( b_k \) coefficients) represent the independent contributions of each in dependent variable to the prediction of the dependent variable. Another way to
express this fact is to say that, for example, variable X1 is correlated with the Y variable, after controlling for all other independent variables. This type of correlation is also referred to as a partial correlation (this term was first used by Yule, 1907). Perhaps the following example will clarify this issue. We would probably find a significant negative correlation between hair length and height in the population (i.e., short people have longer hair). At first this may seem odd; however, if we were to add the variable Gender into the multiple regression equation, this correlation would probably disappear. This is because women, on the average, have longer hair than men; they also are shorter on the average than men. Thus, after we remove this gender difference by entering Gender into the equation, the relationship between hair length and height disappears because hair length does not make any unique contribution to the prediction of height, above and beyond what it shares in the prediction with variable Gender. Put another way, after controlling for the variable Gender, the partial correlation between hair length and height is zero.

The regression surface (a line in simple regression, a plane or higher-dimensional surface in multiple regression) expresses the best prediction of the dependent variable (Y), given the independent variables (X's). However, nature is rarely (if ever) perfectly predictable, and usually there is substantial variation of the observed points from the fitted regression surface. The deviation of a particular point from the nearest corresponding point on the predicted regression surface (its predicted value) is called the residual value. Since the goal of linear regression procedures is to fit a surface, which is a linear function of the X variables, as closely as possible to the observed Y variable, the residual values for the observed points can be used to devise a criterion for the "best fit." Specifically, in regression problems the surface is computed for which the sum of the squared deviations of the observed points from that surface is minimized. Thus, this general procedure is sometimes also referred to as least squares estimation.

The actual computations involved in solving regression problems can be expressed compactly and conveniently using matrix notation. Suppose that there are n observed values of Y and n associated observed values for each of k different X variables. Then Yᵢ, Xᵢk, and eᵢ can represent the ith observation of the Y variable, the ith observation of each of the X variables, and the ith unknown residual value, respectively. Collecting these terms into matrices we have

\[
Y = \begin{bmatrix}
Y_1 \\
\vdots \\
Y_n
\end{bmatrix}, \quad X = \begin{bmatrix}
1 & X_{11} & \cdots & \cdots & \cdots & X_{1k} \\
1 & \vdots & \cdots & \cdots & \cdots & \vdots \\
1 & \vdots & \cdots & \cdots & \cdots & \vdots \\
1 & X_{n1} & \cdots & \cdots & \cdots & X_{nk}
\end{bmatrix}, \quad e = \begin{bmatrix}
e_1 \\
\vdots \\
e_n
\end{bmatrix}
\]

The multiple regression model in matrix notation then can be expressed as
CHAPTER – 2

REGRESSION MODELS

\[ Y = Xb + e \]

where \( b \) is a column vector of 1 (for the intercept) + \( k \) unknown regression coefficients. Recall that the goal of multiple regressions is to minimize the sum of the squared residuals. Regression coefficients that satisfy this criterion are found by solving the set of normal equations

\[ X'Xb = X'Y \]

When the \( X \) variables are linearly independent (i.e., they are non-redundant, yielding an \( X'X \) matrix which is of full rank) there is a unique solution to the normal equations. Pre multiplying both sides of the matrix formula for the normal equations by the inverse of \( X'X \) gives

\[ (X'X)^{-1}X'Xb = (X'X)^{-1}X'Y \]

Or

\[ b = (X'X)^{-1}X'Y \]

This last result is very satisfying in view of its simplicity and its generality. With regard to its simplicity, it expresses the solution for the regression equation in terms just 2 matrices (\( X \) and \( Y \)) and 3 basic matrix operations, (1) matrix transposition, which involves interchanging the elements in the rows and columns of a matrix, (2) matrix multiplication, which involves finding the sum of the products of the elements for each row and column combination of two conformable (i.e., multipliable) matrices, and (3) matrix inversion, which involves finding the matrix equivalent of a numeric reciprocal, that is, the matrix that satisfies

\[ A^{-1}AA = A \]

For a matrix \( A \).

It took literally centuries for the ablest mathematicians and statisticians to find a satisfactory method for solving the linear least square regression problem. But their efforts have paid off, for it is hard to imagine a simpler solution.

With regard to the generality of the multiple regression model, its only notable limitations are that (1) it can be used to analyze only a single dependent variable, (2) it cannot provide a solution for the regression coefficients when the \( X \) variables are not linearly independent and the inverse of \( X'X \) therefore does not exist. These restrictions, however, can be overcome, and in doing so the multiple regression model is transformed into the general linear model.

(c) Extension of Multiple Regressions to the General Linear Model

One way in which the general linear model differs from the multiple regression model is in terms of the number of dependent variables that can be analyzed. The \( Y \) vector of \( n \) observations of a single \( Y \) variable can be replaced by a \( Y \) matrix of \( n \) observations of \( m \) different \( Y \) variables. Similarly, the \( b \) vector of regression coefficients for a single \( Y \) variable can be replaced by a \( b \) matrix of regression coefficients, with one
vector of $b$ coefficients for each of the $m$ dependent variables. These substitutions yield what is sometimes called the multivariate regression model, but it should be emphasized that the matrix formulations of the multiple and multivariate regression models are identical, except for the number of columns in the $Y$ and $b$ matrices. The method for solving for the $b$ coefficients is also identical, that is, $m$ different sets of regression coefficients are separately found for the $m$ different dependent variables in the multivariate regression model.

The general linear model goes a step beyond the multivariate regression model by allowing for linear transformations or linear combinations of multiple dependent variables. This extension gives the general linear model important advantages over the multiple and the so-called multivariate regression models, both of which are inherently univariate (single dependent variable) methods. One advantage is that multivariate tests of significance can be employed when responses on multiple dependent variables are correlated. Separate univariate tests of significance for correlated dependent variables are not independent and may not be appropriate. Multivariate tests of significance of independent linear combinations of multiple dependent variables also can give insight into which dimensions of the response variables are, and are not, related to the predictor variables. Another advantage is the ability to analyze effects of repeated measure factors. Repeated measure designs, or within-subject designs, have traditionally been analyzed using ANOVA techniques. Linear combinations of responses reflecting a repeated measure effect (for example, the difference of responses on a measure under differing conditions) can be constructed and tested for significance using either the univariate or multivariate approach to analyzing repeated measures in the general linear model.

A second important way in which the general linear model differs from the multiple regression model is in its ability to provide a solution for the normal equations when the $X$ variables are not linearly independent and the inverse of $XX$ does not exist. Redundancy of the $X$ variables may be incidental (e.g., two predictor variables might happen to be perfectly correlated in a small data set), accidental (e.g., two copies of the same variable might unintentionally be used in an analysis) or designed (e.g., indicator variables with exactly opposite values might be used in the analysis, as when both Male and Female predictor variables are used in representing Gender). Finding the regular inverse of a non-full-rank matrix is reminiscent of the problem of finding the reciprocal of 0 in ordinary arithmetic. No such inverse or reciprocal exists because division by 0 is not permitted. This problem is solved in the general linear model by using a generalized inverse of the $X'X$ matrix in solving the normal equations. A generalized inverse is any matrix that satisfies

$$AA^{-1}A = A$$
CHAPTER – 2

REGRESSION MODELS

For a matrix A.

A generalized inverse is unique and is the same as the regular inverse only if the matrix A is full rank. A generalized inverse for a non-full-rank matrix can be computed by the simple expedient of zeroing the elements in redundant rows and columns of the matrix. Suppose that an X'X matrix with r non-redundant columns is partitioned as

\[ X'X = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \]

Where \( A_{11} \) is an r by r matrix of rank r. Then the regular inverse of \( A_{11} \) exists and a generalized inverse of \( X'X \) is

\[ (X'X)^{-1} = \begin{bmatrix} A_{11}^{-1} & 0_{12} \\ 0_{21} & 0_{22} \end{bmatrix} \]

Where each 0 (null) matrix is a matrix of 0's (zeroes) and has the same dimensions as the corresponding A matrix.

In practice, however, a particular generalized inverse of \( X'X \) for finding a solution to the normal equations is usually computed using the sweep operator (Dempster, 1960). This generalized inverse, called a g2 inverse, has two important properties. One is that zeroing of the elements in redundant rows is unnecessary. Another is that partitioning or reordering of the columns of \( X'X \) is unnecessary, so that the matrix can be inverted "in place."

There are infinitely many generalized inverses of a non-full-rank \( X'X \) matrix, and thus, infinitely many solutions to the normal equations. This can make it difficult to understand the nature of the relationships of the predictor variables to responses on the dependent variables, because the regression coefficients can change depending on the particular generalized inverse chosen for solving the normal equations. It is not cause for dismay, however, because of the invariance properties of many results obtained using the general linear model.

A simple example may be useful for illustrating one of the most important invariance properties of the use of generalized inverses in the general linear model. If both Male and Female predictor variables with exactly opposite values are used in an analysis to represent Gender, it is essentially arbitrary as to which predictor variable is considered to be redundant (e.g., Male can be considered to be redundant with Female, or vice versa). No matter which predictor variable is considered to be redundant, no matter which corresponding generalized inverse is used in solving the normal equations, and no matter which resulting regression equation is used for computing predicted values on the dependent variables, the predicted values and the corresponding residuals for males and females will be unchanged. In using the general linear model, we must keep in mind that finding a particular arbitrary solution to the
normal equations is primarily a means to the end of accounting for responses on the dependent variables, and not necessarily an end in itself.

(d) Sigma-Restricted and over parameterized Model

Unlike the multiple regression model, which is usually applied to cases where the X variables are continuous, the general linear model is frequently applied to analyze any ANOVA or MANOVA design with categorical predictor variables, any ANCOVA or MANCOVA design with both categorical and continuous predictor variables, as well as any multiple or multivariate regression design with continuous predictor variables.

To illustrate, Gender is clearly a nominal level variable (anyone who attempts to rank order the sexes on any dimension does so at his or her own peril in today's world). There are two basic methods by which Gender can be coded into one or more (non-offensive) predictor variables and analyzed using the general linear model.

Sigma-restricted model (coding of categorical predictors)

Using the first method, males and females can be assigned any two arbitrary, but distinct values on a single predictor variable. The values on the resulting predictor variable will represent a quantitative contrast between males and females. Typically, the values corresponding to group membership are chosen not arbitrarily but rather to facilitate interpretation of the regression coefficient associated with the predictor variable. In one widely used strategy, cases in the two groups are assigned values of 1 and -1 on the predictor variable, so that if the regression coefficient for the variable is positive, the group coded as 1 on the predictor variable will have a higher predicted value (i.e., a higher group mean) on the dependent variable, and if the regression coefficient is negative, the group coded as -1 on the predictor variable will have a higher predicted value on the dependent variable. An additional advantage is that since each group is coded with a value one unit from zero, this helps in interpreting the magnitude of differences in predicted values between groups, because regression coefficients reflect the units of change in the dependent variable for each unit change in the predictor variable. This coding strategy is aptly called the sigma-restricted parameterization, because the values used to represent group membership (1 and -1) sum to zero.

Note that the sigma-restricted parameterization of categorical predictor variables usually leads to X'X matrices which do not require a generalized inverse for solving the normal equations. Potentially redundant information, such as the characteristics of maleness and femaleness, is literally reduced to full-rank by creating quantitative contrast variables representing differences in characteristics.
CHAPTER – 2
REGRESSION MODELS

As further illustration, consider an example where a model is specified that has 1 factor that contains 3 three levels A, B, and C. Under the sigma-restricted parameterization, the factor would be coded as follows:

This parameterization leads to the interpretation that each coefficient estimates the difference between each level and the average of the other 2 levels, i.e., the coefficient for A is the estimate of the difference between level A and the average of levels of B and C.

2.7 TYPES OF ANALYSIS

A wide variety of types of designs can be analyzed using the general linear model. In fact, the flexibility of the general linear model allows it to handle so many different types of designs that it is difficult to develop simple typologies of the ways in which these designs might differ. Some general ways in which designs might differ can be suggested, but keep in mind that any particular design can be a "hybrid" in the sense that it could have combinations of features of a number of different types of designs.

The levels or values of the predictor variables in an analysis describe the differences between the n subjects or the n valid cases that are analyzed. Thus, when we speak of the between subject design (or simply the between design) for an analysis, we are referring to the nature, number, and arrangement of the predictor variables.

Concerning the nature or type of predictor variables, between designs which contain only categorical predictor variables can be called ANOVA (analysis of variance) designs, between designs which contain only continuous predictor variables can be called regression designs, and between designs which contain both categorical and continuous predictor variables can be called ANCOVA (analysis of covariance) designs. Further, continuous predictors are always considered to have fixed values, but the levels of categorical predictors can be considered to be fixed or to vary randomly. Designs which contain random categorical factors are called mixed-model designs. Between designs may involve only a single predictor variable and therefore be described as simple (e.g., simple regression) or may employ numerous predictor variables.

Concerning the arrangement of predictor variables, some between designs employ only "main effect" or first-order terms for predictors, that is, the values for different predictor variables are independent and raised only to the first power. Other between designs may employ higher-order terms for predictors by raising the values for the original predictor variables to a power greater than 1 (e.g., in polynomial regression designs), or by forming products of different predictor variables (i.e., interaction terms). A common arrangement for ANOVA designs is the full-factorial design, in which every combination of levels for each of the categorical predictor variables is
represented in the design. Designs with some but not all combinations of levels for each of the categorical predictor variables are aptly called fractional factorial designs. Designs with a hierarchy of combinations of levels for the different categorical predictor variables are called nested designs.

These basic distinctions about the nature, number, and arrangement of predictor variables can be used in describing a variety of different types of between designs. Some of the more common between designs can now be described.

One-Way ANOVA. A design with a single categorical predictor variable is called a one-way ANOVA design. For example, a study of 4 different fertilizers used on different individual plants could be analyzed via one-way ANOVA, with four levels for the factor Fertilizer.

In general, consider a single categorical predictor variable A with 1 case in each of its 3 categories. Using the sigma-restricted coding of A into 2 quantitative contrast variables, the matrix $X$ defining the between design is

$$X = \begin{bmatrix}
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & -1 & -1
\end{bmatrix}$$

That is, cases in groups A1, A2, and A3 are all assigned values of 1 on $X_0$ (the intercept), the case in group A1 is assigned a value of 1 on $X_1$ and a value 0 on $X_2$, the case in group A2 is assigned a value of 0 on $X_1$ and a value 1 on $X_2$, and the case in group A3 is assigned a value of -1 on $X_1$ and a value -1 on $X_2$. Of course, any additional cases in any of the 3 groups would be coded similarly. If there were 1 case in group A1, 2 cases in group A2, and 1 case in group A3, the $X$ matrix would be

$$X = \begin{bmatrix}
1 & 1 & 0 \\
1 & 0 & 1 \\
1 & -1 & -1 \\
1 & 1 & 0 \\
1 & 0 & 1
\end{bmatrix}$$

where the first subscript for A gives the replicate number for the cases in each group. For brevity, replicates usually are not shown when describing ANOVA design matrices.

Note that in one-way designs with an equal number of cases in each group, sigma-restricted coding yields $X_1 \ldots X_k$ variables all of which have means of 0.

Using the over parameterized model to represent A, the $X$ matrix defining the between design is simply
These simple examples show that the $X$ matrix actually serves two purposes. It specifies (1) the coding for the levels of the original predictor variables on the $X$ variables used in the analysis as well as (2) the nature, number, and arrangement of the $X$ variables, that is, the between design.

**Main Effect ANOVA** Main effect ANOVA designs contain separate one-way ANOVA designs for 2 or more categorical predictors. A good example of main effect ANOVA would be the typical analysis performed on screening designs as described in the context of the Experimental Design section.

Consider 2 categorical predictor variables $A$ and $B$ each with 2 categories. Using the sigma-restricted coding, the $X$ matrix defining the between design is

$$
X = \begin{bmatrix}
X_0 & X_1 & X_2 & X_3 \\
A_1 & 1 & 1 & 0 & 0 \\
A_2 & 1 & 0 & 1 & 0 \\
A_3 & 1 & 0 & 0 & 1 \\
\end{bmatrix}
$$

Note that if there are equal numbers of cases in each group, the sum of the cross-products of values for the $X_1$ and $X_2$ columns is 0, for example, with 1 case in each group $(1*1)+(1*-1)+(-1*1)+(-1*-1)=0$. Using the over parameterized model, the matrix $X$ defining the between design is

$$
X = \begin{bmatrix}
X_0 & X_1 & X_2 \\
A_1B_1 & 1 & 1 & 1 \\
A_1B_2 & 1 & 1 & -1 \\
A_2B_1 & 1 & -1 & 1 \\
A_2B_2 & 1 & -1 & -1 \\
\end{bmatrix}
$$

Comparing the two types of coding, it can be seen that the overparameterized coding takes almost twice as many values as the sigma-restricted coding to convey the same information.

**Factorial ANOVA** Factorial ANOVA designs contain $X$ variables representing combinations of the levels of 2 or more categorical predictors (e.g., a study of boys and girls in four age groups, resulting in a 2 (Gender) x 4 (Age Group) design). In particular, full-factorial designs represent all possible combinations of the levels of the categorical predictors. A full-factorial design with 2 categorical predictor
variables A and B each with 2 levels each would be called a 2 x 2 full-factorial design. Using the sigma-restricted coding, the X matrix for this design would be

$$X = \begin{bmatrix} X_3 & X_1 & X_2 & X_3 \\ A_1B_1 & 1 & 1 & 1 \\ A_1B_2 & 1 & 1 & -1 \\ A_2B_1 & 1 & -1 & 1 \\ A_2B_2 & 1 & -1 & -1 \end{bmatrix}$$

Several features of this X matrix deserve comment. Note that the X1 and X2 columns represent main effect contrasts for one variable, (i.e., A and B, respectively) collapsing across the levels of the other variable. The X3 column instead represents a contrast between different combinations of the levels of A and B. Note also that the values for X3 are products of the corresponding values for X1 and X2. Product variables such as X3 represent the multiplicative or interaction effects of their factors, so X3 would be said to represent the 2-way interaction of A and B. The relationship of such product variables to the dependent variables indicates the interactive influences of the factors on responses above and beyond their independent (i.e., main effect) influences on responses. Thus, factorial designs provide more information about the relationships between categorical predictor variables and responses on the dependent variables than is provided by corresponding one-way or main effect designs.

When many factors are being investigated, however, full-factorial designs sometimes require more data than reasonably can be collected to represent all possible combinations of levels of the factors, and high-order interactions between many factors can become difficult to interpret. With many factors, a useful alternative to the full-factorial design is the fractional factorial design. As an example, consider a 2 x 2 x 2 fractional factorial design to degree 2 with 3 categorical predictor variables each with 2 levels. The design would include the main effects for each variable, and all 2-way interactions between the three variables, but would not include the 3-way interaction between all three variables. Using the overparameterized model, the X matrix for this design is

$$X = \begin{bmatrix} A_1B_1C_1 & A_1B_2C_1 & A_1B_1C_2 & A_1B_2C_2 & A_2B_1C_1 & A_2B_2C_1 & A_2B_1C_2 & A_2B_2C_2 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}$$
The 2-way interactions are the highest degree effects included in the design.

**Nested ANOVA Designs** Nested designs are similar to fractional factorial designs in that all possible combinations of the levels of the categorical predictor variables are not represented in the design. In nested designs, however, the omitted effects are lower-order effects. Nested effects are effects in which the nested variables never appear as main effects. Suppose that for 2 variables A and B with 3 and 2 levels, respectively, the design includes the main effect for A and the effect of B nested within the levels of A. The X matrix for this design using the over parameterized model is

\[
X = \begin{bmatrix}
X_0 & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 \\
A_1B_1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
A_1B_2 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
A_2B_1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
A_2B_2 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
A_3B_1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
A_3B_2 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

Note that if the sigma-restricted coding were used, there would be only 2 columns in the X matrix for the B nested within A effect instead of the 6 columns in the X matrix for this effect when the over parameterized model coding is used (i.e., columns X4 through X9). The sigma-restricted coding method is overly-restrictive for nested designs, so only the over parameterized model is used to represent nested designs.

**Balanced ANOVA** Most of the between designs discussed in this section can be analyzed much more efficiently, when they are balanced, i.e., when all cells in the ANOVA design have equal \( n \), when there are no missing cells in the design, and, if nesting is present, when the nesting is balanced so that equal numbers of levels of the factors that are nested appear in the levels of the factor(s) that they are nested in. In that case, the XX matrix (where X stands for the design matrix) is a diagonal matrix, and many of the computations necessary to compute the ANOVA results (such as matrix inversion) are greatly simplified.

**Analysis of Covariance** In general, between designs which contain both categorical and continuous predictor variables can be called ANCOVA designs. Traditionally, however, ANCOVA designs have referred more specifically to designs in which the first-order effects of one or more continuous predictor variables are taken into account when assessing the effects of one or more categorical predictor variables.

To illustrate, suppose a researcher wants to assess the influences of a categorical predictor variable A with 3 levels on some outcome, and that measurements on a continuous predictor variable P, known to covary with the outcome, are available. If the data for the analysis are
CHAPTER – 2

REGRESSION MODELS

Then the sigma-restricted X matrix for the design that includes the separate first-order effects of P and A would be

\[
\begin{array}{c|c|c|c|c}
 P & \text{Group} \\
 7 & A_1 \\
 4 & A_1 \\
 9 & A_2 \\
 3 & A_2 \\
 5 & A_3 \\
 8 & A_3 \\
\end{array}
\]

The b2 and b3 coefficients in the regression equation
\[
Y = b0 + b1X1 + b2X2 + b3X3
\]
represent the influences of group membership on the A categorical predictor variable, controlling for the influence of scores on the P continuous predictor variable. Similarly, the b1 coefficient represents the influence of scores on P controlling for the influences of group membership on A. This traditional ANCOVA analysis gives a more sensitive test of the influence of A to the extent that P reduces the prediction error, that is, the residuals for the outcome variable.

The X matrix for the same design using the overparameterized model would be

\[
X = \begin{bmatrix}
X_6 & X_1 & X_2 & X_3 \\
1 & 7 & 1 & 0 \\
1 & 4 & 1 & 0 \\
1 & 9 & 0 & 1 \\
1 & 3 & 0 & 1 \\
1 & 6 & -1 & -1 \\
1 & 8 & -1 & -1 \\
\end{bmatrix}
\]

The interpretation is unchanged except that the influences of group membership on the A categorical predictor variables are represented by the b2, b3 and b4 coefficients in the regression equation.
CHAPTER – 2

REGRESSION MODELS

Y = b0 + b1X1 + b2X2 + b3X3 + b4X4

**Separate Slope Designs** The traditional analysis of covariance (ANCOVA) design for categorical and continuous predictor variables is inappropriate when the categorical and continuous predictors interact in influencing responses on the outcome. The appropriate design for modeling the influences of the predictors in this situation is called the separate slope design. For the same example data used to illustrate traditional ANCOVA, the over parameterized X matrix for the design that includes the main effect of the three-level categorical predictor A and the 2-way interaction of P by A would be

\[
X = \begin{bmatrix}
X_0 & X_1 & X_2 & X_4 & X_5 & X_9 \\
1 & 1 & 0 & 0 & 7 & 0 \\
1 & 1 & 0 & 0 & 4 & 0 \\
1 & 0 & 1 & 0 & 0 & 9 \\
1 & 0 & 1 & 0 & 0 & 3 \\
1 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 \\
\end{bmatrix}
\]

The b4, b5, and b6 coefficients in the regression equation

Y = b0 + b1X1 + b2X2 + b3X3 + b4X4 + b5X5 + b6X6

Give the separate slopes for the regression of the outcome on P within each group on A, controlling for the main effect of A.

As with nested ANOVA designs, the sigma-restricted coding of effects for separate slope designs is overly restrictive, so only the over parameterized model is used to represent separate slope designs. In fact, separate slope designs are identical in form to nested ANOVA designs, since the main effects for continuous predictors are omitted in separate slope designs.

**Homogeneity of Slopes** The appropriate design for modeling the influences of continuous and categorical predictor variables depends on whether the continuous and categorical predictors interact in influencing the outcome. The traditional analysis of covariance (ANCOVA) design for continuous and categorical predictor variables is appropriate when the continuous and categorical predictors do not interact in influencing responses on the outcome, and the separate slope design is appropriate when the continuous and categorical predictors do interact in influencing responses. The homogeneity of slopes designs can be used to test whether the continuous and categorical predictors interact in influencing responses, and thus, whether the traditional ANCOVA design or the separate slope design is appropriate for modeling the effects of the predictors. For the same example data used to illustrate the
traditional ANCOVA and separate slope designs, the over parameterized X matrix for
the design that includes the main effect of P, the main effect of the three-level cate-
gorical predictor A, and the 2-way interaction of P by A would be

\[ X = \begin{bmatrix}
1 & 7 & 1 & 0 & 0 & 7 & 0 & 0 \\
1 & 4 & 1 & 0 & 0 & 4 & 0 & 0 \\
1 & 9 & 0 & 1 & 0 & 0 & 9 & 0 \\
1 & 3 & 0 & 1 & 0 & 0 & 3 & 0 \\
1 & 6 & 0 & 0 & 1 & 0 & 0 & 6 \\
1 & 8 & 0 & 0 & 1 & 0 & 0 & 8 
\end{bmatrix} \]

If the b5, b6, or b7 coefficient in the regression equation

\[ Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 \]

is non-zero, the separate slope model should be used. If instead all 3 of these regre-
sion coefficients are zero the traditional ANCOVA design should be used.

The sigma-restricted X matrix for the homogeneity of slopes design would be

\[ X = \begin{bmatrix}
1 & 7 & 1 & 0 & 7 & 0 \\
1 & 4 & 1 & 0 & 4 & 0 \\
1 & 9 & 0 & 1 & 0 & 9 \\
1 & 3 & 0 & 1 & 3 & 0 \\
1 & 6 & -1 & -1 & -6 & -6 \\
1 & 8 & -1 & -1 & -8 & -8 
\end{bmatrix} \]

Using this X matrix, if the b4, or b5 coefficient in the regression equation

\[ Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 \]

is non-zero, the separate slope model should be used. If instead both of these regres-
sion coefficients are zero the traditional ANCOVA design should be used.

**Mixed Model ANOVA and ANCOVA** Designs that contain random effects for one
or more categorical predictor variables are called mixed-model designs. Random ef-
fects are classification effects where the levels of the effects are assumed to be ran-
domly selected from an infinite population of possible levels. The solution for the
normal equations in mixed-model designs is identical to the solution for fixed-effect
designs (i.e., designs which do not contain Random effects. Mixed-model designs dif-
fer from fixed-effect designs only in the way in which effects are tested for signifi-
cance. In fixed-effect designs, between effects are always tested using the mean
squared residual as the error term. In mixed-model designs, between effects are tested using relevant error terms based on the co-variation of random sources of variation in the design. Specifically, this is done using Satterthwaite's method of denominator synthesis (Satterthwaite, 1946), which finds the linear combinations of sources of random variation that serve as appropriate error terms for testing the significance of the respective effect of interest. A basic discussion of these types of designs, and methods for estimating variance components for the random effects can also be found in the Variance Components and Mixed Model ANOVA/ANCOVA topic.

Mixed-model designs, like nested designs and separate slope designs, are designs in which the sigma-restricted coding of categorical predictors is overly restrictive. Mixed-model designs require estimation of the co-variation between the levels of categorical predictor variables, and the sigma-restricted coding of categorical predictors suppresses this co-variation. Thus, only the over parameterized model is used to represent mixed-model designs. It is important to recognize, however, that sigma-restricted coding can be used to represent any between design, with the exceptions of mixed-model, nested, and separate slope designs. Furthermore, some types of hypotheses can only be tested using the sigma-restricted coding (i.e., the effective hypothesis, Hocking, 1996), thus the greater generality of the over parameterized model for representing between designs does not justify it being used exclusively for representing categorical predictors in the general linear model.

2.8 CONCLUSION

Study of present chapter shows overview of Regression Models, fundamental of statistics corresponding to the Regression Models, Meaning and Types of Regression Models, Concept of General Lineal Model and Types of Analysis for Regression Models on the Base of Types of Regression Model. For the creative ones, one can even cook up new regressions, if one feels the need to use a combination of the parameters above, which people haven’t used before.
References

(a) Books


8) **Murphy (1997)**, The American Statistician, 51(2), 155-157 for ‘‘How to read the statistical methods literature: A guide for students”.

(b) Journal and Articles


7) http://www.camo.com/rt/Resources/statistical-regression-analysis.html


9) http://www.analyticsvidhya.com/blog/2015/06/establish-causality-events/