CHAPTER – V
CHAPTER - V
MATHEMATICAL MODEL

MATHEMATICAL MODEL FOR THE PROBLEM

The data taken for the analysis consists of eight different factors causing diabetes in pima Indian population. To give a mathematical study of the relationship between various factors and the total effect in the cause of the disease correlation and regression coefficients can be used. Regression is generally used to predict future values based on past values by fitting a set of points on the curve. Correlation however, is used to examine the degree to which values two or more variables behave in a similar manner. Correlation is exactly the method of determining how much alike the given variables are and linear correlation [54] can be measured using correlation coefficient. Given two variables, for example, X & Y the correlation coefficient real value \( r \in [-1,1] \). A positive number indicates a positive correlation, whereas a negative number indicates a negative correlation. Here negative correlation indicates that one variable increases while the other decreases in value. The closer the value of \( r \) is to 0, the smaller the correlation. A perfect relationship exists with a value of 1 or \(-1\), whereas no correlation exists with a value of 0. The value for \( r \) is defined as

\[
r = \frac{\Sigma (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\Sigma (x_i - \bar{X})^2 \Sigma (y_i - \bar{Y})^2}}
\]

where \( \bar{X} \) and \( \bar{Y} \) are the means for \( X \) and \( Y \), respectively.
When two data variables have a strong correlation, they are similar. Thus, the correlation coefficient can be used to define similarly for clustering or classification. Here, multiple correlation coefficients [54] are evaluated to find the relationship between each diabetes causing variable and their correlation with the variable nine which is the diabetes response (diabetes acquired or not). The SPSS statistical software is used for this and the values are tabulated.

STATISTICAL INVESTIGATION FROM THE COLLECTED DATA SAMPLE

The following statistical coefficients [55] have been calculated for the study from the 8 input parameters.

(1) Spearman’s correlation coefficient & Pearson’s correlation coefficient

(2) Yule’s coefficient for association of attributes

(3) Posterior probabilities of the attributes

5.1. CORRELATION ANALYSIS

Table [5.1] I Correlation Coefficients among Attributes

<table>
<thead>
<tr>
<th>Spearman’s Coefficient</th>
<th>Var 1</th>
<th>Var 2</th>
<th>Var 3</th>
<th>Var 4</th>
<th>Var 5</th>
<th>Var 6</th>
<th>Var 7</th>
<th>Var 8</th>
<th>Var 9 RESPO</th>
<th>DIAB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Var 1</td>
<td>1.000</td>
<td>1.000</td>
<td>.999</td>
<td>.999</td>
<td>1.000</td>
<td>.986</td>
<td>.886</td>
<td>.960</td>
<td>.961</td>
<td></td>
</tr>
<tr>
<td>Var 2</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.986</td>
<td>.886</td>
<td>.961</td>
<td>.960</td>
<td></td>
</tr>
<tr>
<td>Var 3</td>
<td>.999</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.986</td>
<td>.886</td>
<td>.961</td>
<td>.961</td>
<td></td>
</tr>
<tr>
<td>Var 4</td>
<td>.999</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.986</td>
<td>.886</td>
<td>.961</td>
<td>.960</td>
<td></td>
</tr>
<tr>
<td>Var 5</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.986</td>
<td>.886</td>
<td>.961</td>
<td>.961</td>
<td></td>
</tr>
<tr>
<td>Var 6</td>
<td>.986</td>
<td>.986</td>
<td>.986</td>
<td>.986</td>
<td>.987</td>
<td>1.000</td>
<td>.949</td>
<td>.992</td>
<td>.998</td>
<td></td>
</tr>
<tr>
<td>Var 7</td>
<td>.886</td>
<td>.886</td>
<td>.886</td>
<td>.886</td>
<td>.886</td>
<td>.949</td>
<td>1.000</td>
<td>.978</td>
<td>.961</td>
<td></td>
</tr>
<tr>
<td>Var 8</td>
<td>.960</td>
<td>.961</td>
<td>.961</td>
<td>.961</td>
<td>.961</td>
<td>.992</td>
<td>.978</td>
<td>1.000</td>
<td>.988</td>
<td></td>
</tr>
<tr>
<td>Var 9</td>
<td>.961</td>
<td>.960</td>
<td>.961</td>
<td>.960</td>
<td>.961</td>
<td>.984</td>
<td>.961</td>
<td>.988</td>
<td>1.000</td>
<td></td>
</tr>
</tbody>
</table>
The correlation between each of the variable with every other variable has been found, particularly the relation of each variable with variable 9 (diabetic response) is around .9 or unity which shows that

- Each factor contributes to the cause of diabetes.
- Each factor is related to every other factor which shows that all 8 parameters are interdependent for causing diabetes mellitus
- The maximum correlation between variable 8 (AGE) and diabetic response indicates that age factor plays major role in diabetic disorders.

## 5.2. ASSOCIATION OF ATTRIBUTES

### II Theory of Association of Attributes

From the theory of attributes two attributes are said to be associated, if they appear together in a large number of cases than expected of them. When two attributes are present or absent together in the data, the actual frequency is more than the expected frequency it is called positive association and if less
than expected frequency it is called negative association and if no association occurs they are said to be independent.

**Yule’s coefficient** [56] of association is tested in the data if Y is found to be +1 there is perfect association and if \( Y = -1 \) there is perfect negative association. Nine square table can be used.

### Table [5.2] Yule’s Coefficients for Age below 21 years

<table>
<thead>
<tr>
<th>Age</th>
<th>Diabetic</th>
<th>Non-diabetic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age below 22 years</td>
<td>(AB) 5</td>
<td>(aB) 59</td>
<td>(B) 64</td>
</tr>
<tr>
<td>Not below 22 years</td>
<td>(AB) 263</td>
<td>(aB) 441</td>
<td>(B) 704</td>
</tr>
<tr>
<td>Total</td>
<td>(A) 268</td>
<td>(B) 500</td>
<td>(B) 768</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
AB &= 5 \\
\alpha \beta &= 59 \\
\alpha \beta &= 263 \\
\alpha \beta &= 441
\end{align*}
\]

Yule’s coefficient 
\[
Y = \frac{(AB)(\alpha \beta) - (A \beta)(\alpha B)}{(AB)(\alpha \beta) + (A \beta)(\alpha B)}
\]

\( Y = -0.7511 \)
- The attack of diabetes is negatively associated with female of age below 22 years.

### Table [5.3] Yule’s Coefficients for Age above 40 years

<table>
<thead>
<tr>
<th>Age</th>
<th>Diabetic</th>
<th>Non-diabetic</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age above 40 years</td>
<td>107(AB)</td>
<td>98(\alpha B)</td>
<td>205(B)</td>
</tr>
<tr>
<td>Not above 40 years</td>
<td>161(A\beta)</td>
<td>402(\alpha \beta)</td>
<td>563(B)</td>
</tr>
<tr>
<td>Total</td>
<td>268</td>
<td>500</td>
<td>768</td>
</tr>
</tbody>
</table>

Yule’s coefficient 
\[
Y = \frac{(AB)(\alpha \beta) - (A \beta)(\alpha B)}{(AB)(\alpha \beta) + (A \beta)(\alpha B)}
\]

\( Y = 0.46326 \)
- The attack of diabetes is positively associated with female of age above 40 years.
5.3. POSTERIOR PROBABILITY METHOD

5.3.1. CONDITIONAL PROBABILITY

This article defines some terms which characterize probability distributions of two or more variables.

Conditional probability is the probability of some event A, given the occurrence of some other event B. Conditional probability is written \( P(A|B) \), and is read "the probability of A, relative to B".

Joint probability is the probability of two events in conjunction. That is, it is the probability of both events occurring together. Then probability of both A and B to occur is \( P(A \cap B) \) or \( P(A, B) \).

In these definitions, it is noted that there need not be a causal or temporal relation between A and B. A may precede B or vice versa or they may happen at the same time. A may cause B or vice versa or they may have no causal relation at all. Notice, however, that casual and temporal relations are informal notions, not belonging to the probabilistic framework. They may apply in some examples, depending on the interpretation given to events.

**Definition of conditional probability**

Given a probability space \((\Omega,F,P)\) and two events \(A,B \in F\) with \(P(B)>0\), the conditional probability of A given B is defined by

\[
P(A|B) = \frac{P(A \cap B)}{P(B)}
\]

If \(P(B) = 0\) then \(P(A \mid B)\) is undefined.
Conditioning of probabilities, i.e. updating them to take account of (possibly new) information, may be achieved through Bayes' theorem which states that, if \( E_1, E_2, \ldots, E_n \) are any \( n \) mutually disjoint events and \( A \) is any required event then

\[
P\left( \frac{E_1}{A} \right) = \frac{P(E_1)P\left( \frac{A}{E_1} \right)}{\sum_{i=1}^{n} P(E_i)P\left( \frac{A}{E_i} \right)}
\]

### 5.3.2. STATISTICAL INDEPENDENCE

Two random events \( A \) and \( B \) are statistically independent if and only if

\[
P(A \cap B) = P(A) P(B)
\]

Thus, if \( A \) and \( B \) are independent, then the probability of both of them to occur can be expressed as a simple product of their individual probabilities. Equivalently, for two independent events \( A \) and \( B \),

\[
P(A \mid B) = P(A)
\]

and

\[
P(B \mid A) = P(B)
\]

In other words, if \( A \) and \( B \) are independent, then the conditional probability of \( A \), given \( B \) is simply the individual probability of \( A \) alone; likewise, the probability of \( B \) given \( A \) is simply the probability of \( B \) alone.
5.3.3. MUTUAL EXCLUSIVITY

Two events A and B are mutually exclusive if and only if

\[ P(A \cap B) = 0 \]

as long as

\[ P(A) \neq 0 \]

and

\[ P(B) \neq 0 \]

Then

\[ P(A \mid B) = 0 \]

and

\[ P(B \mid A) = 0 \]

\[ P(B \mid A) = P(A \mid B) \frac{P(B)}{P(A)} \]

In other words, the probability of A happening, given that B happens, is nil since A and B cannot both happen in the same situation; likewise, the probability of B happening, given that A happens, is also nil.

5.3.4. APPLICATION TO PROBLEM OF DIABETES

In the following constructed but realistic situation, the difference between \( P(A \mid B) \) and \( P(B \mid A) \) may be surprising, but is at the same time obvious.

In order to identify individuals having a serious disease in an early curable form, one may consider screening a large group of people. While the
benefits are obvious, an argument against such screenings is the disturbance caused by false positive screening results: If a person not having the disease is incorrectly found to have it by the initial test, they will most likely be quite distressed until a more careful test shows that they do not have the disease. Even after being told they are well, their lives may be affected negatively.

The magnitude of this problem is best understood in terms of conditional probabilities.

Suppose 1% of the group suffer from the disease, and the rest are well. Choosing an individual at random,

\[ P(\text{disease-DIABETES MELLITUS}) = 1\% = 0.01 \text{ and } P(\text{well}) = 99\% = 0.99. \]

Suppose that when the screening test is applied to a person not having the disease, there is a 1% chance of getting a false positive result, i.e.

\[ P(\text{positive} \mid \text{well}) = 1\%, \text{ and } P(\text{negative} \mid \text{well}) = 99\%. \]

Finally, suppose that when the test is applied to a person having the disease, there is a 1% chance of a false negative result, i.e.

\[ P(\text{negative} \mid \text{disease-DIABETES MELLITUS}) = 1\% \text{ and } P(\text{positive} \mid \text{disease-DIABETES MELLITUS}) = 99\%. \]

Now, calculation shows that:

\[ P(\text{well} \land \text{negative}) = P(\text{well}) \times P(\text{negative} \mid \text{well}) = 99\% \times 99\% = 98.01\% \]

is the fraction of the whole group being well and testing negative.

\[ P(\text{disease} \land \text{positive}) = P(\text{disease}) \times P(\text{positive} \mid \text{disease}) = 1\% \times 99\% = 0.99\% \]

is the fraction of the whole group being ill and testing positive.
\[ P(\text{well} \cap \text{positive}) = P(\text{well}) \times P(\text{positive} | \text{well}) = 99\% \times 1\% = 0.99\% \]
is the fraction of the whole group having false positive results.

\[ P(\text{disease} \cap \text{negative}) = P(\text{disease}) \times P(\text{negative} | \text{disease}) = 1\% \times 1\% = 0.01\% \]
is the fraction of the whole group having false negative results.

Furthermore,

\[ P(\text{positive}) = P(\text{well} \cap \text{positive}) + P(\text{disease} \cap \text{positive}) = 0.99\% + 0.99\% = 1.98\% \]
is the fraction of the whole group testing positive.

\[ P(\text{disease} | \text{positive}) = \frac{P(\text{disease} \cap \text{positive})}{P(\text{positive})} = \frac{0.99\%}{1.98\%} = 50\% \]
is the probability that you actually have the disease if you tested positive.

In this example, it should be easy to relate to the difference between

\[ P(\text{positive} | \text{disease-DIABETES MELLITUS}) = 99\% \text{ and } P(\text{disease} | \text{positive-DIABETES MELLITUS}) = 50\% \]: The first is the conditional probability that you test positive if you have the disease; the second is the conditional probability that you have the disease if you test positive. With the numbers chosen here, the last result is likely to be deemed unacceptable: Half the people testing positive are actually false positives.

**III POSTERIOR PROBABILITY ANALYSIS**

**Hypothesis Testing**

Hypothesis testing attempts to find a model that explains the observed data by first creating a hypothesis and then testing the hypothesis with the data. The actual data itself drive the model creation. The hypothesis usually is
verified by examining a data sample. If the hypothesis holds for the sample, it is assumed to hold for the population in general. Given a population, the initial (assumed) hypothesis to be tested, $H_0$, is called the null hypothesis. Rejection of the null hypothesis causes another hypothesis, $H_1$, called the alternative hypothesis, to be made. Here null hypothesis is used. In this data sample, 8 attributes influence the diabetes. So $2^8$ subsets can be formed. Let the NULL HYPOTHESIS be formulated as follows.

“All the eight factors contribute to the existence of diabetes and they equally contribute to the cause of Diabetes”

$$P(x_1).P(x_2)\ldots\ldots P(x_8) = P(x_1 \cap x_2 \cap x_3\ldots\ldots \cap x_8)$$

$P(E_i|A)$ be the posterior probability of a diabetic patient due to one of the factors. Let $E_i$ where $E_1, E_2, E_3, \ldots, E_8$ be a set of independent events which are factors causing diabetes in PIMA INDIANS like PRG(X1), PLASMA(X2), BP(X3), THICK(X4), INSULIN(X5), BMI(X6), DPF(X7), AGE(X8), and DIABETIC RESPONSE(X9). Let $A$ be the main event under consideration. Now our aim is to check by Bayesian principle.

As a first step the mean for all 768 values have been found

<table>
<thead>
<tr>
<th>Table [5.4] Probability Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARAMETERS</td>
</tr>
<tr>
<td>MEAN</td>
</tr>
<tr>
<td>PROBABILITIES</td>
</tr>
</tbody>
</table>

Let $A$ be the main event under consideration (diabetic or not) now to check by Bayesian Principle.
The conditional probabilities are,

\[
P\left( \frac{E_i}{A} \right) = \frac{P(E_i)P(A)}{\sum_{i=1}^{8} P(E_i)P(A)}
\]

Similarly,

Table [5.5] Conditional Probability Parameters

<table>
<thead>
<tr>
<th>PARAMETERS (X)</th>
<th>X1</th>
<th>X2</th>
<th>X3</th>
<th>X4</th>
<th>X5</th>
<th>X6</th>
<th>X7</th>
<th>X8</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN x</td>
<td>3.84</td>
<td>120.8</td>
<td>69.1</td>
<td>20.5</td>
<td>79.9</td>
<td>31.9</td>
<td>.47</td>
<td>33.2</td>
</tr>
<tr>
<td>CONDITIONAL PROBABILITIES P(X)</td>
<td>0.0106</td>
<td>.3367</td>
<td>.1923</td>
<td>.0570</td>
<td>.2224</td>
<td>.0891</td>
<td>.00128</td>
<td>.0923</td>
</tr>
</tbody>
</table>

\[
P[x_1 \cap x_2 \cap x_3 \ldots \cap x_n] \Rightarrow x_1 \cdot x_2 \cdot x_3 \ldots x_n = 0 \text{ (independent)}
\]

Product \( = p(x_1) \times p(x_2) \times p(x_3) \ldots \ldots p(x_8) = 5.959 \times 10^{-10} \sim = 0 \)

Sum \( = p(x_1) + p(x_2) + p(x_3) + \ldots \ldots + p(x_8) = .99 \)

From last parameter of data table (all factors causing diabetes)

\[
p(x_1) \cap p(x_2) \cap p(x_3) \ldots \cap p(x_8) = 0
\]

5.4. RESULTS AND DISCUSSIONS

The work aimed at a clinical study of actual patients acquiring diabetes mellitus both by hereditary factors as well as by phenotypic characteristics. Data have been collected from about 768 Indian Origin females who were tested for the presence of diabetes mellitus of which 268 were tested to be positive. The eight major factors which were reasonable for the cause of diabetes were experimentally measured. They are PRG (No of times pregnant),
PLASMA (Plasma glucose concentration in Salvia), BP (Diastolic blood pressure), THICK (Forceps skin fold thickness), INSULIN (Two hours serum insulin), Body (Body mass index (weight/height), PEDIGREE (Diabetes pedigree function), AGE (in years), RESPONSE (1 : Diabetic 0 : Non Diabetic). A statistical model [57] has been developed as explained below and significant conclusions about the dominance of each factor in the cause of the disorder and statistical coefficients are calculated to substantiate the conclusions.

The correlation between each of the variable with every other variable has been found, particularly the relation of each variable with variable 9 (diabetic response) is around .9 or unity which shows that

- Each factor contributes to the cause of diabetes.
- Each factor is related to every other factor as per the findings of statistical modeling and all the eight parameters are interdependent for causing diabetes mellitus
- The maximum correlation between variable 8 (AGE) and diabetic response factor was found. Hence age is a major factor in diabetes.

5.4.1. NULL HYPOTHESIS WAS FORMULATED AS FOLLOWS

“All the eight factors contribute to the existence of diabetes and they equally contribute to the cause of Diabetes”.
• No individual has all 8 factors. So every factor is a positive contributor to the diabetes disorder independently which proves our Null hypothesis.

• From Yule’s coefficient, the attack of diabetes is negatively associated with females of age below 22 years, and positively associated with age above 40 years.

• From theory of probabilities

\[
\text{product} = p(x_1) \times p(x_2) \times p(x_3) \cdots \cdots \cdots \times p(x_8) = 5.959 \times 10^{-10} = 0
\]

\[
\text{sum} = p(x_1) + p(x_2) + p(x_3) + \cdots + p(x_8) = .99
\]

From last parameter of data table (all factors causing diabetes)

\[
p(x_1) \cap p(x_2) \cap p(x_3) \cap \cdots \cap p(x_8) = 0
\]

Since \(p(e_1|a) \cap (e_2|a) \cdots \cdots \cap (e_8|a) = 0\) no individual has all 8 factors. So, every factor is a positive contributor to the diabetes disorder independently, which proves our null hypothesis and also interdependently by correlation analysis except for the case of an individual having all the eight factors.