Diabetic Neuropathy Prediction by Logistic Model

P. Sankar¹, T. Subhramaniyan², M. S. Paulraj³, J. Jayanthi⁴, M. G. Ragunathan⁴

¹Assistant Professor, Department of Mathematics, Guru Nanak College, Chennai – 42, Tamil Nadu, India
²Associate Professor, Department of Mathematics, Guru Nanak College, Chennai – 42, Tamil Nadu, India
³Associate Professor, Department of Mathematics, A. M. Jain College, Chennai – 114, Tamil Nadu, India
⁴Associate Professor, Department of Advanced Zoology and Biotechnology, Guru Nanak College, Chennai – 42, Tamil Nadu, India

Abstract - Diabetic peripheral neuropathy (DPN) is the most common complication of diabetes mellitus, and can be related to Type1 as well as Type2 diabetes. This neurological impairment is not known or understood, hence the treatments are still empirical and not efficient as stated by the physicians. Logistic modelling of clinical DPN offers a power tool in order to understand diabetes-mediated peripheral nerve injury. Logistic modelling clinical relevance has two draw backs – Firstly the prevalence of Type2 as compared to Type1 in adults and Secondly the lack of morphological changes in peripheral nerves. Many studies have contributed to a better pathophysiological and pharmacological understanding of the DPN.

Keywords - Diabetes, Diabetic Neuropathy, Mathematical models, Odds Ratio, Logistic Distribution.

Overview

Binary logistic regression is a method which estimates the probability of a character given the values of explanatory variables, in this case a single categorical variable: \( p = \Pr(Y = 1/X = x) \). Suppose a physician is interested in estimating the proportion of neuropathy diabetic persons in a population; one should know that all sections of the population do not have equal probability of ‘success’, i.e. being neuropathy diabetic. Older population, population with hypertension, individuals with diabetes incidence in family are more likely to have neuropathy diabetes. If the predictor variable X to be any of the risk factor that might contribute to the disease. Probability of success will depend on levels of the risk factor.

I. INTRODUCTION

Insulin is a hormone that is needed to convert sugar, starches and other food into energy needed for daily life. But if the body does not produce or properly use insulin a disease diabetes is attained. The prevalence of non-insulin dependent diabetes mellitus is known to be very high among Asian Indians [1,2] as well as within the Indian sub-continent [3,4]. Non-insulin dependent diabetes mellitus in India differs from that seen among Europeans in several aspects [5]: (i) the onset of diabetes occurs at a younger age [6]. (ii) obesity is less common [4] and (iii) genetic factors appear to be stronger [7,8]. Patients with diabetes may experience a wide range of neurological disorders that can involve different types of sensory and motor nerves. Diabetes peripheral neuropathy (DPN) is one neuromuscular disorder that can occur in patients with diabetes [9]. This usually occurs within 10 years of the onset of the disease [10]. Diabetic foot ulcers are largely dependent on the sensory impairment caused by diabetic neuropathy, which is a serious complication that can lead to amputation. Other problems in diabetic neuropathy and neuropathic pain are caused by dysfunction of the sympathetic nervous system and can cause many problems for the patient and therapist [11]. Early in my studies neuropathy detection and prediction [12] is of great importance in the prevention of complications such as pain, loss of sensation, foot ulcers, gangrene, and amputations.

A. Design and Data base of the Study

The data used in the secondary form of medical records of patients with diabetes mellitus in Dr. Mohan diabetes specialities centre, Chennai. This centre has admitting patients with diabetes for treatment and providing prevalence care for complications. In addition, this centre carries out diabetes research. This research uses one response variable the NDP status(Y) and 10 variables predictors i.e. the age of patients(\( x_1 \)), gender(\( x_2 \)), BMI(\( x_3 \)), father(\( x_4 \)), mother(\( x_5 \)), sibling(\( x_6 \)), smoking(\( x_7 \)), alcohol(\( x_8 \)), hypertension(\( x_9 \)) and activity last visit(\( x_{10} \)). In this paper logistic regression equation and numerical approximations are used to monitor the size of population of neuropathy diabetes with and without complications. SPSS version 24 (IBM Corporation) was used for all statistical analyses. A two sided ‘p’ value of 0.05 was considered statistically significant.
B. Likelihood Function

Consider a family of probability distributions defined by a set of parameters $\phi$. The distributions may be probability mass functions or probability density function. Suppose that we have a random sample drawn from a fixed but unknown member of this family. The random sample is a training set of $n$ examples $x_1, x_2, \cdots, x_n$. An example may also be an observation, an outcome, an instance, or a data point.

We assume that $x_1, x_2, \cdots, x_n$ are independent, therefore the probability of the set is the product of the probabilities of the individual examples:

$$f(x_1, x_2, \cdots, x_n; \phi) = \prod_{i=1}^{n} f_i(x_i; \phi)$$

Here $\phi$ as fixed and the examples $x_i$ as unknown, or varying. However, we can think of the training data as fixed and consider alternative parameter values. This is the point of view behind the definition of likelihood function:

$$L(\phi; x_1, x_2, \cdots, x_n) = f(x_1, x_2, \cdots, x_n; \phi).$$

Note that if $f(x; \phi)$ is a probability mass function, then the likelihood is always less than one, but if $f(x; \phi)$ is probability density function, then the likelihood can be greater than one, since densities can be greater than one. The principle of maximum likelihood says that given the training data, we use the distribution $f(x; \hat{\phi})$ that gives the greatest possible probability to the training data. Therefore, $\hat{\phi} = \arg \max_\phi L(\phi; x_1, x_2 \cdots x_n).$

The value $\hat{\phi}$ is called the maximum likelihood estimator (MLE) of $\phi$.

C. Maximizing likelihood function

For finding a maximum likelihood estimator, consider estimating the parameter of a Bernoulli distribution. A random variable with this distribution is a formalisation of a coin toss. The value of the random variable is 1 with probability $\phi$ and 0 with probability $1 - \phi$. Let $Y$ be a Bernoulli random variable and let $X$ be an outcome of $Y$. We have

$$P(Y = x) = \begin{cases} \phi & \text{if } x = 1 \\ 1 - \phi & \text{if } x = 0 \end{cases}$$

For mathematical convenience we write $P(Y)$ as

$$P(Y = x) = \phi^x (1 - \phi)^{1-x}.$$ 

Suppose that the training data are $x_1, x_2, \cdots, x_n$ where each $x_i \in \{0,1\}$.

The likelihood function is

$$L(\phi; x_1, x_2, \cdots, x_n) = f(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} P(Y = x_i) = \phi^m (1 - \phi)^{n-m} \quad \text{where} \quad m = \sum_{i=1}^{n} x_i.$$ 

The maximization is performed over the possible values $0 \leq \phi \leq 1$.

We can get the maximization by setting the derivatives with respect to $\phi$ equal to zero. The derivative is

$$\frac{d}{d\phi} \left[ \phi^m (1 - \phi)^{n-m} \right] = m\phi^{m-1}(1 - \phi)^{n-m} + \phi^m(n-m)(1 - \phi)^{n-m-1}(-1) = \phi^{m-1}(1 - \phi^{n-m-1})[m(1 - \phi) - (n-m)\phi]$$

Which has solutions $\phi = 0$, $\phi = 1$, and $\phi = \frac{m}{n}$.

The solution which is a maximum is clearly $\phi = \frac{m}{n}$ while $\phi = 0$ and $\phi = 1$ are minima. So we have the maximum likelihood estimate $\hat{\phi} = \frac{m}{n}$. The log likelihood function is simply the logarithm of the likelihood function. Because logarithm is a monotonically strictly increasing function, maximizing the log likelihood is precisely equivalent to maximizing the likelihood, and also to minimizing the negative log likelihood.

D. Conditional likelihood

The notation $p(y/x)$ is the conditional probabilities $p(Y = y | X = x)$ where $Y$ and $X$ are random variables. The conditional likelihood of $\phi$ given data $x$ and $y$ is $L(\phi; y/x) = p(y/x)$

$$= f(y/x; \phi)$$

Intuitively, $Y$ follows a probability distribution that is different for different $x$. Given training data consisting of $(x_i, y_i)$, the principle of maximum conditional likelihood says to choose a parameter estimate $\hat{\phi}$ that maximizes the
product \( \prod_{i=1}^{n} f(y_i / x_i; \theta) \). Assume that \( y_i \) are independent and \( x_i \) are no need to independent. For a fixed value of \( x \), \( \hat{\phi} \) can be used to compute probabilities for alternative values \( y \) of \( Y \).

Then the conditional likelihood function is

\[
L(\phi; y / x) = p(y / x) = \prod_{i=1}^{n} P(Y = y_i / X = x_i) = \prod_{i=1}^{n} p(x_i; \phi)^y_i (1 - p(x_i; \phi))^{1-y_i}
\]

### E. Odds Ratio

An odds ratio (OR) is a measure of association between a certain property \( A \) and a second property \( B \) in a population. In particular, it tells you how the presence or absence of property \( A \) has an effect on the presence or absence of property \( B \). The Odds ratio is also used to figure out if a particular exposure is a risk factor for a particular outcome and to compare the various risk factors for that outcome. The calculation of the odds ratio is quite simple. The formula is as follows:

<table>
<thead>
<tr>
<th>Event Happen</th>
<th>Standard Treatment</th>
<th>New Treatment</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td></td>
<td>d</td>
</tr>
</tbody>
</table>

Odds Ratio = \( \frac{(a / b)}{(c / d)} = \frac{ad}{bc} \).

OR=1, Exposure does not affect odds of outcome.
OR>1, Exposure associated with higher odds of outcome.
OR<1, Exposure associated with lower odds of outcome.

The great value of the odds ratio is that it is simple to calculate, very easy to interpret, and provides results upon which clinical decisions can be made. Furthermore, it is sometimes helpful in clinical situations to be able to provide the patient with information on the odds of one outcome versus another.

### F. Logistic Regression

To sum up, assume that \( Y \) is a binary (Bernoulli) outcome and \( X \) is a real valued vector, we want to model the conditional probability \( p(Y = 1 / x) \) as a linear function of \( X \) and is defined by

\[
p(Y = 1 / x; \alpha, \beta) = \sigma(\alpha + \sum_{i=1}^{n} \beta_i x_i)
\]

Where \( \sigma(x) = \frac{1}{1 + e^{-x}} \) is the non linear function.

This model is called logistic regression.

The logistic regression is also of the form

\[
\log \left( \frac{p}{1 - p} \right) = \alpha + \sum_{i=1}^{n} \beta_i x_i
\]

Where \( p \) is a probability that a case is in a particular category, \( e \) is a base of natural logarithms (approx 2.72), \( \alpha \) is a constant of the equation and \( \beta \) is a coefficient of the predictor variables.

The ratio \( \frac{p}{1 - p} \) is called the odds of the event \( Y = 1 \) given \( X = x \), and \( \log \left( \frac{p}{1 - p} \right) \) is called the log odds. Since the probabilities range between 0 and 1, odds range between 0 and \( +\infty \) and log odds range unbounded between \(-\infty \) and \( +\infty \).

### G. Likelihood Function for Logistic Regression

Since logistic regression predicts probabilities, rather than just classes, we can fit it using likelihood. For each training data point \( \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \), we have a vector of features, \( x_i \), and an observed class, \( y_i \). The probability of that class was either \( p \), if \( y_i = 1 \), or \( 1 - p \), if \( y_i = 0 \).

The likelihood is then

\[
L(\beta_0, \beta) = \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}
\]

The log likelihood turns products into sums

\[
\log L = \ell = \sum_{i=1}^{n} y_i \log p(x_i)
\]
+ \sum_{i=1}^{n} (1 - y_i) \log(1 - p(x_i))
= \sum_{i=1}^{n} \log(1 - p(x_i)) + \sum_{i=1}^{n} y_i \log \frac{p(x_i)}{1 - p(x_i)}
= \sum_{i=1}^{n} \log(1 - p(x_i)) + \sum_{i=1}^{n} y_i (\beta_0 + x_i \beta)

\ell = \sum_{i=1}^{n} \log(1 + e^{\beta_0 + x_i \beta}) + \sum_{i=1}^{n} y_i (\beta_0 + x_i \beta)

Typically, to find the maximum likelihood estimates, take the partial derivative of $\log L$ with respect to one component of $\beta$, say $\beta_j$

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} e^{\beta_0 + x_i \beta} x_{ij} + \sum_{i=1}^{n} y_i x_{ij}
= \sum_{i=1}^{n} (y_i - p(x_i; \beta_0, \beta)) x_{ij}
$$

where $x_{ij}$ is the value of $j^{th}$ feature of the $i^{th}$ training example.

We are not going to set this to zero and solve exactly. We can however approximately solve it numerically.

### H. Results and Discussion

As mentioned earlier, the dataset used in the study was obtained from the Dr. Mohan diabetes specialties centre database. From the baseline survey, it was found that the mean age of the neuropathic diabetes was 64.93 years. 33% were smokers, 24.55% were alcoholic during the first visit of the treatment. Also, it is inferred that mean time staying at hospital for treatment was 14.01 months, height 163 cm, weight 69.9 kg. Body mass index (BMI) was calculated as weight in kilograms divided by the square of height in meters ($kg/m^2$), Waist 26.3 inches, Hip 98.61 cms, Diastolic Blood Pressure 81.01 and Systolic Blood pressure 134.33. The following results shows that before(first visit) and after treatment (last visit), HbA1c in first visit was 9.08 and in last visit was 8.23, Cholesterol in first visit was 175.99 and in last visit 149.70, HDL cholesterol in first visit was 40.36 and in last visit 38.30, LDL cholesterol in first visit was 102.5 and in last visit 82.10, Urea in first visit was 26.64 and in the last visit 27.85, Creatinine in first visit was 0.88 and in the last visit 0.93. It is medically proved that if Great toe right or left is greater than are equal to 20 then the patient is called diabetic under neuropathy. In our study, there were 1017 patients without neuropathy, 376 patients with neuropathy. We chose the “important” feature sets and ordered the feature sets by the mean of importance. After fixing the tentative attributes, the selected features were age, gender, diabetic father, diabetic mother, sibling, smoking habit, alcohol consuming habit, hypertension and activity. These features were ordered by the mean of importance.

### II. Determination of results of a Neuropathy Diabetic

#### A. Odds Ratio

![Table 1](http://www.ijmttjournal.org)

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Neuropathy</th>
<th>Non-Neuropathy</th>
<th>Total</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Male</td>
<td>255</td>
<td>645</td>
<td>900</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>121</td>
<td>372</td>
<td>493</td>
</tr>
<tr>
<td>Father’s Diabetics</td>
<td>Yes</td>
<td>122</td>
<td>443</td>
<td>565</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>254</td>
<td>574</td>
<td>828</td>
</tr>
<tr>
<td>Mother’s Diabetics</td>
<td>Yes</td>
<td>112</td>
<td>411</td>
<td>523</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>264</td>
<td>606</td>
<td>870</td>
</tr>
<tr>
<td>Sibling</td>
<td>Yes</td>
<td>202</td>
<td>476</td>
<td>678</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>174</td>
<td>541</td>
<td>715</td>
</tr>
<tr>
<td>Smoking</td>
<td>Yes</td>
<td>100</td>
<td>203</td>
<td>303</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>276</td>
<td>814</td>
<td>1090</td>
</tr>
<tr>
<td>Alcohol</td>
<td>Yes</td>
<td>82</td>
<td>252</td>
<td>334</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>294</td>
<td>765</td>
<td>1059</td>
</tr>
</tbody>
</table>

The result of an odds ratio is interpreted as follows

The male patients who will get the neuropathy diabetics 1.21 times more often than female respondents. Based on these results to conclude that all males patients diagnosed by neuropathy diabetic be prescribed the new drug. Also the result conclude that the odds ratio of smoking is 1.45 times greater for men than for women. The hypothesis that cigarette smoking is associated with diabetic neuropathy should be investigated further, both prospectively and in a more representative population. Based on this study, smoking programs be targeted toward men. The result yields that the odds ratio of alcohol is 0.846 times greater for men than for women. Alcohol is a readily available toxic chemical that can yield pleasurable experience or disastrous effects that can cause enormous suffering. Based on this
study, perhaps alcohol cessation programs should be targeted toward men.

B. Model Estimation Method

<table>
<thead>
<tr>
<th>Step</th>
<th>Observed</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Neuro</td>
<td>Percentage correct</td>
</tr>
<tr>
<td></td>
<td>Others</td>
<td>Neuropathy</td>
</tr>
<tr>
<td>1</td>
<td>1017</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Neuropathy</td>
<td>376</td>
</tr>
<tr>
<td></td>
<td>Over all Percentage</td>
<td>73.0</td>
</tr>
</tbody>
</table>

This yields the results when the predictors Neuropathy and others are included. Later a classification table which shows how the classification error rate has changed from the original 73.0%. By adding the variables we can now predict with 76.8% accuracy.

C. Omnibus Tests of Model Coefficient

<table>
<thead>
<tr>
<th>Variables</th>
<th>Co Efficiencies (β)</th>
<th>S.E Of Estimates (β)</th>
<th>Wald Test df</th>
<th>Significance (p)</th>
<th>exp (β)</th>
<th>95% C.I for exp (β)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age(z)</td>
<td>0.112</td>
<td>0.009</td>
<td>166.471</td>
<td>0.000</td>
<td>1.118</td>
<td>1.100 - 1.138</td>
</tr>
<tr>
<td>Gender</td>
<td>0.329</td>
<td>0.163</td>
<td>4.070</td>
<td>0.044</td>
<td>1.390</td>
<td>1.009 - 1.914</td>
</tr>
<tr>
<td>BMI</td>
<td>0.025</td>
<td>0.017</td>
<td>2.329</td>
<td>0.127</td>
<td>1.026</td>
<td>0.993 - 1.059</td>
</tr>
<tr>
<td>Father</td>
<td>-0.209</td>
<td>0.145</td>
<td>2.097</td>
<td>0.148</td>
<td>0.811</td>
<td>0.614 - 1.077</td>
</tr>
<tr>
<td>Mother</td>
<td>-0.347</td>
<td>0.151</td>
<td>5.296</td>
<td>0.021</td>
<td>0.707</td>
<td>0.526 - 0.950</td>
</tr>
<tr>
<td>Smoking</td>
<td>0.626</td>
<td>0.192</td>
<td>10.680</td>
<td>0.001</td>
<td>1.870</td>
<td>1.285 - 2.722</td>
</tr>
<tr>
<td>Alcohol</td>
<td>-0.443</td>
<td>0.195</td>
<td>5.150</td>
<td>0.023</td>
<td>0.642</td>
<td>0.438 - 0.941</td>
</tr>
<tr>
<td>Hypertension</td>
<td>0.279</td>
<td>0.140</td>
<td>3.952</td>
<td>0.047</td>
<td>1.322</td>
<td>1.004 - 1.740</td>
</tr>
<tr>
<td>Activity L.V</td>
<td>-0.328</td>
<td>0.191</td>
<td>2.955</td>
<td>0.086</td>
<td>0.721</td>
<td>0.496 - 1.047</td>
</tr>
<tr>
<td>Constant</td>
<td>0.8463</td>
<td>0.080</td>
<td>109.621</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000 - 0.000</td>
</tr>
</tbody>
</table>

Interpretation1: Association between Gender and Alcohol:

Let \( A \) be the set of all neuropathy diabetes and \( B = \{ x / x \text{ is a alcoholist} \} \) is a subset of \( A \).

Where

\[
\text{Alcoholist} = \begin{cases} 
1 & \text{if } x \text{ is a man} \\
0 & \text{if } x \text{ is a women}
\end{cases}
\]
Define a function \( f : A \rightarrow \{0,1\} \) such that
\[
f(x) = \begin{cases} 
1 & \text{if } x \in B \\
0 & \text{if } x \in B^c
\end{cases}
\]
Now
\[
\log\left(\frac{p}{1-p}\right) = \alpha + \beta f(x)
\]
This shows that the odds of alcoholicity about 1.5 times greater for men than for women. Alcoholic neuropathy presents with a gradual decrease in sensory and motor peripheral nerve function in patients who chronically consume excessive amounts of alcohol. Based on this study, perhaps alcohol cessation programs should be targeted toward men.

**Interpretation2: Association between Age, Gender and Smoking Status:**

Let \( C \) be the set of all neuropathy diabetes and \( D = \{x| x \text{ is a smoker}\} \) is a subset of \( C \). Where
\[
\text{Smoker} = \begin{cases} 
1 & \text{if } x \text{ is a men} \\
0 & \text{if } x \text{ is a women}
\end{cases}
\]
Define a function \( g : C \rightarrow \{0,1\} \) such that
\[
g(x) = \begin{cases} 
1 & \text{if } x \in D \\
0 & \text{if } x \in D^c
\end{cases}
\]
\[
\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1(z) + \beta_2 g(x)
\]
\[
\log\left(\frac{p}{1-p}\right) = -8.463 + 0.112(54) + 0.329(1)
\]
if \( x \in D \), \( p = 0.1105 \)
\[
\log\left(\frac{p}{1-p}\right) = -8.463 + 0.112(54) + 0.329(0)
\]
if \( x \in D^c \), \( p = 0.0821 \)

The present study concludes that smoking affects 1.35 times greater for men than women. The possibility of neuropathic pain increases as the duration of smoking increase with diabetes even more. It is extremely important that the smokers should be informed regarding these facts and possibilities. Based on this study, perhaps smoking cessation programs should be targeted toward men.

**III. CONCLUSIONS**

The present study confirms that if administration of neuropathy was interrupted in diabetic peripheral neuropathy patients it affects the health by causing a decrease of insulin and increase of glucose. Hence this study supports that DNP can put patients at risk. Thus the present study suggests that the good quality of life is significantly related to good diabetes self-management and glucose control in Type2 diabetic mellitus especially for patients with diabetic neuropathy(DN). Therefore, these patients should be encouraged to perform self-management for controlling their blood glucose levels and improving their quality of life. Therefore, in order to preserve a good health-related quality life, it is obviously important to prevent diabetes complications and properly manage concomitant chronic diseases even when the patient already has DN.

**ACKNOWLEDGMENT:**

We gratefully acknowledge the Dr. Mohan diabetes speciality centre database for providing the data for the research and the help, support extended by Dr. S. Saravanan, Madras School of Economics, Chennai in compilation of this article.

**REFERENCES**


