A Neural Network model to analyse how intervention of BMI would affect incidence of diabetes

Thomas Richardson, Yuan Zhang
Operations Research
North Carolina State University

December 11, 2010

Introduction

There are about 23.6 million children and adults in the United States 7.8% of the population living with diabetes, and 1.6 million new cases of diabetes are diagnosed in people aged 20 years and older each year [**?**]. Complications of diabetes including heart disease, stroke, blindness, kidney disease and etc. Overall, the risk for death among people with diabetes is about twice that of people without diabetes of similar age. [**?**].

U.S. is spending $174 billion for patients with diabetes, combining $116 billion for direct medical costs and $58 billion for indirect costs (disability, work loss, premature mortality) [**?**]. Today diabetes prevention and early treatment are under-emphasized, if adults received their recommended diabetes screenings and early lifestyle intervention of medicine treatment, complications and disability could be avoided and billions of dollars can be saved [**?**].

SAS has created a contest in which the United States Health Expense Think Tank (USHETH), a fictional group, wants to know the impact of preventive measures on diabetes. Furthermore, it would like to know how an increase or decrease in diabetes prevalence would affect the amount of money spent on health care. Specifically, USHETH wants to research on the following two questions:

* USHETH wants to know diabetes influences a patient total health care expense. There are many estimates of the cost of diabetes care that have focused on diabetes-specific expenses, but it is known that diabetes might affect total health care expenses in many ways. There are many factors in the patient database that need to be taken into account so we propose to build a model to know the difference in health care costs for a person with diabetes from the costs for a person without the disease. It acknowledges that there are many factors that need to be taken into account.

* USHETH needs to know how many people have diabetes by age and perhaps gender. This will be critical for the analysis. We know that BMI is established as one of the significant risk factors for diabetes [**?**], and studies show that lifestyle intervention can reduce the incidence of diabetes in persons at high risk [**?**]. USHETH wants to measure the impact of diabetes preventive measure. The measure focuses on people who have a BMI (body mass index) larger than 25 in adults. In children the threshold BMI varies by age linearly: at 5 years old, the threshold BMI is 17 and at 20 years old, the threshold BMI is the same for adults. These people will be enrolled in special programs to reduce their BMI by ten percent. USHETH then wants to know how many people would contract diabetes.

In order to complete this analysis, SAS has provided a 50,000+ patient data set with 42 parameters for each patient, as listed in Table **??**. We classify the parameters provided into four categories: demographics, general physical characteristics, current diseases and health care characteristics. The sample provided is representative of the population and provides a snapshot at a point in time.

In determining how effective the proposed measure will be, we must analyse the onset of diabetes in the stated BMI categories, and determine relevant correlations. Notice that the data provides only a snapshot, and not a time series of data points. Therefore we have little knowledge of how these parameters changed for individuals over time, though we can extrapolate the onset of particular things (diabetes, heart conditions, BMI changes) by trends in the data. Using statistics

Table 1: **Parameters in The Data Set**

DEMOGRAPHICS	**CURRENT DISEASES**
SEX	HIGH BLOOD PRESSURE DIAGNOSIS
CENSUS REGION	CORONARY HRT DISEASE DIAGNOSIS
AGE	ANGINA DIAGNOSIS
MARITAL STATUS	HEART ATTACK DIAGNOSIS
YEARS OF EDUCATION	OTHER HEART DISEASE DIAGNOSIS
EVER SERVED IN ARMED FORCES	STROKE DIAGNOSIS
DID ANYONE PURCHASE FOOD STAMPS	JOINT PAIN LAST 12 MNTH
TOTAL INCOME	ASTHMA DIAGNOSIS
HAS MORE THAN ONE JOB	DIABETES_DIAG_BINARY
GENERAL PHYSICAL CHARACTERISTICS	**HEATH CARE CHARACTERISTICS**
WEARS EYEGLASSES	DENTAL CHECK-UP
PERSON IS BLIND	HOW LONG CHOLEST LAST CHECK
PERSON WEARS HEARING AID	HOW LONG LAST ROUTNE CHECKUP
PERSON IS DEAF	HOW LONG LAST FLU SHOT
PERSON WEIGHT	NUM OFFICE-BASED PROVIDER VISITS
LOST ALL UPPR AND LOWR TEETH	HOW LONG SINCE LAST PSA
ADULT BMI	HOW LONG LAST PAP SMEAR TEST
CHILD BMI	HOW LONG SNCE LAST BREAST EXAM
CURRENTLY SMOKE	HOW LONG SNCE LAST MAMMOGRAM
	BLOOD STOOL TEST
	SIGMOIDOSCOPY/COLONOSCOPY
	WEARS SEAT BELT

methods, we can calculate the proportion of diabetes patients in a specified population (in demographics, physical characteristics, and etc). How would the proportion of diabetes change if the specified population reduce BMI by 10%?

In order to answer the proposed questions, we have decided to construct a neural network to parse out the relationship between population characteristics (input) and the binary parameter diabetes diagnosis (output). Many neural network structures have been used in medical analysis [**?**] [**?**] [**?**]. Specifically, Park, et al. [**?**] built a neural network to evaluate the Heath Risk Appraisal data for diabetes prediction. They claimed that the use of neural network can enhance the identification of individuals who are at high risk for specific diseases in a time-sensitive manner. We propose to identify the high risk population for diabetes, then measure the reduction of diabetes risk by reducing BMI by a certain level for the high risk population. In addition, we propose to examine the threshold of BMI, currently set as 25, that is used to identify the high risk population.

Model

We propose to create a three layers feedforward neural network which take the binary parameter of diabetes diagnosis as the output Y and selected parameters (X_{BMI}, X) in Table **??** as input.

We propose to use the Neural Network Toolbox of Matlab to build, train and validate the model.

Methodology

The basic procedure of our analysis is outlined below:

Step 1: We propose to filter the data set by eliminating incomplete data. The target data set is called A.

Step 2: Construct neural network(NN), $N1$, and train via A or some random sub-population from A. The main point is that the proportion of non-diabetes to diabetes patients is equal in the sub-population to the ratio in the data-set.

Step 3: Validate $N1$ with A; select some sub-set of A such that $Y1$ (the output of $N1$ from the input A) is between some range around 0.5. This data set is $A2$.

Step 4: Construct and train NN, $N2$, with $A2$. Validate $N2$ with $A2$; record output $Y2$, which is a decimal value between 0 and 1.

Step 5: Note: let Y_{21} be output of $N2$ from A_{21}, which is a subset of $A2$ with diabetes, and Y_{20} be output of $N2$ from A_{20}, which is the complimentary of A_{21}. Examine if there is a statistical difference between Y_{21} and Y_{20}. Construct a function $f(Y_{21}, Y_{20})$ to measure the statistical difference.

Step 6: Construct $A3 = \{(X_{BMI} * (1-p), X, Y) | (X_{BMI}, X, Y) in A2, and X_{BMI} >= Tar_{BMI}\}$., given policy $P = \{p, Tar_{BMI}\}$. Our first test case is $p = 10\%, Tar_{BMI} = 25$.

Step 7: Validate $N2$ with $A3$; record output $Y3$.

Step 8: Calculate $diff_Y = f(Y2, Y3)$. $f(Y2, Y3)$ gives some indication as to how effective the reduction of BMI at reducing the occurrence of diabetes in an individual, given policy $P = \{p, Tar_{BMI}\}$.

Step 9: Run different policies to examine if $Tar_{BMI} = 25$ is a good policy value.

Results

Performance Analysis

Conclusion and Discussion