Northeastern University
Mechanical and Industrial Engineering Department
OR 6205: Deterministic Operations Research
Prof. Md. Noor E Alam
Fall 2023

Determining Influential Factors in Diabetes: A Comprehensive Analysis using Mc. Nemar's Test

Shalini Dutta | Siddhi Yeshwant Sonwalkar | Shweta Shinde | Esha Joshi | Andy Bai

PROBLEM STATEMENT

Recently, more people have been diagnosed with diabetes, and it's become a significant health concern. To figure out why this is happening, our study explores the factors that could be linked to diabetes in individuals. We're looking into things like smoking, high blood pressure, high cholesterol, eating habits, physical activity, fruit intake, age, BMI (Body Mass Index) etc.

By closely examining these factors, we aim to understand why there's an increase in diabetes cases. Our goal is to uncover how these elements interact and contribute to more people developing diabetes. This insight will help us come up with practical ideas to create targeted plans for preventing and managing diabetes.

The main purpose of our study is to contribute to finding effective ways to prevent and address diabetes. We hope that by gaining more knowledge about what causes diabetes, we can improve the overall health of the public. Ultimately, the information from our study may lead to better strategies for addressing the growing issue of diabetes in our communities.

DATASET

The dataset is found on Kaggle.com and called “Diabetes Health Indicators”. It is the collection of survey that responses from over 400,000 Americans on health-related behaviors and conditions. The survey has been collected by the CDC since 1984 until today, and the dataset we are using for this particular project is from 2015 to today. According to Kaggle, the initial dataset contains surveys from around 440,000 users and 330 feature columns, those features are either directly answered by the individuals or calculated based on participant’s responses. The updated dataset which we are using at this moment contains 22 feature columns and 253,681 records. Because of the large dataset that consists huge amounts of records.

We shall look at each column of the reduced dataset below-

Diabetes_binary
0 = no diabetes 1 = prediabetes 2 = diabetes

HighBP
0 = no high BP 1 = high BP

HighChol
0 = no high cholesterol 1 = high cholesterol

CholCheck
0 = no cholesterol check in 5 years 1 = yes cholesterol check in 5 years

BMI
Body Mass Index

Smoker
Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] 0 = no 1 = yes

Stroke
(Ever told) you had a stroke. 0 = no 1 = yes

HeartDiseaseorAttack
coronary heart disease (CHD) or myocardial infarction (MI) 0 = no 1 = yes

PhysActivity
physical activity in past 30 days - not including job 0 = no 1 = yes

Fruits
Consume Fruit 1 or more times per day 0 = no 1 = yes

Veggies
Consume Vegetables 1 or more times per day 0 = no 1 = yes

HvyAlcoholConsump
(adult men >=14 drinks per week and adult women>=7 drinks per week) 0 = no 1 = yes

AnyHealthcare
Have any kind of health care coverage, including health insurance, prepaid plans such as HMO, etc. 0 = no 1 = yes

NoDocbcCost

Was there a time in the past 12 months when you needed to see a doctor but could not because of cost? 0 = no 1 = yes

GenHlth
Would you say that in general your health is: scale 1-5 1 = excellent 2 = very good 3 = good 4 = fair 5 = poor

MentHlth
days of poor mental health scale 1-30 days

PhysHlth
physical illness or injury days in past 30 days scale 1-30

DiffWalk
Do you have serious difficulty walking or climbing stairs? 0 = no 1 = yes

Sex
0 = female 1 = male

Age
13-level age category 1 = 18-24 … 9 = 60-64 … 13 = 80 or older

Education
Education level scale 1-6 1 = Never attended school or only kindergarten 2 = elementary etc.

Income
Income scale 1-8 1 = less than $10,000 5 = less than $35,000 8 = $75,000 or more

HYPOTHESIS

STUDENT 1 : SHALINI DUTTA (002235745)

In this study, we are exploring the hypothesis that smoking is closely tied to the occurrence of diabetes. The null hypothesis suggests that there is no substantial association between smoking and the presence of diabetes in the population. Conversely, the alternative hypothesis asserts

that there is a significant relationship between smoking and the occurrence of diabetes. Through our investigation, we aim to discern whether smoking plays a noteworthy role in influencing the likelihood of individuals developing diabetes.

Null Hypothesis (H0):
"There is no significant association between smoking and the occurrence of diabetes in the population."

Alternative Hypothesis (H1):
"There is a significant association between smoking and the occurrence of diabetes in the population."

ANALYSIS

The study investigates whether smoking is significantly associated with the occurrence of diabetes in the population. For this analysis, four additional covariates, namely **Age, BMI, High Blood Pressure, and Physical Activity**, are included, where Age and BMI contain numerical values and High BP And Physical Activity are binary variables.

We first segregate the data into Treatment Group and Control Group. The treatment group gets the experiment, and the control group, without the experiment, gives a baseline for comparison to see the impact of the treatment.

In our study analyzing the impact of smoking on diabetes:

Treatment Group: Individuals who smoke or are exposed to smoking (experimental condition), i.e., column value = 1.

Control Group: Individuals who do not smoke or are not exposed to smoking (baseline or comparison condition), i.e., column value = 0.

Our Treatment and Control Groups are as follows –

	A	B	C	D	E	F	G
1	**Sno**	**HighBP**	**BMI**	**Age**	**PhysActivity**	**Smoker**	**Diabetes_binary**
2	1	0	27	2	1	1	**0**
3	2	0	31	8	0	1	**0**
4	3	0	29	10	1	1	**0**
5	4	0	27	10	0	1	**0**
6	5	0	33	8	1	1	**0**
7	6	0	27	10	1	1	**0**
8	7	0	36	7	0	1	**0**
9	8	0	33	1	1	1	**0**
10	9	1	22	11	0	1	**0**
11	10	0	26	3	1	1	**0**

Fig 1. Treatment Group

	A	B	C	D	E	F	G
1	**SNo**	**HighBP**	**BMI**	**Age**	**PhysActivity**	**Smoker**	**Diabetes_binary**
2	1	0	21	7	0	0	**0**
3	2	1	28	13	1	0	**0**
4	3	0	24	1	1	0	**0**
5	4	1	33	7	0	0	**0**
6	5	1	25	8	1	0	**0**
7	6	0	31	4	0	0	**0**
8	7	0	20	4	1	0	**0**
9	8	1	28	12	1	0	**0**
10	9	0	29	5	0	0	**0**
11	10	0	23	10	1	0	**0**

Fig 2. Control Group

Note that all values in the column named Smoker are 1 for Treatment Group and 0 for Control Group.

Now we load this data into AMPL to perform the McNemar's test on our dataset.

Data files –

We have 3 .dat files namely : **Diabetes.dat** which includes the number of rows in the Treatment and Control Groups and the Covariates set, **Smoking_Covariates.dat** which includes the covariates data for the Treatment and Control Groups, and **Outcome_Diabetes.dat** which includes the outcome(does or does not have Diabetes) of the Treatment and Control Groups.

```
#Other data (total # of treatment and control units, covariates set)

param T2:=536;
param T1:=464;

set C1:=  HighBP    BMI Age PhysActivity;


#param n:=30;
```

Fig 3. Snippet of Diabetes.dat

```
#Covariates data in AMPL format for case study – Diabetes vs Smoking

param CC:HighBP BMI Age PhysActivity :=
1   0   21  7   0
2   1   28  13  1
3   0   24  1   1
4   1   33  7   0
5   1   25  8   1


param CT: HighBP    BMI Age PhysActivity :=
1   0   27  2   1
2   0   31  8   0
3   0   29  10  1
4   0   27  10  0
5   0   33  8   1
6   0   27  10  1
7   0   36  7   0
```

Fig 4. Snippet of Smoking_Covariates.dat

```
#Outcome data in AMPL format for case study – Diabetes vs Smoking

param OC:=
1   0
2   0
3   0
4   0
5   0
6   0
7   0
```

Fig 5. Snippet of Outcome_Diabetes.dat

Model file -

The model file **Diabetes.mod** is used to define the discordant pairs using McNemar's test. The algorithm finds pairs of one unit each from the Treatment and Control Groups, where the covariate values are similar, such that the only difference is in the value of the Treatment or Control group variable (in our case, Smoker). This is to determine whether Smoking can be considered as the only differentiating factor between people with and without diabetes, consequently implying whether or not smoking causes diabetes.

Run file –

The run file **Diabetes.run** uses the above model and data files to determine the Z-values from which we can draw our hypothesis' conclusion. Here, we set the permissible difference between the values of the covariates in the Treatment and Control Groups. Thus, pairs are made in such a way that we keep the covariate values uniform whilst only having differing values of Smoking (0 for non-smoker, 1 for smoker)

Age: Ages 18 to 80 and above are divided into 13 groups. We can consider people from the age of 18 to 30(18-24 and 25-30) as being somewhat in the same age-group, and so on. Therefore, age-groups of Treatment and Control group units can have a maximum permissible difference of 2, to be grouped as discordant pairs.

BMI: The permissible difference in BMI of Treatment and Control Group units is considered to be 5.

HighBP: This is a binary variable so the only way that there is no difference between the HighBP value in units of Treatment and Control groups is if they are exactly equal(either 0 or 1). Hence the permissible difference is 0.

PhysActivity: This is a binary variable so the only way that there is no difference between the PhysActivity value in units of Treatment and Control groups is if they are exactly equal(either 0 or 1). Hence the permissible difference is 0.

Number of iterations = 10

```
for{1..10}{

for {i in 1..T1, j in 1..T2}

       {

           if (CT[i,"Age"]-CC[j,"Age"])^2 <=4 then

               {let  NW[i,j]:=0;}
           else
               {let  NW[i,j]:=1;}
       }

   for {i in 1..T1, j in 1..T2}

       {
           if (CT[i,"BMI"]-CC[j,"BMI"])^2<=25 then

               {let  NH[i,j]:=0;}
           else
               {let  NH[i,j]:=1;}
       }

   for {i in 1..T1, j in 1..T2}

       {
           if CT[i,"HighBP"]==CC[j,"HighBP"] then

               {let  NT[i,j]:=0;}
           else
               {let  NT[i,j]:=1;}
       }
   for {i in 1..T1, j in 1..T2}

       {
           if CT[i,"PhysActivity"]==CC[j,"PhysActivity"] then

               {let  NB[i,j]:=0;}
           else
               {let  NB[i,j]:=1;}
       }
```

Fig 6. Snippet of Diabetes.run

Output –

On running Diabetes.run file, we generate two files with the minimum and maximum Z-values respectively.

```
AMPL
ampl: include '/Users/shalinidutta/Desktop/AMPL/Diabetes.run';
CPLEX 22.1.1.0: optimal integer solution; objective 5.294651389
8 MIP simplex iterations
0 branch-and-bound nodes
CPLEX 22.1.1.0: optimal integer solution; objective -5.659799761
0 MIP simplex iterations
0 branch-and-bound nodes
CPLEX 22.1.1.0: optimal integer solution; objective 5.388159061
11 MIP simplex iterations
0 branch-and-bound nodes
CPLEX 22.1.1.0: optimal integer solution; objective -5.747369665
0 MIP simplex iterations
0 branch-and-bound nodes
CPLEX 22.1.1.0: optimal integer solution; objective 5.480077554
12 MIP simplex iterations
0 branch-and-bound nodes
CPLEX 22.1.1.0: optimal integer solution; objective -5.833630945
0 MIP simplex iterations
0 branch-and-bound nodes
CPLEX 22.1.1.0: optimal integer solution; objective 5.570484991
12 MIP simplex iterations
0 branch-and-bound nodes
CPLEX 22.1.1.0: optimal integer solution; objective -5.918640302
0 MIP simplex iterations
0 branch-and-bound nodes
CPLEX 22.1.1.0: optimal integer solution; objective 5.65945331
12 MIP simplex iterations
0 branch-and-bound nodes
```

Fig 7. Snippet of Output on AMPL

30.000	5.295	30.000	−5.660
31.000	5.388	31.000	−5.747
32.000	5.480	32.000	−5.834
33.000	5.570	33.000	−5.919
34.000	5.659	34.000	−6.002
35.000	5.747	35.000	−6.085
36.000	5.833	36.000	−6.167
37.000	5.918	37.000	−6.247
38.000	6.002	38.000	−6.327
39.000	6.085	39.000	−6.405

Fig 8. Max and Min Z-values

We use these Z-values and find the respective p-values using the 1-NORMSDIST() function on excel and plot these against the Total number of untied responses.

n	Min P-value	Max P-value
30	5.95082E-08	0.999999992
31	3.5623E-08	0.999999995
32	2.12663E-08	0.999999997
33	1.2737E-08	0.999999998
34	7.61288E-09	0.999999999
35	4.54204E-09	0.999999999
36	2.72198E-09	1
37	1.6294E-09	1
38	9.74508E-10	1
39	5.82457E-10	1

	min	max	Max *P* -value	Min *P* -value
30	-5.66	5.295	0.999999992	5.9508E-08
31	-5.747	5.388	0.999999995	3.5623E-08
32	-5.834	5.48	0.999999997	2.1266E-08
33	-5.919	5.57	0.999999998	1.2737E-08
34	-6.002	5.659	0.999999999	7.6129E-09
35	-6.085	5.747	0.999999999	4.542E-09
36	-6.167	5.833	1	2.722E-09
37	-6.247	5.918	1	1.6294E-09
38	-6.327	6.002	1	9.7451E-10
39	-6.405	6.085	1	5.8246E-10

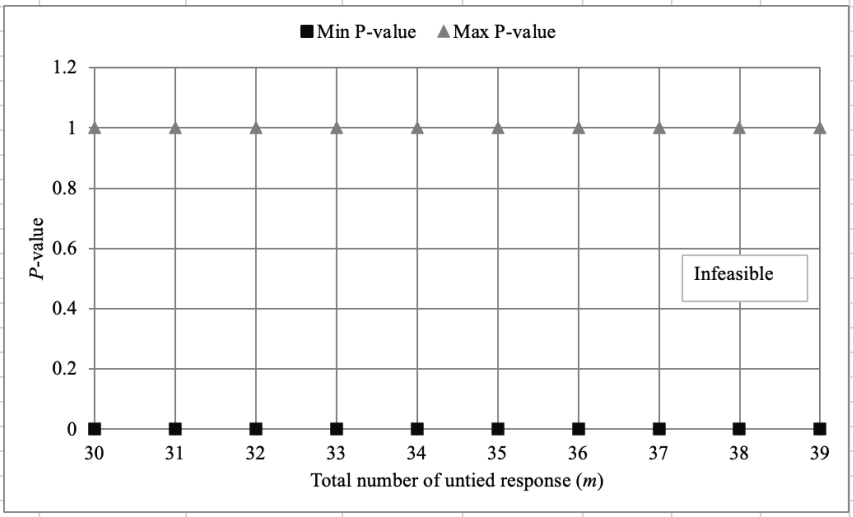


Fig 9. Calculating p_max and p_min for Z-values and plotting it against Total Number of untied Responses

FINDINGS

From our analysis, we clearly identify that the p_min and p_max values are divergent and do not converge into a common p-value that could be used to reject or fail to reject our Null Hypothesis($p>0.05$ implies we fail to reject the Null Hypothesis and $p<0.05$ implies we reject the Null Hypothesis). p_min keeps decreasing and p_max keeps increasing as we increase the total number of untied responses. Hence, we can infer that our dataset does not provide the opportunity for drawing robust conclusions. We cannot confirm or deny that smoking causes diabetes.

STUDENT 2: ANDY BAI

Null Hypothesis (H0):
"There is no significant association between eating vegetables and having diabetes."

Alternative Hypothesis (H1):
"Not eating vegetables will have potential risk of having diabetes"

I randomly select the 20,000 records from the original dataset. From that point, I separated my control and treatment group as shown below, simply splitting the data where veggies is equal to 1 (control group) and 1 (treatment group).

```
control_df=  df_sample[df_sample['Veggies']==1].reset_index(drop = True)


control_df
```

	BMI	Smoker	PhysActivity	Sex	Fruits	HighBP	Stroke	Diabetes_binary	Veggies
0	21	0	0	0	1	0	0	0	1
1	28	0	1	0	1	1	0	0	1
2	24	0	1	1	1	0	0	0	1
3	27	1	1	1	0	0	0	0	1
4	31	1	0	0	1	0	0	0	1
...	...	...	...	...	...	...	...	...	...
1619	40	1	1	0	0	1	0	0	1
1620	21	1	0	0	1	1	0	0	1
1621	25	1	1	0	0	0	1	0	1
1622	25	1	1	1	1	1	0	0	1
1623	29	0	1	0	1	0	0	0	1

Fig 1. Control group dataset

```
treat_df= df_sample[df_sample['Veggies']==0].reset_index(drop = True)
treat_df
```

	BMI	Smoker	PhysActivity	Sex	Fruits	HighBP	Stroke	Diabetes_binary	Veggies
0	27	1	0	1	0	0	0	0	0
1	36	1	0	1	0	0	0	0	0
2	27	1	1	1	0	1	1	1	0
3	29	0	0	0	0	0	0	0	0
4	26	1	0	1	1	1	0	0	0
...	...	...	...	...	...	...	...	...	...
371	25	1	0	0	1	1	0	0	0
372	28	0	0	1	0	1	0	0	0
373	30	0	1	0	1	1	0	0	0
374	25	0	1	1	0	0	0	0	0
375	44	0	0	0	1	1	0	0	0

Fig 2. Treatment group dataset

ANALYSIS

In my control group, I filtered the dataset where veggies = 1 (eating veggies). This way, I can compare the diabetes outcomes between those who consume vegetables and those who do not, and my treatment group is on the other hand where veggies = 0 (not eating veggies). As mentioned earlier, due to the large dataset and I reduced the dataset records in order to reduce the computation time. As ready to do the analysis, I have 1624 records in my pre-cleaned dataset and a total of 376 records in my treatment group. There are 5 files for this case study. The first project_covariates.dat contains the pre-processed data records for control group and treatment group. The treatment is *Veggies***.** I chose several feature columns which serve as pretreatment covariates including *BMI (continuous value), Smoker (binary format), Physical activity (binary), Gender (binary), Fruits (binary), High blood pressure (binary), Stroke (binary)*. Most importantly, our target variable is the Diabetes column which is also in binary form. 0 means individual does not have diabetes, vice versa. In the project_fracture.dat file, I am just setting up the parameters as well as the feature columns for my study.

Then, in project_fracture.mod file, we started to formulate the constraints that can help us to perform the **Robust McNemar's test**. McNemar's test is applied when we have two related groups, and each subject is measured or classified twice. This often occurs in "before-and-after" studies or situations where the same individuals are exposed to different conditions. The way that Mcnemar's test works is basically the paired data would consist of the number of individuals who tested positive before and after treatment (cell a), the number who tested negative before but positive after treatment (cell b), those who tested positive before but

negative after treatment (cell c), and those who tested negative both times (cell d). Additionally, since we do have binary outcomes (0 or 1), and McNemar's test is suited for this situation. The test is useful in non-randomized studies or situations where randomization is not feasible. In such cases, pairing helps control individual variability.

In our objective function, $Max/Min\ z(a) = \frac{\mathrm{B-C-1}}{\sqrt{B+C}}$ is designed to quantify the difference between treatment (B) and control (C) groups in a manner that emphasizes untied responses, where outcomes differ. The optimization process aims to either maximize or minimize this expression, exploring extreme scenarios. The focus is on understanding how variations in decision variables impact the objective function, providing insights into the sensitivity of the model and identifying optimal configurations that maximize or minimize the specified difference between the treatment and control outcomes. The objective is to gain a nuanced understanding of the data, with a particular emphasis on untied responses, without explicit constraints on the number of pairs with the same outcome or the total number of pairs. The following equation $\mathrm{aij} \in \{0,1\}$ defines the binary variable as well as serve for constraints where aij = either 0 or 1. For continuous covariates such as BMI, I set the caliper difference to 6^2 = 36 to separate the patients into different groups.

```
for {i in 1..T1, j in 1..T2}
    {
        if (CT[i,"Smoker"]==CC[j,"Smoker"]) then
            {let  NS[i,j]:=0;}
        else
            {let  NS[i,j]:=1;}
    }
for {i in 1..T1, j in 1..T2}
    {
        if (CT[i,"PhysActivity"]==CC[j,"PhysActivity"]) then
            {let  NA[i,j]:=0;}
        else
            {let  NA[i,j]:=1;}
    }
for {i in 1..T1, j in 1..T2}
    {
        if (CT[i,"Gender"]==CC[j,"Gender"]) then
            {let  NG[i,j]:=0;}
        else
            {let  NG[i,j]:=1;}
    }
```

```
for {i in 1..T1, j in 1..T2}
    {
        if (CT[i,"Fruits"]==CC[j,"Fruits"]) then
            {let  NF[i,j]:=0;}
        else
            {let  NF[i,j]:=1;}
    }
for {i in 1..T1, j in 1..T2}
    {
        if (CT[i,"HighBP"]==CC[j,"HighBP"]) then
            {let  NH[i,j]:=0;}
        else
            {let  NH[i,j]:=1;}
    }
for {i in 1..T1, j in 1..T2}
    {
        if (CT[i,"Stroke"]==CC[j,"Stroke"]) then
            {let  NK[i,j]:=0;}
        else
            {let  NK[i,j]:=1;}
    }
```

Fig 3. AMPL binary constraints

FINDINGS:

In this study, I investigated the relationship between specific variables and their associated p-values within the framework of our optimization model. As shown in the below graph, the variables under consideration exhibited a range of values, with a minimum observed

value of 5.295 and a maximum of 5.659. Conversely, the associated p-values demonstrated a broader spectrum, ranging from a minimum of 5.95082E-08 to a maximum of 0.999999999.

	min	max	Max *P*-value	Min *P*-value
30	-5.66	5.295	0.999999992	5.95082E-08
31	-5.747	5.388	0.999999995	3.5623E-08
32	-5.834	5.48	0.999999997	2.12663E-08
33	-5.919	5.57	0.999999998	1.2737E-08
34	-6.002	5.659	0.999999999	7.61288E-09

Fig 4. Min and Max values

The minimum and maximum values of the variables suggest a constrained range, indicating potential boundaries within which these variables operate in the optimization context. These boundaries play a crucial role in guiding the optimization process and ensuring that the model adheres to predefined limits.

Simultaneously, the diversity in p-values suggests a variability in the statistical significance of certain parameters within our optimization framework. The minimum p-values, ranging from 5.95082E-08 to 7.61288E-09, indicate a high level of statistical significance, while the maximum p-values approach unity, suggesting lower levels of certainty associated with certain parameters.

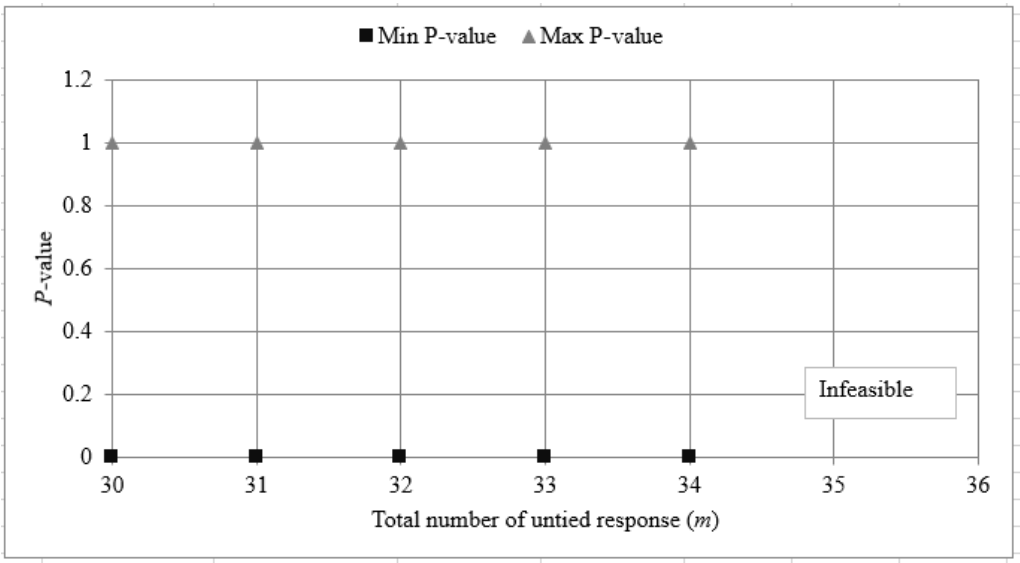


Fig 5. Min P-value vs Max P-value

As the graph shown on the above, each data point in our study represents an individual, and our investigation centers around the development of diabetes as an outcome. The primary treatment under consideration is the consumption of vegetables. To ensure a fair comparison, we've matched individuals on various pretreatment covariates, including BMI, smoker, physical

activity, gender, stroke, high blood pressure, fruits. As a result, this lack of connection tells us that the things we're studying don't seem to influence the scores they get in our optimization model. This surprising result suggests that not eating vegetables has no potential relationship with getting diabetes. Additionally, in the graph shown above, the min p value lies in 0, and max p value lies in 1, where min p value is suggesting significant reject the null hypothesis and the max p value is suggesting accepting the null hypothesis at the same time. Hence, we could not draw any robust conclusion for this particular research because of the uncertainty.

STUDENT 3: SIDDHI YESHWANT SONWALKAR (NEU ID: 002857261)

This study investigates the potential link between high blood pressure (HBP) and diabetes. The null hypothesis states that HBP and diabetes are not meaningfully related in the population. By contrast, the alternative hypothesis claims that HBP and diabetes are significantly associated. We are conducting this research to determine if HBP has a major influence on a person's chance of getting diabetes. In other words, we want to find out if HBP makes people more prone to developing diabetes or if these two conditions are unrelated. Through our analysis, we will either reject the null hypothesis, lending support to the alternative hypothesis, or fail to reject the null hypothesis, suggesting HBP is not an important determinant of diabetes onset. Ultimately, we hope to gain clarity on whether or not HBP plays a substantive role in predisposing people to diabetes.

Null Hypothesis (H0):
"There is no significant association between High BP and the occurrence of diabetes in the population."

Alternative Hypothesis (H1):
"There is a significant association between High BP and the occurrence of diabetes in the population."

ANALYSIS

This study explores if high blood pressure (HBP) is meaningfully linked to developing diabetes. We incorporate four additional factors into the analysis that may influence this relationship: **age, BMI, fruit consumption, and heavy alcohol use**. Age and BMI are numerical variables, while fruit intake and heavy drinking are binary (yes/no) variables.

To assess if HBP impacts diabetes risk, we separate participants into a treatment group with HBP and a control group without HBP. Comparing diabetes rates between these groups reveals the effect of HBP, using the control group as a baseline without the influence of HBP.

Specifically, we categorize participants based on:

Treatment Group: Individuals with HBP, i.e., column value = 1

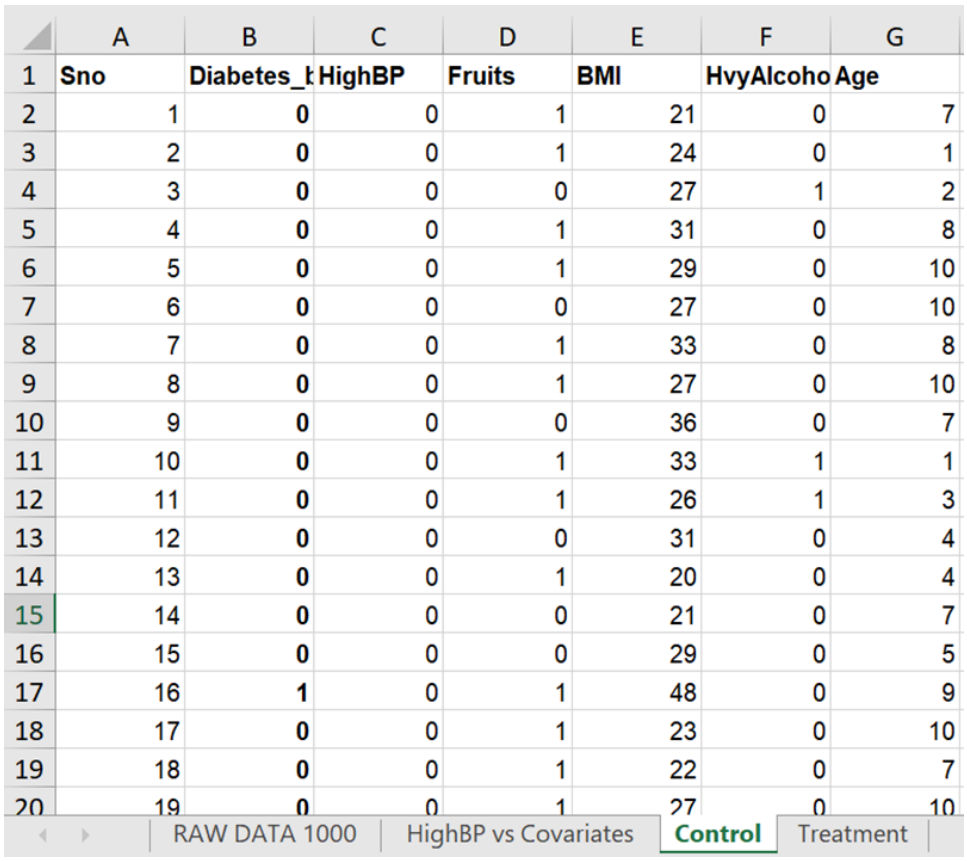
Control Group: Individuals without HBP, i.e., column value = 0.
By evaluating and contrasting diabetes occurrence between the treatment and control groups, we can determine whether having HBP makes one more prone to developing diabetes versus those with normal blood pressure. Our analysis aims to clarify if HBP substantially raises one's likelihood of acquiring diabetes.

Our Treatment and Control Groups are as follows –

	A	B	C	D	E	F	G
1	**Sno**	**Diabetes_b**	**HighBP**	**Fruits**	**BMI**	**HvyAlcoho**	**Age**
2	1	**0**	1	1	28	0	13
3	2	**0**	1	1	33	0	7
4	3	**0**	1	1	25	0	8
5	4	**0**	1	1	22	1	11
6	5	**1**	1	0	27	0	10
7	6	**0**	1	1	26	0	2
8	7	**0**	1	1	28	0	12
9	8	**0**	1	1	36	0	10
10	9	**0**	1	0	30	0	11
11	10	**0**	1	1	25	0	8
12	11	**0**	1	1	43	0	10
13	12	**0**	1	0	33	0	7
14	13	**0**	1	1	26	0	3
15	14	**0**	1	1	28	0	9
16	15	**1**	1	1	33	0	12
17	16	**0**	1	1	24	0	5
18	17	**0**	1	1	32	0	3
19	18	**1**	1	0	40	0	11
20	19	**0**	1	0	16	0	8

RAW DATA 1000 | HighBP vs Covariates | Control | **Treatment**

Fig 1. Treatment Group



	A	B	C	D	E	F	G
1	Sno	Diabetes_b	HighBP	Fruits	BMI	HvyAlcoho	Age
2	1	0	0	1	21	0	7
3	2	0	0	1	24	0	1
4	3	0	0	0	27	1	2
5	4	0	0	1	31	0	8
6	5	0	0	1	29	0	10
7	6	0	0	0	27	0	10
8	7	0	0	1	33	0	8
9	8	0	0	1	27	0	10
10	9	0	0	0	36	0	7
11	10	0	0	1	33	1	1
12	11	0	0	1	26	1	3
13	12	0	0	0	31	0	4
14	13	0	0	1	20	0	4
15	14	0	0	0	21	0	7
16	15	0	0	0	29	0	5
17	16	1	0	1	48	0	9
18	17	0	0	1	23	0	10
19	18	0	0	1	22	0	7
20	19	0	0	1	27	0	10

RAW DATA 1000 | HighBP vs Covariates | Control | Treatment

Fig 2. Control Group

Note that all values in the column named HighBP are 1 for Treatment Group and 0 for Control Group.

Now we load this data into AMPL to perform the McNemar's test on our dataset.

Data files –

We utilize three data files: **Diabetes.dat** contains participant counts and covariate info; **HighBP.dat** provides the covariate measurements; and **Outcome_Diabetes.dat** indicates diabetes outcome for all participants. Together these enable analysis of the relationship between HBP and diabetes occurrence.

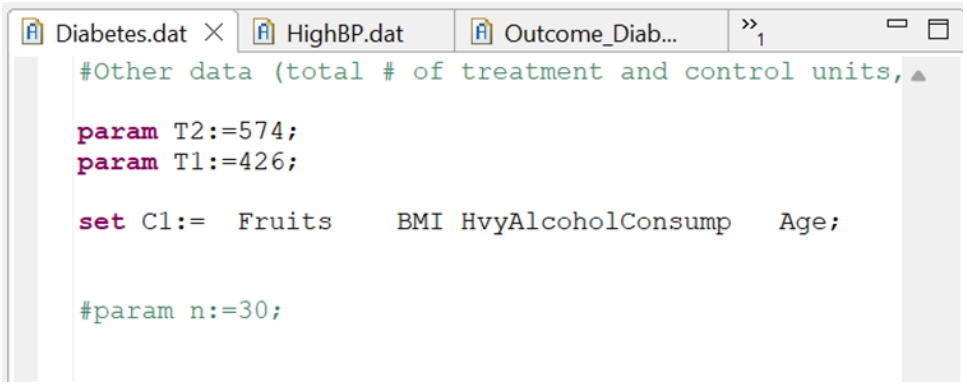


Fig 3. Snippet of Diabetes.dat

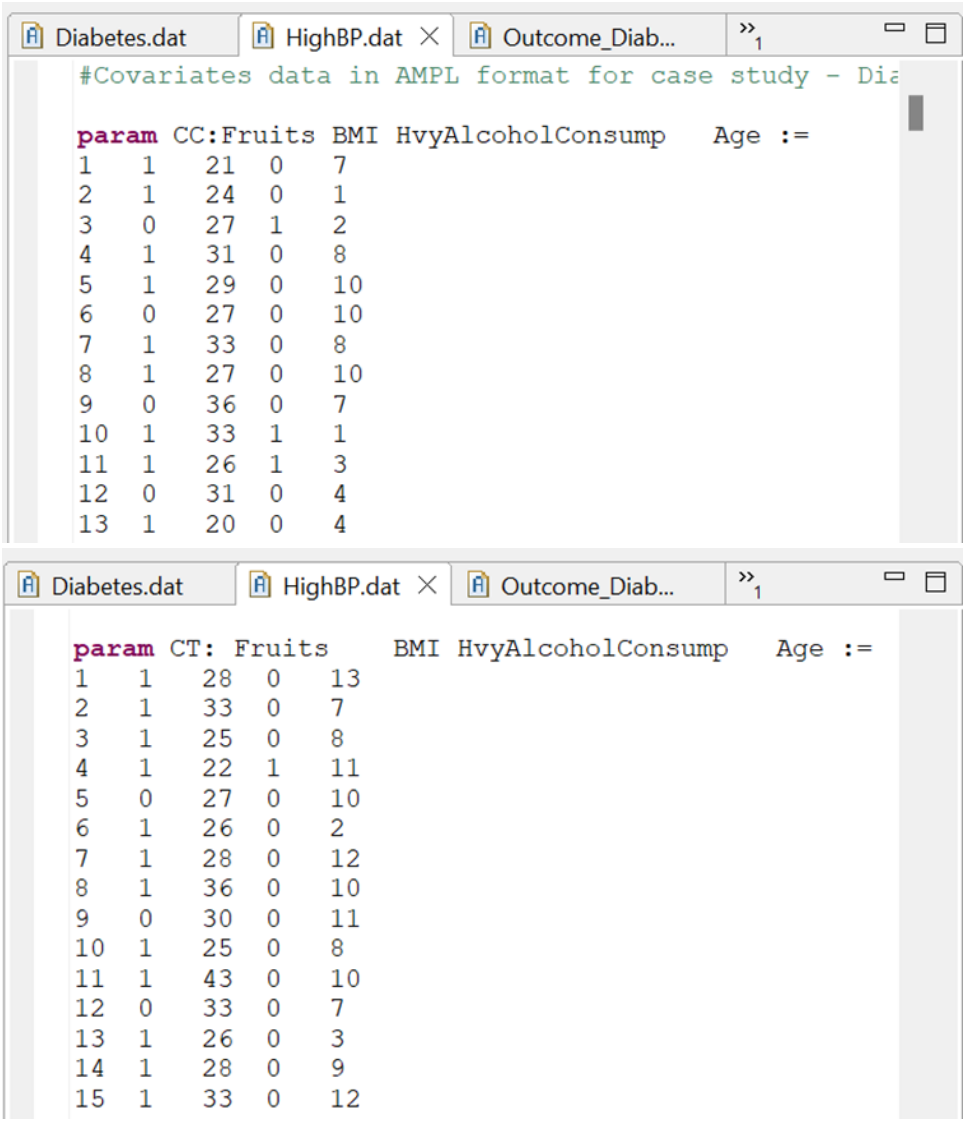


Fig 4. Snippet of HighBP.dat

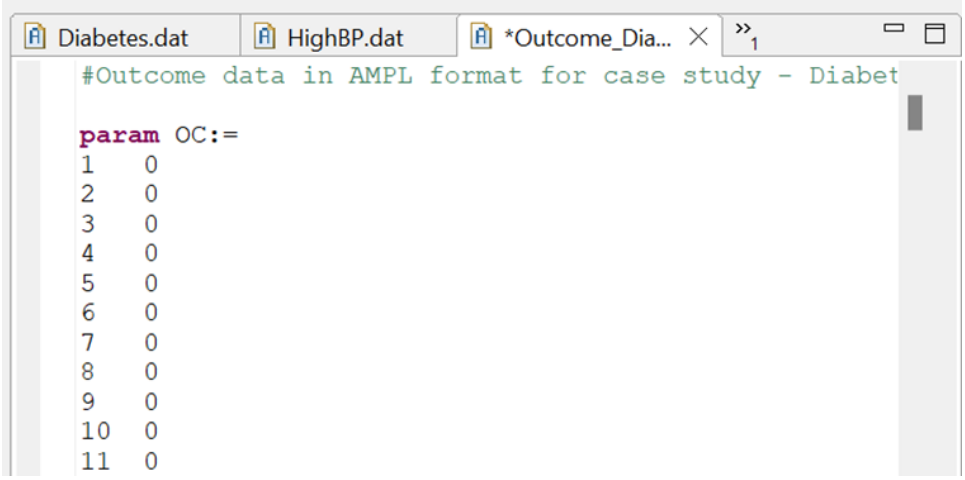


Fig 5. Snippet of Outcome_Diabetes.dat

Diabetes.mod Model File:

This file defines the discordant pairs to be used in the McNemar's test. Specifically, it identifies pairs consisting of one participant from the treatment group (with high blood pressure/HBP) and one from the control group (without HBP), where the covariates values are as similar as possible between the paired participants except for their HBP status. Finding these matched pairs with maximal covariate similarity helps isolate HBP as the only major difference to determine if it may directly cause higher diabetes occurrence.

Diabetes.run Run File:

This file implements the model by running the analysis using the identified paired units across the two groups. It calculates the z-values based on the difference in diabetes outcomes between treatment and control groups, from which conclusions about the study hypothesis can be drawn. Additionally, this run file enforces permissible differences in covariate values when matching treatment and control units into pairs. Allowing some small differences enables more pairs to be created while still keeping covariates relatively consistent. The maximum differences set are:

Age: Participants ages 18-80+ years old are separated into 13 age groups. People with up to 2 age group difference (e.g. 18-24 years and 25-30 years) are considered similar enough in age to be paired.

BMI: BMI can differ by up to 5 units between matched pairs. For example, BMIs of 28 and 33 could be paired.

Fruit Intake: This yes/no variable must be exactly the same (0 or 1 difference) for matched pairs since any difference would contrast fruit consumption behavior.

Heavy Alcohol Use: This yes/no variable must also be identical (difference of 0), since a difference would represent a significant contrast in drinking levels.

In summary, age and BMI allow small differences between pairs, enabling more matches, while the yes/no variables require an exact match between treatment and control groups since they naturally fully differentiate behavior.

The algorithm makes **10 iterations** to identify discordant pairs that best fit these parameters in order to maximize detection of an HBP-specific effect on diabetes likelihood.

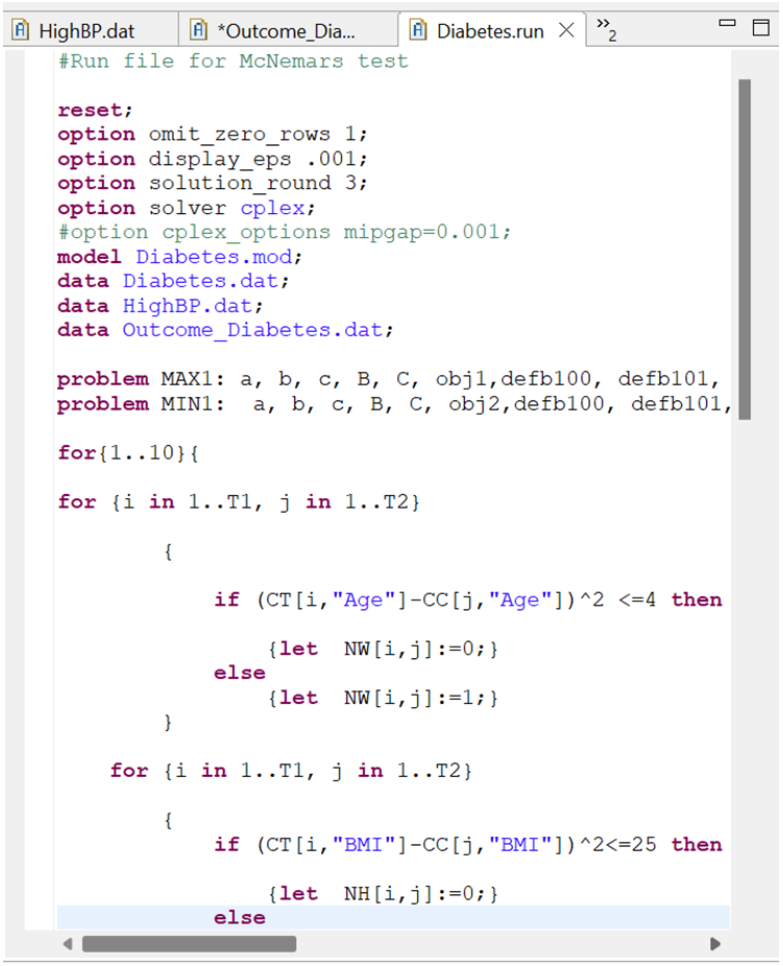


Fig 6. Snippet of Diabetes.run

Output –

On running Diabetes.run file, we generate two files with the minimum and maximum Z-values respectively.

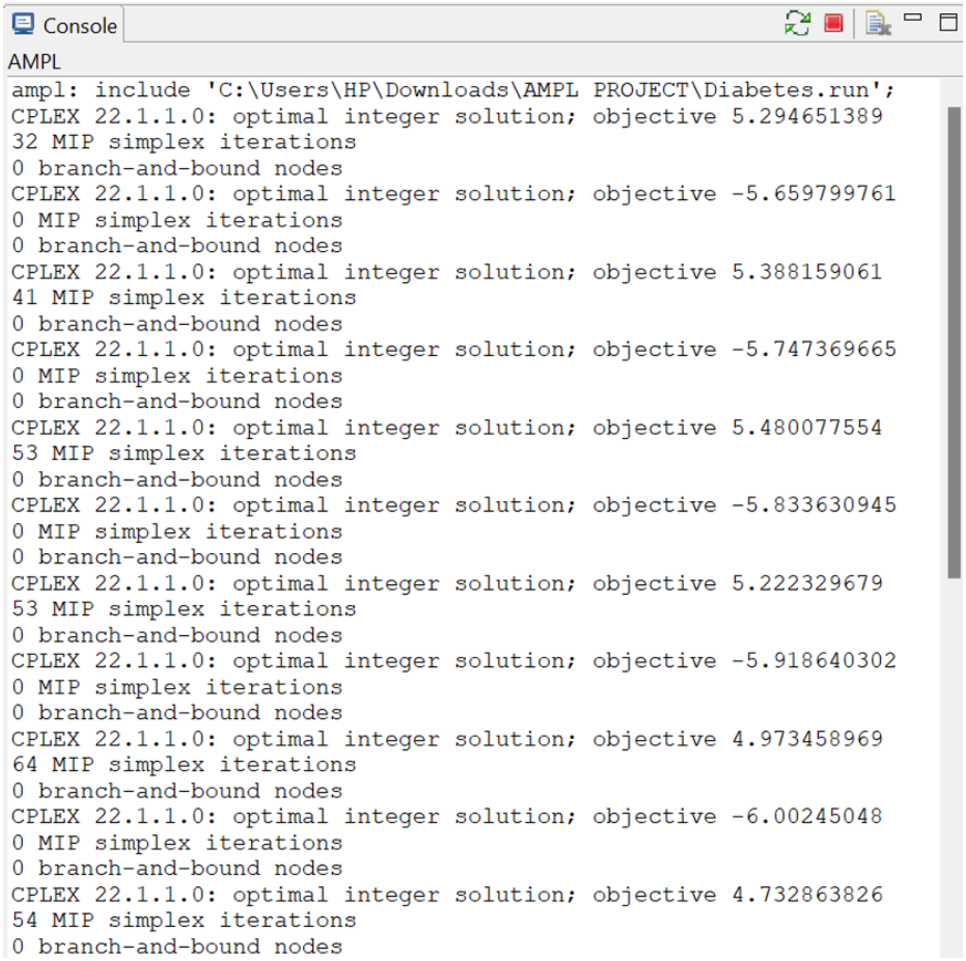
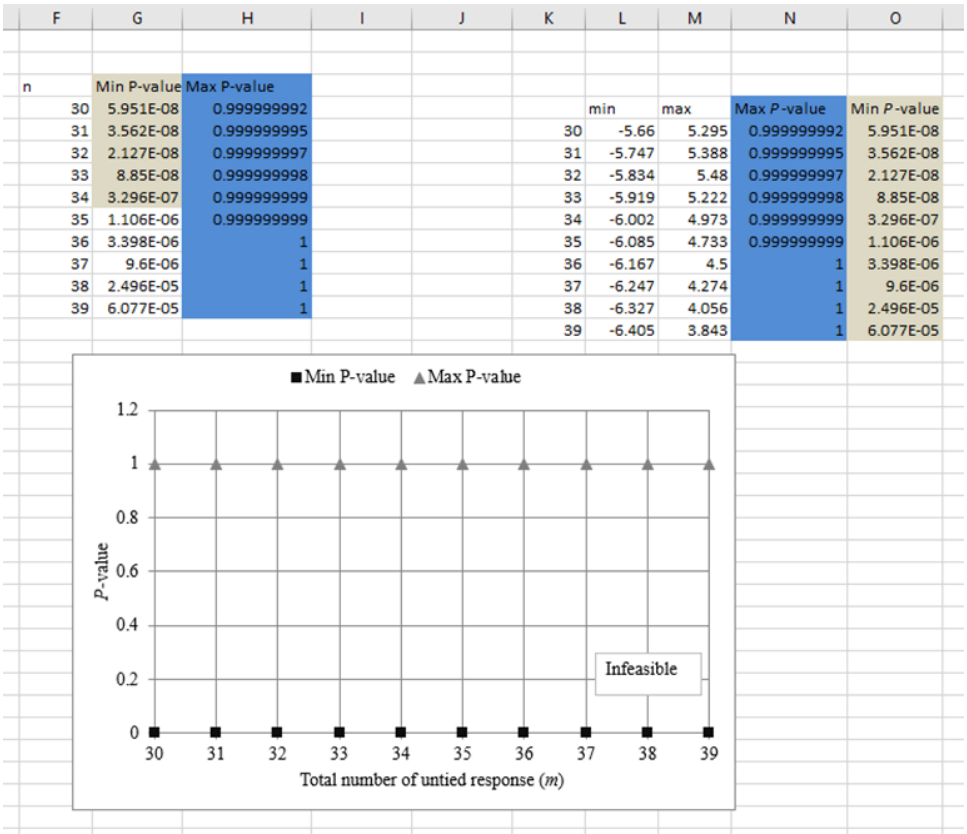


Fig 7. Snippet of Output on AMPL

30.000	-5.660	30.000	5.295
31.000	-5.747	31.000	5.388
32.000	-5.834	32.000	5.480
33.000	-5.919	33.000	5.222
34.000	-6.002	34.000	4.973
35.000	-6.085	35.000	4.733
36.000	-6.167	36.000	4.500
37.000	-6.247	37.000	4.274
38.000	-6.327	38.000	4.056
39.000	-6.405	39.000	3.843

Fig 8. Max and Min Z-values

We use these Z-values and find the respective p-values using the 1-NORMSDIST() function on excel and plot these against the Total number of untied responses.



n	Min P-value	Max P-value
30	5.951E-08	0.999999992
31	3.562E-08	0.999999995
32	2.127E-08	0.999999997
33	8.85E-08	0.999999998
34	3.296E-07	0.999999999
35	1.106E-06	0.999999999
36	3.398E-06	1
37	9.6E-06	1
38	2.496E-05	1
39	6.077E-05	1

	min	max	Max *P*-value	Min *P*-value
30	-5.66	5.295	0.999999992	5.951E-08
31	-5.747	5.388	0.999999995	3.562E-08
32	-5.834	5.48	0.999999997	2.127E-08
33	-5.919	5.222	0.999999998	8.85E-08
34	-6.002	4.973	0.999999999	3.296E-07
35	-6.085	4.733	0.999999999	1.106E-06
36	-6.167	4.5	1	3.398E-06
37	-6.247	4.274	1	9.6E-06
38	-6.327	4.056	1	2.496E-05
39	-6.405	3.843	1	6.077E-05

Fig 9. Calculating p_max and p_min for Z-values and plotting it against Total Number of untied Responses

FINDINGS

Our analysis yields a decreasing p_min value and increasing p_max value as unmatched responses rise. This divergence prevents determining a clear p-value threshold to reject or fail to reject the null hypothesis. Thus, our data lacks adequate sensitivity to conclusively assess if high blood pressure causes diabetes when accounting for confounds. We require more robust data to define the specific effect of high blood pressure on diabetes risk.

STUDENT 4 : ESHA JOSHI (002208382)

This study aims to investigate the potential correlation between physical activity and the incidence of diabetes. The null hypothesis posits that there is no substantial association between physical activity and the presence of diabetes in the population. Conversely, the alternative hypothesis contends that there exists a significant relationship between physical activity and the occurrence of diabetes. Our objective is to discern whether the level of physical activity significantly influences the likelihood of individuals developing diabetes.

Null Hypothesis (H0):
"There is no significant association between physical activity and the occurrence of diabetes in the population."

Alternative Hypothesis (H1):
"There is a significant association between physical activity and the occurrence of diabetes in the population."

ANALYSIS

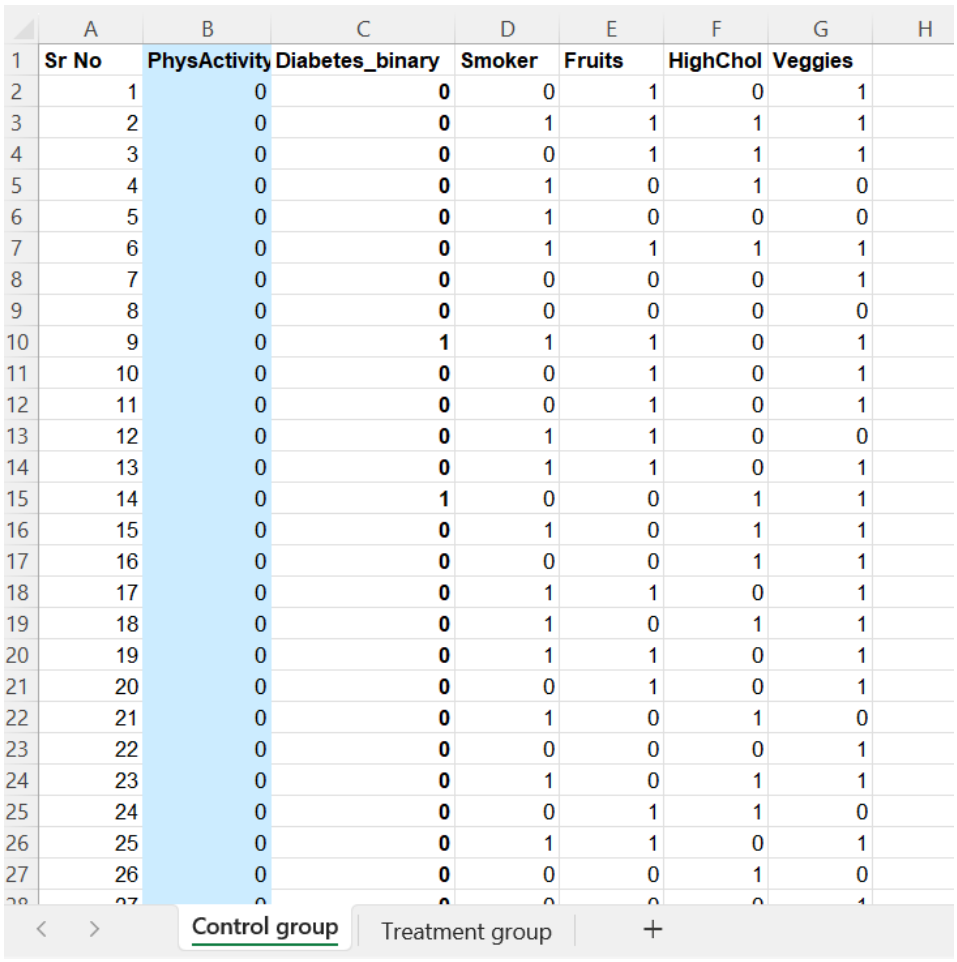
The study investigates whether physical activity is significantly associated with the occurrence of diabetes in the population. For this analysis, four additional covariates, namely smoking, high cholesterol and consumption of fruits and vegetables are included, where all contain binary variables.

We first segregate the data into Treatment Group and Control Group. The treatment group gets the experiment, and the control group, without the experiment, gives a baseline for comparison to see the impact of the treatment. In our study analyzing the impact of physical activity on diabetes:

Treatment Group: Individuals who engage in physical activity (experimental condition), i.e., value = 1.

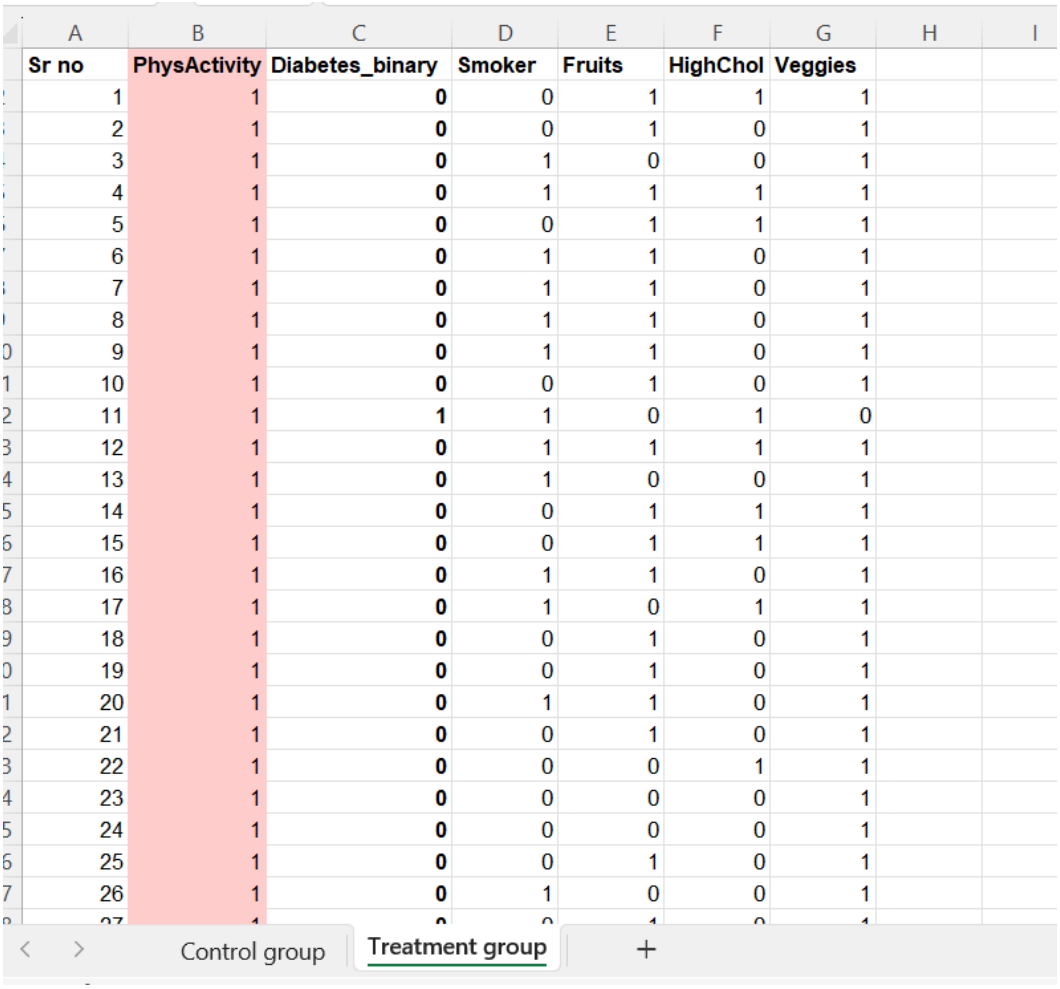
Control Group: Individuals who do not engage in physical activity (baseline or comparison condition), i.e., value = 0.

Here all values in the column named physical activity have 1 for Treatment Group and 0 for Control Group. Now we load this data into AMPL to conduct the McNemar's test on our dataset.



	A	B	C	D	E	F	G	H
1	Sr No	PhysActivity	Diabetes_binary	Smoker	Fruits	HighChol	Veggies	
2	1	0	0	0	1	0	1	
3	2	0	0	1	1	1	1	
4	3	0	0	0	1	1	1	
5	4	0	0	1	0	1	0	
6	5	0	0	1	0	0	0	
7	6	0	0	1	1	1	1	
8	7	0	0	0	0	0	1	
9	8	0	0	0	0	0	0	
10	9	0	1	1	1	0	1	
11	10	0	0	0	1	0	1	
12	11	0	0	0	1	0	1	
13	12	0	0	1	1	0	0	
14	13	0	0	1	1	0	1	
15	14	0	1	0	0	1	1	
16	15	0	0	1	0	1	1	
17	16	0	0	0	0	1	1	
18	17	0	0	1	1	0	1	
19	18	0	0	1	0	1	1	
20	19	0	0	1	1	0	1	
21	20	0	0	0	1	0	1	
22	21	0	0	1	0	1	0	
23	22	0	0	0	0	0	1	
24	23	0	0	1	0	1	1	
25	24	0	0	0	1	1	0	
26	25	0	0	1	1	0	1	
27	26	0	0	0	0	1	0	

Fig- Control group for physical activity



Sr no	PhysActivity	Diabetes_binary	Smoker	Fruits	HighChol	Veggies
1	1	**0**	0	1	1	1
2	1	**0**	0	1	0	1
3	1	**0**	1	0	0	1
4	1	**0**	1	1	1	1
5	1	**0**	0	1	1	1
6	1	**0**	1	1	0	1
7	1	**0**	1	1	0	1
8	1	**0**	1	1	0	1
9	1	**0**	1	1	0	1
10	1	**0**	0	1	0	1
11	1	**1**	1	0	1	0
12	1	**0**	1	1	1	1
13	1	**0**	1	0	0	1
14	1	**0**	0	1	1	1
15	1	**0**	0	1	1	1
16	1	**0**	1	1	0	1
17	1	**0**	1	0	1	1
18	1	**0**	0	1	0	1
19	1	**0**	0	1	0	1
20	1	**0**	1	1	0	1
21	1	**0**	0	1	0	1
22	1	**0**	0	0	1	1
23	1	**0**	0	0	0	1
24	1	**0**	0	0	0	1
25	1	**0**	0	1	0	1
26	1	**0**	1	0	0	1

Control group | Treatment group

Fig- Treatment group for physical activity

Data files –

We have three .dat files namely: Diabetes.dat which includes the number of rows in the Treatment and Control Groups and the Covariates set, Physicalactivity.dat which includes the covariates data for the Treatment and Control Groups, and Outcome_Diabetes.dat which includes the outcome (does or does not have Diabetes) of the Treatment and Control Groups.

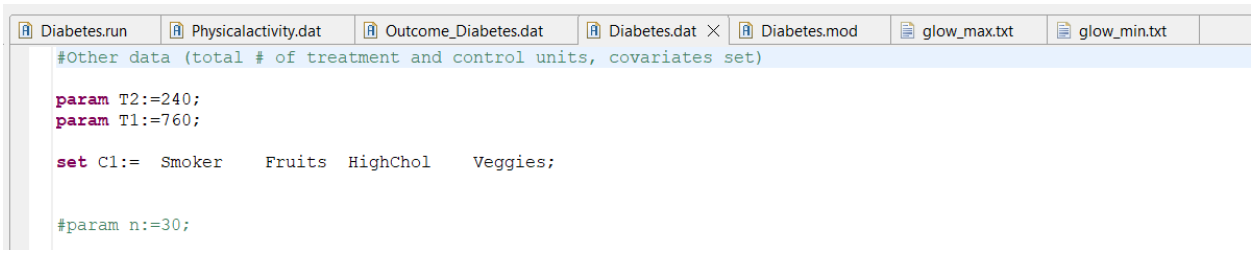


Fig- Diabetes.dat file

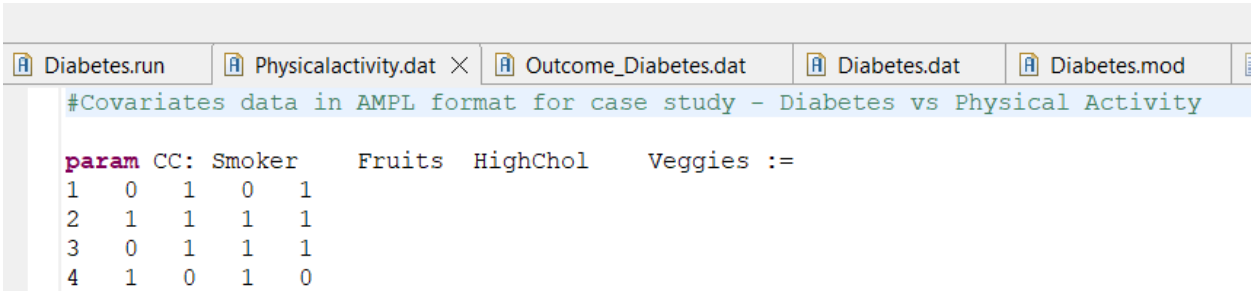


Fig- Physical activity.dat file

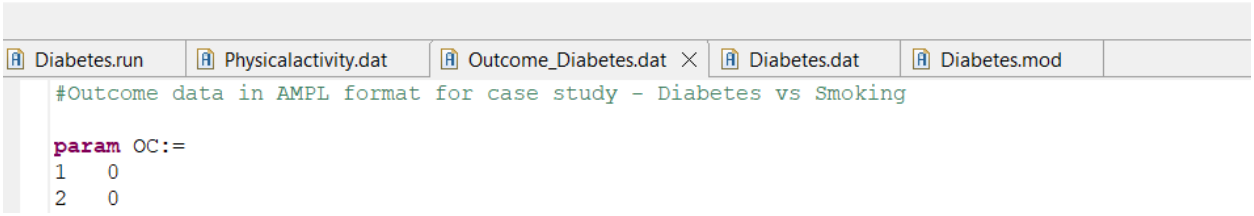


Fig- Outcome_diabetes.dat file

Model file -

The model file Diabetes.mod is used to define the discordant pairs using McNemar's test. The algorithm finds pairs of one unit each from the Treatment and Control Groups, where the covariate values are similar, such that the only difference is in the value of the Treatment or Control group variable (in our case, Physical Activity). This is to determine whether Physical Activity can be considered as the only differentiating factor between people with and without diabetes.

Run file –

The run file Diabetes.run uses the above model and data files to determine the Z-values from which we can draw our hypothesis' conclusion. Here, we set the permissible difference between the values of the covariates in the Treatment and Control Groups. Thus, pairs are made in such a way that we keep the covariate values uniform whilst only having differing values of Physical Activity.

High Cholesterol: This is a binary variable this means that the variables are described in the form 0 and 1 so there are no discrepancies in the data.1 indicates high cholesterol and 0 represents no cholesterol.
Veggies: Consumption of one or more veggies in a day is 1 and 0 is no vegetable consumption.
Fruits: the consumption of fruits is represented in 0 and 1. 0 indicating no fruit consumption and 1 indicating fruit consumption of 1 or more times per day

Smoking: 0 indicates no smoking and 1 indicates smoking. This is a binary variable hence it allows the analysis to take place without inconsistency.

Output –
We create two files with the minimum and maximum Z-values, respectively, when we execute the Diabetes.run file.

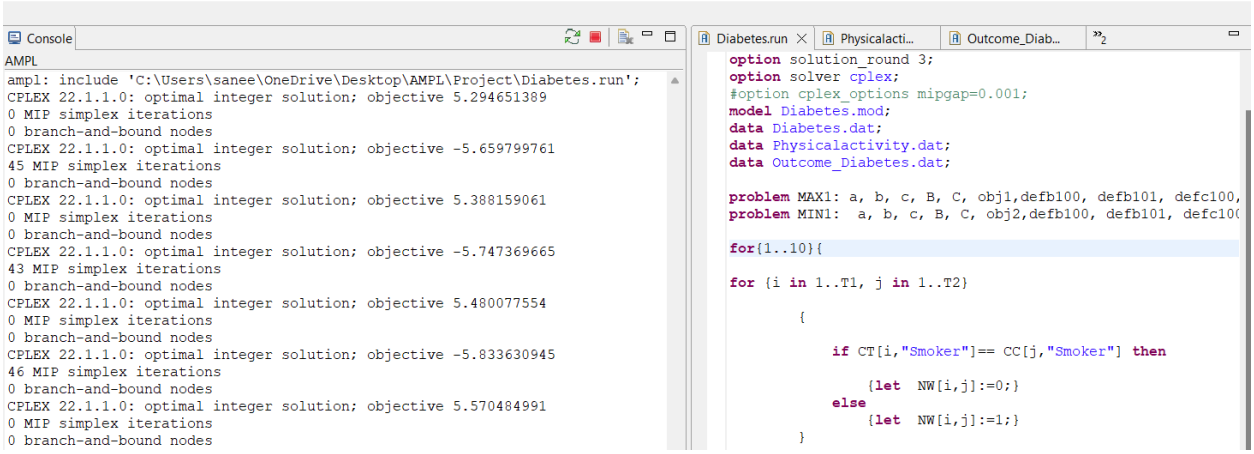


Fig- Run file and outcome.

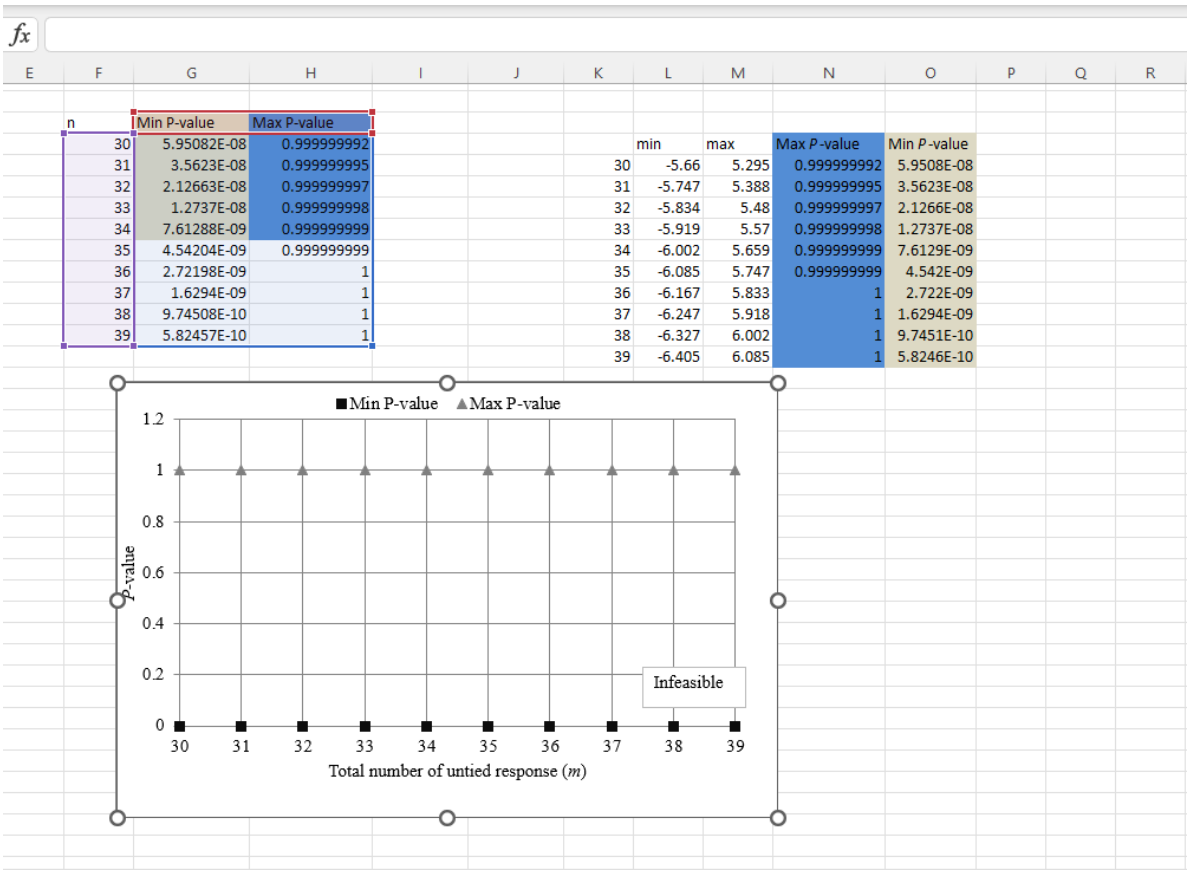


Fig- Max and Min P value with graph

FINDINGS

From our research we can conclude that the p_min and p_max values are divergent and do not converge into a common p-value that would indicate whether our null hypothesis should be rejected or not (p>0.05 indicates that we do not reject the null hypothesis, and p<0.05 indicates that we do reject the null hypothesis). As we raise the total number of unbound responses, p_min continues to decrease and p_max continues to climb. Therefore, it can be concluded that our dataset does not offer sufficient information to draw firm conclusions We can therefore conclude that physical activity has no effect on the prevalence of diabetes.

STUDENT 5: SHWETA SHINDE (002819510)

In this research, we are exploring the hypothesis that high cholesterol is closely tied to the occurrence of diabetes. According to the null hypothesis, there isn't a strong correlation between high cholesterol and the prevalence of diabetes in the general population. On the other hand, the alternate theory claims that there is a strong correlation between excessive cholesterol and the development of diabetes. Our goal in doing this research is to determine whether having high cholesterol significantly affects a person's risk of developing diabetes.

Null Hypothesis (H0):
"There is no significant association between high cholesterol and the occurrence of diabetes in the population."

Alternative Hypothesis (H1):
"There is a significant association between high cholesterol and the occurrence of diabetes in the population."

ANALYSIS
The purpose of the study is to determine whether a population's risk of developing diabetes is significantly correlated with having high cholesterol. Four other factors are included in this analysis: **High BP, BMI, Fruits, and Veggies**. BMI is a numerical variable, and High BP Physical Fruits and Veggies are binary variables.

Data is first divided into two groups: the Treatment Group and the Control Group. The control group provides a baseline for comparison to assess the treatment's impact, while the treatment group receives the experiment.

In our study analyzing the impact of high cholesterol on diabetes:

Treatment Group: Individuals who have high cholesterol i.e., column value = 1.

Control Group: Individuals who do not have high cholesterol i.e., column value = 0.

Our Treatment and Control Groups are as follows –

F	G	H	I	J	K	L	M
Sno	**Diabetes_bin**	**HighChol**	**HighBP**	**BMI**	**Fruits**	**Veggies**	
1	**0**	1	1	28	1	1	
2	**0**	1	0	31	1	1	
3	**0**	1	1	33	1	1	
4	**0**	1	0	29	1	1	
5	**0**	1	0	27	0	0	
6	**0**	1	1	25	1	1	
7	**0**	1	1	22	1	1	
8	**1**	1	1	27	0	0	
9	**0**	1	1	26	1	1	
10	**0**	1	1	28	1	1	
11	**0**	1	0	23	1	1	
12	**0**	1	1	30	0	1	
13	**0**	1	0	47	0	1	
14	**0**	1	1	28	1	1	
15	**0**	1	0	26	1	1	
16	**1**	1	1	33	1	1	
17	**0**	1	1	32	1	1	
18	**1**	1	1	40	0	1	
19	**0**	1	1	29	0	1	
20	**0**	1	0	47	0	1	
21	**0**	1	1	24	1	1	
22	**0**	1	0	77	1	1	
23	**0**	1	1	27	0	1	

Fig 1. Treatment Group

	F	G	H	I	J	K	L	N
	Sno	Diabetes_bir	HighChol	HighBP	BMI	Fruits	Veggies	
	1	0	0	0	21	1	1	
	2	0	0	0	24	1	1	
	3	0	0	0	27	0	1	
	4	0	0	0	33	1	1	
	5	0	0	0	27	1	1	
	6	0	0	0	36	0	0	
	7	0	0	0	33	1	1	
	8	0	0	0	26	1	1	
	9	0	0	0	31	0	1	
	10	0	0	0	20	1	1	
	11	0	0	0	21	0	1	
	12	0	0	0	29	0	0	
	13	1	0	0	48	1	1	
	14	0	0	1	36	1	1	
	15	0	0	0	22	1	1	
	16	0	0	0	27	1	1	
	17	0	0	1	25	1	1	
	18	0	0	0	26	1	1	
	19	0	0	0	20	1	1	
	20	0	0	0	24	0	1	
	21	0	0	0	26	0	1	
	22	0	0	1	43	1	1	
	23	0	0	0	27	1	1	
	24	0	0	1	33	0	1	
	25	0	0	0	21	1	1	
	26	0	0	1	26	1	0	
	27	0	0	0	25	1	1	

Fig 2. Control Group

The values in the "high cholesterol" column are all 1 for the Treatment Group and 0 for the Control Group, as you can see.

We now put this data into AMPL so that we may run our dataset through the McNemar's test.

Data files –

We have three.dat files: Outcome_Diabetes.dat contains the outcome (does not have diabetes) of the Treatment and Control Groups; Diabetes.dat contains the number of rows in the Treatment and Control Groups and the Covariates set; and HighCh_covariates.dat contains the covariates data for the Treatment and Control Groups.

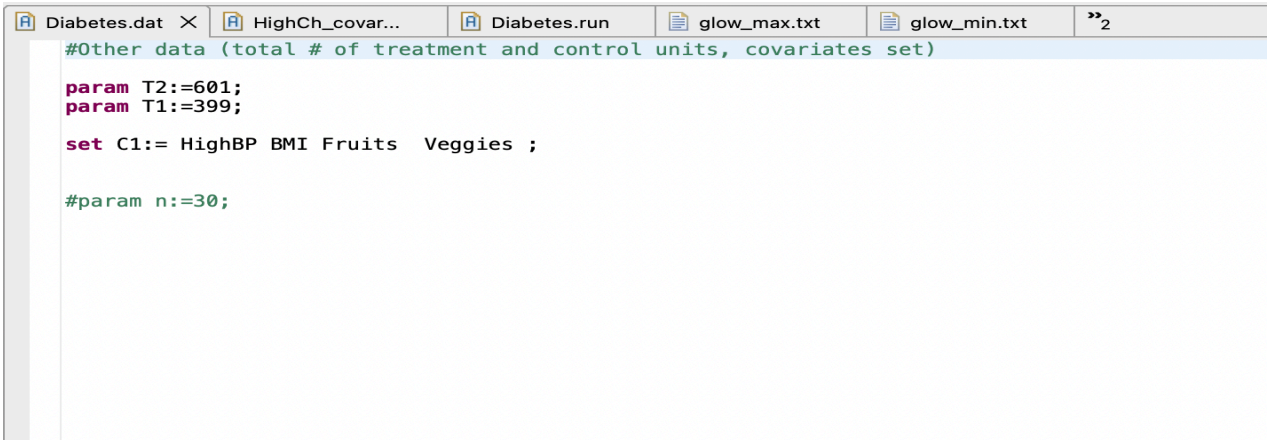


Fig 3. Snippet of Diabetes.dat

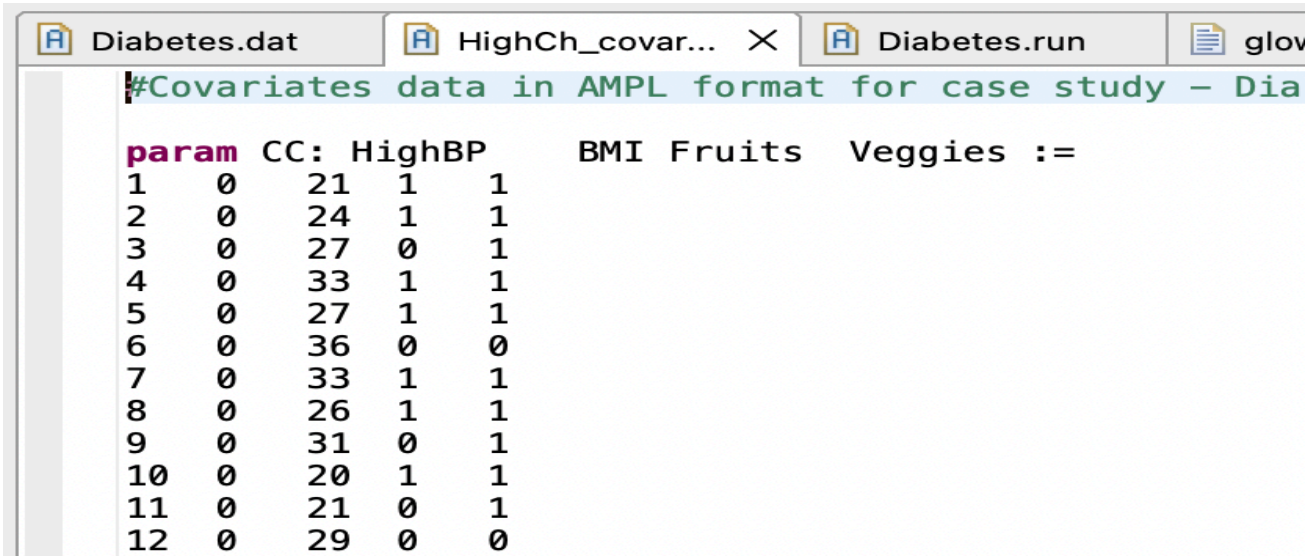


Fig 4. Snippet of HighCh_covariates*.dat*

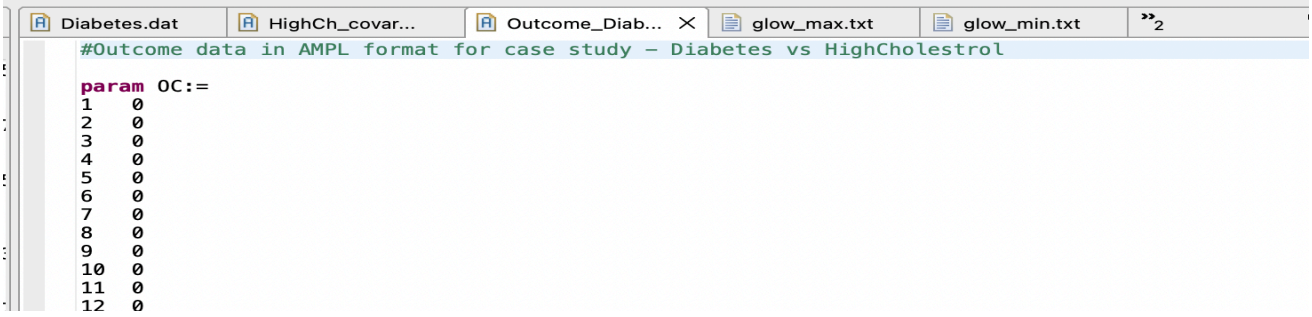


Fig 5. Snippet of Outcome_Diabetes.dat

Model file -

Using McNemar's test, the model file Diabetes.mod is utilized to define the discordant pairs. The method identifies pairs of one unit each from the Treatment and Control Groups when the covariate values are comparable, leaving the Treatment or Control group variable's value (high cholesterol, in this case) as the only difference. The purpose of this is to ascertain whether or not high cholesterol is the sole factor that distinguishes individuals with diabetes from those who do not, and whether or not high cholesterol contributes to the development of diabetes.

Run file –

The model and data files mentioned above are used by the run file Diabetes.run to calculate the Z-values, which allow us to conclude our hypothesis. The allowable difference between the covariate values in the Treatment and Control Groups is determined here. As a result, pairs are created with simply varying values of high cholesterol (0 for no high cholesterol, 1 for high cholesterol), maintaining the covariate values uniform.

BMI: The permissible difference in BMI of Treatment and Control Group units is considered to be 5.

High BP: This is a binary variable so the only way that there is no difference between the HighBP value in units of Treatment and Control groups is if they are exactly equal (either 0 or 1). Hence the permissible difference is 0.

Fruits: This is a binary variable so the only way that there is no difference between the Fruits value in units of Treatment and Control groups is if they are exactly equal (either 0 or 1). Hence the permissible difference is 0.

Vegetables: This is a binary variable so the only way that there is no difference between the Vegetables value in units of Treatment and Control groups is if they are exactly equal (either 0 or 1). Hence the permissible difference is 0.

Number of iterations = 10

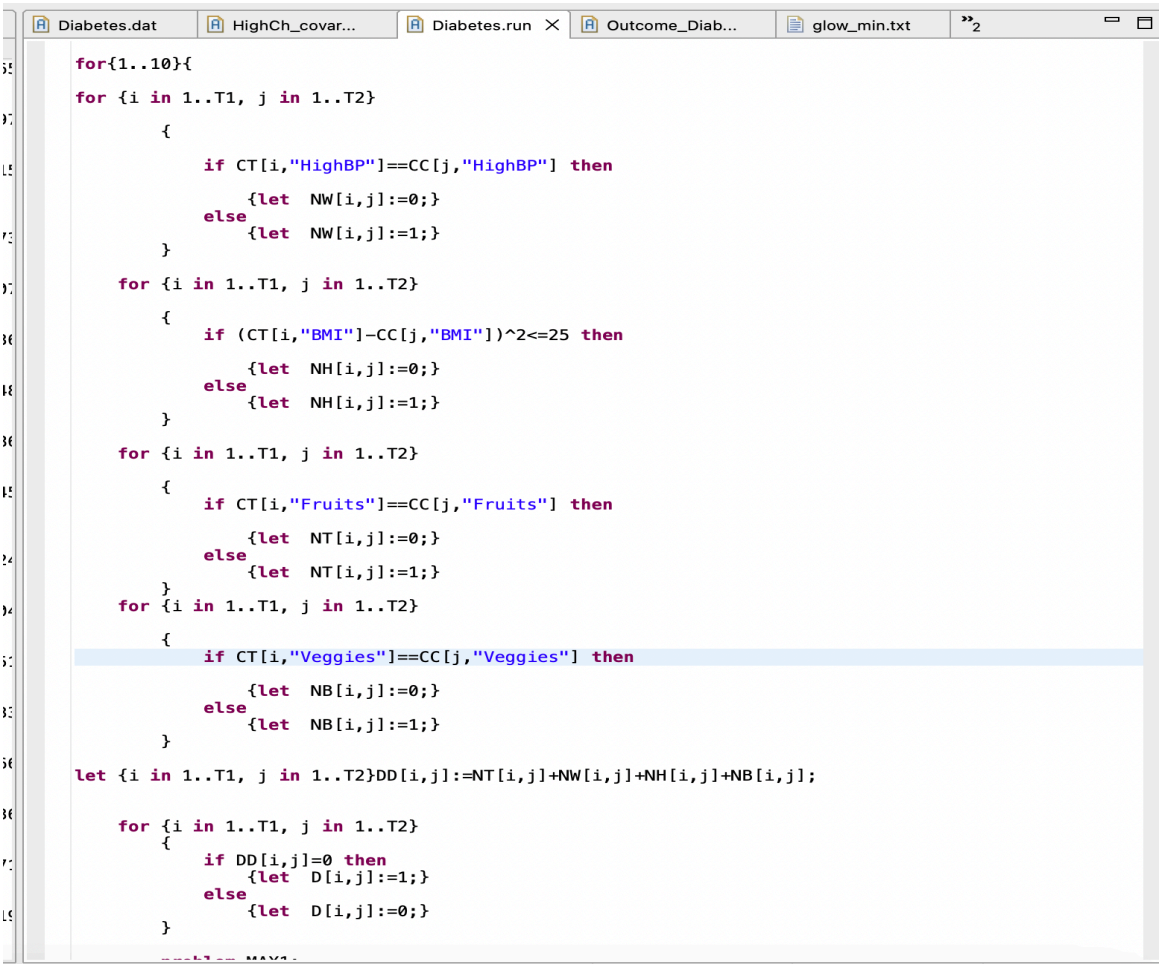


Fig 6. Snippet of Diabetes.run

Output –

We create two files with the minimum and maximum Z-values, respectively, when we execute the Diabetes.run file.

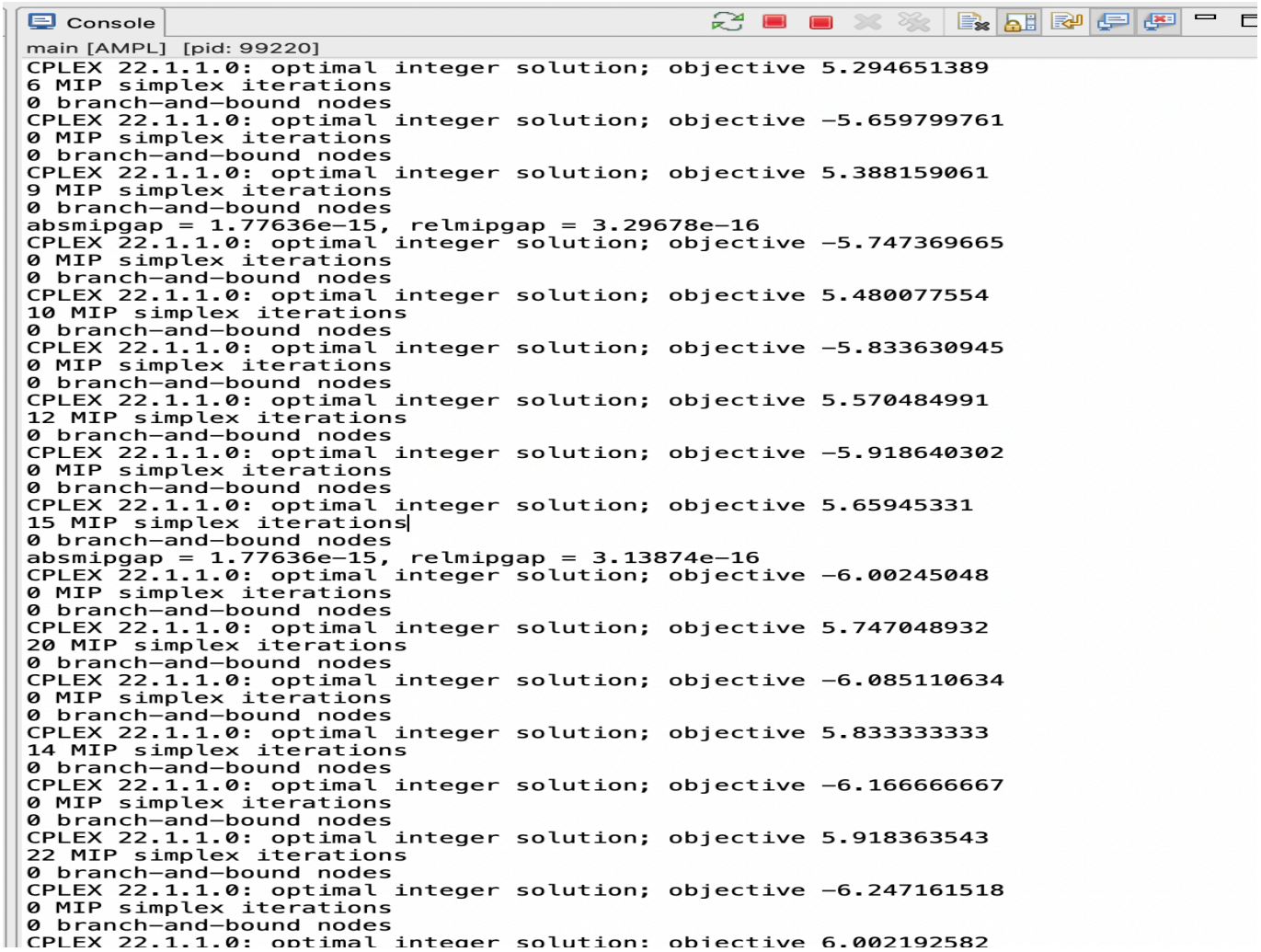
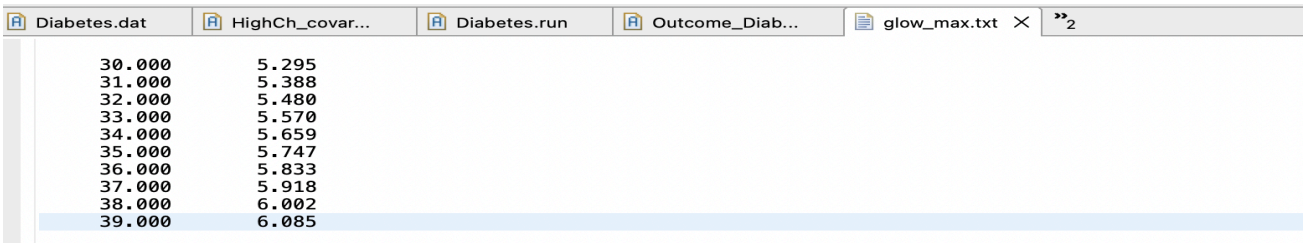


Fig 7. Snippet of Output on AMPL



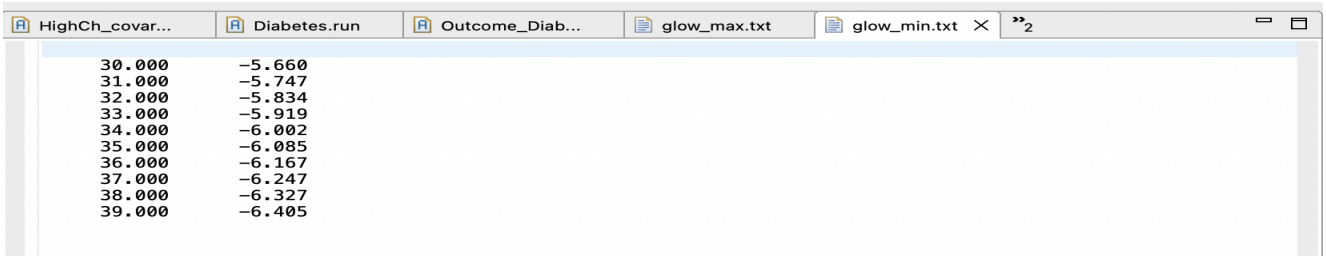


Fig 8. Max and Min Z-values

Using Excel's 1-NORMSDIST() function, we utilize these Z-values to calculate the corresponding p-values, which we then plot against the total number of untied responses.

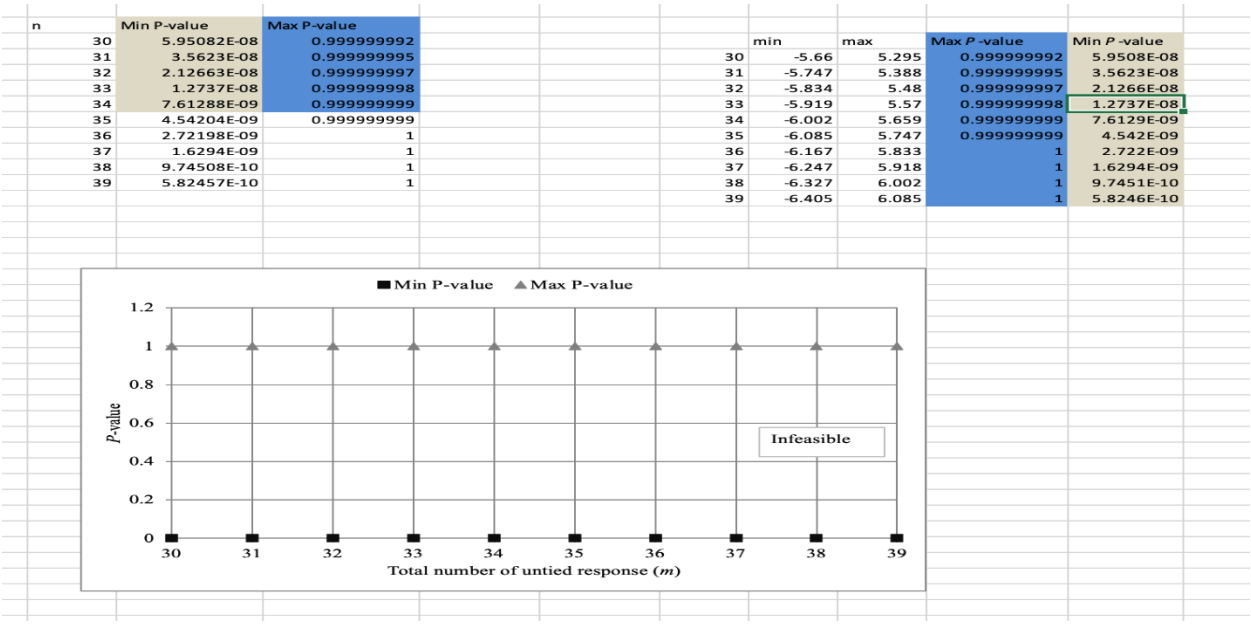


Fig 9. Calculating p_max and p_min for Z-values and plotting it against Total Number of untied Responses

FINDINGS

We may conclude from our research that the p_min and p_max values are divergent and do not converge into a common p-value that would indicate whether our null hypothesis should be rejected or not (p>0.05 indicates that we do not reject the null hypothesis, and p<0.05 indicates

that we do reject the null hypothesis). As we raise the total number of unbound responses, p_min continues to decrease and p_max continues to climb. Therefore, it can be concluded that our dataset does not offer sufficient information to draw firm conclusions. Diabetes is not proven to be caused by excessive cholesterol.

CONCLUSION

In our analyses exploring potential factors associated with diabetes risk, we examined the impacts of smoking, vegetable consumption, high blood pressure, physical activity levels, and high cholesterol. However, across all five of our investigation areas, we encountered limitations with inconclusive results. Specifically, the divergence of minimum and maximum p-values as untied responses increased prevented us from determining clear thresholds to reject or fail to reject the null hypotheses. Essentially, the data available to us proved inadequate in terms of sensitivity and robustness for making definitive determinations. This highlights the complex, multi-faceted nature of diabetes, which can be influenced by a variety of elements including family history. While we established exploratory optimization models and conducted preliminary statistical tests, the quality and depth of data constrained our ability to draw firm conclusions. Our findings indicate that to comprehensively understand the factors causing increased diabetes prevalence, we need to investigate other potential variables beyond those examined here. Moving forward, access to more extensive, higher fidelity datasets will likely be necessary for us to complete conclusive analyses that substantiate relationships between hypothesized influencing variables and diabetes outcomes. In summary, our findings were ultimately limited across all analyses, emphasizing the need to explore additional factors and secure robust data to uncover possible causes for this disease.