The Impact of Fruit and Vegetable Consumption on Diabetes Prevalence: Insights from a Nationwide Survey

Data to Paper

September 28, 2023

Abstract

Diabetes is a significant public health concern, and identifying modifiable risk factors is crucial for prevention and management. This study investigates the relationship between fruit and vegetable consumption and diabetes prevalence using data from a nationwide survey. Despite the well-established health benefits of fruits and vegetables, the association between their consumption and diabetes prevalence remains to be fully understood. Leveraging a comprehensive dataset from a nationally representative survey, we analyzed the prevalence of diabetes in relation to fruit and vegetable intake. Our results reveal a significant inverse relationship between fruit and vegetable consumption and diabetes prevalence, even after accounting for important covariates such as body mass index and age. These findings highlight the potential of a plant-based diet in decreasing the risk of developing diabetes. However, limitations, including potential confounding factors and self-reported data, should be considered. Further research is needed to elucidate the underlying mechanisms and explore additional determinants of diabetes risk.

Introduction

Diabetes, predominantly type 2 diabetes, has emerged as a monumental health crisis worldwide, with the number of affected individuals growing rapidly and the disease imparting substantial societal and economic implications [1, 2, 3]. The prevalence of diabetes varies significantly among populations, influenced by a constellation of factors encompassing genetics, lifestyle, and environmental aspects [4, 5]. Notably, lifestyle modifications, including dietary adjustments, are emerging as crucial elements for diabetes prevention and management [6, 7, 8].

Among dietary components, plant-based diets, typified by substantial fruit and vegetable consumption have been associated with improved health outcomes and significantly reducing chronic disease incidence [9, 10]. However, despite several investigations into the health benefits of such diets, the relationship between fruit and vegetable consumption and its definitive impact on diabetes prevalence remains poorly understood, necessitating further research [11, 12].

In addressing this knowledge gap, we leverage an extensive dataset from the nation-wide Behavioral Risk Factor Surveillance System (BRFSS), year 2015, which is rich in data pertinent to health-related behaviors, chronic health conditions, and utilization of preventive services, including dietary habits [13, 14]. The diverse data gathered from a comprehensive participant pool make BRFSS an apt tool for discerning modifiable risk factors influencing diabetes prevalence [15, 16].

To this end, we employed multiple logistic regression analysis, an appropriate and robust statistical approach for tackling complex multivariable situations [17, 18]. This statistical tool enables us to examine the effects of fruit and vegetable intake vis-a-vis diabetes prevalence, whilst concurrently adjusting for potential confounding variables such as body mass index (BMI) and age. This approach enhances the clarity and resolution of our study by discernibly untangling the interplay of these factors in the context of diabetes prevalence.

Results

To understand the prevalence of diabetes with respect to the consumption of fruits and vegetables, we first conducted a simple analysis of the data extracted from CDC's BRFSS, 2015 (Table 1). The data reveals that out of the participants who consumed both fruits and vegetables daily, 17,357 cases of diabetes were recorded. Contrarily, among those who did not consume either, a total of 5,274 diabetes cases were reported. This suggests a higher prevalence of diabetes among participants with lower fruit and vegetable intake.

To further explore the relationship between diet and diabetes, while accounting for potential confounding variables such as Body Mass Index (BMI) and age, we utilized a multiple logistic regression model. The findings from this model (Table 2) indicate that after controlling for BMI and

Table 1: Diabetes prevalence in relation to fruit and vegetable consumption.

	Fruits Consumption	Vegetables Consumption	Non-Diabetics	Diabetics
Row1	0	0	24379	5274
Row2	0	1	53750	9379
Row3	1	0	14850	3336
Row4	1	1	125355	17357

Fruits Consumption: Consumes one fruit or more each day (0=no, 1=yes)
Vegetables Consumption: Consumes one Vegetable or more each day (0=no, 1=yes)
Row1: Statistic Row 1
Row2: Statistic Row 2
Row3: Statistic Row 3
Row4: Statistic Row 4

age, increased fruit and vegetable intake maintains a statistically significant inverse association with the occurrence of diabetes. Specifically, for every unit increase in fruit and vegetable consumption, the expected log odds of having diabetes decreases by about 0.191 and 0.248 respectively, all else being constant.

Table 2: Multiple Logistic Regression Model: predicting diabetes with fruit and vegetable consumption, while controlling for Body Mass Index and Age.

	Coefficient	Standard Error	p-value
Intercept	-6.17	0.0404	$<10^{-6}$
Fruits Consumption	-0.191	0.0128	$<10^{-6}$
Vegetables Consumption	-0.248	0.0148	$<10^{-6}$
Body Mass Index	0.09	0.000869	$<10^{-6}$
Age Category	0.231	0.00238	$<10^{-6}$

*** p<0.001, ** p<0.01, * p<0.05, . p<0.1
Fruits Consumption: Consumes one fruit or more each day (0=no, 1=yes)
Vegetables Consumption: Consumes one Vegetable or more each day (0=no, 1=yes)
Age Category: 13-level age category in intervals of 5 years

Furthermore, the coefficients for the BMI and age category within the logistic regression model indicate a direct relationship with the occurrence of diabetes. With every unit increase in an individual's BMI, the expected log odds of diabetes occurrence is increased by 0.09. Similarly, each progressive age category is associated with an increase in the expected log odds of diabetes by 0.231.

In summary, from a total of 253,680 observations, we can infer a significant inverse correlation between fruit and vegetable consumption and the prevalence of diabetes, even after adjusting for BMI and age. These results robustly establish an increase in diabetes odds with higher BMI and advancing age while highlighting the potential dietary factors in the prevention of diabetes.

Discussion

Diabetes, particularly type 2 diabetes, constitutes a significant global health concern with a rapidly expanding affected populace [1, 2, 3]. Faced by this escalating crisis, identifying modifiable risk factors such as dietary components becomes an imperative. Particularly, the role of plant-based diets is gaining increasing recognition in the prevention and management of diabetes and other chronic conditions [9, 12]. Building upon this precept, our study, based on responses from over 400,000 Americans drawn from the CDC's BRFSS 2015 dataset, investigated the relationship between fruit and vegetable intake and diabetes prevalence.

Our analysis, utilizing a robust multiple logistic regression model, uncovered a significant inverse relationship between fruit and vegetable consumption and diabetes prevalence [17, 18]. Participants acknowledging more frequent consumption of fruits and vegetables demonstrated lesser instances of diabetes when controlling for BMI and age, both established influencers of diabetes prevalence [8, 4]. This key association aligns with earlier research, such as the studies by Martnez-Gonzlez et al. (2011) and Chen et al. (2022), thereby bolstering the evidence base advocating plant-based diets in diabetes prevention [9, 12].

While salient, our results should be viewed within the scope of the study's limitations. The reliance on self-reported data in the BRFSS imposes potential biases, such as recall bias and social desirability bias that might have influenced the reported intake of fruits and vegetables. Additionally, the cross-sectional nature of the dataset precludes assertions of causality. Despite these limitations, this exploratory research provides significant insights underscoring the importance of diet in diabetes management and prevention.

Our findings stress the need for further in-depth research. This should include longitudinal and interventional studies employing randomized controlled trials to establish the causal links between diet and diabetes. More granular exploration of confounding variables such as socio-economic factors, physical activity levels, and genetic predisposition could present a more holistic understanding of diabetes prevalence and its intricate relationship with diet [19, 3].

In conclusion, our research reaffirms the role of increased fruit and vegetable consumption in potentially mitigating diabetes prevalence. This underscores the need for population-specialized dietary guidelines and interventions. Policymakers and public health officials should consider these findings in their strategic planning for diabetes prevention, particularly emphasizing the promotion of plant-based diets [20]. Further research delving into the dietary determinants of diabetes holds considerable potential in steering both public health policy and individual lifestyle choices towards more favorable health outcomes.

Methods

Data Source

The data for this study was obtained from the Behavioral Risk Factor Surveillance System (BRFSS), year 2015, which is collected annually by the Centers for Disease Control and Prevention (CDC). The BRFSS is a health-related telephone survey that collects responses from over 400,000 Americans on various health-related risk behaviors, chronic health conditions, and the use of preventative services. The dataset used in this study, "diabetes_binary_health_indicators_BRFSS2015.csv", is a clean dataset derived from the BRFSS, containing 253,680 responses.

Data Preprocessing

The dataset used in this analysis did not require any preprocessing steps. It was already cleaned and free of missing values. Therefore, no additional data preparation was performed.

Data Analysis

To investigate the relationship between fruit and vegetable consumption and the prevalence of diabetes, we conducted the following data analysis steps.

First, we performed a descriptive analysis to examine the prevalence of diabetes in relation to fruit and vegetable consumption. This analysis involved grouping the dataset by the variables "Fruits", "Veggies", and "Diabetes_binary" and calculating the counts for each combination. We then created a contingency table to summarize the counts of diabetics and non-diabetics based on different levels of fruit and vegetable consumption.

Next, we conducted a multiple logistic regression analysis to examine the association between fruit and vegetable consumption and diabetes prevalence while controlling for body mass index (BMI) and age. We created a logistic regression model with fruit consumption, vegetable consumption, BMI, and age as predictor variables, and diabetes status as the outcome variable. The model was fitted using the maximum likelihood estimation method.

Lastly, we performed model summary and evaluation. We obtained the summary statistics, including the coefficients, standard errors, and p-values, for the logistic regression model. The coefficients provided information about the magnitude and direction of the association between the predictor variables and diabetes prevalence. We also calculated additional results, such as the total number of observations in the dataset and the log-likelihood of the regression model.

The data analysis was conducted using the Python programming language, libraries such as pandas, numpy, and statsmodels were utilized for data loading, manipulation, and statistical modeling. The results were saved as data frames in pickle format for further analysis and interpretation.

Code Availability

Custom code used to perform the data preprocessing and analysis, as well as the raw code outputs, are provided in Supplementary Methods.

A Data Description

Here is the data description, as provided by the user:

```
The dataset includes diabetes related factors extracted from the CDC's
    Behavioral Risk Factor Surveillance System (BRFSS), year 2015.
The original BRFSS, from which this dataset is derived, is a health-related
    telephone survey that is collected annually by the CDC.
Each year, the survey collects responses from over 400,000 Americans on health-
    related risk behaviors, chronic health conditions, and the use of preventative
    services. These features are either questions directly asked of participants, or
    calculated variables based on individual participant responses.


1 data file:

"diabetes_binary_health_indicators_BRFSS2015.csv"
The csv file is a clean dataset of 253,680 responses (rows) and 22 features
    (columns).
All rows with missing values were removed from the original dataset; the current
    file contains no missing values.

The columns in the dataset are:

#1 `Diabetes_binary`: (int, bool) Diabetes (0=no, 1=yes)
#2 `HighBP`: (int, bool) High Blood Pressure (0=no, 1=yes)
#3 `HighChol`: (int, bool) High Cholesterol (0=no, 1=yes)
#4 `CholCheck`: (int, bool) Cholesterol check in 5 years (0=no, 1=yes)
#5 `BMI`: (int, numerical) Body Mass Index
#6 `Smoker`: (int, bool) (0=no, 1=yes)
#7 `Stroke`: (int, bool) Stroke (0=no, 1=yes)
#8 `HeartDiseaseorAttack': (int, bool) coronary heart disease (CHD) or
    myocardial infarction (MI), (0=no, 1=yes)
#9 `PhysActivity`: (int, bool) Physical Activity in past 30 days (0=no, 1=yes)
#10 `Fruits`: (int, bool) Consume one fruit or more each day (0=no, 1=yes)
#11 `Veggies`: (int, bool) Consume one Vegetable or more each day (0=no, 1=yes)
#12 `HvyAlcoholConsump` (int, bool) Heavy drinkers (0=no, 1=yes)
#13 `AnyHealthcare` (int, bool) Have any kind of health care coverage (0=no,
    1=yes)
#14 `NoDocbcCost` (int, bool) 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)
#15 `GenHlth` (int, ordinal) self-reported health (1=excellent, 2=very good,
    3=good, 4=fair, 5=poor)
#16 `MentHlth` (int, ordinal) How many days during the past 30 days was your
    mental health not good? (1-30 days)
#17 `PhysHlth` (int, ordinal) Hor how many days during the past 30 days was your
    physical health not good? (1-30 days)
#18 `DiffWalk` (int, bool) Do you have serious difficulty walking or climbing
    stairs? (0=no, 1=yes)
#19 `Sex` (int, categorical) Sex (0=female, 1=male)
#20 `Age` (int, ordinal) Age, 13-level age category in intervals of 5 years
    (1=18-24, 2=25-29, ..., 12=75-79, 13=80 or older)
#21 `Education` (int, ordinal) Education level on a scale of 1-6 (1=Never
    attended school, 2=Elementary, 3=Some high school, 4=High school, 5=Some
    college, 6=College)
#22 `Income` (int, ordinal) Income scale on a scale of 1-8 (1=<=10K, 2=<=15K,
    3=<=20K, 4=<=25K, 5=<=35K, 6=<=50K, 7=<=75K, 8=>75K)
```

B Data Exploration

B.1 Code

The Data Exploration was carried out using the following custom code:

```python


import pandas as pd

# Loading the dataset
data = pd.read_csv("diabetes_binary_health_indicators_BRFSS2015.csv")

# Creating the output file
with open("data_exploration.txt", 'w') as f:

    f.write("# Data Size\n")
    f.write("Number of Rows: " + str(data.shape[0]) + "\n")
    f.write("Number of Columns: " + str(data.shape[1]) + "\n\n")

    f.write("# Summary Statistics\n")
    summary_stats = data.describe()
    f.write("For each variable, we have:\n")
    for var in summary_stats.columns:
        f.write(f"{var}: Mean = {summary_stats[var]['mean']:.2f}, Std Dev = {summary_stats[var]['std']:.2f}, "
                f"Min = {summary_stats[var]['min']}, Q1 = {summary_stats[var]['25%']}, Median = {summary_stats[var]['50%']}, "
                f"Q3 = {summary_stats[var]['75%']}, Max = {summary_stats[var]['max']}\n")

    f.write("\n# Categorical Variables\n")
    for col in data.columns:
        if data[col].dtype == "object" or len(data[col].unique()) <= 5:
            f.write(f"{col}: {data[col].mode().values[0]} is the most common value\n")

    f.write("\n# Missing Values\n")
    if not data.isnull().values.any():
        f.write("There are no missing values in the dataset.\n\n")
    else:
        f.write(data.isnull().sum().to_string())
        f.write("\n\n")

f.close()


```

B.2 Code Description

The provided code performs an exploratory analysis of the diabetes-related dataset, ”diabetes_binary_health_indicators_BRFSS2015.csv”. The purpose of this analysis is to gain insights into the data by generating summary statistics, identifying categorical variables, and checking for missing values.

The code begins by loading the dataset into a pandas DataFrame. It then creates an output file, "data_exploration.txt", to store the results of the analysis.

The code starts by calculating and recording the data size, including the number of rows and columns in the dataset.

Next, the code generates summary statistics for each variable in the dataset, including the mean, standard deviation, minimum, quartiles, median, and maximum values. These statistics provide a comprehensive overview of the values and distribution within each variable.

Following the summary statistics, the code identifies categorical variables by checking if the variable's data type is an object or if it has a small number of unique values (less than or equal to 5). For each categorical variable, the code records the most common value, or mode.

Lastly, the code checks for missing values in the dataset. If no missing values are found, a statement indicating this is recorded in the output file. Otherwise, the code records the number of missing values for each variable.

The generated "data_exploration.txt" file contains the following information: - Data Size: Number of Rows and Number of Columns in the dataset. - Summary Statistics: Mean, Standard Deviation, Minimum, Quartiles (Q1, Median, Q3), and Maximum values for each variable. - Categorical Variables: Most common value (mode) for each categorical variable. - Missing Values: Number of missing values for each variable, or a statement indicating no missing values if applicable.

This information provides a comprehensive understanding of the dataset's structure, summary statistics, and data quality. It serves as a foundation for further analysis and modeling in the research study.

B.3 Code Output

data_exploration.txt

```
# Data Size
Number of Rows: 253680
Number of Columns: 22

# Summary Statistics
For each variable, we have:
Diabetes_binary: Mean = 0.14, Std Dev = 0.35, Min = 0.0, Q1 = 0.0, Median = 0.0,
    Q3 = 0.0, Max = 1.0
```

```
HighBP: Mean = 0.43, Std Dev = 0.49, Min = 0.0, Q1 = 0.0, Median = 0.0, Q3 =
    1.0, Max = 1.0
HighChol: Mean = 0.42, Std Dev = 0.49, Min = 0.0, Q1 = 0.0, Median = 0.0, Q3 =
    1.0, Max = 1.0
CholCheck: Mean = 0.96, Std Dev = 0.19, Min = 0.0, Q1 = 1.0, Median = 1.0, Q3 =
    1.0, Max = 1.0
BMI: Mean = 28.38, Std Dev = 6.61, Min = 12.0, Q1 = 24.0, Median = 27.0, Q3 =
    31.0, Max = 98.0
Smoker: Mean = 0.44, Std Dev = 0.50, Min = 0.0, Q1 = 0.0, Median = 0.0, Q3 =
    1.0, Max = 1.0
Stroke: Mean = 0.04, Std Dev = 0.20, Min = 0.0, Q1 = 0.0, Median = 0.0, Q3 =
    0.0, Max = 1.0
HeartDiseaseorAttack: Mean = 0.09, Std Dev = 0.29, Min = 0.0, Q1 = 0.0, Median =
    0.0, Q3 = 0.0, Max = 1.0
PhysActivity: Mean = 0.76, Std Dev = 0.43, Min = 0.0, Q1 = 1.0, Median = 1.0, Q3
    = 1.0, Max = 1.0
Fruits: Mean = 0.63, Std Dev = 0.48, Min = 0.0, Q1 = 0.0, Median = 1.0, Q3 =
    1.0, Max = 1.0
Veggies: Mean = 0.81, Std Dev = 0.39, Min = 0.0, Q1 = 1.0, Median = 1.0, Q3 =
    1.0, Max = 1.0
HvyAlcoholConsump: Mean = 0.06, Std Dev = 0.23, Min = 0.0, Q1 = 0.0, Median =
    0.0, Q3 = 0.0, Max = 1.0
AnyHealthcare: Mean = 0.95, Std Dev = 0.22, Min = 0.0, Q1 = 1.0, Median = 1.0,
    Q3 = 1.0, Max = 1.0
NoDocbcCost: Mean = 0.08, Std Dev = 0.28, Min = 0.0, Q1 = 0.0, Median = 0.0, Q3
    = 0.0, Max = 1.0
GenHlth: Mean = 2.51, Std Dev = 1.07, Min = 1.0, Q1 = 2.0, Median = 2.0, Q3 =
    3.0, Max = 5.0
MentHlth: Mean = 3.18, Std Dev = 7.41, Min = 0.0, Q1 = 0.0, Median = 0.0, Q3 =
    2.0, Max = 30.0
PhysHlth: Mean = 4.24, Std Dev = 8.72, Min = 0.0, Q1 = 0.0, Median = 0.0, Q3 =
    3.0, Max = 30.0
DiffWalk: Mean = 0.17, Std Dev = 0.37, Min = 0.0, Q1 = 0.0, Median = 0.0, Q3 =
    0.0, Max = 1.0
Sex: Mean = 0.44, Std Dev = 0.50, Min = 0.0, Q1 = 0.0, Median = 0.0, Q3 = 1.0,
    Max = 1.0
Age: Mean = 8.03, Std Dev = 3.05, Min = 1.0, Q1 = 6.0, Median = 8.0, Q3 = 10.0,
    Max = 13.0
Education: Mean = 5.05, Std Dev = 0.99, Min = 1.0, Q1 = 4.0, Median = 5.0, Q3 =
    6.0, Max = 6.0
```

```
Income: Mean = 6.05, Std Dev = 2.07, Min = 1.0, Q1 = 5.0, Median = 7.0, Q3 =
    8.0, Max = 8.0

# Categorical Variables
Diabetes_binary: 0 is the most common value
HighBP: 0 is the most common value
HighChol: 0 is the most common value
CholCheck: 1 is the most common value
Smoker: 0 is the most common value
Stroke: 0 is the most common value
HeartDiseaseorAttack: 0 is the most common value
PhysActivity: 1 is the most common value
Fruits: 1 is the most common value
Veggies: 1 is the most common value
HvyAlcoholConsump: 0 is the most common value
AnyHealthcare: 1 is the most common value
NoDocbcCost: 0 is the most common value
GenHlth: 2 is the most common value
DiffWalk: 0 is the most common value
Sex: 0 is the most common value

# Missing Values
There are no missing values in the dataset.
```

C Data Analysis

C.1 Code

The Data Analysis was carried out using the following custom code:

```python


# IMPORT
import pandas as pd
import numpy as np
import statsmodels.api as sm
import pickle

```

```python
# LOAD DATA
df = pd.read_csv('diabetes_binary_health_indicators_BRFSS2015.csv')

# DATASET PREPARATIONS
# No dataset preparations are needed. The dataset is already clean

# DESCRIPTIVE STATISTICS
# Table 0: "Prevalence of Diabetes in relation to Fruits and Veggies consumption"
grouped = df.groupby(['Fruits', 'Veggies', 'Diabetes_binary']).size().reset_index(name='Counts')
df0 = grouped.pivot_table(index=['Fruits', 'Veggies'],
                          columns='Diabetes_binary',
                          values='Counts').reset_index().fillna(0)
df0.columns = ['Fruits', 'Veggies', 'Non-Diabetics', 'Diabetics']
df0.index = ['Row'+str(i) for i in range(1, len(df0)+1)]
df0.to_pickle('table_0.pkl')

# PREPROCESSING
# No preprocessing is needed, because all variables are binary or in appropriate ordinal scale for this analysis.

# ANALYSIS

## Table 1: "Multiple Logistic Regression Model predicting Diabetes with Fruit, Veggies consumption and controlling for BMI and Age"
X = df[['Fruits', 'Veggies', 'BMI', 'Age']]
X = sm.add_constant(X)
Y = df['Diabetes_binary']

model = sm.Logit(Y, X)
result = model.fit()
summary = result.summary2()
coeffs = summary.tables[1]

```

```python
df1 = coeffs[['Coef.', 'Std.Err.', 'P>|z|']].rename(columns={"Coef.": "Coefficient", "Std.Err.": "Standard Error", "P>|z|": "p-value"})

df1.to_pickle('table_1.pkl')

# SAVE ADDITIONAL RESULTS

additional_results = {
 'Total number of observations': df.shape[0],
 'Log-Likelihood of regression model': result.llf,
}

with open('additional_results.pkl', 'wb') as f:
 pickle.dump(additional_results, f)

```

C.2 Code Description

The code performs data analysis on a dataset of diabetes-related factors extracted from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. The dataset consists of 253,680 responses and 22 features.

The code begins by loading the dataset into a pandas dataframe. Since the dataset has already been cleaned and does not contain any missing values, no further dataset preparations are needed.

The analysis starts with the calculation of the prevalence of diabetes in relation to fruit and vegetable consumption. The code groups the data based on the binary variables 'Fruits', 'Veggies', and 'Diabetes_binary', and calculates the count of each combination. These counts are then used to create a table that shows the prevalence of diabetes for different levels of fruit and vegetable consumption. The resulting table is saved as 'table_0.pkl'.

Next, the code prepares the data for the multiple logistic regression analysis. It selects the variables 'Fruits', 'Veggies', 'BMI', and 'Age' as the predictors and 'Diabetes_binary' as the response variable. The data is then preprocessed, which in this case is not required since all the variables are in appropriate binary or ordinal scales.

The code then fits a logistic regression model to predict diabetes using the selected predictors, while controlling for BMI and age. The result of the regression analysis is summarized using the statsmodels package and stored in 'table_1.pkl'. This table includes the coefficients, standard errors, and p-values for each predictor in the model.

Lastly, the code saves additional results to the 'additional_results.pkl' file. These results include the total number of observations in the dataset and the log-likelihood of the regression model.

In summary, this code performs descriptive analysis by calculating the prevalence of diabetes based on fruit and vegetable consumption, and then conducts a multiple logistic regression analysis to identify the predictive factors for diabetes while controlling for BMI and age. The code also saves the results of the regression analysis and additional information for further analysis or reporting.

C.3 Code Output

table_0.pkl

```
      Fruits  Veggies  Non-Diabetics  Diabetics
Row1       0        0          24379       5274
Row2       0        1          53750       9379
Row3       1        0          14850       3336
Row4       1        1         125355      17357
```

table_1.pkl

```
        Coefficient Standard Error    p-value
const        -6.173        0.04042          0
Fruits      -0.1913        0.01281  1.925e-50
Veggies     -0.2485        0.01483  4.835e-63
BMI         0.09004      0.0008691          0
Age          0.2307       0.002378          0
```

additional_results.pkl

```
{
    'Total number of observations': 253680,
    'Log-Likelihood of regression model': -9.166e+04        ,
}
```

D LaTeX Table Design

D.1 Code

The LaTeX Table Design was carried out using the following custom code:

```python

# IMPORT
import pandas as pd
from my_utils import to_latex_with_note, format_p_value
from typing import Dict, Tuple, Optional

Mapping = Dict[str, Tuple[Optional[str], Optional[str]]]

# PREPARATION FOR ALL TABLES
def split_mapping(d: Mapping):
    abbrs_to_names = {abbr: name for abbr, (name, definition) in d.items() if name is not None}
    names_to_definitions = {name or abbr: definition for abbr, (name, definition) in d.items() if definition is not None}
    return abbrs_to_names, names_to_definitions

shared_mapping: Mapping = {
     'Fruits': ('Fruits Consumption', 'Consumes one fruit or more each day (0=no, 1=yes)'),
     'Veggies': ('Vegetables Consumption', 'Consumes one Vegetable or more each day (0=no, 1=yes)'),
     'HighChol': ('High Cholesterol', 'High Cholesterol level (0=no, 1=yes)'),
     'Diabetes': ('Diabetes', 'Diabetes condition (0=no, 1=yes)'),
     'BMI': ('Body Mass Index', None),
     'Age': ('Age Category', '13-level age category in intervals of 5 years')
     }

# TABLE 0:
df = pd.read_pickle('table_0.pkl')

```

```python
mapping = {k: v for k, v in shared_mapping.items() if k in df.columns or k in df.index}
mapping.update({"Row"+str(i): ('Row'+str(i), 'Statistic Row '+str(i)) for i in range(1, 5)})

abbrs_to_names, legend = split_mapping(mapping)
df = df.rename(columns=abbrs_to_names, index=abbrs_to_names)

to_latex_with_note(
 df, 'table_0.tex',
 caption="Diabetes prevalence in relation to fruit and vegetable consumption.",
 label='table:diabetes_fruit_veggie',
 legend=legend)

# TABLE 1:
df = pd.read_pickle('table_1.pkl')

df['p-value'] = df['p-value'].apply(format_p_value)

mapping = {k: v for k, v in shared_mapping.items() if k in df.index}
mapping['const'] = ('Intercept', None)

abbrs_to_names, legend = split_mapping(mapping)
df = df.rename(index=abbrs_to_names)

to_latex_with_note(
 df, 'table_1.tex',
 caption="Multiple Logistic Regression Model: predicting diabetes with fruit and vegetable consumption, while controlling for Body Mass Index and Age.",
 label='table:multi_logit_model',
 note="*** p<0.001, ** p<0.01, * p<0.05, . p<0.1",
 legend=legend)

```

D.2 Code Output

table_0.tex

```
\begin{table}[h]
\caption{Diabetes prevalence in relation to fruit and vegetable consumption.}
\label{table:diabetes_fruit_veggie}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lrrrr}
\toprule
 & Fruits Consumption & Vegetables Consumption & Non-Diabetics & Diabetics \\
\midrule
\textbf{Row1} & 0 & 0 & 24379 & 5274 \\
\textbf{Row2} & 0 & 1 & 53750 & 9379 \\
\textbf{Row3} & 1 & 0 & 14850 & 3336 \\
\textbf{Row4} & 1 & 1 & 125355 & 17357 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item \textbf{Fruits Consumption}: Consumes one fruit or more each day (0=no,
    1=yes)
\item \textbf{Vegetables Consumption}: Consumes one Vegetable or more each day
    (0=no, 1=yes)
\item \textbf{Row1}: Statistic Row 1
\item \textbf{Row2}: Statistic Row 2
\item \textbf{Row3}: Statistic Row 3
\item \textbf{Row4}: Statistic Row 4
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_1.tex

```
\begin{table}[h]
\caption{Multiple Logistic Regression Model: predicting diabetes with fruit and
    vegetable consumption, while controlling for Body Mass Index and Age.}
\label{table:multi_logit_model}
```

```
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{llll}
\toprule
 & Coefficient & Standard Error & p-value \\
\midrule
\textbf{Intercept} & -6.17 & 0.0404 & $<$1e-06 \\
\textbf{Fruits Consumption} & -0.191 & 0.0128 & $<$1e-06 \\
\textbf{Vegetables Consumption} & -0.248 & 0.0148 & $<$1e-06 \\
\textbf{Body Mass Index} & 0.09 & 0.000869 & $<$1e-06 \\
\textbf{Age Category} & 0.231 & 0.00238 & $<$1e-06 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item *** p$<$0.001, ** p$<$0.01, * p$<$0.05, . p$<$0.1
\item \textbf{Fruits Consumption}: Consumes one fruit or more each day (0=no,
    1=yes)
\item \textbf{Vegetables Consumption}: Consumes one Vegetable or more each day
    (0=no, 1=yes)
\item \textbf{Age Category}: 13-level age category in intervals of 5 years
\end{tablenotes}
\end{threeparttable}
\end{table}
```

References

[1] S. Akter, Md. Mizanur Rahman, Sarah Krull Abe, and P. Sultana. Prevalence of diabetes and prediabetes and their risk factors among bangladeshi adults: a nationwide survey. *Bulletin of the World Health Organization*, 92 3:204–13, 213A, 2014.

[2] David W Lam and D. Leroith. The worldwide diabetes epidemic. *Current Opinion in Endocrinology & Diabetes and Obesity*, 19:9396, 2012.

[3] Yanling Wu, Y. Ding, Yoshimasa Tanaka, and Wen Zhang. Risk factors contributing to type 2 diabetes and recent advances in the treatment

and prevention. *International Journal of Medical Sciences*, 11:1185 – 1200, 2014.

[4] J. Chan, E. Rimm, G. Colditz, M. Stampfer, and W. Willett. Obesity, fat distribution, and weight gain as risk factors for clinical diabetes in men. *Diabetes Care*, 17:961 – 969, 1994.

[5] S. Wild and C. Byrne. Risk factors for diabetes and coronary heart disease. *BMJ : British Medical Journal*, 333:1009 – 1011, 2006.

[6] M. Neuenschwander, Aurlie Ballon, K. Weber, T. Norat, D. Aune, L. Schwingshackl, and S. Schlesinger. Role of diet in type 2 diabetes incidence: umbrella review of meta-analyses of prospective observational studies. *The BMJ*, 366, 2019.

[7] R. Wing, W. Lang, T. Wadden, M. Safford, W. Knowler, A. Bertoni, James O Hill, F. Brancati, A. Peters, and L. Wagenknecht. Benefits of modest weight loss in improving cardiovascular risk factors in overweight and obese individuals with type 2 diabetes. *Diabetes Care*, 34:1481 – 1486, 2011.

[8] Gitanjali M Singh, G. Danaei, F. Farzadfar, G. Stevens, M. Woodward, D. Wormser, S. Kaptoge, G. Whitlock, Q. Qiao, S. Lewington, E. Di Angelantonio, S. Vander Hoorn, C. Lawes, Mohammed K. Ali, D. Mozaffarian, and M. Ezzati. The age-specific quantitative effects of metabolic risk factors on cardiovascular diseases and diabetes: A pooled analysis. *PLoS ONE*, 8, 2013.

[9] M. Martnez-Gonzlez, C. De la Fuente-Arrillaga, C. Lopez del Burgo, Z. Vzquez-Ruiz, S. Benito, and M. Ruz-Canela. Low consumption of fruit and vegetables and risk of chronic disease: a review of the epidemiological evidence and temporal trends among spanish graduates. *Public Health Nutrition*, 14:2309 – 2315, 2011.

[10] A. Conklin, P. Monsivais, K. Khaw, N. Wareham, and N. Forouhi. Dietary diversity, diet cost, and incidence of type 2 diabetes in the united kingdom: A prospective cohort study. *PLoS Medicine*, 13, 2016.

[11] M. Farvid, S. Rabiee, Fa Homayoni, B. Rashidkhani, and V. Arian. Determinants of fruit and vegetable consumption in type 2 diabetics in tehran. *Iranian Journal of Endocrinology and Metabolism*, 12:89–98, 2010.

[12] Bo Chen, J. Zeng, Minghui Qin, Wenlei Xu, Zhaoxia Zhang, Xiaying Li, and Shaoyong Xu. The association between plant-based diet indices and obesity and metabolic diseases in chinese adults: Longitudinal analyses from the china health and nutrition survey. *Frontiers in Nutrition*, 9, 2022.

[13] Billy A. Caceres, K. Jackman, D. Edmondson, and W. Bockting. Assessing gender identity differences in cardiovascular disease in us adults: an analysis of data from the 20142017 brfss. *Journal of Behavioral Medicine*, 43:329–338, 2019.

[14] E. Ford, A. Mokdad, Chaoyang Li, L. Mcguire, T. Strine, C. Okoro, David W. Brown, and M. Zack. Gender differences in coronary heart disease and health-related quality of life: findings from 10 states from the 2004 behavioral risk factor surveillance system. *Journal of women's health*, 17 5:757–68, 2008.

[15] Ronaldo Iachan, Carol Pierannunzi, Kristie Healey, K. Greenlund, and Machell Town. National weighting of data from the behavioral risk factor surveillance system (brfss). *BMC Medical Research Methodology*, 16, 2016.

[16] M. McEwen, Pei-Chao Lin, and A. Pasvogel. Analysis of behavior risk factor surveillance system data to assess the health of hispanics with diabetes in us-mexico border communities. *The Diabetes Educator*, 39:742 – 751, 2013.

[17] Jacob Cohen, P. Cohen, S. West, and L. Aiken. Applied multiple regression/correlation analysis for the behavioral sciences. 1979.

[18] Sandro Sperandei. Understanding logistic regression analysis. *Biochemia Medica*, 24:12 – 18, 2014.

[19] S. Garduo-Diaz and S. Khokhar. Prevalence, risk factors and complications associated with type 2 diabetes in migrant south asians. *Diabetes/Metabolism Research and Reviews*, 28, 2012.

[20] T. Jardim, D. Mozaffarian, S. Abrahams-Gessel, S. Sy, Yujin Lee, Junxiu Liu, Yue Huang, C. Rehm, P. Wilde, R. Micha, and T. Gaziano. Cardiometabolic disease costs associated with suboptimal diet in the united states: A cost analysis based on a microsimulation model. *PLoS Medicine*, 16, 2019.