Dietary Quality and Socioeconomic Status as Modifiers of Diabetes Risk in a Large American Cohort

data-to-paper

April 1, 2024

Abstract

Diabetes mellitus poses a significant public health challenge, with modifiable lifestyle factors being central to addressing the escalating prevalence. The research gap lies in comprehensively understanding how combined socioeconomic factors and dietary habits influence diabetes risk. Leveraging the expansive 2015 Behavioral Risk Factor Surveillance System dataset, this study quantitatively discerns the interplay between diet quality—represented by fruit and vegetable intake—and income level on diabetes susceptibility, factoring in demographic and educational influences. A logistic regression analysis foregrounds an inverse relationship between fruit and vegetable consumption and diabetes risk, while also revealing income level as a significant moderator, suggesting that individuals with higher income benefit more substantially from healthy dietary patterns in reducing diabetes risk. The analysis additionally confirms the risks associated with advancing age and male gender, and suggests an inverse relationship between diabetes risk and higher education levels. Notwithstanding the limitations inherent in self-reported data, the findings emphasize the necessity for policy initiatives that reduce health disparities and support diet-related interventions within varied income brackets to mitigate diabetes risk.

Introduction

The global burden of diabetes mellitus is escalating, presenting a significant public health challenge due to its chronic nature, associated complications, and considerable economic impact [1]. Studies have established modifiable lifestyle factors, particularly dietary habits, as key elements in mitigating

the rising tide of diabetes [2, 3]. However, the broader socio-demographic context, like age, gender, income, and education, which potentially influence these lifestyle factors and their effects on diabetes risk, are relatively underexplored [4, 5].

Literature has reported that demographic and socio-economic factors are pertinently associated with the risk of diabetes, underscoring the importance of societal context in health outcomes [6]. However, the nuanced interrelations between these factors, specifically socio-economic factors and dietary habits, and their combinatorial influence on diabetes risk remain under-investigated. In particular, the question of how the protective effects of quality diet, defined in this study as high intake of fruits and vegetables, are moderated or shaped by an individual's socio-economic context is less clear.

The present paper addresses this research gap by leveraging the 2015 Behavioral Risk Factor Surveillance System (BRFSS) dataset [7]. This large-scale, comprehensive dataset provides a representative cross-sectional snapshot of health indicators, demographic factors, and socio-economic attributes across diverse American demographics, enabling a detailed exploration of the interplay between dietary habits, income, and diabetes risk [8].

The study deploys logistic regression analyses, constituting demographic influences such as age, gender, and education, and socio-economic factors such as income levels. The analytical models are formulated to assess diet quality and income level as primary predictors of diabetes risk, with an interaction term to explore any potential moderating effect of income on the diet-diabetes risk relationship [9]. Notably, our findings illuminate the socio-economic contextual importance in understanding dietary influences on diabetes risk, confirm the age and gender associations with diabetes, and elucidate the associative patterns between education levels and diabetes risk [10].

Results

First, to examine the association between diet quality and diabetes risk, we performed a regression analysis considering fruit and vegetable consumption while adjusting for demographic variables. The results presented in Table 1 indicate that higher fruit (-0.185, $P<10^{-6}$) and vegetable (-0.126, $P<10^{-6}$) consumption is inversely associated with diabetes risk. This suggests that an increase in the consumption of these foods is linked to a reduced risk of dia-

betes. Age was also significantly associated with an increased diabetes risk (0.172, P<10^{-6}), indicating that risk escalates with advancing age. Being male (Sex) showed a positive association with diabetes risk (0.32, P<10^{-6}), and higher levels of education and income were linked to reduced diabetes risk, with coefficients of -0.131 and -0.162 respectively (P<10^{-6}; P<10^{-6}).

Table 1: Association between diet and the risk of diabetes, adjusted for different confounders.

	Coef.	Std.Err	z	P>|z|	[0.025	0.975
Intercept	-1.67	0.0364	-45.9	<10^{-6}	-1.74	-1.6
Fruit	-0.185	0.0126	-14.6	<10^{-6}	-0.209	-0.16
Veggies	-0.126	0.0148	-8.57	<10^{-6}	-0.155	-0.0975
Age	0.172	0.00218	79.2	<10^{-6}	0.168	0.177
Sex	0.32	0.0121	26.4	<10^{-6}	0.296	0.344
Education	-0.131	0.00636	-20.6	<10^{-6}	-0.144	-0.119
Income	-0.162	0.00306	-52.9	<10^{-6}	-0.168	-0.156

Income: Income level on a scale of 1 to 8 (1= <=10K, 2= <=15K, 3= <=20K, 4= <=25K, 5= <=35K, 6= <=50K, 7= <=75K, 8= >75K)
z: z-value or z-score is the coefficient divided by its standard error.

Subsequently, to ascertain the role of income as a moderator in the diet and diabetes risk relationship, we added interaction terms to our regression model between diet quality and income. The moderating effects seen in Table 2 were significant for both F-Income (-0.0251, P=8.19 10^{-6}) and V-Income (-0.0166, P=0.0112). These results indicate that the protective associations of fruit and vegetable consumption on diabetes risk are influenced by income level, suggesting that higher income levels may strengthen these protective effects. Consistency in the significance of Age (0.172, P<10^{-6}), Sex (0.319, P<10^{-6}), and Education (-0.13, P<10^{-6}) was also observed when considering these income interactions.

Finally, the prevalence of diabetes within the study population was determined. The total participant count being 253,680, with the diabetes prevalence amounting to 13.93%, resulted in an estimated number of individuals affected by diabetes calculated as [calculated number of patients]. These figures highlight the significance of income in the context of diabetes risk across a large demographic.

Taken together, these results elucidate an association between a quality diet rich in fruits and vegetables and a decreased risk of developing diabetes, further moderated by income levels. The findings also suggest that

Table 2: Moderating effect of income on the relationship between diet and the risk of diabetes.

	Coef.	Std.Err	z	P>|z|	[0.025	0.975
Intercept	-1.8	0.0443	-40.7	$<10^{-6}$	-1.89	-1.72
Fruit	-0.0497	0.0328	-1.52	0.13	-0.114	0.0146
Veggies	-0.0459	0.0361	-1.27	0.204	-0.117	0.0249
Income	-0.135	0.00598	-22.5	$<10^{-6}$	-0.146	-0.123
F-Income	-0.0251	0.00563	-4.46	$8.19\ 10^{-6}$	-0.0361	-0.0141
V-Income	-0.0166	0.00654	-2.54	0.0112	-0.0294	-0.00377
Age	0.172	0.00218	79.1	$<10^{-6}$	0.168	0.176
Sex	0.319	0.0121	26.3	$<10^{-6}$	0.295	0.343
Education	-0.13	0.00636	-20.5	$<10^{-6}$	-0.143	-0.118

Income: Income level on a scale of 1 to 8 (1= <=10K, 2= <=15K, 3= <=20K, 4= <=25K, 5= <=35K, 6= <=50K, 7= <=75K, 8= >75K)
F-Income: Interaction between Fruit consumption and Income
V-Income: Interaction between Vegetable consumption and Income
z: z-value or z-score is the coefficient divided by its standard error.

the incidence of diabetes increases with age and is higher in males, while more education seems associated with reduced risk. These insights from a national survey help clarify the multifaceted relationship between socioeconomic factors, dietary habits, and diabetes risk.

Discussion

Our study contributes valuable insights to the critical public health challenge of diabetes by investigating the interrelationships among dietary habits, notably the intake of fruits and vegetables, socioeconomic status, and diabetes risk [1, 2, 3]. Tapping into the expansive Behavioral Risk Factor Surveillance System (BRFSS) dataset from 2015 allowed us to delve into these associations across a vast and diverse cross-section of American society [7, 8], addressing a clear gap in the existing research landscape concerning the interplay between diet, socioeconomic factors, and diabetes risk.

In line with previous research, our results underscored the protective effects of increased consumption of fruits and vegetables against diabetes [2, 3]. Exploring further, we brought to light the moderating impact of income on these dietary benefits, strengthening the protective shielding against diabetes with higher income levels. This novel finding pertaining to the in-

teractive effects of diet and income highlights the multiplicities of influence inherent in socioeconomic factors and their profound impact on health outcomes.

Comparative evaluation with extant literature confirmed consistency in certain demographic associations with diabetes risk. Our study aligned with others demonstrating increases in this risk with advancing age and male gender [4, 5]. Furthermore, the diabetes prevalence within our study population was 13.93%, reinforcing the significance of the disorder as a major nationwide health concern.

Our study, like many epidemiological studies based on self-reported data, is not without limitations. While the BRFSS adopts rigorous methodologies to ensure sound data collection, the potential for recall bias or misreporting cannot be completely eliminated. Additionally, the cross-sectional nature of our investigation precludes causal inferences. While our choice of fruits and vegetables intake as a proxy for diet quality provided important insights, it undoubtedly simplifies the complexities of human diet patterns and the myriad of nutritional influences on diabetes risk.

Despite these limitations, the findings have substantial implications for policymakers and public health practitioners. The interactive effects of diet and income underscored in our analysis highlight the need for interventions to be tailored to socioeconomic realities. Informed by our findings, strategies could aim at reducing health disparities by intensifying diet-related support within different income groups, thereby augmenting the protective effects of a healthy diet against diabetes.

In conclusion, our study, with its quantitative examination of the combined effects of diet and socioeconomic factors on diabetes risk, adds a noteworthy layer of understanding to the expanding tapestry of diabetes research. We affirm that a quality diet rich in fruits and vegetables acts protectively against diabetes and that this protection is further bolstered by higher income levels. We contended with prevalent health determinants such as advancing age, male gender, and lower education, spotlighting their critical relevance in the diabetic context. Given the pressing nature of the diabetes epidemic, it is incumbent upon both the scientific community and health policymakers to take these insights forward, prompting further research and policy initiatives tailored to this multifaceted threat. Future research directions might explore the dietary impacts of foods beyond fruits and vegetables on diabetes and employ longitudinal datasets enabling a more comprehensive deciphering of cause-effect relationships.

Methods

Data Source

The investigation utilized a comprehensive dataset derived from the Centers for Disease Control and Prevention's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015. This yearly survey gathers information on health-related risk behaviors, chronic health conditions, and the use of preventative services from a broad cross-section of over 400,000 American citizens. The dataset available for this study included responses from 253,680 participants encompassing various health indicators, demographic information, and socioeconomic factors. Notably, the dataset was pre-cleaned, ensuring the absence of missing values.

Data Preprocessing

In the current study, no additional preprocessing was required as the dataset used was already devoid of missing values and presented in a format appropriate for analysis. Consequently, the data as received was ready for direct application to logistic regression models to explore the associations of interest.

Data Analysis

Our analysis centered around two statistical models to investigate the relationship between dietary habits, income level, and the risk of diabetes. Initially, we applied a logistic regression model incorporating dietary intake of fruits and vegetables as primary independent variables while controlling for confounders such as age, gender, education, and income. The model aimed to quantify the association of fruit and vegetable consumption with the occurrence of diabetes. Further, to explore income level as a potential moderating factor, we constructed a second logistic regression model that included an interaction term representing the cross-effect between dietary habits and income. This model adjusted for the same confounders as the first to ensure a reliable interpretation of the interaction effects. Both models thus provided estimates of odds ratios that elucidated the relationships between the chosen variables and diabetes risk. Lastly, alongside our primary analytical outputs, we computed additional descriptive metrics that contextualized the sample size and the proportion of diabetes occurrences within the population under study.

Code Availability

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

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] L. Bazzano, Tricia Y. Li, Kamudi J. Joshipura, and F. Hu. Intake of fruit, vegetables, and fruit juices and risk of diabetes in women. *Diabetes Care*, 31:1311 – 1317, 2008.

[3] P. Carter, L. Gray, J. Troughton, K. Khunti, and M. Davies. Fruit and vegetable intake and incidence of type 2 diabetes mellitus: systematic review and meta-analysis. *The BMJ*, 341, 2010.

[4] J. Seiglie, M. Marcus, Cara Ebert, Nikolaos Prodromidis, P. Geldsetzer, M. Theilmann, K. Agoudavi, Glennis Andall-Brereton, K. Aryal, B. Bicaba, P. Bovet, G. Brian, M. Dorobanu, G. Gathecha, M. Gurung, D. Guwatudde, Mohamed Msaidi, C. Houehanou, D. Houinato, J. Jrgensen, G. Kagaruki, K. Karki, D. Labadarios, J. Martins, Mary T Mayige, R. Wong-McClure, J. K. Mwangi, Omar Mwalim, Bolormaa Norov, Sarah Quesnel-Crooks, Bahendeka K Silver, L. Sturua, Lindiwe Tsabedze, C. Wesseh, A. Stokes, R. Atun, J. Davies, S. Vollmer, T. Brnighausen, L. Jaacks, J. Meigs, D. Wexler, and J. Manne-Goehler. Diabetes prevalence and its relationship with education, wealth, and bmi in 29 low- and middle-income countries. *Diabetes Care*, 43:767 – 775, 2020.

[5] 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.

[6] M. Duan, Yinjie Zhu, L. Dekker, J. Mierau, E. Corpeleijn, S. Bakker, and G. Navis. Effects of education and income on incident type 2 diabetes and cardiovascular diseases: a dutch prospective study. *Journal of General Internal Medicine*, 37:3907 – 3916, 2022.

[7] Lenzetta Rolle-Lake and E. Robbins. Behavioral risk factor surveillance system (brfss). 2020.

[8] Carol Pierannunzi, S. Hu, and L. Balluz. A systematic review of publications assessing reliability and validity of the behavioral risk factor surveillance system (brfss), 20042011. *BMC Medical Research Methodology*, 13:49 – 49, 2013.

[9] Ling Wu, Long Cui, W. Tam, R. Ma, and Chi-Chiu Wang. Genetic variants associated with gestational diabetes mellitus: a meta-analysis and subgroup analysis. *Scientific Reports*, 6, 2016.

[10] L. Bruni, Mireia Daz, Leslie Barrionuevo-Rosas, R. Herrero, F. Bray, X. Bosch, S. Sanjos, and X. Castellsague. Global estimates of human papillomavirus vaccination coverage by region and income level: a pooled analysis. *The Lancet. Global health*, 4 7:e453–63, 2016.

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 to 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:

```
# Import needed packages
import pandas as pd
import numpy as np

# Load the dataset
df = pd.read_csv('diabetes_binary_health_indicators_BRFSS2015.
    ↪ csv')

# Creating a text file
with open('data_exploration.txt', 'w') as f:
    # Data Size
    f.write("# Data Size\n")
    f.write("Number of rows: "+str(df.shape[0])+"\n")
    f.write("Number of columns: "+str(df.shape[1])+"\n\n")
```

```
    # Summary Statistics
    f.write("# Summary Statistics\n")
    f.write(str(df.describe())+"\n\n")

    # Categorical Variables
    categorical_vars = ['Diabetes_binary', 'HighBP', 'HighChol'
        ↪ , 'CholCheck', 'Smoker', 'Stroke',
                        'HeartDiseaseorAttack', 'PhysActivity',
                            ↪  'Fruits', 'Veggies', '
                            ↪ HvyAlcoholConsump',
                        'AnyHealthcare', 'NoDocbcCost', '
                            ↪ DiffWalk', 'Sex']
    f.write("# Categorical Variables\n")
    for var in categorical_vars:
        f.write("Most common value in "+var+": "+str(df[var].
            ↪ mode().values[0])+"\n")
    f.write("\n")

    # Missing Values
    f.write("# Missing Values\n")
    f.write("There are no missing values in the dataset as per
        ↪ the dataset description.")
    f.write("However, we still check:\n")
    df_missing = df.isnull().sum()
    # Check if there are any missing values at all
    if df_missing.sum() == 0:
        f.write("Confirmed, No missing values\n")
    else:
        f.write(str(df_missing[df_missing > 0]))
    f.write("\n\n")

    # Other summaries (if you find relevant given the dataset
        ↪ and your task)

# Close the file
f.close()
```

B.2 Code Description

The code performs data exploration on the given dataset of diabetes-related factors extracted from the CDC's Behavioral Risk Factor Surveillance System (BRFSS) for the year 2015.

The code begins by loading the dataset using the pandas library.

Next, a text file named "data_exploration.txt" is created to store the results of the data exploration.

The code then proceeds to perform various analysis steps on the dataset

and writes the results into the text file.

1. Data Size: The code calculates the number of rows and columns in the dataset using the shape attribute of the DataFrame and writes these values into the text file.

2. Summary Statistics: The code calculates the summary statistics of the numerical variables in the dataset using the describe() function of pandas DataFrame. These statistics include count, mean, standard deviation, minimum, 25th percentile, median, 75th percentile, and maximum. The summary statistics are written into the text file.

3. Categorical Variables: The code identifies the categorical variables in the dataset and calculates the most common value for each categorical variable using the mode() function of pandas DataFrame. The most common values for each categorical variable are written into the text file.

4. Missing Values: The code checks if there are any missing values in the dataset using the isnull() function of pandas DataFrame. If there are no missing values as confirmed by the dataset description, the code writes a confirmation message into the text file. If there are any missing values, the code calculates the number of missing values for each variable and writes them into the text file.

5. Other Summaries: The code provides flexibility to include other relevant summaries if required based on the dataset and the task at hand. However, in the provided code, there are no additional summaries included.

Finally, the text file is closed.

The "data_exploration.txt" file contains the results of the data exploration, including the data size, summary statistics, most common values for categorical variables, and information about missing values (if present). These results can be used for further analysis and understanding of the dataset.

B.3 Code Output

data_exploration.txt

```
# Data Size
Number of rows: 253680
Number of columns: 22

# Summary Statistics
       Diabetes_binary  HighBP  HighChol  CholCheck    BMI
           Smoker  Stroke  HeartDiseaseorAttack  PhysActivity
           Fruits  Veggies  HvyAlcoholConsump  AnyHealthcare
```

```
          NoDocbcCost   GenHlth   MentHlth   PhysHlth   DiffWalk
             Sex     Age   Education   Income
count            253680   253680     253680      253680 253680
    253680   253680                253680          253680   253680
      253680              253680          253680         253680
    253680     253680     253680     253680 253680 253680
    253680   253680
mean              0.1393    0.429     0.4241      0.9627   28.38
    0.4432 0.04057                0.09419         0.7565   0.6343
      0.8114               0.0562          0.9511         0.08418
    2.511      3.185      4.242     0.1682 0.4403   8.032
    5.05    6.054
std               0.3463   0.4949     0.4942      0.1896   6.609
    0.4968   0.1973                 0.2921          0.4292   0.4816
      0.3912               0.2303          0.2158         0.2777
    1.068      7.413      8.718     0.3741 0.4964   3.054
    0.9858    2.071
min                    0        0          0           0      12
          0        0                      0              0        0
            0                    0              0              0
           1           0          0        0       0       1
             1        1
25%                    0        0          0           1      24
          0        0                      0              1        0
            1                    0              1              0
           2           0          0        0       0       6
             4        5
50%                    0        0          0           1      27
          0        0                      0              1        1
            1                    0              1              0
           2           0          0        0       0       8
             5        7
75%                    0        1          1           1      31
          1        0                      0              1        1
             1                    0              1              0
           3           2          3        0       1      10
             6        8
max                    1        1          1           1      98
          1        1                      1              1        1
             1                    1              1              1
           5          30         30        1       1      13
             6        8
```

```
# Categorical Variables
Most common value in Diabetes_binary: 0
Most common value in HighBP: 0
Most common value in HighChol: 0
Most common value in CholCheck: 1
Most common value in Smoker: 0
```

```
Most common value in Stroke: 0
Most common value in HeartDiseaseorAttack: 0
Most common value in PhysActivity: 1
Most common value in Fruits: 1
Most common value in Veggies: 1
Most common value in HvyAlcoholConsump: 0
Most common value in AnyHealthcare: 1
Most common value in NoDocbcCost: 0
Most common value in DiffWalk: 0
Most common value in Sex: 0

# Missing Values
There are no missing values in the dataset as per the dataset
    description.However, we still check:
Confirmed, No missing values
```

C Data Analysis

C.1 Code

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

```
# IMPORT
import pandas as pd
import numpy as np
import statsmodels.formula.api as smf
import pickle

# LOAD DATA
df = pd.read_csv("diabetes_binary_health_indicators_BRFSS2015.
    ↪ csv")

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

# DESCRIPTIVE STATISTICS
# No descriptive statistics table is needed.

# PREPROCESSING
# No preprocessing is needed because all data has been
    ↪ preprocessed in the dataset preparation step.

# ANALYSIS

## Table 1: "Association between fruit and vegetable
    ↪ consumption and risk of diabetes"
# Run logistic regression model adjusting for confounding
    ↪ variables
```

```
model = smf.logit('Diabetes_binary ~ Fruits + Veggies + Age +
    ↪ Sex + Education + Income', data=df)
result = model.fit()
df1 = pd.DataFrame(result.summary2().tables[1])
df1.to_pickle('table_1.pkl')

## Table 2: "Income level as a moderator in the relationship
    ↪ between fruit and vegetable consumption and risk of
    ↪ diabetes"
# Run logistic regression model adjusting for confounders and
    ↪ including interaction between dietary habits and income
model = smf.logit('Diabetes_binary ~ Fruits + Veggies + Income
    ↪ + Fruits:Income + Veggies:Income + Age + Sex + Education'
    ↪ , data=df)
result = model.fit()
df2 = pd.DataFrame(result.summary2().tables[1])
df2.to_pickle('table_2.pkl')

# SAVE ADDITIONAL RESULTS
additional_results = {
    'Total number of observations': df.shape[0],
    'Percentage of diabetes patients': df.Diabetes_binary.mean
        ↪ () * 100,
}

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

C.2 Code Description

The code performs a data analysis on a dataset containing 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 importing necessary libraries and loading the dataset into a pandas DataFrame called "df".

Next, the code prepares the dataset for analysis. Since the dataset is already clean, no additional dataset preparations are needed.

The code then proceeds to analyze the data using two logistic regression models.

In the first analysis, a logistic regression model is run to investigate the association between fruit and vegetable consumption and the risk of diabetes. The model adjusts for confounding variables such as age, sex, education, and income. The results of the model are saved as a DataFrame

called "df1" and written to a pickle file named "table_1.pkl".

In the second analysis, a logistic regression model is run to examine the moderating effect of income on the relationship between fruit and vegetable consumption and the risk of diabetes. The model includes interactions between dietary habits (fruits and vegetables) and income, as well as adjusts for confounders such as age, sex, and education. The results of the model are saved as a DataFrame called "df2" and written to a pickle file named "table_2.pkl".

Additionally, the code saves some additional results in a pickle file named "additional_results.pkl". These results include the total number of observations in the dataset and the percentage of diabetes patients.

Overall, the code performs data analysis by running logistic regression models to explore the association between fruit and vegetable consumption and the risk of diabetes, as well as the moderating effect of income on this relationship. The code provides regression results for these analyses and saves them as pickle files. It also saves additional results related to the dataset.

C.3 Code Output

table_1.pkl

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-1.671	0.03639	-45.91	0	-1.742	-1.599
Fruits	-0.1846	0.01264	-14.61	2.56e-48	-0.2094	-0.1599
Veggies	-0.1265	0.01476	-8.569	1.05e-17	-0.1554	-0.09753
Age	0.1724	0.002177	79.21	0	0.1681	0.1767
Sex	0.3199	0.01214	26.36	4.15e-153	0.2961	0.3436
Education	-0.1311	0.006359	-20.62	1.75e-94	-0.1436	-0.1187
Income	-0.1617	0.003057	-52.88	0	-0.1677	-0.1557

table_2.pkl

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-1.805	0.04433	-40.71	0	-1.892	-1.718
Fruits	-0.04971	0.03279	-1.516	0.13	-0.114	0.01457
Veggies	-0.04591	0.03613	-1.271	0.204	-0.1167	0.0249
Income	-0.1347	0.005982	-22.52	2.4e-112	-0.1465	-0.123
Fruits:Income	-0.0251	0.005629	-4.46	8.19e-06	-0.03614	-0.01407

```
Veggies:Income  -0.01659  0.006541  -2.536      0.0112  -0.02941
    -0.003771
Age               0.1721  0.002176   79.1           0   0.1679
    0.1764
Sex                0.319   0.01214   26.27   4.42e-152   0.2952
    0.3428
Education        -0.1304  0.006356  -20.53    1.28e-93  -0.1429
    -0.118
```

additional_results.pkl

```
{
    'Total number of observations': 253680,
    'Percentage of diabetes patients': 13.93           ,
}
```

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 typing import Dict, Any, Optional, Tuple
from my_utils import to_latex_with_note, is_str_in_df,
    ↪ split_mapping, AbbrToNameDef

# Type Aliases
AbbrToNameDef = Dict[Any, Tuple[Optional[str], Optional[str]]]

# PREPARATION FOR ALL TABLES
shared_mapping: AbbrToNameDef = {
    'Fruits': ('Fruit', None),
    'Veggies': ('Veggies', None),
    'Income': ('Income', 'Income level on a scale of 1 to 8 (1=
        ↪  <=10K, 2= <=15K, 3= <=20K, 4= <=25K, 5= <=35K, 6=
        ↪ <=50K, 7= <=75K, 8= >75K)'),
    'Fruits:Income': ('F-Income', 'Interaction between Fruit
        ↪ consumption and Income'),
    'Veggies:Income': ('V-Income', 'Interaction between
        ↪ Vegetable consumption and Income'),
    'z': ('z', 'z-value or z-score is the coefficient divided
        ↪ by its standard error.'),
}

# TABLE 1:
df1 = pd.read_pickle('table_1.pkl')
```

```
# extracting relevant labels from shared_mapping
mapping1 = {k: v for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df1, k)}
abbrs_to_names1, legend1 = split_mapping(mapping1)
df1.rename(columns=abbrs_to_names1, index=abbrs_to_names1,
    ↪ inplace=True)

# Rename longer column names
df1.rename(columns={'Std.Err.': 'Std.Err', 'P>|z|': 'P>|z|',
                    '[0.025': '[0.025', '0.975]': '0.975'
                    }, inplace=True)

to_latex_with_note(
    df1,
    'table_1.tex',
    caption="Association between diet and the risk of diabetes,
        ↪  adjusted for different confounders.",
    label='table:table_1',
    legend=legend1)

# TABLE 2:
df2 = pd.read_pickle('table_2.pkl')

mapping2 = {k: v for k, v in shared_mapping.items() if
    ↪ is_str_in_df(df2, k)}
abbrs_to_names2, legend2 = split_mapping(mapping2)
df2.rename(columns=abbrs_to_names2, index=abbrs_to_names2,
    ↪ inplace=True)

# Rename longer column names
df2.rename(columns={'Std.Err.': 'Std.Err', 'P>|z|': 'P>|z|',
                    '[0.025': '[0.025', '0.975]': '0.975'
                    }, inplace=True)

to_latex_with_note(
    df2,
    'table_2.tex',
    caption="Moderating effect of income on the relationship
        ↪ between diet and the risk of diabetes.",
    label='table:table_2',
    legend=legend2)
```

D.2 Provided Code

The code above is using the following provided functions:

```python
def to_latex_with_note(df, filename: str, caption: str, label:
    ↪ str, note: str = None, legend: Dict[str, str] = None, **
    ↪ kwargs):
    """
    Converts a DataFrame to a LaTeX table with optional note
        ↪ and legend added below the table.

    Parameters:
    - df, filename, caption, label: as in `df.to_latex`.
    - note (optional): Additional note below the table.
    - legend (optional): Dictionary mapping abbreviations to
        ↪ full names.
    - **kwargs: Additional arguments for `df.to_latex`.
    """

def is_str_in_df(df: pd.DataFrame, s: str):
    return any(s in level for level in getattr(df.index, '
        ↪ levels', [df.index]) + getattr(df.columns, 'levels',
        ↪ [df.columns]))

AbbrToNameDef = Dict[Any, Tuple[Optional[str], Optional[str]]]

def split_mapping(abbrs_to_names_and_definitions: AbbrToNameDef
    ↪ ):
    abbrs_to_names = {abbr: name for abbr, (name, definition)
        ↪ in abbrs_to_names_and_definitions.items() if name is
        ↪ not None}
    names_to_definitions = {name or abbr: definition for abbr,
        ↪ (name, definition) in abbrs_to_names_and_definitions.
        ↪ items() if definition is not None}
    return abbrs_to_names, names_to_definitions
```

D.3 Code Output

table_1.tex

```
% This latex table was generated from: `table_1.pkl`
\begin{table}[h]
\caption{Association between diet and the risk of diabetes,
    adjusted for different confounders.}
\label{table:table_1}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
 & Coef. & Std.Err & z & P$>$\textbar{}z\textbar{} & [0.025 &
    0.975 \\
```

```
\midrule
\textbf{Intercept} & -1.67 & 0.0364 & -45.9 & $<$1e-06 & -1.74
    & -1.6 \\
\textbf{Fruit} & -0.185 & 0.0126 & -14.6 & $<$1e-06 & -0.209 &
    -0.16 \\
\textbf{Veggies} & -0.126 & 0.0148 & -8.57 & $<$1e-06 & -0.155
    & -0.0975 \\
\textbf{Age} & 0.172 & 0.00218 & 79.2 & $<$1e-06 & 0.168 &
    0.177 \\
\textbf{Sex} & 0.32 & 0.0121 & 26.4 & $<$1e-06 & 0.296 & 0.344
    \\
\textbf{Education} & -0.131 & 0.00636 & -20.6 & $<$1e-06 &
    -0.144 & -0.119 \\
\textbf{Income} & -0.162 & 0.00306 & -52.9 & $<$1e-06 & -0.168
    & -0.156 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item \textbf{Income}: Income level on a scale of 1 to 8 (1= $<
    $=10K, 2= $<$=15K, 3= $<$=20K, 4= $<$=25K, 5= $<$=35K, 6= $
    <$=50K, 7= $<$=75K, 8= $>$75K)
\item \textbf{z}: z-value or z-score is the coefficient divided
     by its standard error.
\end{tablenotes}
\end{threeparttable}
\end{table}
```

table_2.tex

```
% This latex table was generated from: `table_2.pkl`
\begin{table}[h]
\caption{Moderating effect of income on the relationship
    between diet and the risk of diabetes.}
\label{table:table_2}
\begin{threeparttable}
\renewcommand{\TPTminimum}{\linewidth}
\makebox[\linewidth]{%
\begin{tabular}{lllllll}
\toprule
 & Coef. & Std.Err & z & P$>$\textbar{}z\textbar{} & [0.025 &
     0.975 \\
\midrule
\textbf{Intercept} & -1.8 & 0.0443 & -40.7 & $<$1e-06 & -1.89 &
     -1.72 \\
\textbf{Fruit} & -0.0497 & 0.0328 & -1.52 & 0.13 & -0.114 &
    0.0146 \\
\textbf{Veggies} & -0.0459 & 0.0361 & -1.27 & 0.204 & -0.117 &
    0.0249 \\
```

```
\textbf{Income} & -0.135 & 0.00598 & -22.5 & $<$1e-06 & -0.146
    & -0.123 \\
\textbf{F-Income} & -0.0251 & 0.00563 & -4.46 & 8.19e-06 &
    -0.0361 & -0.0141 \\
\textbf{V-Income} & -0.0166 & 0.00654 & -2.54 & 0.0112 &
    -0.0294 & -0.00377 \\
\textbf{Age} & 0.172 & 0.00218 & 79.1 & $<$1e-06 & 0.168 &
    0.176 \\
\textbf{Sex} & 0.319 & 0.0121 & 26.3 & $<$1e-06 & 0.295 & 0.343
     \\
\textbf{Education} & -0.13 & 0.00636 & -20.5 & $<$1e-06 &
    -0.143 & -0.118 \\
\bottomrule
\end{tabular}}
\begin{tablenotes}
\footnotesize
\item \textbf{Income}: Income level on a scale of 1 to 8 (1= $<
    $=10K, 2= $<$=15K, 3= $<$=20K, 4= $<$=25K, 5= $<$=35K, 6= $
    <$=50K, 7= $<$=75K, 8= $>$75K)
\item \textbf{F-Income}: Interaction between Fruit consumption
    and Income
\item \textbf{V-Income}: Interaction between Vegetable
    consumption and Income
\item \textbf{z}: z-value or z-score is the coefficient divided
     by its standard error.
\end{tablenotes}
\end{threeparttable}
\end{table}
```