Associations between fasting blood glucose and diet, physical activity and sleep: A regression analysis

Introduction

As the prevalence of diabetes and insulin resistance continues to increase in Americans, understanding the impact of lifestyle factors on metabolic health becomes important in modern healthcare. This study aims to explore the relationship between physical activity, dietary habits, sleep duration, and fasting blood glucose levels — a marker of metabolic stability. Employing a multivariate linear regression analysis, this investigation seeks to elucidate potential associations between these lifestyle factors and fasting blood glucose levels. By uncovering these connections, the study aims to contribute valuable insights into preventive healthcare strategies and lifestyle interventions for improvement of metabolic health.

Methods

The data for this study comes from the National Health and Nutrition Examination Survey (NHANES), a program designed to assess health and nutritional status of people of all ages in the United States by combining both interview questions and physical examination results. NHANES examines a nationally representative sample of about 5,000 people from counties all across the country per year. The data I used for my analysis comes from 2017 to pre-pandemic 2020. Due to the COVID-19 pandemic, field operations were suspended and data for the year 2018-2019 were not able to be completed, so the data was combined with 2017-2018 data.

Data Information

The NHANES data is separated in many individual files. For this investigation, I downloaded the following files from the CDC website: demographics, plasma fasting glucose, alcohol use, body measures, physical activity, diet behavior, and sleep. The variables of interest were extracted from each of these files, and then joined to make the final data set. Gender and age were pulled from the demographics file. Glucose, measured in mg/dL, was pulled from the glucose file. BMI was extracted from the body measures file. Physical activity was determined by calculating a score from multiple variables in the physical activity file. The equation, published in the International Physical Activity Questionnaire, assigns different weights for activities of different intensities. For this analysis, a multiplier of 4 was used for average minutes of moderate activity and a multiplier of 8 for average minutes of vigorous activity. The average minutes of physical activity are a daily estimate, and are multiplied by number of days per week of activity to get a combined total physical activity MET-min/week. The equation is:

$$\text{Moderate MET-minutes/week} = 4 * \text{moderate-intensity activity minutes } * \text{moderate days}$$

$$\text{Vigorous MET-minutes/week} = 8 * \text{vigorous-intensity activity minutes } * \text{vigorous-intensity days}$$

$$\text{total physical activity MET-mins/week } = \text{Moderate MET-min/week} + \text{Vigorous MET-min/week}$$

It it worth noting that NHANES asks participants to distinguish between physical activity that is for recreation and physical activity that is for work or employment. For the purposes of this analysis, these MET-mins/week were combined.

Diet score is a self-assigned score ranging from one to five based on how a participant views their eating habits overall, with one being the healthiest and five being the unhealthiest score.

Code or Value	Value Description	Count
1	Excellent	827
2	Very good	2048
3	Good	3966
4	Fair	2637
5	Poor	712
7	Refused	2
9	Don't know	3
.	Missing	5365

Alcohol frequency was a question that asked about how often participants consumed alcohol in the past 12 months. The table below details the values.

Code or Value	Value Description	Count
0	Never in the last year	1638
1	Every day	240
2	Nearly every day	273
3	3 to 4 times a week	514
4	2 times a week	586
5	Once a week	562
6	2 to 3 times a month	1042
7	Once a month	570
8	7 to 11 times in the last year	457
9	3 to 6 times in the last year	795
10	1 to 2 times in the last year	822
77	Refused	2
99	Don't know	2
.	Missing	1462

Finally, sleep hours represents the average number of hours of sleep per night that a participant gets, on weekday or weekend.

Data Preprocessing and Univariate Analyses

Participants with missing data in any of the columns were removed from the data set. Since no variables had a high amount of missing data, not much information was lost by excluding missing values. An additional filtering step included filtering out participants who did not meet the criteria of having fasted for at least 8 hours prior to examination, as this may have a large impact on the fasting blood glucose levels. After filtering, the sample size remaining was 3,315. Approximately half of the sample were male, and the average age was 50. Distributions of the covariates are detailed below in Table 1.

Table 1

variables	*n*	*mean*	*sd*	*median*	*min*	*max*
Glucose (mg/dL)	4158	112.86	37.48	103.00	47.00	451.00
Gender (1 = Male, 2 = Female)	4423	1.52	0.50	2.00	1.00	2.00
Age	4423	49.62	18.29	51.00	18.00	80.00
BMI (kg/m^2)	4333	29.82	7.73	28.50	14.90	92.30
Alcohol Frequency	3748	4.83	3.51	5.00	0.00	10.00
Diet Health Score	4423	3.06	1.03	3.00	1.00	5.00
Sleep Hours (Weekdays or Weekends)	4387	7.60	1.69	7.50	2.00	14.00
MET (activity score)	4423	4115.26	6976.88	1080.00	0.00	51840.00

Univariate analyses were performed between each of the covariates of interest and fasting blood glucose levels. First, the functional form of each of the covariates of interest were determined by both the fractional polynomials method and by converting the variable to categorical. Diet health and sleep hours were kept as linear variables, and MET score was log-transformed for better model fit, as suggested by fractional polynomials. The results from univariates analyses are depicted below in Table 2.

Table 2

	Fasting Glucose (mg/dL)			*Fasting Glucose (mg/dL)*			*Fasting Glucose (mg/dL)*		
Predictors	Estimates	CI	p	Estimates	CI	p	Estimates	CI	p
(Intercept)	108.78	106.99 – 110.57	**<0.001**	103.91	99.82 – 108.00	**<0.001**	115.43	109.32 – 121.54	**<0.001**
log(MET)	-1.30	-1.70 – -0.90	**<0.001**						
Diet Health				2.90	1.64 – 4.16	**<0.001**			
Sleep Hours							-0.34	-1.14 – 0.45	0.400
Observations	3315			3315			3315		
R^2 / R^2 adjusted	0.012 / 0.012			0.006 / 0.006			0.000 / -0.000		

Next, the functional form for the other variables were determined. Age was converted into quintiles with the ranges being [18-31], (31-44], (44, 56], (56, 67], (67, 80] as the extra sums of squares test showed that this offered better model fit $(F = 58.2, p < 0.001)$. Gender was converted into a categorical variable, as it is dichotomous. Alcohol frequency was also treated as categorical variable $(F = 3.46, p < 0.001)$. The fractional polynomials method suggested a $\frac{1}{x^2} + \sqrt{x}$ transformation for alcohol frequency, but they were just converted to categorical for more interpretability. Finally, BMI was transformed as $\frac{1}{\sqrt{x}}$, as suggested by fractional polynomials.

Results

Frequency of alcohol consumption was suspected to be an effect modifier for diet health and sleep hours. However, the extra sums of squares test revealed that neither of these interaction terms were significant at an alpha level of 0.15, $(F = 1.3, p = 0.23)$ for diet health and $(F = 1.1, p = 0.37)$ for sleep hours. Additionally, it was suspected that BMI was an effect modifier for physical activity, but this interaction was not significant either $(F = 1.2, p = 0.27)$.

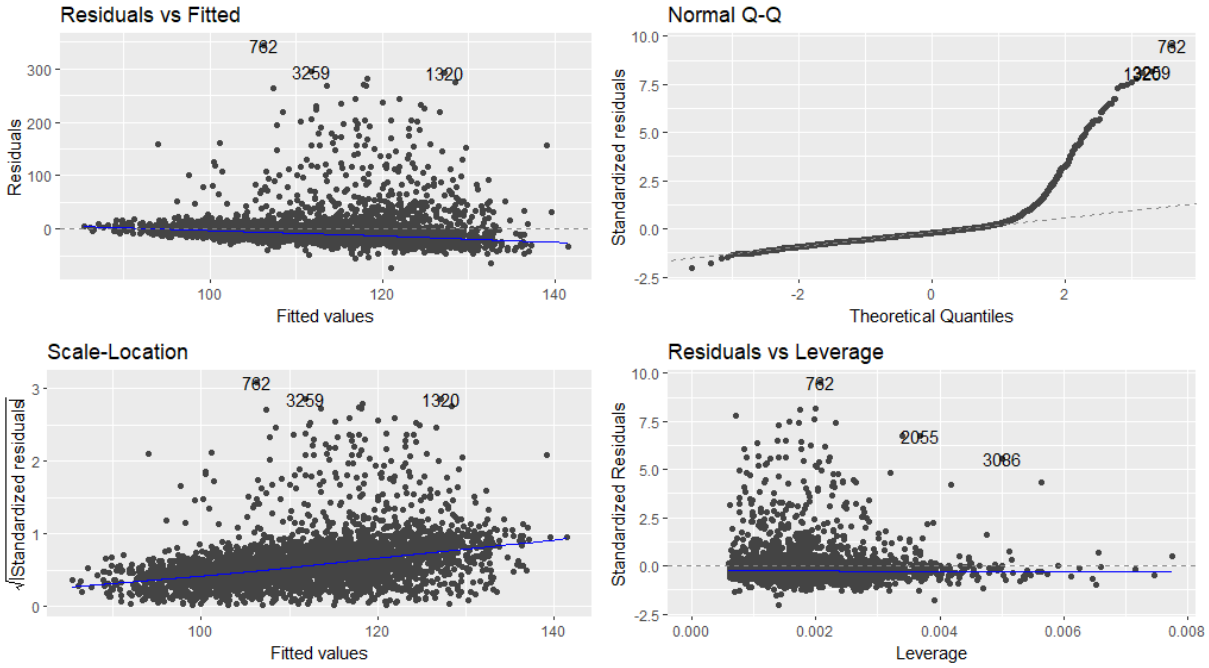
Age and gender were suspected to be confounding variables for all three covariates of interest. Age and gender could both have an effect on diet health, physical activity, and hours of sleep as well as on fasting blood glucose levels. When

added into the model, age and gender both change the parameter estimates for all of the covariates by over 15 percent, so they were both left in the model. The parameter estimates are shown below in Supplementary Table 1.

Supplementary Table 1

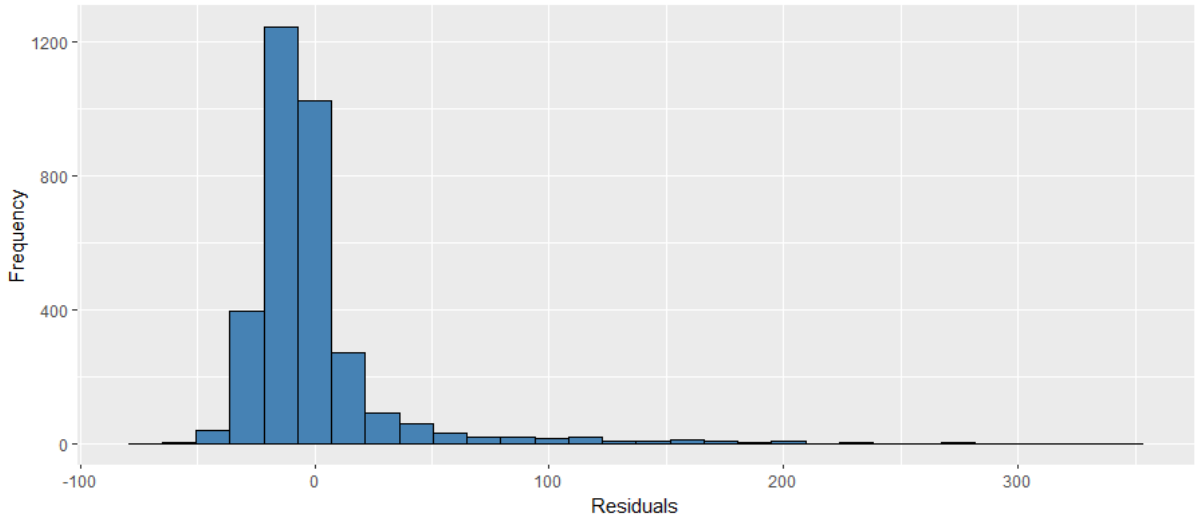
	Unadjusted			Age Adjusted			Gender Adjusted			Age and Gender Adjusted		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
Diet Health Score	2.70	1.45 – 3.96	**<0.001**	4.18	2.94 – 5.43	**<0.001**	2.91	1.66 – 4.16	**<0.001**	4.31	3.07 – 5.55	**<0.001**
Sleep Hours	-0.43	-1.23 – 0.36	0.282	-0.38	-1.16 – 0.39	0.332	-0.33	-1.11 – 0.46	0.419	-0.29	-1.06 – 0.48	0.456
log(MET)	-1.28	-1.68 – -0.88	**<0.001**	-0.48	-0.89 – -0.07	**0.022**	-1.44	-1.85 – -1.04	**<0.001**	-0.64	-1.05 – -0.22	**0.003**
Observations	3315			3315			3315			3315		
R^2 / R^2 adjusted	0.018 / 0.017			0.065 / 0.064			0.027 / 0.026			0.071 / 0.070		

A preliminary final model was fit using the variables diet health, sleep hours, logarithm of physical activity score, age, and gender. Upon assessing the assumptions of linear regression, it appears that normality and homoscedasticity were violated, as can be seen in the Normal Q-Q plot and Scale-Location Plot in Figure 1 below.

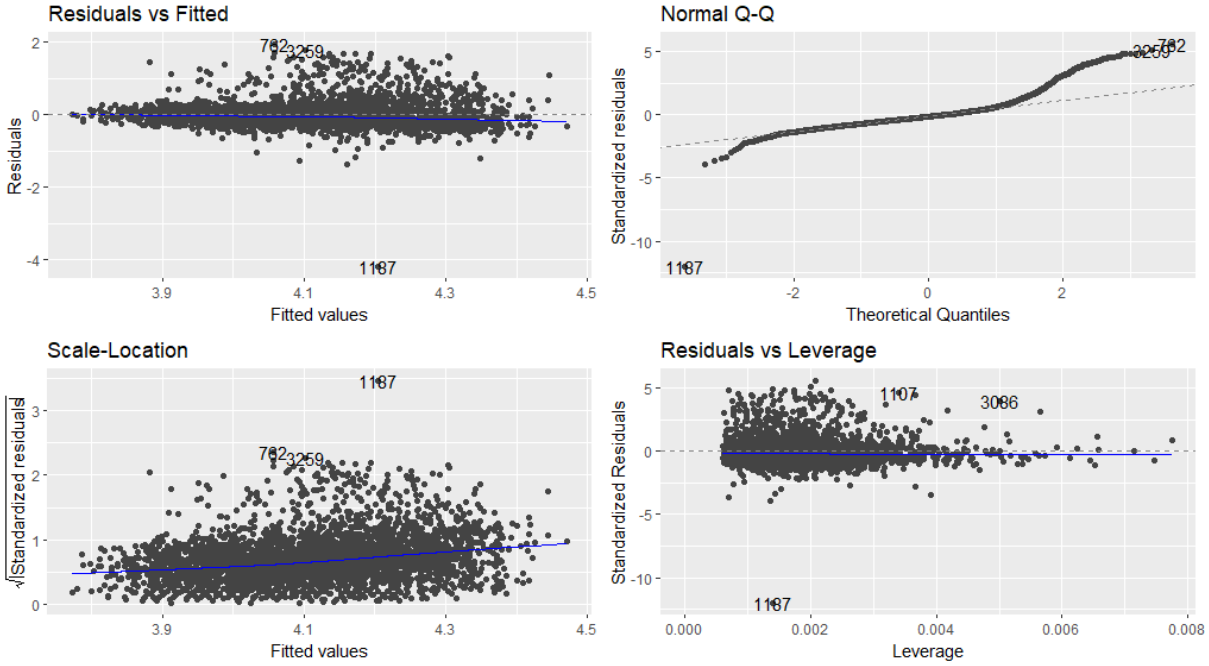


Indeed a histogram of the residuals revealed that the residuals were strongly positively skewed.

Figure 2



In order to better satisfy the assumptions of linear regression, glucose measurements was log-transformed. The formula for the transformation is: $y \log(y + c)$, where $c = 1 - \min(y)$. After transforming glucose, the assumptions of linear regression appear to hold. Figure 3 below shows the diagnostics plots post-log transformation.



The Normal Q-Q plot still shows some slight curvature at the higher quantiles, but since the sample size is in the thousands, the Central Limit Theorem makes inference robust to non-normality of residuals. Homoscedasticity has also improved as a result of the log transformation.

Figure 4

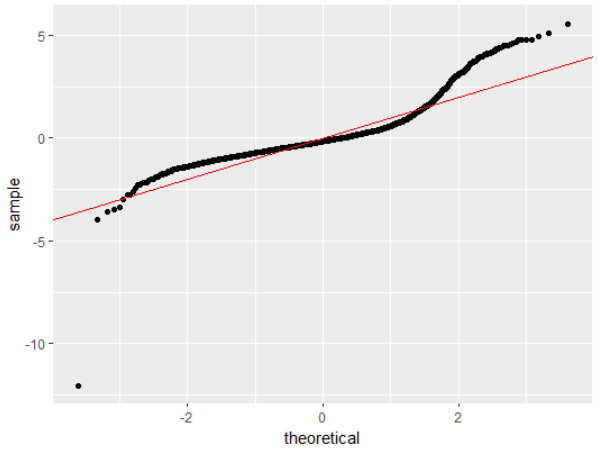
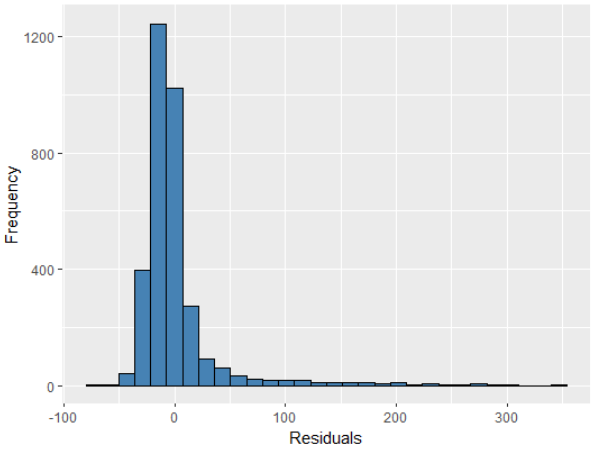
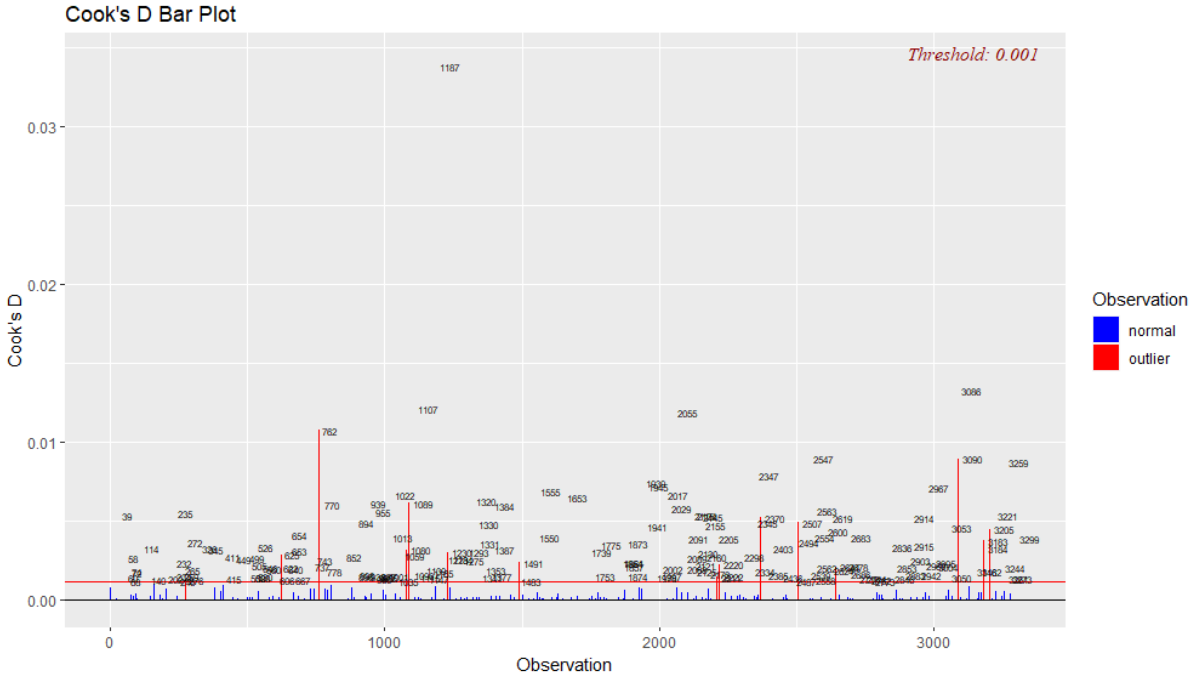


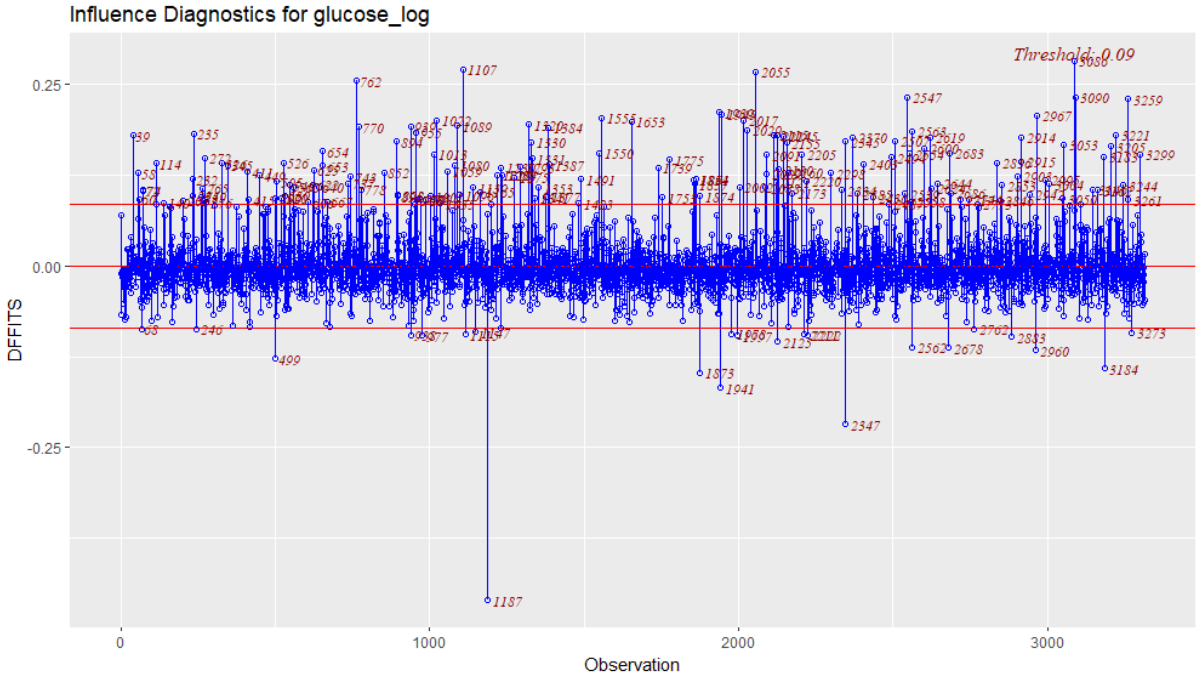
Figure 5



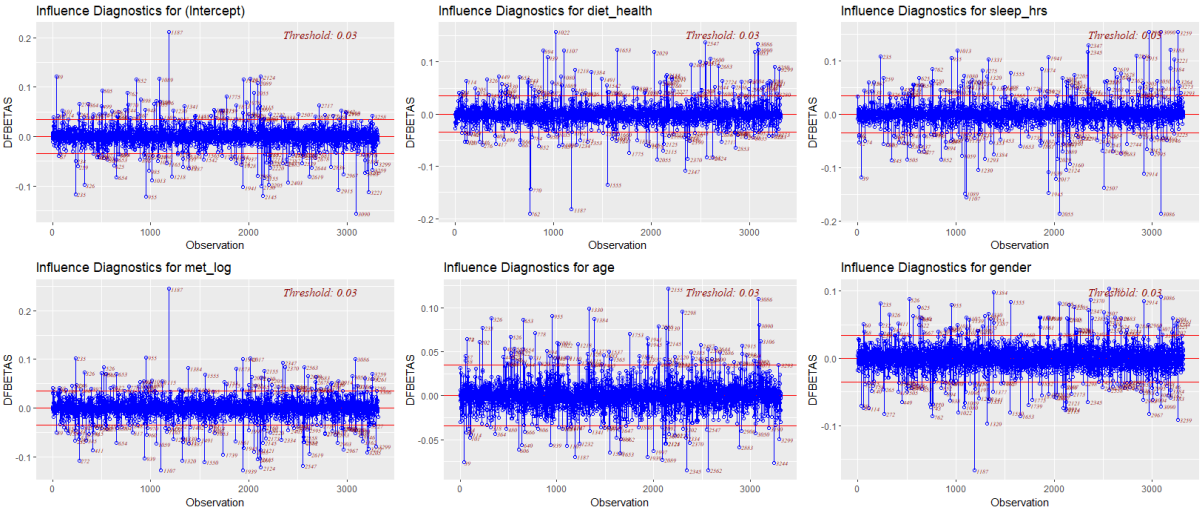
Finally, influential points affecting model parameter estimates were checked using Cook's Distance, DFFITS, and DFBETAS. Cook's Distance reveled that observation 1187 (participant id 115089) had a high Cook's Distance of about 0.034 relative to other observations, but was still below 0.5.



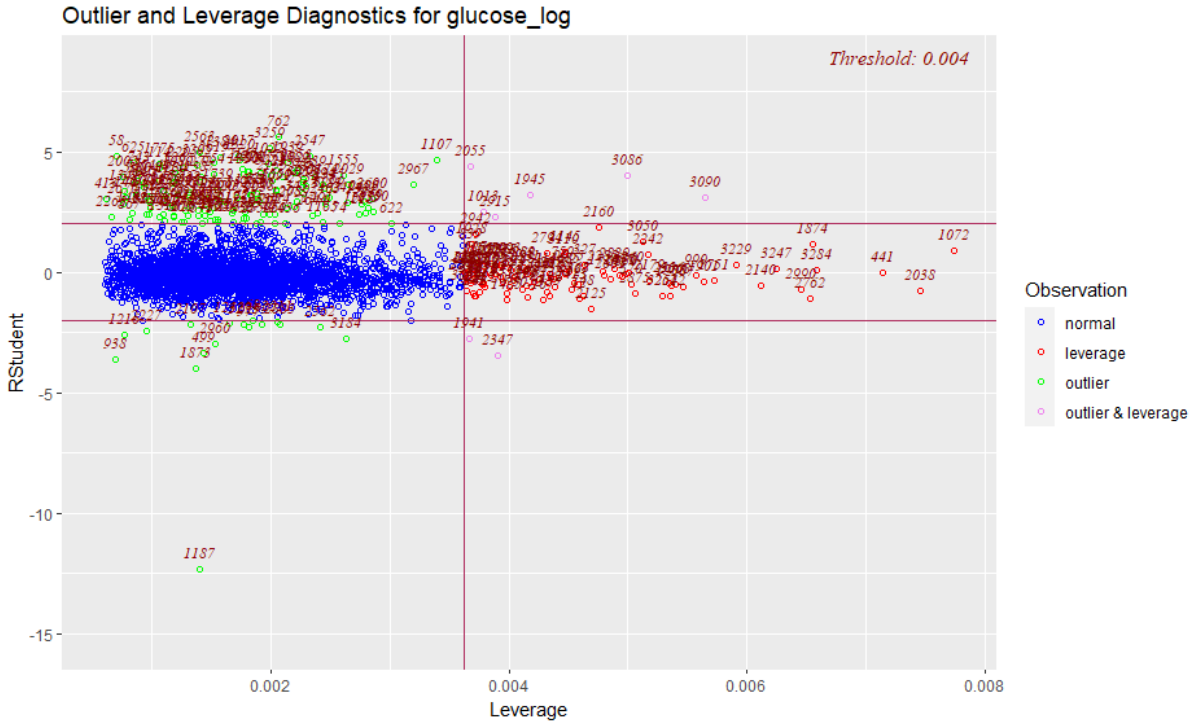
The DFFITS plot revealed that observation 1187 had a large negative DFFITS value. Additionally, observations 2347, 1107, 2055, and 3086 show large influences on model parameter estimates.



As for the DFBETAs, similar points were shown to be poorly fit including 1187, 1107, 2155, and 3086.



Lastly, a Studentized residuals vs. leverage plot was constructed, shown in Figure 9 below.



While 1187 was shown to have high leverage, it was not a strong outlier. Observations that are shown to be influential are 1013, 2915, 2055, 1945, 3086, 3090, 1941, and 2347. These points were both high-leverage and outliers. A sensitivity analysis was performed to see if removal of these observations resulted in different parameter estimates.

Supplementary Table 2

	Original Model			Influential Points Removed		
Predictors	*Estimates*	*CI*	*p*	*Estimates*	*CI*	*p*
(Intercept)	3.79	3.70 – 3.87	**<0.001**	3.80	3.72 – 3.89	**<0.001**
Diet Health Score	0.05	0.03 – 0.06	**<0.001**	0.05	0.04 – 0.06	**<0.001**
Sleep Hours	-0.00	-0.01 – 0.00	0.459	-0.00	-0.01 – 0.00	0.399
log(MET)	-0.01	-0.01 – -0.00	**0.001**	-0.01	-0.01 – -0.00	**0.001**
Age	0.01	0.01 – 0.01	**<0.001**	0.01	0.01 – 0.01	**<0.001**
Gender	-0.09	-0.11 – -0.06	**<0.001**	-0.09	-0.11 – -0.07	**<0.001**
Observations	3315			3307		
R^2 / R^2 adjusted	0.121 / 0.119			0.120 / 0.119		

As can be seen in Supplementary Table 2, the parameter estimates did not change much when the influential points were removed. Additionally, since there were no obvious data errors in the influential observations, there was no compelling reason to remove them from the model. In light of this, all observations were kept in the final model.

Table 3, which gives the parameter estimates for the final model is shown below:

Table 3

	log(Fasting Glucose (mg/dL))		
Predictors	Estimates	CI	p
(Intercept)	3.79	3.70 – 3.87	**<0.001**
Diet Health Score	0.05	0.03 – 0.06	**<0.001**
Sleep Hours - Weekdays or Weekends	-0.00	-0.01 – 0.00	0.459
log(MET)	-0.01	-0.01 – -0.00	**0.001**
Age	0.01	0.01 – 0.01	**<0.001**
Gender	-0.09	-0.11 – -0.06	**<0.001**
Observations		3315	
R^2 / R^2 adjusted		0.121 / 0.119	

The equation for the best fit model is:

$$\log(\hat{Y}) = 3.79 + 0.05X_{\text{diet health}} - 0.003X_{\text{sleep hours}} - 0.001\log(X_{\text{MET-mins/week}}) + 0.01X_{\text{age}} - 0.09X_{\text{gender}}$$

A one point increase in diet health score is associated with a 4.79 percent increase in fasting blood glucose level holding all other variables constant, and this association is statistically significant $(t = 7.6, p < 0.001)$. A one hour increase in sleep is associated with a 2.81 percent decrease in fasting blood glucose level holding all other variables constant, and this association is not statistically significant $(t = -0.7, p = 0.46)$. A one percent increase in MET physical activity score is associated with a 0.67 percent decrease in fasting blood glucose level holding all other variables constant, and this association is statistically significant $(t = -3.3, p = 0.001)$. Overall, these variables explain approximately 12 percent of the variation in fasting blood glucose levels.

This construction of this model involved meticulously processing the NHANES data set, ensuring quality and relevance. Methodical variable selection, validation against regression assumptions, and handling influential points all underpin the model's accuracy.

Addenda

The link to the code written for this analysis is available on GitHub:
https://github.com/flemm0/PM592/tree/main/Final_Project

The majority of the data preprocessing was done in the R script `data_preprocessing.R` and the analysis is in the notebook `final_project.Rmd`. The link to the XPT data files used in the analysis are located under the `data` folder in the same repository: https://github.com/flemm0/PM592/tree/main/Final_Project/data

Also located in the `data` folder are the shell script used to download the files is `download_data.sh` and the merged data file used for analysis is `nhanes.parquet`. The merged data file is stored in Apache Parquet file format and requires the `arrow` library to open.

The link to where the data was downloaded from the CDC website is:
https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020