Inference and Prediction on Crude Diabetes Prevalence in U.S. States Based on Vegetable Consumption

Data 102: Data, Inference, and Decisions
Christina Đặng, Tetsuro Escudero, Conan Minihan, Yehchan Yoo

Data Overview

Diabetes is a pressing issue in the United States as a precursor to many illnesses. In fact, 34.2 million Americans have diabetes (Halvorsen et al., 2021). In past research, frequent consumption of vegetables was inversely associated with the risk of having non-insulin-dependent diabetes mellitus (Williams et al., 1999). Also, diets high in fruit and vegetables were found to help with prevention of type 2 diabetes (Ford & Mokdad, 2001). So, we decided to see how vegetable consumption could impact diabetes prevalence in American states using datasets from Centers for Disease Control and Prevention (CDC) and the United States Census.

CDC Datasets

CDC's Chronic Disease dataset contains indicators from many public health data sources on diseases, the source of our diabetes and vegetable consumption data (Centers for Disease Control and Prevention, 2023a). Behavioral Risk Factor Surveillance System (BRFSS) is a parent source for this dataset and is the source of our diabetes and vegetable consumption data. BRFSS data came from a sample, collected via telephone surveys with respondents being aware of their data being collected. BRFSS addresses the population of Americans 18 years or older, so the scope of the data will not impact the generalizability of our analysis. No differential privacy methods were used. In addition, the data has a state-year granularity, and individual responses were aggregated before reported (Centers for Disease Control and Prevention, 2019).

Self-selection bias and measurement error are concerns. Some respondents refused to participate or responded only to certain questions, self-selection bias (Centers for Disease Control and Prevention, 2019). Respondents also may have overestimated their vegetable consumption because of perceived societal pressures to make healthier decisions.

We also would have liked to have segregated information on crude prevalence of type 1 and type 2 diabetes instead of an aggregated crude prevalence. We also did not have data on diabetes for 2012, 2014, 2016, and 2018, as the BRFSS collected vegetable consumption information biennially. There was missing data on New Jersey in 2019, as New Jersey did not collect enough data to be included in the 2019 BRFSS dataset (Centers for Disease Control and Prevention, 2019). We imputed the missing data with mean crude prevalence and vegetable consumption from New Jersey in 2011, 2013, 2015, and 2017.

We removed territories in the dataset, as our income data from CPS lacked information about the territories (as will be discussed later). We also selected crude prevalence versus age-adjusted prevalence in order to keep age as a separate feature. We also did not include data beyond 2019 for concerns that COVID-19 would heavily impact our analysis.

U.S. Census Datasets

We used datasets from the U.S. Census to gather information about our covariates that could be related to vegetable consumption and diabetes prevalence. Age, education, and race data came from the American Community Survey (ACS); income data from the Current Population Survey (CPS).

ACS

ACS data comes from mail or Internet responses from a sample of approximately 3.5 million addresses and provides a variety of demographic estimates (The United States Census Bureau, 2022b). We downloaded age and race data from KFF (previously known as Kaiser Family Foundation) and not directly from ACS, since KFF broke down the data nicely by year and state – reducing our cleaning work (The United States Census Bureau, 2022a). We did gather our education data from the Census directly, as we found the Census's education data satisfactory (The United States Census Bureau, n.d.-c).

The ACS sample was representative of the greater American population, since household samples were chosen randomly from across the country and stratified by region. The Census attempts to minimize nonresponse rate via "Computer Assisted Personal Interviewing" for those who were initially picked for the random sample but do not respond. Selection bias exists in ACS, since the data was collected only from geographic areas with 65,000 people or more. Respondents were also aware of data collection and use (The United States Census Bureau, 2022b). While the Census Bureau is planning to adopt differential privacy in ACS in 2025 at the earliest, the ACS data we used did not adopt differential privacy techniques (Daily, 2022).

We also should note that ACS excluded certain groups such as soup kitchens, mobile food vans, natural disaster shelters, and commercial maritime vessels due to limitations in data collection (The United States Census Bureau, 2022b) There were also no relevant columns with missing data in the ACS dataset..

Since ACS comprises economic, demographic, social, and housing data, we would have liked ACS to contain information on exercise. We might have used this data to see the relationship between exercise and diabetes. Other columns we wish we had in the ACS dataset are cost of living and access to a dietitian or nutritionist, as we could have used these columns to see how living cost and access to a dietitian or nutritionist may motivate or discourage vegetable consumption for American adults.

Also, as with the CDC data, we removed U.S. territories data in the dataset. In the Age dataset, we binarized age to focus on the age group greater or equal to 45 years, since type 2 diabetes becomes much more prevalent in adults more than 45 years old (Centers for Disease Control and Prevention, 2023b). In the Education dataset, we also preprocessed the data to

standardize percentages to proportions for the years 2011-2014, since they did not match the proportions given to us for years 2015-2019. We also made corrections in the educational dataset for inconsistencies (e.g. certain year ranges were reported as percentages, others as raw counts; age bins needed to be aggregated to fit our research question pertaining to all adults). We binarized education to focus on those who had achieved a bachelor's degree or higher, as we found that bachelor's degree is a common and widely recognized indicator of educational attainment.. This decision could have affected our model and inferences, as it is possible for someone with a high school diploma to be equally informed about the relationship between nutrition and health and make similar choices as someone with an advanced degree. In the race dataset, we were given the White proportion of our US population by state, so we used the dataset to compare the white demographic to the non-white demographic.

CPS

The Census Bureau's Current Population Survey (CPS) is a regularly held survey conducted jointly with the Bureau of Labor Statistics for gathering employment-related data (The United States Census Bureau, n.d.-b). CPS gathered data on approximately 60,000 households, only interviewing people older than 14, not in the armed forces or certain institutions like prisons, hospitals, and nursing homes (The United States Census Bureau, n.d.-e). CPS gathers data through in-person and telephone interviews (The United States Census Bureau, n.d.-a). Due to the use of a probability sample over the entire American population, we find that the sample is generalizable, but we also find that the omission of individuals stated above may lead to overestimation of median income by state. Also, CPS data we used were collected before 2020, not involving the use of differential privacy (Daily, 2022).

The CPS income data was downloaded directly from the Census website and only included data from the 50 American states, which is why we did not include data from American territories for data analysis (The United States Census Bureau, n.d.-d). The data was in Excel format and was cleaned and formatted as a Pandas dataframe. Also, the data was originally in pivot table format, showing median household income for a state-year pair. So, we made the table tidy with each row representing a state-year pair. In addition, we used median household income in 2021 U.S. dollars instead of nominal dollars to ensure that our findings will not be impacted by inflation.

Income distribution in our dataset is same as the one expected in the U.S. population. According to the Census Bureau, the median household income in the U.S. increased 0.8% from 2017 to 2018 and 6.8% from 2018 to 2019 (Guzman, 2019; Rothbaum, 2020). Since this is consistent with the increase in distribution of income from 2017 to 2019 in our dataset, our results are generalizable based on income.

Research Questions

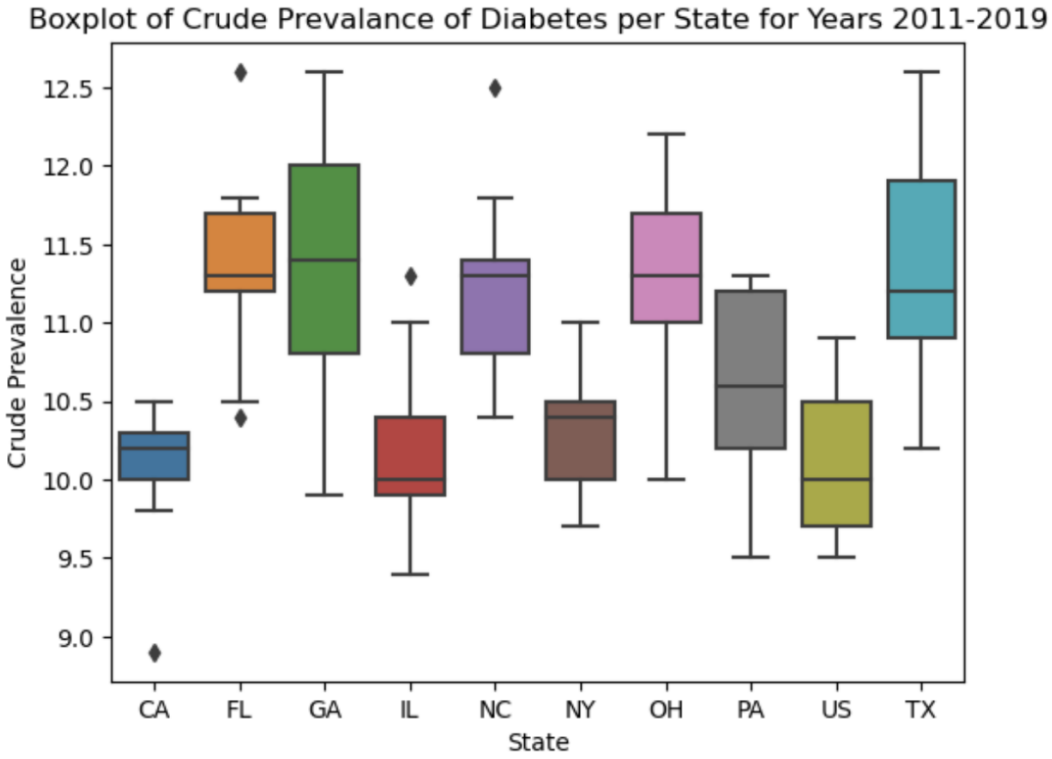
1. From 2011 to 2019, did consumption of vegetables have a causal impact on diabetes prevalence for American adults by state?
2. How effectively can vegetable consumption and demographic features predict diabetes prevalence by state per year for American adults?

Question 1 can be used to decide whether local and federal governments should encourage vegetable consumption to lower diabetes prevalence. Question 2 can be used to help public health systems prepare for potential increases or decreases in the number of diabetes patients.

We plan to use causal inference for question 1, since we seek to analyze the impact vegetable consumption by itself might have on diabetes prevalence while accounting for any other impactful variables. This is so that we could potentially look into what the U.S. society can do in regards to vegetable consumption to reduce the number of Americans who develop diabetes. A major limitation is that our data all come from observational studies, so the SUTVA conditions may not be satisfied, especially if there is geographical spillover between states.

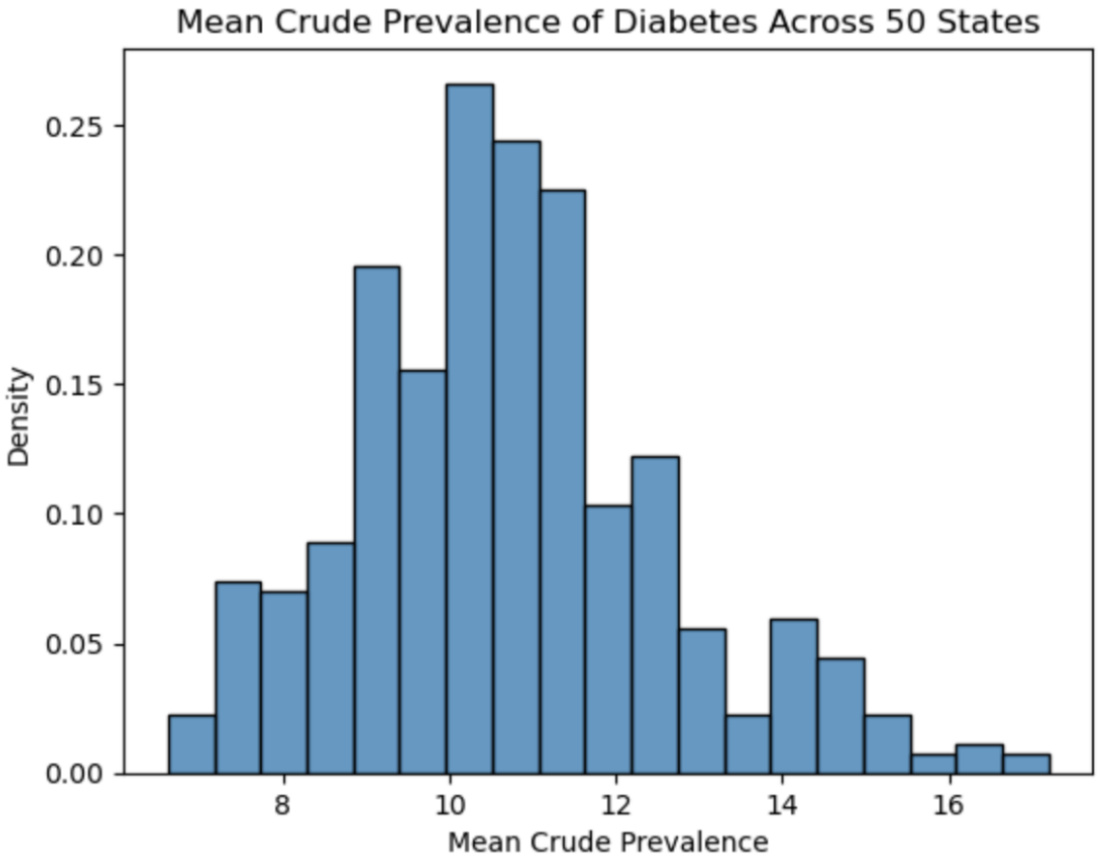
For Question 2, we plan to use general linearized models (GLMs) and random forest for predicting diabetes prevalence. GLMs assume a linear relationship between the logit of the response variable and the features, but the assumption may not hold up for any prominent logit function we choose – a limitation of our method. We also used random forest, because it does not make many assumptions and therefore performs well when we do not know much about our data or domain knowledge. Random forest randomly selects subsets of features and bootstrapped samples, which reduces overfitting without need for feature engineering. However, random forest does have the limitation of low interpretability.

Exploratory Data Analysis (EDA)



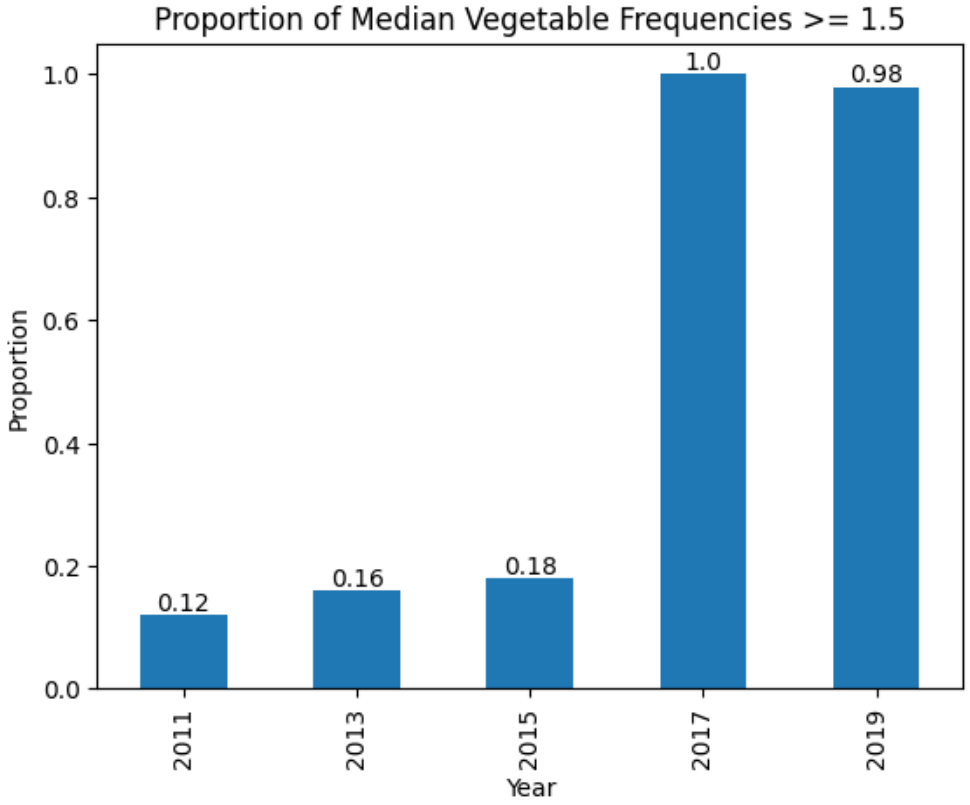
Above, we visualized the distribution of diabetes prevalence over 2011, 2013, 2015, 2017, and 2019 for nine states. These states were chosen because they represent about half of the US population and regionally represent the entirety of the United States. We also included a boxplot for the entire U.S. (labeled “US”) so we can see how each state compares to the national aggregate.

The West, Northeast coast, and Midwest states have lower medians (i.e. California, New York) for crude prevalence of diabetes for adults, while the Southern states (i.e. Georgia, Texas) have high medians. Based on these trends, we see significant variation in crude prevalence distribution from state to state, which makes crude diabetes prevalence useful as a response. We could use region as a covariate for our prediction models as there appears to be a relationship between region and crude prevalence.

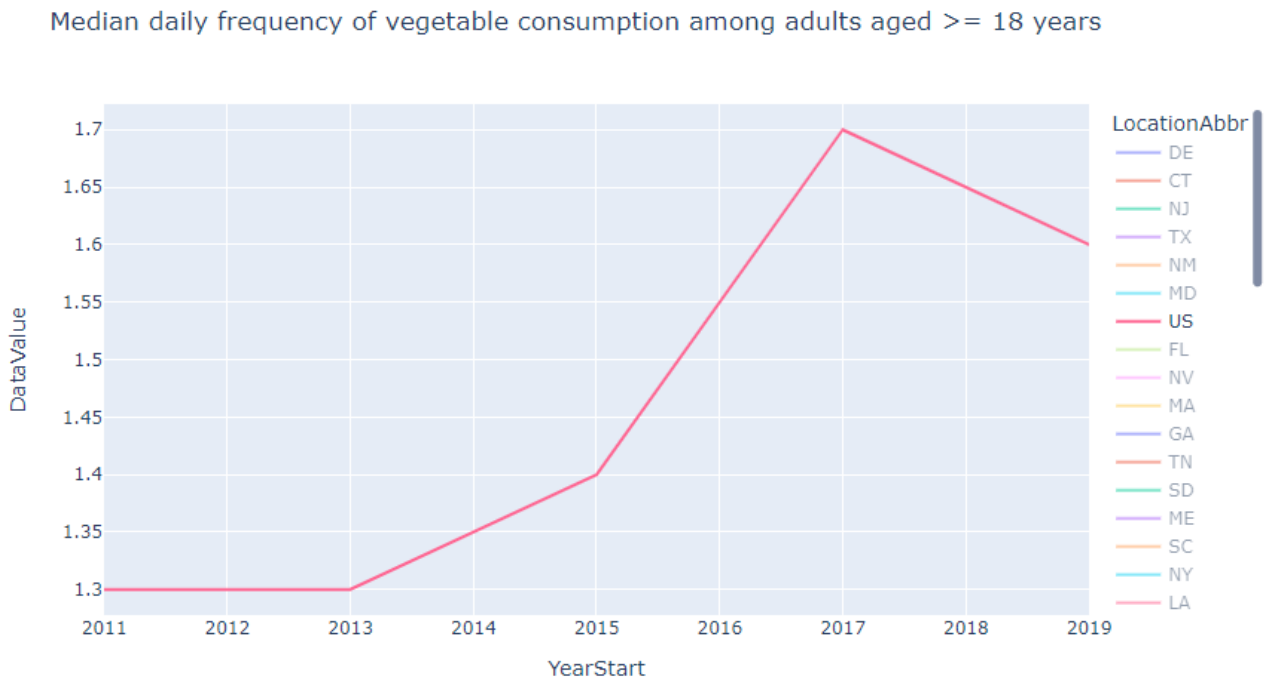


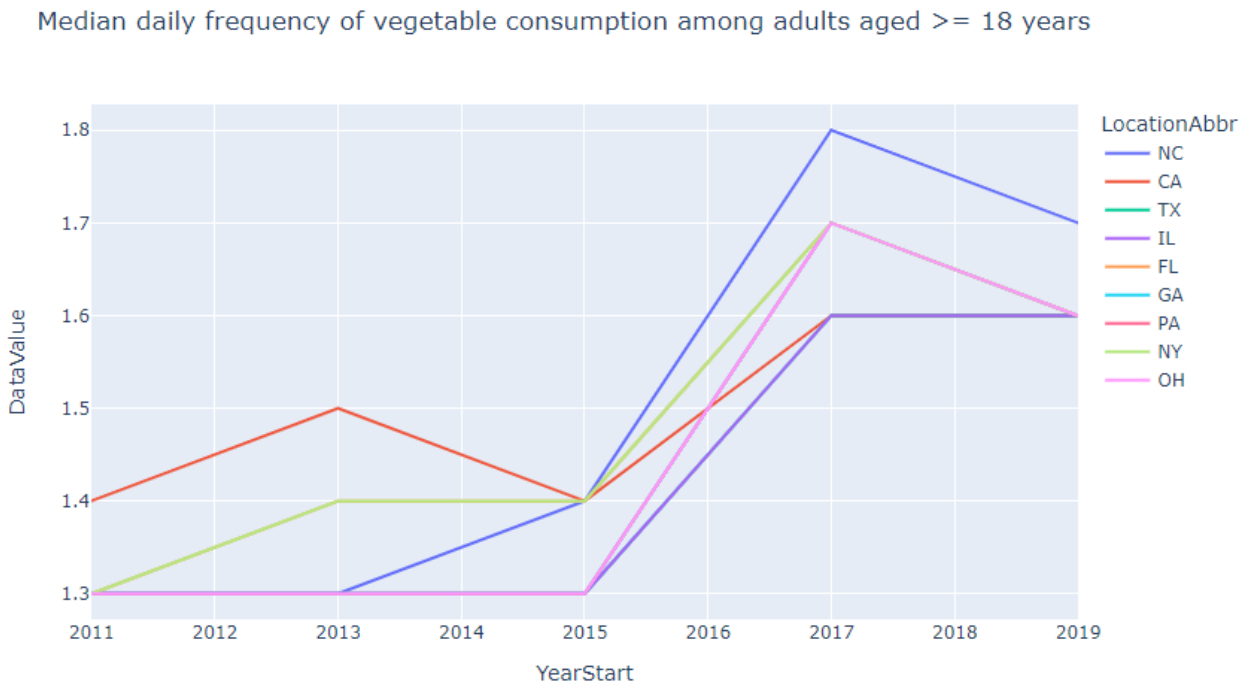
This visualization displays the distribution of the mean crude prevalence for each state over years 2011, 2013, 2015, 2017, and 2019; and is relevant to our causal inference work, since any unusual characteristic of this distribution (e.g. bimodal) could indicate potential presence of a confounding variable. We see that the distribution is roughly symmetric and bell-shaped without any particularly unusual characteristics. We also see from this histogram that the most frequent mean crude prevalence is 10% and the distribution has a right tail starting from 14%. Most of the data values are 8%-13%. A relationship we want to follow up on is

whether the state or time is more highly correlated with the outliers in the right tail. We would also like to see if the same state appears multiple times in the most frequent bins.



This visualization is relevant to our research question because – for causal inference – we can observe for each year how many of the states were classified as either high vegetable consumption states (i.e. states in the treatment group and above or equal to the threshold of in terms of median daily frequency of vegetable consumption among all of our state-years, which was 1.5) or low vegetable consumption states. The trends we observed are, in 2017 and 2019, the proportion of states that are at or above the threshold in daily vegetable consumption frequency is close to or equal to 1, making at least almost all states “high” consumption states in 2017 and 2019. These proportions are significantly higher than those in the previous years (2011, 2013, 2015). This was due to a change in data collection, BRFSS changed its questionnaire regarding vegetable consumption in 2017 (Lee & Moore, 2017).





These line plots are relevant to our research questions because they display the trends in our treatment and motivate the question, how the consumption of vegetables per state changes over time for causal inference. We observed that the median daily frequency of vegetable consumption generally increased between 2011 and 2019 not only at the national level but also at the state level. The first line plot shows that, while the nationwide frequency went down from 2017 to 2019, the nationwide frequency increased from 2011 to 2017 and still

remains high in 2019 compared to in 2011. The second line plot looks at state-level frequencies in the same states that are looked at in the Crude Prevalence boxplot. The second line plot shows that the vegetable consumption frequencies by state either increased or stayed the same in 2015 compared to 2011. But the state frequencies drastically went up between 2015 and 2017 and either stayed the same or decreased between 2017 and 2019.

Causal Inference

Methods

Variables

We obtained data on median vegetable consumption from each state-year – that is, median vegetable consumption among adults in each American state in each odd year from 2011 to 2019 – from the Chronic Disease Indicator dataset (Centers for Disease Control and Prevention, 2023a). We will call the median vegetable consumption in year *i* and state *j* as $\text{veg}_{(i,j)}$. We calculated the mean of the median state vegetable consumptions in each year – that is,

$$\frac{1}{50}(\text{veg}_{(i,\text{Alabama})} + \text{veg}_{(i,\text{Alaska})} + ... + \text{veg}_{(i,\text{Wyoming})})$$

for each year $i \in \{2011, 2013, 2015, 2017, 2019\}$. We will call the yearly mean of median state vegetable consumptions $\text{veg}_{(i,\text{ mean})}$ for year *i*. For each year *i* and state *j*, we calculated $\text{veg\_residual}_{(i,j)} = \text{veg}_{(i,j)} - \text{veg}_{(i,\text{ mean})}$, the residual from the mean of median state consumptions from that year. We then calculated $\text{veg\_residual}_{\text{median}}$, the median of all $\text{veg\_residual}_{(i,j)}$'s.

We then tried to first formulate our variables as follows:

- **Treatment:** High residual from the yearly mean of median state vegetable consumptions ($\text{veg\_residual}_{(i,j)} > \text{veg\_residual}_{\text{median}}$)
- **Control:** Low residual from the yearly mean of median vegetable consumption ($\text{veg\_residual}_{(i,j)} \leq \text{veg\_residual}_{\text{median}}$)
- **Outcome:** Crude prevalence of diabetes (denoted as $y_{(i,j)}$ for year *i* and state *j*)

It would be ideal to define the treatment variable based on whether the median vegetable consumption in a state-year is over a certain threshold or not (that is, whether $\text{veg}_{(i,j)} > t$ for some reasonable value of *t* for each year *i* and state *j*). However, the median vegetable consumption is measured in frequencies (that is, how many times the respondent ate vegetables per day) and not on the amount of vegetables eaten (Lee & Moore, 2017). But the recommended vegetable consumption is generally given in cups, not frequencies (Lee et al., 2022). So, it was difficult to find a reasonable threshold with domain knowledge. Another problem we faced was that BRFSS changed its questionnaire regarding vegetable consumption in 2017, so vegetable consumption data from 2011, 2013, and 2015 were not directly comparable to that from 2017 and 2019 (Lee & Moore, 2017). So, we used $\text{veg\_residual}_{\text{median}}$ as

our threshold to simply see which state-years have unusually high or low vegetable consumptions compared to other states in that year.

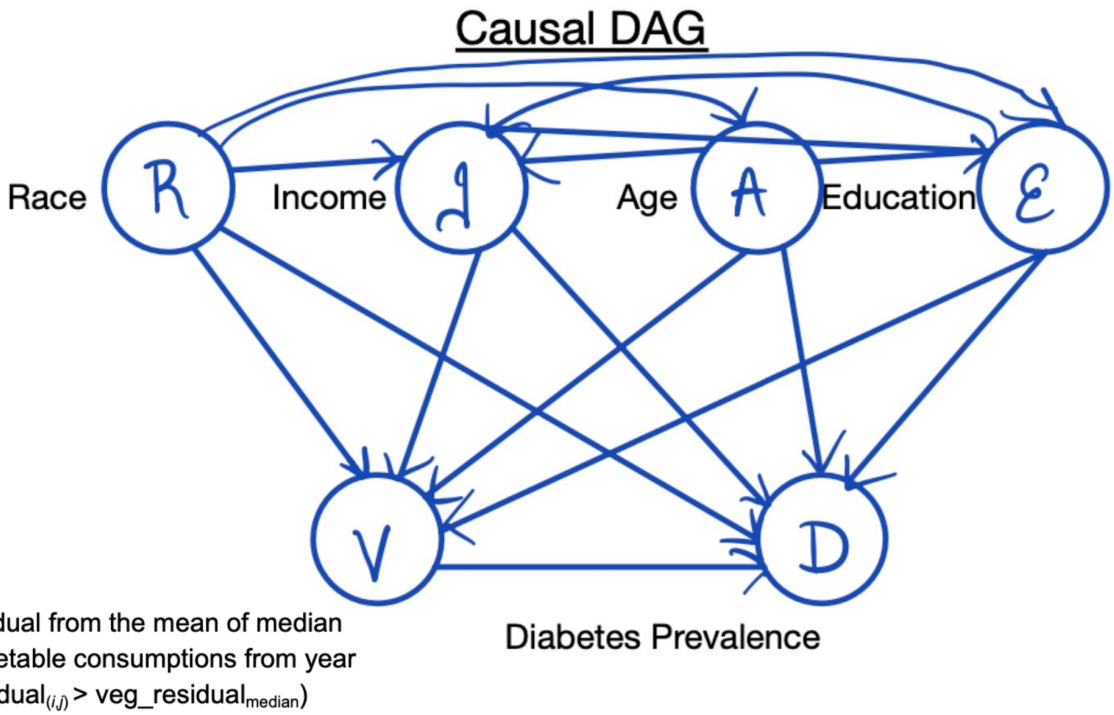
Confounders

We also used age, education, race, and income as confounders. The confounder data is also organized in state-years.

- **Age** (≥ 45 years): We hypothesized based on our experiences that both vegetable consumption and diabetes prevalence is influenced by age, partially due to the impact of age on income. This is especially true for diabetes, as diabetes prevalence spikes after the age of 45 years (Basina, 2023).
- **Education** (Bachelor's degree): We believe that education affects vegetable consumption and diabetes prevalence because of its impact on socioeconomic status and health knowledge
- **Income:** We believe that income has an effect on vegetable consumption and diabetes prevalence, as a healthy diet is often too expensive for the low income earners to afford (Caporuscio, 2020).
- **Race** (White or Non-White): We believe that race has an effect on vegetable consumption and diabetes prevalence due to its potential impact on income, age, and education. In fact, minorities have higher rates of diabetes in the United States (Rodríguez & Campbell, 2017).

The unconfoundedness assumption holds because, within each strata of the confounders, it is reasonable to believe that consumption of vegetables is randomly distributed. We believe that adjusting for age, education, race, and income are sufficient to account for the various confounding effects of these demographic and socioeconomic factors.

Causal Directed Acyclic Graph



As can be seen above, there are no colliders in our dataset.

Adjusting for Confounders: Inverse Propensity Score Weighting (IPW)

We used IPW to account for confounders to directly assess the relationship between vegetable consumption and diabetes prevalence.

We applied inverse propensity score weighting first by training a logistic regression on the existing covariates and treatment data and then using the logistic regression to calculate the probability of each state-year receiving the treatment given the confounders as if the treatments were not given.

Mathematically speaking, say $x_{(i,j)} \in \mathbb{R}^4$ is a vector containing the four confounder values for year *i* and state *j*. Let $z_{(i,j)}$ be the treatment variable for year *i* and state *j*. Then, we estimated the coefficient parameters $\beta \in \mathbb{R}^4$ for predicting $\hat{z}_{(i,j)}$, the probability that $z_{(i,j)}$ is 1, as if $z_{(i,j)}$ is unknown and with the model…

$$\hat{z}_{(i,j)} = \sigma(\beta^T x_{(i,j)}), \text{ where } \sigma(t) = \frac{1}{1+e^{-t}} \text{ is the sigmoid function.}$$

…by minimizing overall binary cross-entropy loss over β to calculate $\hat{\beta}$:

$$\begin{aligned} l(z_{(i,j)}, \hat{z}_{(i,j)}; \beta) &= -[z_{(i,j)} \log(\hat{z}_{(i,j)}) + (1 - z_{(i,j)}) \log(1 - \hat{z}_{(i,j)})] \\ &= -[z_{(i,j)} \log(\sigma(\beta^T x_{(i,j)})) + (1 - z_{(i,j)}) \log(1 - \sigma(\beta^T x_{(i,j)}))] \end{aligned}$$

$$\hat{\beta} = \text{argmin}_\beta \sum_{i=1}^{n} l(z_{(i,j)}, \hat{z}_{(i,j)}; \beta)$$

After doing so, we calculated the propensity score of each state-year by using this logistic regression model: $e(x_{(i,j)}) = \hat{z}_{(i,j)}$.

Then, we weighed each state-year by the inverse of its propensity score depending on treatment status and then calculated the average treatment effect (ATE) for each year *i* (or $\hat{\tau}_i$) using weighted difference in means…

$$\hat{\tau}_i = \frac{1}{n_i} \sum_{j:z_{(i,j)}=1} y_{(i,j)} \frac{1}{e\left(x_{(i,j)}\right)} - \frac{1}{n_i} \sum_{j:z_{(i,j)}=0} y_{(i,j)} \frac{1}{1 - e\left(x_{(i,j)}\right)}$$

…where n_i = 50 is the number of state-years in year *i*.

Accounting for Spillover

From here, instead of calculating one ATE for all five years, we estimated the overall causal effect (denoted by $\hat{\tau}$) by averaging the five ATEs from the five years:

$$\hat{\tau} = \frac{1}{5}\left(\hat{\tau}_{2011} + \hat{\tau}_{2013} + \hat{\tau}_{2015} + \hat{\tau}_{2017} + \hat{\tau}_{2019}\right)$$

This was because using the full data by itself came with the risk of temporal spillover effect (with, for instance, the treatment/vegetable consumption of California in 2011 potentially affecting that of California in 2013) and therefore the risk of violation of Stable Unit Treatment Value Assumption (SUTVA). Since IPW operates under SUTVA, we had to make this adjustment to account for the spillover.

Bootstrapping

To estimate uncertainty in our IPW estimator, we used the bootstrap method. By taking 1000 bootstrap samples, we estimated the uncertainty for each year separately. We ran our samples in parallel for computational efficiency. Bootstrapping allowed us to calculate an overall ATE uncertainty estimate for all the years combined.

Results

Year	ATE	95% CI
2011	-2.899	(-4.034, -0.309)
2013	-3.169	(-4.680, -0.433)
2015	-3.293	(-4.953, -1.555)
2017	-1.445	(-3.163, 0.132)
2019	-2.223	(-3.779, -0.644)
Overall	-2.596	(-3.305, -1.473)

The ATE for each year is given by the table above. The year with the lowest ATE magnitude was 2015; the year with the highest ATE magnitude was 2017. The overall ATE across all five years in our dataset can be interpreted as meaning that the high consumption of vegetables led to an approximately 2.6% decrease in the crude prevalence for diabetes in a state-year between 2011 and 2019. Our overall ATE (-2.596) suggests a strong causal effect for high vegetable consumption on decreasing the crude prevalence for diabetes. Even though the confidence interval for 2017 contains 0, because our confidence intervals for overall ATE and the other four years do not contain 0, there is strong evidence that higher vegetable consumption decreases diabetes prevalence.

Discussion

There were limitations in our data that could not be accounted for in our analysis. One limitation came from how we defined the treatment based on our data constraints. Our treatment definition involved a comparative analysis of the residual of state-level treatment (median consumption of vegetables) to that at a nationwide-level. Since vegetable consumption is measured in frequencies, we could not apply domain knowledge to set a reasonable threshold for vegetable consumption treatment variable. The data collection method for vegetable consumption changed in 2017 (Lee & Moore, 2017). So, we had to rely on the residuals from yearly average median vegetable consumption to determine the threshold. For a state to have been classified as having the treatment, it must have had a high vegetable consumption relative to the nationwide average in that year, but the state would not necessarily have a high vegetable consumption in absolute terms.

In our causal DAG, we accounted for important demographic factors; however, while we believe that the unconfoundedness assumption holds reasonably well, we believe we may have missed other potential confounders such as state and accessibility to vegetables. For example, state could be a confounder if different state policies emphasized nutrition education differently. Accessibility could be a confounder because cost and physical access limitations could affect the consumption of vegetables.

As mentioned in the Data Overview section, additional data we wished we had included more detailed information on nutritional intake, Type I and Type II diabetes segregation, levels of exercise, cost of living, and access to a dietitian/nutritionist. This data could have increased our certainty in our causal effect estimates. Our findings are also broad due to a high level of granularity (state-year instead of, say, county-year) in our data. Plus, our findings are also confined to the United States and will not be generalizable for regions outside of the United States. We are reasonably confident of a causal relationship between our treatment and outcome, even though we recognize that there are other potential confounders that could alter our findings.

Prediction with GLMs and Nonparametric Method

Methods

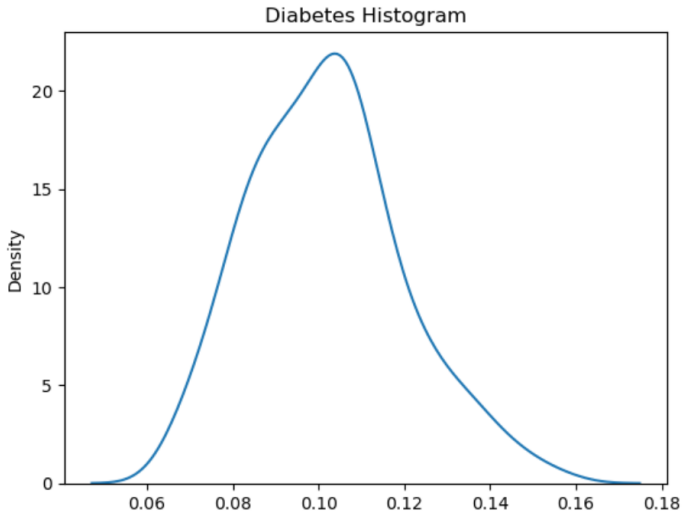
We will use the same features for each model for fair comparison. We will predict adult diabetes crude prevalence, where each data point is a state-year pair in 2011, 2013, 2015, 2017, or 2019.

- **Year**: Data collection process, diabetes onset, and population demographic densities change over time, so year will affect prevalence. Year scale differs, so it is normalized.
- **Age**: Diabetes prevalence spikes after 45 years of age, so proportion of adults at least 45 years old is an appropriate age feature (Basina, 2023).
- **Education**: We believe that diabetes results from poor diet and exercise, which depend on socioeconomic status and health knowledge, which are further influenced by education. Education is measured by proportion with bachelor's degree.
- **Income**: Income affects access to healthcare, healthy food, healthy markets, etc. (Caporuscio, 2020). Median income is normalized due to scale.
- **Race**: Minorities have higher rates of diabetes in the United States (Rodríguez & Campbell, 2017). Race is represented as non-white proportion per state-year.
- **Regions**: Different regions have different cultures and cuisines, which impacts health. Regions are one-hot-encoded as the four regions from the US Census classifications (Midwest, Northeast, South, and West) (The United States Census Bureau, n.d.-f).
- **Vegetable Consumption**: Vegetable consumption is essential to overall health.

Feature	Correlation Coefficient with Response
Year	0.201485
Vegetable Consumption	-0.067563
Race	-0.201687
Income	0.761838
Age	0.201485
Education	-0.492967
Midwest	-0.128374
Northeast	-0.201687
South	0.690467
West	-0.419151

Our Frequentist GLM used the Beta likelihood to model prevalence. Beta regression is ideal because its output matches the range of prevalence (a proportion). Logistic regression also fits, but responses are not binary. Beta regression assumes prevalence follows a Beta distribution, which may not be the case. However, beta distributions are highly flexible and can appropriately model the shape of prevalence, which is roughly normal-shaped in [0, 1]. One limitation is that GLM assumes independent datapoints, so each state-year is independent from another. However, this assumption may be false because a state's prevalence in one year may affect the state's prevalence for the next year. Also, states may affect nearby states via cultural influence and migration. We assume there is a linear relationship between logit of prevalence and the features because inverse link function is sigmoid. We exclude bias terms to emphasize coefficients; this will not hurt performance because features are near zero.

Our Bayesian GLM will use N(0, 1) prior for each coefficient. We use zero-centered priors due to uncertainty of coefficient signs, and to also avoid overfitting with strong priors. We



use uninformative priors due to uncertainty in relationships between features and response because of limited domain knowledge. We use HalfNormal(0.01) prior for variance of output because Beta distributions and prevalence rates have low variance, and to ensure variance is positive. Inverse link function is sigmoid because mean of the Beta distribution needs to be within [0, 1].

Random forest was chosen for nonparametric prediction. This is because it performs well without extensive domain knowledge, since it randomly selects subsets of features and uses bootstrapped samples, limiting overfitting. Random forest, which limits variance via bagging, was chosen over decision trees, which are prone to overfitting. Random forests do not assume anything about the distribution of data, unlike GLMs.

To compare models, we computed root mean square error (RMSE) to measure loss, since accuracy is undefined for regression. We computed RMSEs on training and test data, with 80/20 training/test split, to assess models performance and extent of overfitting. Also, we computed log-likelihood to compare goodness of fit between frequentist and Bayesian GLMs. For Bayesian GLM, we checked 94% credible intervals to determine parameter uncertainty. We also plotted posterior predictive distribution against observed values to assess model uncertainty. Additionally, we inspected R-hat values to determine if MCMC coefficient sampling has converged. For frequentist GLM, coefficient values were evaluated by 95% confidence intervals and corresponding two-tailed p-values. We computed R-squared values (coefficient of determination) for random forest to assess train and test fit.

Results

Frequentist GLM

BetaModel Results

Dep. Variable:	y	Log-Likelihood:	699.76
Model:	BetaModel	AIC:	-1378.
Method:	Maximum Likelihood	BIC:	-1341.
Date:	Tue, 09 May 2023		
Time:	18:46:29		
No. Observations:	200		
Df Residuals:	189		
Df Model:	10		

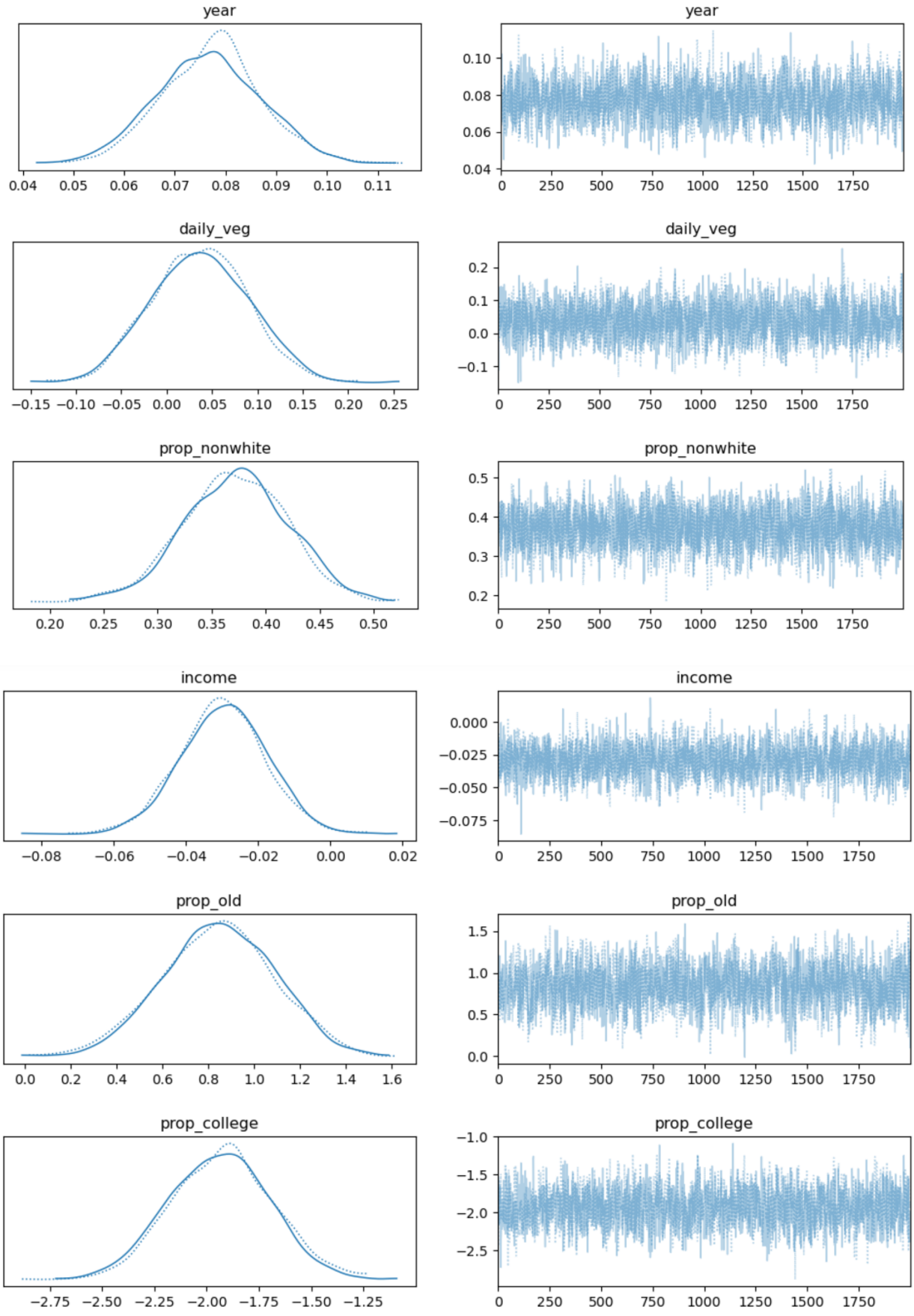
	coef	std err	z	P>|z|	[0.025	0.975]
Year	0.0692	0.010	6.935	0.000	0.050	0.089
Vegetable_Consumption_per_Day	0.0766	0.053	1.453	0.146	-0.027	0.180
Proportion_Non_White	0.4224	0.047	8.895	0.000	0.329	0.515
Median_Income_Current_Dollars	-0.0292	0.012	-2.515	0.012	-0.052	-0.006
Proportion_of_State__45_yrs	1.1297	0.248	4.548	0.000	0.643	1.617
Proportion_of_Bachelors_or_Higher	-1.9426	0.235	-8.283	0.000	-2.402	-1.483
Midwest	-2.3686	0.124	-19.092	0.000	-2.612	-2.125
North_East	-2.3499	0.136	-17.329	0.000	-2.616	-2.084
South	-2.2580	0.125	-18.008	0.000	-2.504	-2.012
West	-2.5210	0.124	-20.273	0.000	-2.765	-2.277
precision	7.4281	0.100	74.277	0.000	7.232	7.624

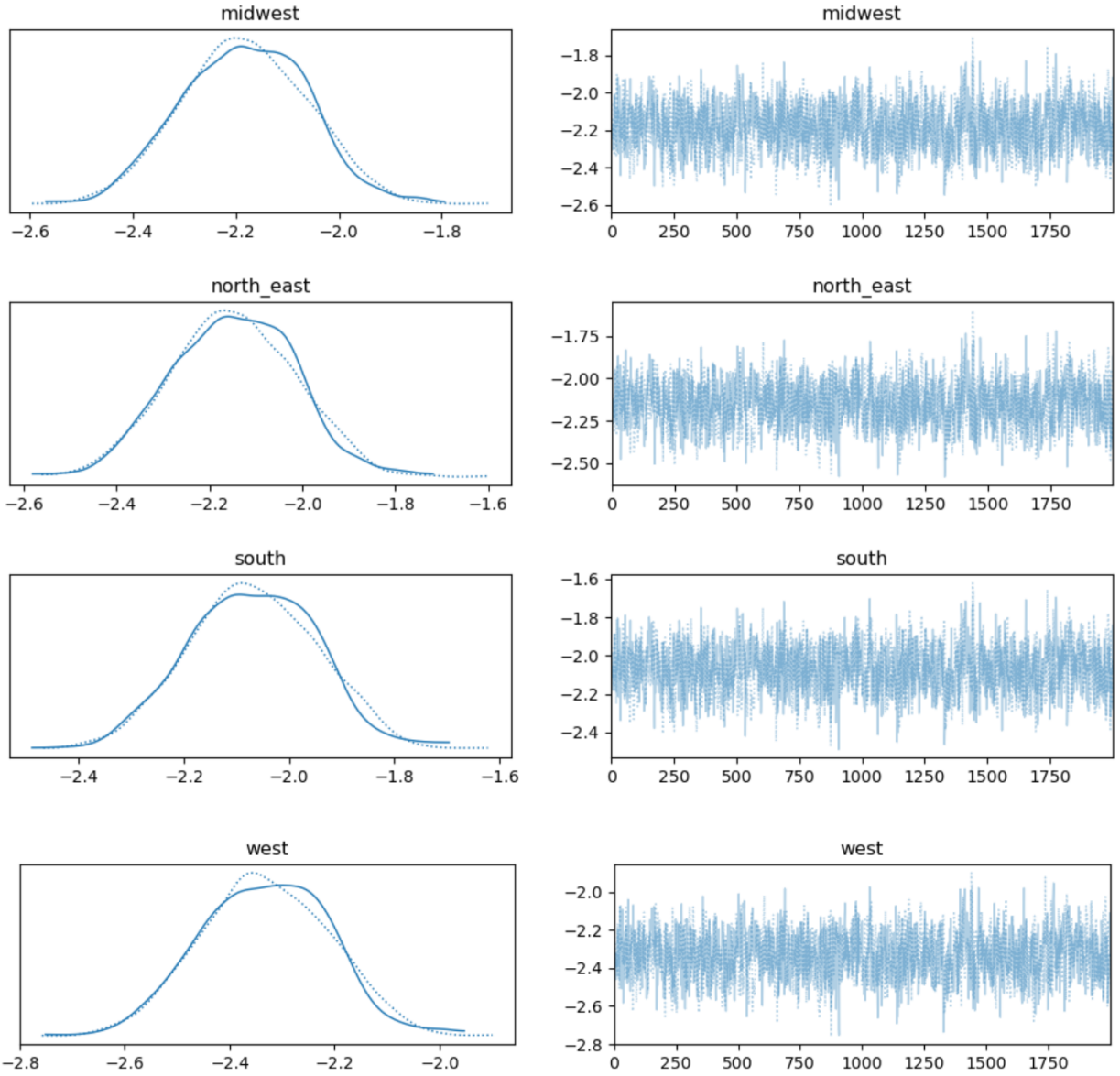
Frequentist GLM log-likelihood is 699.76; average log-likelihood is 3.499. Fitted parameters are shown above. Vegetable consumption includes zero in the 95% confidence interval, e.g. we cannot reject the null hypothesis that the true value of the coefficient is zero; there is no statistically significant evidence that vegetable consumption has an association with diabetes prevalence. Every other variable's *p*-value is less than 0.05, without zero in the confidence interval, meaning they each have a statistically significant relationship with diabetes prevalence in this model. The training RMSE is 0.00734; test RMSE is 0.0104.

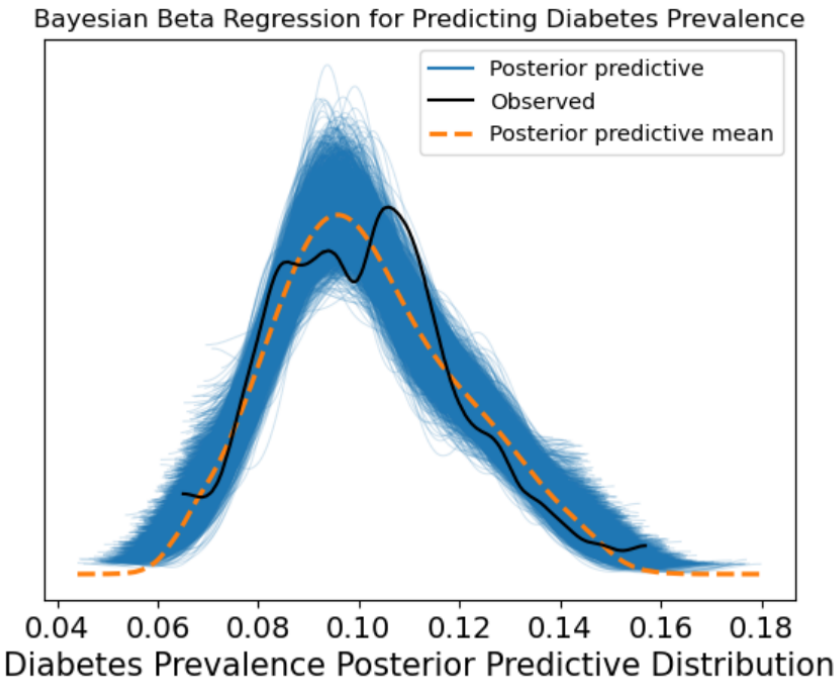
Bayesian GLM

	mean	sd	hdi_3%	hdi_97%	mcse_mean	mcse_sd	ess_bulk	ess_tail	r_hat
year	0.077	0.010	0.058	0.096	0.000	0.000	1913.0	2099.0	1.0
daily_veg	0.037	0.053	-0.062	0.135	0.001	0.001	2209.0	2287.0	1.0
prop_nonwhite	0.372	0.048	0.284	0.466	0.001	0.001	2551.0	2069.0	1.0
income	-0.030	0.012	-0.051	-0.007	0.000	0.000	2456.0	2516.0	1.0
prop_old	0.844	0.247	0.394	1.302	0.006	0.004	1521.0	1809.0	1.0
prop_college	-1.932	0.241	-2.403	-1.505	0.005	0.004	2072.0	2831.0	1.0
midwest	-2.183	0.122	-2.413	-1.965	0.004	0.002	1172.0	1644.0	1.0
north_east	-2.149	0.133	-2.412	-1.919	0.004	0.003	1227.0	1534.0	1.0
south	-2.070	0.123	-2.291	-1.837	0.004	0.003	1125.0	1601.0	1.0
west	-2.333	0.123	-2.551	-2.098	0.004	0.003	1198.0	1569.0	1.0
sigma	0.008	0.000	0.007	0.008	0.000	0.000	3246.0	2253.0	1.0

Bayesian GLM average log-likelihood is 3.463. Similarly, vegetable consumption is the only feature with zero in the 94% credible interval, which means there is 94% probability the true coefficient lies within this interval. Hence we cannot reject the null hypothesis that vegetable consumption has a statistically significant relationship with diabetes prevalence. The opposite is true for every other variable, which excludes 0 in their credible intervals. Each variable's R-hat value is 1.0, meaning the chains have converged and are sampling from the target distribution effectively, creating reliable estimates of the posterior distribution. Sigma (variance) is very close to 0.008 with almost no variance. Training RMSE is 0.00737; test RMSE is 0.0102. Below are credible intervals, showing coefficient variance in the 2000 samples. Also shown is the posterior predictive distribution, comparing the GLM with observed data.


year
0.04 0.05 0.06 0.07 0.08 0.09 0.10 0.11
year
0.10 0.08 0.06 0.04
0 250 500 750 1000 1250 1500 1750
daily_veg
−0.15 −0.10 −0.05 0.00 0.05 0.10 0.15 0.20 0.25
daily_veg
0.2 0.1 0.0 −0.1
0 250 500 750 1000 1250 1500 1750
prop_nonwhite
0.20 0.25 0.30 0.35 0.40 0.45 0.50
prop_nonwhite
0.5 0.4 0.3 0.2
0 250 500 750 1000 1250 1500 1750
income
−0.08 −0.06 −0.04 −0.02 0.00 0.02
income
0.000 −0.025 −0.050 −0.075
0 250 500 750 1000 1250 1500 1750
prop_old
0.0 0.2 0.4 0.6 0.8 1.0 1.2 1.4 1.6
prop_old
1.5 1.0 0.5 0.0
0 250 500 750 1000 1250 1500 1750
prop_college
−2.75 −2.50 −2.25 −2.00 −1.75 −1.50 −1.25
prop_college
−1.0 −1.5 −2.0 −2.5
0 250 500 750 1000 1250 1500 1750


midwest
−2.6
−2.4
−2.2
−2.0
−1.8
midwest
−1.8
−2.0
−2.2
−2.4
−2.6
0
250
500
750
1000
1250
1500
1750
north_east
−2.6
−2.4
−2.2
−2.0
−1.8
−1.6
north_east
−1.75
−2.00
−2.25
−2.50
0
250
500
750
1000
1250
1500
1750
south
−2.4
−2.2
−2.0
−1.8
−1.6
south
−1.6
−1.8
−2.0
−2.2
−2.4
0
250
500
750
1000
1250
1500
1750
west
−2.8
−2.6
−2.4
−2.2
−2.0
west
−2.0
−2.2
−2.4
−2.6
−2.8
0
250
500
750
1000
1250
1500
1750



Random Forest

Random forest train RMSE is 0.0029; test RMSE is 0.0087. After random grid search hyperparameter tuning, test RMSE decreased to 0.0084. The optimal hyperparameters are shown below:

Hyperparameter	Value
Number of Trees	2000
Minimum Samples per Leaf Node	1
Minimum Samples Required to Split	2
Max Features	Auto (square root of # features)
Max Depth	80

Training R-squared value is 0.975, and test R-squared value is 0.734, which means 97.5% and 73.3% of the variance in the response variable is explained via the model by the features in the training and test data, respectively.

Discussion

The random forest model outperformed both GLMs. The results are replicated for convenience:

Model	Train RMSE	Test RMSE	Average Log-Likelihood
Frequentist GLM	0.0073370059308775 1	0.0103602613319844 4	3.4987971142464893
Bayesian GLM	0.0073676573286971 594	0.0102338660232479 59	3.463049014178649
Random Forest	0.0029	0.0084	N/A

The RMSEs can be interpreted as the Euclidean distance between predictions and reality. Random forest model's predictions are closer to reality because it makes fewer assumptions about relationships between features and response. GLMs assume linear relationships with respect to the link function between the variables, and Bayesian GLMs additionally assume priors, but random forests do not make assumptions about the data. Frequentist GLM is better at predicting observed data due to higher average log-likelihood and slightly lower training RMSE, but this indicates slightly higher overfitting compared to Bayesian GLM. Hence although RMSEs are almost identical, Bayesian GLM performed slightly better than frequentist GLM, despite the vague standard normal priors. High training R-squared value indicates random forest fit training data exceptionally well, but sharp decrease in test R-squared indicates overfitting. Due to lack of structural assumptions, random forest outperformed GLMs. We are confident in applying each model to future datasets in a similar setting (United States, 21st century) because of relatively low overfitting and RMSEs. Outside of similar settings, generalizability is uncertain.

Bayesian and frequentist GLM coefficients are similar, but Bayesian coefficients are smaller. Interestingly, Bayesian GLM performed better on test data. This can be explained by the weak zero-centered Bayesian priors, which moves coefficients closer to zero. This is equivalent to introducing weak regularization to Bayesian GLM, which penalizes larger coefficients, increasing bias (e.g. higher train set loss) while reducing variance (e.g. lower test set loss). Both GLMs suggest only vegetable consumption is an ineffective predictor. GLM coefficients are repeated below:

Feature	Frequentist	Bayesian
Year	0.0692	0.077
Vegetable Consumption	0.0766	0.037
Race	0.4224	0.372
Income	-0.0292	-0.030
Age	1.1297	0.844
Education	-1.9426	-1.932
Midwest	-2.3686	-2.183
North East	-2.3499	-2.149
South	-2.2580	-2.070
West	-2.5210	-2.333

Each coefficient can be interpreted as follows: conditioned on every other feature, the coefficient for a given feature reflects its association with diabetes prevalence. Negative coefficients indicate a negative association with diabetes, and magnitude indicates size of the effect. The only exception is vegetable consumption, which is not significant at 5% and 6% for frequentist and Bayesian, respectively. So given every other feature, the following are negatively associated with diabetes: income, education, Midwest, Northeast, South, and West; the following are positively associated: year, percent nonwhite, age. Random forest is non-interpretable because it is difficult to understand how the model makes predictions since it is composed of an ensemble of decision trees, each of which makes a prediction based on a subset of the features and bootstrapped samples with unique splits.

A limitation is that GLMs assume a linear relationship between the features and the link function applied to diabetes prevalence, which is not certain. Another limitation is an assumption of independence between datapoints, which is false. Furthermore, it assumes diabetes prevalence follows a Beta distribution, which may be false. These limitations may cause GLMs to underfit, and underperform on test data. Furthermore, our features are highly collinear, so GLMs may cause unstable predictions. Random forests lack these limitations due to flexibility and lack of assumptions, but a limitation is lack of interpretability. Furthermore, random forests may still overfit despite bagging, especially with deep trees and noisy data.

Relationships between features and prevalence may be nonlinear, so adding polynomial or transformed features may improve each model. Additional features that may also improve models include diabetic factors: exercise, smoking, alcohol, etc. Furthermore, increasing granularity instead of statewide statistics may increase data variance and make features more prominent. Finally, dimensionality reduction like PCA may minimize impact of multicollinearity in GLMs.

The training data set is small (200), so there is variance from small sample size. Furthermore, datapoints are not independent, which amplifies data variance. Despite this, each model has low uncertainty. Uncertainty in GLMs is quantitatively low for each feature except vegetable consumption, which lacks significance at the 5%-6% level. Since the scale of credible and confidence intervals are small, the coefficients are stable and consistent. Bayesian GLM predictions have low uncertainty because the posterior distribution roughly matches the shape of observations, although it does not capture its full complexity. Frequentist GLM predictions have even less uncertainty due to higher average log-likelihood. Random forest has relatively higher uncertainty due to drop in R-squared from training to test data, but better performance overall. The GLMs' low uncertainty stems from strong assumptions about the distribution of the data, as well as zero-centered priors for Bayesian. Random forest makes fewer assumptions about data, at the cost of higher uncertainty. However, random forest uncertainty is mitigated by the hyperparameters; an ensemble of 2000 decision trees has relatively low variance. Overall, each model has low uncertainty because the data consists of statewide statistics, so our data overall has low variance (due to the Central Limit Theorem) and random noise is mitigated by large sample sizes in census data collection. Additionally, the scale of features and prevalence are close to zero, which makes model estimation low-variance.

Even though the random forest model performed better than the other models it is less interpretable. Since many of the features the random forest model is using to precut crude prevalence of diabetes, it could be harmful to applying the to future dataset for real world problems. This is because most of the features like race, age, income, education, and region are effects of social and historical trends. Predicting on these features could reinforce those trends.

Conclusions

Causal Inference

We found that higher consumption of vegetables causes a decrease in the crude prevalence of diabetes. However, our findings are not as generalizable as we would have liked. Our findings are broad, given that our data had a high-level of granularity (state-year). Our findings are also confined to the United States and will not be generalizable for regions outside of the United States. We also could not use domain knowledge to set a threshold for our treatment variable due to limitations in data collection for vegetable consumption. Further research is needed to assess causal effects of more specified nutrition intake on diabetes prevalence.

Based on our results, we suggest further research on assessing causal effects of more specified nutrition intake on diabetes prevalence. More specific the recommendations are, more helpful they can be for making individualized health decisions. Further researchwould be helpful for dietitians to weigh in on nutrition policies.

Prediction with GLMs and Random Forest

Random Forest outperformed both the Frequentist and Bayesian GLMs when measured by RMSE. The Bayesian GLM had a lower test RMSE than the frequentist GLM. However, the frequentist GLM had a slightly larger log-likelihood meaning that it fit the data better. These findings are confined to the United States and will not be generalizable for regions outside of the United States. Also, the predictions must be used as an aggregate measure for centralized planning, as the socioeconomic natures of the variables used could be harmful if applied too narrowly.

These predictions can be helpful for state health agencies to plan for the state diabetes care initiatives. However, the predictions must be used as an aggregate measure for centralized planning, as the socioeconomic consequences of using the variables we used for prediction could be harmful if applied too narrowly. Additionally, the American government should consider providing additional funding to the Southern states for diabetes care, as our GLM's show there is likely to be a higher prevalence of diabetes if in the South.

Overall

The analysis involved merging data from the CDC (Chronic Disease Indicator dataset) and the Census (ACS, CPS). The benefit of combining these sources was to obtain a variety of confounders/covariates without overly relying on one data source, although combining these sources meant we had to deal with inconsistencies in data collection.

Future studies could build on the work by holding longitudinal studies, adding in more confounding variables, expanding the geographical range of the work, or holding a randomized controlled experiment on the relationship between vegetable consumption and diabetes.

Having done both causal inference and predictions, we learned that we have to be careful in distinguishing predictive and causal relationships. For instance, being in the South has relatively high predictive power compared to other regions in both of our GLMs, but that does not mean that being in the South causes an increase in crude diabetes prevalence. There may be socioeconomic factors that are unaccounted for and are causing this high predictive power, and we would not want to risk creating a false conception about the South as a region from our study.

References

Basina, M. (2023, April 28). *Type 2 diabetes: Average age of onset, risk factors, prevention*. Medical News Today. https://www.medicalnewstoday.com/articles/317375

Caporuscio, J. (2020, June 22). *Food deserts: Definition, effects, and solutions*. Medical News Today. https://www.medicalnewstoday.com/articles/what-are-food-deserts

Centers for Disease Control and Prevention. (2019). *BRFSS Overview*. https://www.cdc.gov/brfss/annual_data/2019/pdf/overview-2019-508.pdf

Centers for Disease Control and Prevention. (2023a). *U.S. Chronic Disease Indicators*. https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-CDI-/g4ie-h725

Centers for Disease Control and Prevention. (2023b, April 18). *Type 2 Diabetes*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/type2.html

Daily, D. (2022, December 14). *Disclosure Avoidance Protections for the American Community Survey*. The United States Census Bureau. https://www.census.gov/newsroom/blogs/random-samplings/2022/12/disclosure-avoidance-protections-acs.html

Ford, E. S., & Mokdad, A. H. (2001). Fruit and vegetable consumption and diabetes mellitus incidence among U.S. adults. *Preventive Medicine*, *32*(1), 33–39. https://doi.org/10.1006/pmed.2000.0772

Guzman, G. (2019). New Data Show Income Increased in 14 States and 10 of the Largest Metros. The United States Census Bureau. https://www.census.gov/library/stories/2019/09/us-median-household-income-up-in-2018-from-2017.html#:~:text=Median%20Household%20Income%3A%20Historical%20Comparisons,than%20the%20prior%20three%20years.

Halvorsen, R. E., Elvestad, M., Molin, M., & Aune, D. (2021). Fruit and vegetable consumption and the risk of type 2 diabetes: a systematic review and dose–response meta-analysis of prospective studies. *BMJ Nutrition, Prevention & Health*, *4*(2), 519–531. https://doi.org/10.1136/bmjnph-2020-000218

Lee, S. H., & Moore, L. V. (2017). *A Data Users Guide to the BRFSS Fruit and Vegetable Questions: How to Analyze Consumption of Fruits and Vegetables*. Centers for Disease Control and Prevention. https://www.cdc.gov/nutrition/downloads/Data-Users-Guide-BRFSS-Fruit-and-Vegetable-Questions-508.pdf

Rodríguez, J. E., & Campbell, K. M. (2017). Racial and Ethnic Disparities in Prevalence and Care of Patients With Type 2 Diabetes. *Clinical Diabetes*, *35*(1), 66–70. https://doi.org/10.2337/cd15-0048

Rothbaum, J. (2020). Was Household Income the Highest Ever in 2019? The United States Census Bureau. https://www.census.gov/library/stories/2020/09/was-household-income-the-highest-ever-in-2019.html

The United States Census Bureau. (n.d.-a). *Collecting Data*. Census.Gov. Retrieved May 11, 2023, from https://www.census.gov/programs-surveys/cps/technical-documentation/methodology/collecting-data.html

The United States Census Bureau. (n.d.-b). *Current Population Survey (CPS)*. Census.Gov. Retrieved May 9, 2023, from https://www.census.gov/programs-surveys/cps.html

The United States Census Bureau. (n.d.-c). *Educational Attainment*. Retrieved May 9, 2023, from https://data.census.gov/table?q=education&tid=ACSST1Y2021.S1501

The United States Census Bureau. (n.d.-d). *Historical Income Tables: Households*. Retrieved May 11, 2023, from

https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-income-households.html

The United States Census Bureau. (n.d.-e). *Methodology*. Census.Gov. Retrieved May 9, 2023, from https://www.census.gov/programs-surveys/cps/technical-documentation/methodology.html

The United States Census Bureau. (n.d.-f). *Statistical Groupings of States and Counties*. https://www2.census.gov/geo/pdfs/reference/GARM/Ch6GARM.pdf

The United States Census Bureau. (2022a). *Population Distribution by Race/Ethnicity*. KFF. https://www.kff.org/other/state-indicator/distribution-by-raceethnicity/

The United States Census Bureau. (2022b). *Sample Design and Selection*. https://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/2022/acs_design_methodology_ch04_2022.pdf

Williams, D. E. M., Wareham, N. J., Cox, B. D., Byrne, C. D., Hales, C. N., & Day, N. E. (1999). Frequent Salad Vegetable Consumption Is Associated with A Reduction in the Risk of Diabetes Mellitus. *Journal of Clinical Epidemiology*, *52*(4), 329–335. https://doi.org/10.1016/S0895-4356(99)00006-2