Causal Inference Final Project

Joshua Hug & Matthew Martinez

December 2020

Abstract

With the pandemic causing financial and food insecurities around the world, a stronger understanding of the ramifications of food insecurity is needed if the effects of the pandemic are to be measured fully. With this in mind, we attempt to measure any causal effect of proximity to food on average life expectancy. We carry out this analysis by attempting several observational causal inference methods. We begin by conducting a traditional regression outcome method, followed by several propensity score variants. Our final method is an attempt at a more recent machine learning approaches to causal inference. To conclude we review the covariate balance across our methods to measure our risk to biased treatment effect estimates.

1 Introduction

For our final project we focus on data from the USDA Food Research Atlas data set USDA (2017) as well as life expectancy data from the CDC (2019). This data set contains demographic, socioeconomic and life expectancy information from over 65,000 US census tracts. The focus of this data set is to break down the various measures of food accessibility of each tract as well as highlight other demographic information in the tract. In particular, we will focus on areas that are in food deserts, which are defined as:

1) Areas without access to healthy foods as measured by distance to food stores.
2) Individual-level of resources, such as income or access to a vehicle
3) Neighborhood level of resources and access to public transport.

While there are many ways to evaluate how healthy diets are in certain areas, one such factor would be proximity to food access. We have decided to conduct an observational study on the causal effects of living in a tract with low access to food on

the average life expectancy of the tract. As it stands, over 23.5 million Americans live in food deserts, with nearly half of them being classified as low income. This means residents living in food deserts also have a hard time finding foods that are culturally relevant and that meet their dietary restrictions. Consequently people living in these areas have roughly 2.5 times the exposure to fast-food restaurants, a traditionally unhealthy subset of food (dosomething.org). This problem is likely to only be exacerbated by the recent global pandemic. As such we feel a causal analysis on the accessibility to food in a tract on the average life expectancy of the tract is a worth while analysis.

1.1 Data Overview

The gathered data consists of several binary indicator level variables which reflect varying levels of accessibility to food. We intend on analyzing the Low access tract 1 and 10 binary treatment variable. This variable can be described as follows: it takes the value 1 if a tract with at least 500 people, or 33% of the population, living more than 1 mile (urban areas) or 10 miles (rural areas) from the nearest supermarket,supercenter or large grocery store and 0 otherwise. We hope that this variable is able to successfully capture a tracts availability to food and therefore a potentially healthy diet.

In order to ensure that we can infer a causal effect from our observational study we include pre-treatment variables in our study. We assume that conditioning on our observed pre-treatment covariates, the selection bias is zero:

$$E[Y(0)|Z=1,X] = E[Y(0)|Z=0,X]$$

$$E[Y(1)|Z=1,X] = E[Y(1)|Z=0,X]$$

Further, many of our analysis' we impose a stronger assumption, known as strong ignorability (Rubin 1978),which states that our potential outcomes are independent of our treatment variable given our observed covariates:

$$(Y(1),Y(0)) \perp\!\!\!\perp Z|X$$

For the purposes of this project we have decided to include the following tract level pre-treatment variables in our analysis. The tract level pre-treatment variables are as follows:

- Poverty Rate
- Median Family Income
- Percentage of households receiving supplemental nutrition assistance program (SNAP) benefits
- Percentage of households with no vehicles
- Binary indicator if the tract is in an urban area or not
- Percentage of tract population that is Hispanic
- Percentage of tract population that is Black

Each of these covariates was chosen because they are most likely confounders, and they were available to us. As will be explained later, finding data at the census tract level is rather difficult. Poverty rate and by extension median family income, have been shown to have an association with life expectancy and it is not unreasonable to assume that lower access to food occurs in tracts with greater poverty level (or lower median family income) (Chetty et al., 2016). Adjusting for the percentage of households receiving SNAP benefits is also interesting since this is the percentage of households receiving aid to be able to purchase food. Since our analysis is on food deserts this is an important variable. The indicator of whether or not the tract was urban or not, is interesting since there is an association between life expectancy and living in an urban or rural environment, although the food desert definition accounts for urban or rural, it appears more likely that a food desert occurs in rural areas (Singh and Siahpush, 2014).

Finally we have included two variables on the percentage of Hispanic people in the tract as well as the percentage of Black people in the tract. According to the CDC there is a multiple year difference in life expectancy between Black, Hispanic and non-Hispanic White people (CDC).

Some of our data contains missing values, about 7% of the data set was simply omitted due to our lack of knowledge on data imputation methods, however this can be improved on and we do not believe that there is a systematic problem with the missing data so our analysis should still hold. Some of this data is missing because the life expectancy dataset we used from the CDC does not include Maine or Wisconsin as States. It is reasonable to assume that this exclusion should not have a large impact on the overall results.

Having discussed the set up of our problem, we will carry out the analysis in the following way. In Section 2 we will highlight the various methodologies we will use to estimate the average causal treatment effect. This will include the discussion of any necessary assumptions and justifications of the models. In Section 3 we will go over the results of the varying methodologies, Moving on we then discuss and cover the necessary sensitivity analysis and covariate balance checks in Section 4, where we analyze the assumption of strong ignorability imposed above. We discuss possible extensions of the project in Section 5. We will finish our analysis with a brief conclusion in Section 6.

2 Methods

All of the information on these methods is from the Causal Inference lecture notes provided by Professor Ding.

2.1 Regression outcome

As our first method we attempt a Regression outcome method. At the suggestions of Professor Ding, we use Lin's estimator and include interaction terms within our linear model. Our understanding is that if we omit the interaction between the treatment assignment and the pre-treatment covariates we may end up with a suboptimal estimator as it ignores the non-constant effect induced on the treatment by the pre-treatment covariates. Thus the set up of this method is as follows:

$$E(Y|Z,X) = \beta_0 + \beta_Z + \beta_X^T X + \beta_{ZX}^T XZ$$

Which implies that the average treatment effect estimator is as follows:

$$\tau = \beta_Z + \beta_{ZX}^T \bar{X}$$

As has been noted by professor Ding, if we have scaled our covariates X, this treatment effect reduces to the treatment coefficient β_Z. As stated previously, in order for this method to be valid we need to assume strong ignorability of our covariates as well as a linear outcome model. Since a closed form solution of the variance is not yet well defined, we will go about inference using the bootstrap method. We believe that making use of the outcome regression model point estimate and standard error can perform as a benchmark for the subsequent methodologies. Given the simplistic nature of this model specification, it may be of interest to see how more involved or complex models fare in terms of point estimates and standard error.

2.2 Propensity Score Weighting

Our next method of conducting observational based causal inference is via the propensity score. In particular we will use the Horowitz-Thompson and Hajek methods of inverse propensity weighting. In essence, these methods attempt to control for the effect that the pre-treatment variables have on receiving the treatment. The two estimators can be described in the following way:

$$\hat{\tau}^{HT} = \frac{1}{n}\sum_{i=1}^{n} \frac{Z_i Y_i}{\hat{e}(X_i)} - \frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)}$$

$$\hat{\tau}^{Hajek} = \frac{\sum_{i=1}^{n} \frac{Z_i Y_i}{\hat{e}(X_i)}}{\sum_{i=1}^{n} \frac{Z_i}{\hat{e}(X_i)}} - \frac{\sum_{i=1}^{n} \frac{(1-Z_i)Y_i}{1-\hat{e}(X_i)}}{\sum_{i=1}^{n} \frac{1-Z_i}{1-\hat{e}(X_i)}}$$

Were $e(X_i) = P(Z = 1|X)$, which is unknown and must be estimated from the data. We elected to fit our propensity score estimate using logistic regression, which imposes the assumption that:

$$P(Z = 1|X) = \frac{1}{1 + \exp(-\beta^T X)}$$

We include both methods of propensity score weighting in our analysis to potentially highlight the shortcomings of these methods. As evidenced in our results section, we see that the Horowitz-Thompson method behaves in an unstable manner and lends

evidence to previous theory that the Hajek IPW may be a more robust alternative. Additionally, as recommended by professor Ding as well as Crump et al. (2009) and Kurth et al. (2006), we conduct a truncated based analysis on the propensity score. The truncation scheme goes as following:

$$\hat{e}^{trunc}(X_i) = \max[\alpha_L, \min[\hat{e}(X_i), \alpha_U]]$$

Where, α_L, α_U are the lower and upper truncation bounds respectively.
By truncating the propensity score we are able to further address the sensitivity of the average treatment effect estimate to potentially extreme propensity scores. This would be a primary concern if our groups have litte overlap, meaning we receive propensity scores very close to 0 or 1. As we summarize in Section 3, this could very well be the case with this data, leading to otherwise unstable results. As a final precaution, we conduct a covariate balance check which we will go into further in Section 4.

In addition to the above propensity score based analysis, we will also attempt a similar analysis based on the concept of propensity score stratification. In essence this approach will allow us to discretize the propensity score into several bins, reducing the need for correct model specification. In this case only the ordering of the propensity scores matters not the actual values. This analysis will also be useful in potentially improving any covariate imbalance seen in the standard, non-stratified, IPW case. The exact stratification process goes as follows, we can discretize the estimated propensity score by its K quantiles to obtain $\hat{e}^{strat}(X) : \hat{e}^{strat}(Xi) = e_k$, the k^{th} quantile of $\hat{e}^{strat}(X)$.

2.3 Doubly Robust Estimator

The doubly robust estimator gets its name from the fact that it is robust to either a misspecified propensity score model or a misspecified model for the conditional means of the outcome. The motivation for this method is that we are essentially combining both regression adjustment and IPW methods to correct for possible bias introduced by either, hence we have double robustness. The conditional models for the outcome are indexed by some parameter β as below:

$$\mu_1(X, \beta_1) = E\{Y|Z = 1, X\}$$
$$\mu_0(X, \beta_0) = E\{Y|Z = 0, X\}$$

and they are indexed with some parameters (β_0, β_1). After the estimation of both these and the propensity score, the former of which we estimate with linear regression and the latter with logistic regression we obtain the doubly robust estimator with

$$\hat{\mu}_1^{dr} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Z_i Y_i}{e(X_i, \hat{\alpha})} - \frac{Z_i - e(X_i, \hat{\alpha})}{e(X_i, \hat{\alpha})}\mu_1(X_i, \hat{\beta}_1)\right]$$
$$\hat{\mu}_0^{dr} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{(1 - Z_i)Y_i}{1 - e(X_i, \hat{\alpha})} - \frac{e(X_i, \hat{\alpha}) - Z_i}{e(X_i, \hat{\alpha})}\mu_0(X_i, \hat{\beta}_0)\right]$$

And we get that $\hat{\tau}^{dr} = \hat{\mu}_1^{dr} - \hat{\mu}_0^{dr}$. The doubly robust estimator has the property of essentially forcing covariate balance.

2.4 Regression Trees

For our final method we will attempt to use a more modern approach of finding regression outcomes and we reference (Hill, 2011). This approach essentially relies upon a tree based model to predict causal outcomes. The benefit of this model is that it allows for non-linearity's and interaction terms without needing direct specification. Even though this model is non-parametric we still need ignorability to hold in order for us to interpret the results as causal, so it is still limited in regards to unobserved confounders.

As Hill notes this method can handle very large numbers of predictors. If a variable is not useful it simply does not get used, reflecting the signal processing nature of regression trees. Therefore we can include a greater numbers of potential covariates then other methods that suffer from issues in high dimensions. The ability to include many potential confounding covariates as predictors can be quite helpful when trying to satisfy ignorability. The set up for the average treatment effect using regression trees is as follows:

$$\frac{1}{n}\sum_{i=1}^{n} E(Y_i(1)|X_i) - E(Y_i(0)|X_i) = \frac{1}{n}\sum_{i=1}^{n} f(1, X_i) - f(0, X_i)$$

Where $f(z, x_i)$ is the output of a sum of regression trees assembled by the model. Inline with Hill we will be using Bayesian Regression trees. This method resembles that of gradient boosting, however it focuses a prior to avoid over fitting and uses a technique called Bayesian back-fitting. The prior accomplishes three things; 1) it has a preference for trees with only a few bottom nodes, 2) shrink the mean response on the bottom nodes of the tree to zero and 3) a prior which suggests a variance that is smaller than that of least squares. Of course if the data suggests, the trees can be grown larger however this regularizing effect allows for a more robust answer. For more information on this method and construction of the prior please see Hill (2011) and Chipman et al. (2007).

3 Results

3.1 Regression outcome

As stated in Section 2, our primary method of analyzing the causal effect of geographic access to food on average life expectancy is the regression outcome model. On the recommendation of Professor Ding, we have decided to use Lin's method and include interaction terms with the treatment and covariates. Using this method, we observed the below results:

Outcome Regression Analysis	
Est	Boot strap SE
0.053	0.03

This essentially tells us when conditioning on our covariates the causal effect of access to food on the average life expectancy is close to zero. If we construct a confidence interval we will have zero in the coverage range so it will be difficult for us to infer any causal effect based on this outcome model. It may be interesting to use this method of measuring casual effect as a benchmark to more advanced methods.

3.2 Propensity Score Weighting

Moving forward we conduct causal analysis via the inverse propensity score weighting method. For this initial method we use the logistic regression to fit our propensity scores, the histogram of our scores is as follows:

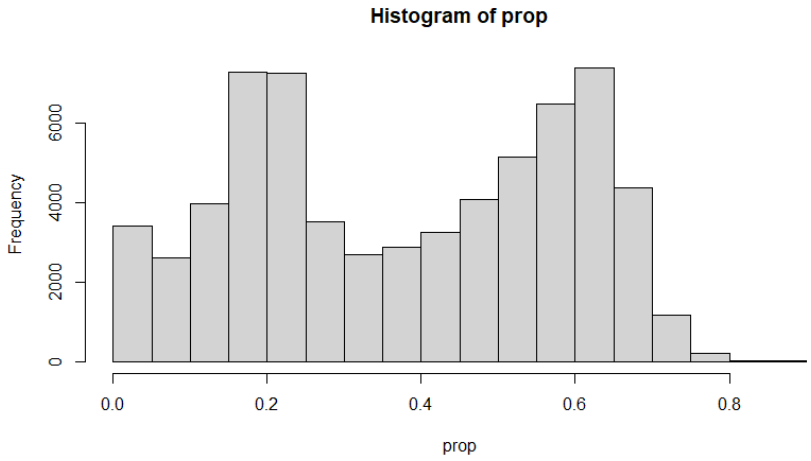

Figure 1: Propensity Scores

We can see that our propensity scores are bi-modal, which is to be expected given this reflects a binary treatment. However, what is interesting is that our propensity scores are skewed towards values closer to 0. The maximum propensity score is $\approx .8$, where as there are many propensity scores very close to 0. This means that we are never overly certain that a tract will receive the treatment, however there are some tracts where we are very confident that they will not receive the treatment. Therefore our logistic regression is much more confident in classifying as a non food desert, rather than classifying as a food desert. Intuitively this makes more sense since certain factors such as low poverty rate or high median income should strongly indicate a non food desert. These small propensity scores can lead to unstable treatment effect estimation and therefore indicate need for truncation to stabilize the results.

As described in Section 2, we conducted a IPW based method using both the Hajek and the Horowitz-Thompson methods. The resulting point estimates and boot strap standard errors can be seen below:

Inverse Propensity Score Weighting		
Method	Est	Boot strap SE
HT	52.7	25.3
Hajek	0.045	0.74

This results confirm some of the concerns we have with IPW based estimators. As expected we see that the Horowitz-Thompson estimator is very unstable and sensitive to extreme propensity score values. In the context of our question, it would not make sense to achieve an average causal effect of 52.7 as this would reflect an increase in average life expectancy of 52.7 years, a non-sensical answer. The Hajek method acts as an improvement on this model and can be seen to provide more robust results. Nonetheless, there is still the possibility of extreme propensity score values effecting the estimator's stability.

As previously stated, to account for these concerns we carried out a truncation based analysis as well. It should be noted in our case that truncation is only important for the low propensity score values, since our max propensity score is $\approx .8$, as such we will focus on the lower bound truncation. Below find our results from re-running the analysis on various truncation scores:

Inverse Propensity Score Weighting				
Truncation Level	HT Est	HT Bootstrap SE	Hajek Est	Hajek Bootstrap SE
.01	4.262	0.513	-0.173	0.051
.05	0.490	0.232	-0.058	0.03
.1	-2.521	0.179	0.032	0.029

We can see that the truncated estimators are much more stable than the untruncated estimators. In particular we see the largest effect on the HT estimator, further underscoring the potential unreliability of this estimator. Given that our data contains nearly deterministic control assignment probability for some counties we most likely need to make this correction to maintain the validity of our results.

The final method we considered with propensity scores, is propensity score stratification. As noted before, this reduces the need for correct model specification and could allow us to make more robust estimations. Below find the results from several different stratification's:

Number of strata					
	5	10	20	50	80
est	0.001	-0.042	-0.068	-0.073	0.064
se	0.038	0.039	0.042	-0.050	0.050

We can see above that using the stratification method we produce relatively similar results across the varying number of strata. Furthermore, we can see the point estimates and standard errors are relatively similar to the truncated HT and Hajek methods. This means that both truncation and stratification make the propensity scores more robust and reduce the impact of model mis-specification. However, this all means that all confidence intervals intersect zero and we cannot determine a causal treatment effect from this analysis. One benefit of the stratification method versus the other propensity score methods is that is does a better job at balancing our covariates, we will elaborate on this in Section 4.

3.3 Doubly Robust Estimator

We use the doubly robust estimator with our outcome model being fitted with a Gaussian response, and with our propensity score being fit with logistic regression. As described above it is possible to use other forms of estimation for either the outcome model or the propensity score model. This is explored in the next section on regression trees. Using the doubly robust method we get the following average treatment results and standard errors:

Doubly Robust Estimator		
Truncation Level	DR Est	DR Bootstrap SE
0	2.771	1.723
.01	0.091	0.046
.05	0.046	0.031
.1	0.038	0.029

Interestingly enough, using the doubly robust method we get a larger, positive average treatment effect then what we see in the other methods as well as a larger

standard error. This inconsistency could be due to the possibility of either miss-specification of the outcome model or the propensity model. Phenomena such as this are responsible for the nickname of this estimator being the doubly sensitive estimator. We also see some sensitivity when we do not truncate the propensity score, as the non truncated propensity score has a many magnitudes larger causal effect and SE compared to the even slightly truncated estimator. Therefore although the DR is not as sensitive as the HT estimator, it is still sensitive to near deterministic propensity scores.

3.4 Regression Trees

Looking at our supplementary analysis using regression trees, we used the bartc function in the bartCause package in R. This function will run 10 separate chains, with 500 posterior samples, and will then average them for inference. We can compare the following results to that of the linear regression outcome model:

Regression Analysis		
Method	Est	Boot strap SE
Linear Regression	-0.005	0.113
Tree Regression	-0.063	0.039

We can see from the above results that the tree based analysis provides roughly the same point estimate as our linear regression but reduces our standard error significantly. While the confidence interval for our tree method still intersects 0, it is a benefit to see that we may be experiencing regularizing effects of the Bayesian nature of this tree method. The prior allows us to regularize the results and therefore potentially reduce the SE of the point estimate. Additionally, this may also lead us to believe that the our regression model is in-fact non-linear and that the tree does a better job at capturing these non-linearity's.

4 Covariate Balance and Sensitivity analysis

Since our analysis relies on the assumption of strong ignorability, or no unmeasured confounders we would like to measure the validity of this assumption, by testing the magnitude of an unmeasured confounder needed to explain away our causal effect. However we have seen no significant causal effect in any of the methods used above in the process. A potential issue that we have observed in a majority of our methods is covariate imbalance. This of course draws into question the risk to biased treatment effect on our obtained estimates. As we can see in figure 2, we were able to get better balance when we perform propensity stratification which suggests that we may have a miss-specified propensity score model. Nonetheless, even when performing the stratification method the balance check leaves much to be desired. On a final note, we can see that as expected the doubly robust estimator forced covariate balance, a beneficial result of the this model.

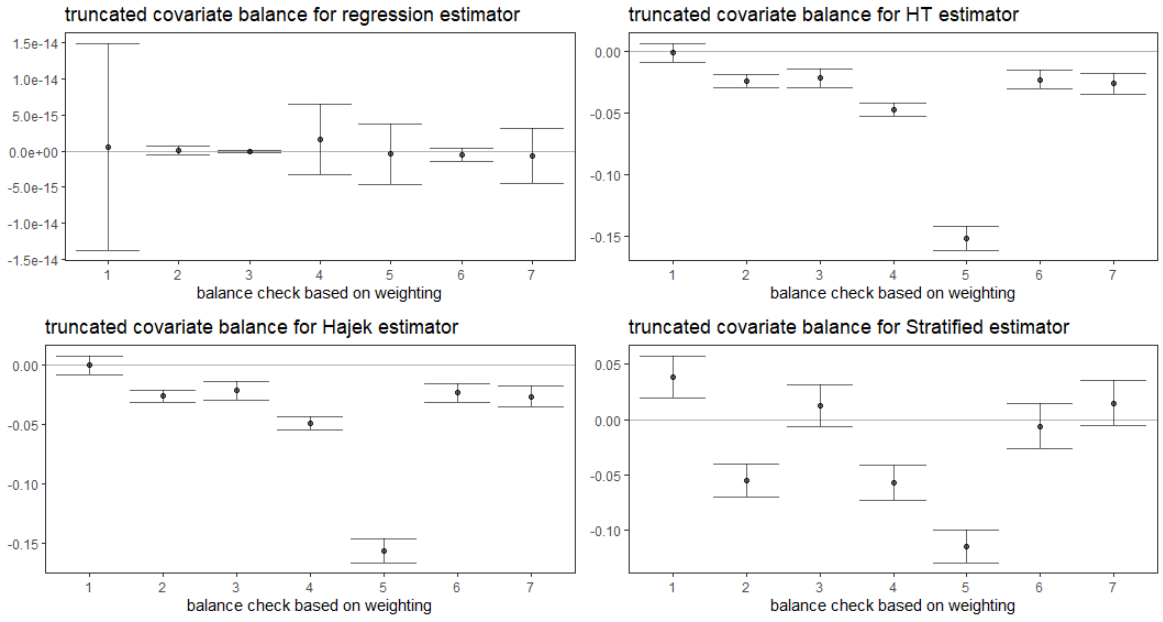

Figure 2: Regression and Propensity Balance Check

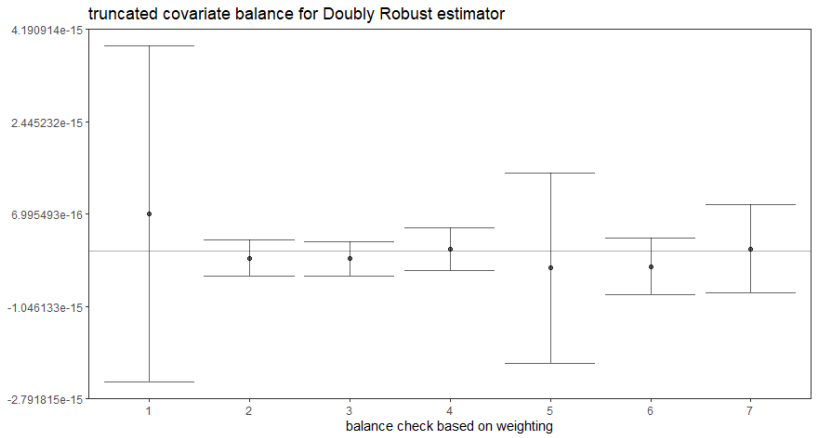

Figure 3: Doubly Robust Balance Check

An interesting note here is the strong influence of race on the causal effect, and it is likely that this is a rather strong confounder. In fact, if we omit race from the model we observe significant causal effect estimates that are negative. However race is a strong enough confounder to reduce our causal effect to essentially 0. There are previous studies that show that race is associated with living in a food desert, and as we described above race is also associated with different life expectancy's (Baker et al., 2006). We ran the estimators without the race variable included and noticed that we estimated a negative causal effect that was significant (indicating that living in a food deset decreased life expectancy) however when including the two races we included we noticed that this causal effect disappears and almost all of our CI's overlap 0.

5 Possible Extensions

One of the hardest parts of this project is its very large scale, in this we have that every census tract is a data point. Therefore our covariates are heavily limited because we need covariates for every tract. Most data is only available at the county level rather than the tract level. Counties contain many more people than census tracts so using them may not be as informative since we are aggregating very different groups, typical counties may contain many cities within them for instance. Census tracts however are much more focused but also much more difficult to obtain data to help analysis. Examples of a covariate that may have been useful for the analysis may be average education level (for instance percentages of those with certain degrees).

Besides data there are many other avenues for estimation here, this paper is mostly

focused on exploring multiple avenues of estimating causal effects. For instance the doubly robust estimator depends on the estimation of some model for the propensity score as well as some model for the conditional means. We restricted ourselves to regression with Gaussian errors, as well as logistic regression. A possible extension may be analyzing with other methods for estimating either of these quantities. This may include non linear estimators or Bayesian methods. Additionally, as noted in the previous sections, we may have suffered from a mis-specified propensity score model. To resolve these concerns we can attempt a observational matching study where we are able to avoid any parametric assumptions and can match on the Mahalanobis distance of the covariates. Furthermore, in our study we have roughly 1:2 ratio of treated to untreated potentially allowing for close matches.

6 Conclusion

Our results, which applied a large variety of estimation techniques for an observational study, almost universally point to little to no causal effect of living in a food desert on life expectancy at the census tract level. This is in contrast to a reasonable hypothesis that would be that living in a food desert would have a negative effect on life expectancy. We do not wish to make claims of certainty since certain aspects of our work lead to problems. For instance, most of our methods do not seem to have an ideal covariate balance, this indicates that it is possible that some of the (non)effect is being driven by covariates rather than the actual treatment. The problem of confounding is always a large assumption that needs to be assessed and there is possibility for unmeasured confounders due to the scope of our study and the overall lack of data of this magnitude. However one thing we can note is that many theoretical properties did seem to pan out in an applied context as they should. In addition we did not receive wildly varying results from the methods, they nearly universally point to limited to no causal effect. If nothing at least this consistency is present and although we leave the door open for further more complicated methods, most of our current methods point to the same result.

7 Code

The code containing all of the analysis performed as well as the data is in this linked GitHub repository. All analysis can be replicated there. Alternatively can access at `https://github.com/guhauhsoj/causal`.

References

Changes in life expectancy by race and hispanic origin in the united states, 2013–2014. `https://www.cdc.gov/nchs/products/databriefs/db244.htm`. Accessed: 2020-12-13.

E. A. Baker, M. Schootman, E. Barnidge, and C. Kelly. The role of race and poverty in access to foods that enable individuals to adhere to dietary guidelines. *Preventing Chronic Disease*, 3(3), Jun 2006. ISSN 1545-1151. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1636719/`.

CDC. U.s. life expectancy at birth by state and census tract - 2010-2015. `https://healthdata.gov/dataset/us-life-expectancy-birth-state-and-census-tract-2010-2015`, 2019. Accessed: 2020-12-13.

R. Chetty, M. Stepner, S. Abraham, S. Lin, B. Scuderi, N. Turner, A. Bergeron, and D. Cutler. The association between income and life expectancy in the united states, 2001–2014. *JAMA*, 315(16):1750–1766, Apr 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.4226.

H. Chipman, E. George, and R. McCulloch. *Bayesian Ensemble Learning*. The MIT Press, 2007. ISBN 9780262256919. doi: 10.7551/mitpress/7503.003.0038. URL `https://direct.mit.edu/books/book/3168/chapter/87401/bayesian-ensemble-learning`.

R. K. Crump, V. J. HOTZ, G. W. IMBENS, and O. A. MITNIK. Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, 96(1):187–199, 2009. ISSN 0006-3444.

dosomething.org. 11 facts about food deserts. https://www.dosomething.org/us/facts/11-facts-about-food-deserts. Accessed: 2020-12-13.

J. L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162.

T. Kurth, A. M. Walker, R. J. Glynn, K. A. Chan, J. M. Gaziano, K. Berger, and J. M. Robins. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *American Journal of Epidemiology*, 163(3):262–270, Feb 2006. ISSN 0002-9262. doi: 10.1093/aje/kwj047.

G. K. Singh and M. Siahpush. Widening rural-urban disparities in life expectancy, u.s., 1969-2009. *American Journal of Preventive Medicine*, 46(2):e19–29, Feb 2014. ISSN 1873-2607. doi: 10.1016/j.amepre.2013.10.017.

USDA. Usda ers - download the data. https://www.ers.usda.gov/data-products/food-access-research-atlas/download-the-data/, 2017. Accessed: 2020-12-13.