Final Project: Walkability and Public Health in the US

Lou Godmer, Kienan Battin, Divakar Mehta

April 17, 2023

Contents

Objective

The objective is to quantify the causal effect that the "walkability" of a region has on the public health indicators: obesity and physical activity. The original data comes from two sources: 1. The U.S. Chronic Disease Indicators provides reported cases of a set of 124 indicators that are important to public health, and the geographic location of the case. 2. The Walkability Index quantifies every Census 2019 block group's relative "Walkability" as defined by the EPA based on characteristics such as easy walking access to public transit, jobs, stores and services. Quantifying the causal effect of walkability on public health can help policy makers understand how community planning measures that may improve or degrade the walkability of the region will impact public health.

The appendix of this document describes the pre-processing methodology that was used combine the two data sets to enable the quantitative analysis. Because the pre-processing methodology can take an hour or more to execute, we exported the pre-processed data and made it available for download in a publicly accessible location. The beginning of this document imports the pre-processed data and the rest of the analysis is done based on the pre-processed data.

Load necessariy libararies

```
rm(list=ls())

options(repos = list(CRAN="http://cran.rstudio.com/"))

if (!require('NHANES')) install.packages('NHANES')
library('openxlsx')

if (!require('ggplot2')) install.packages('ggplot2')
library('ggplot2')

if (!require('dplyr')) install.packages('dplyr')
library('dplyr')

if (!require('GGally')) install.packages('GGally')
library('GGally')

if (!require('tableone')) install.packages('tableone')
library(tableone)

if (!require('pROC')) install.packages('pROC')
library(pROC)

if (!require('tidyr')) install.packages('tidyr')
library(tidyr)

if (!require('tidycensus')) install.packages('tidycensus')
library(tidycensus)

if (!require('tigris')) install.packages('tigris')
library(tigris)

if (!require('sf')) install.packages('sf')
library(sf)
```

```r
if (!require('stringr')) install.packages('stringr')
library(stringr)

if (!require('dplyr')) install.packages('dplyr')
library(dplyr)
```

Load the data

Download the data which has already undergone the pre-processing methodology described in the appendix. WARNING: this may take several minutes. To avoid unnecessary downloads, the commands are commented out. Un-comment and execute the commands to download the data.

```r
#download.file("https://walkabilityandhealth.blob.core.windows.net/walkabilityandhealth/disease_with_wa
#unzip("disease_with_walkability.zip", "disease_with_walkability.csv")
```

```r
disease_with_walkability <- read.csv("disease_with_walkability.csv")
```

Shape the data set for our analysis purposes

The walkability data has observations with information about a BlockGroup. BlockGroup is a geographic unit used by the US Census Bureau and is a subdivision of a census tract, which in turn is a subdivision of a county. Block groups usually contain between 600 and 3000 people and are the smallest standardized unit of geography for which the Walkability score is measured. All of our analysis will be based on block group as the observation unit.

The disease indicator data has a column for "Question" which contains the details of the disease indicator that was measured. Of the 203 total questions we have narrowed down to 12 that will be used in our analysis. Questions were selected which have the same unit of measure - Crude Prevalence - which means that the value is the percentage of the overall population. For simplicity we have also selected the more general questions rather than inter-sectional questions. For example we chose "Arthritis among adults aged >= 18 years" rather than "Arthritis among adults aged >= 18 years who have heart disease". The result of the selection process, along with how we intend to use each variable in our analysis, is in the table below:

Topic	Question	QuestionID	Intended Usage
Nutrition, Physical Activity and Weight Status	Overweight or obesity among adults aged >= 18 years	NPAW2_1	Dependent Var
Alcohol	Heavy drinking among adults aged >= 18 years	ALC5_1	Independent Var
Arthritis	Arthritis among adults aged >= 18 years	ART1_1	Independent var
Asthma	Current asthma prevalence among adults aged >= 18 years	AST1_1	Independent var
Chronic Obstructive Pulmonary Disease	Prevalence of chronic obstructive pulmonary disease among adults >= 18	COPD2_0	Independent var
Oral Health	Visits to dentist or dental clinic among adults aged >= 18 years	ORH1_1	Independent var

Topic	Question	QuestionID	Intended Usage
Overarching Conditions	High school completion among adults aged 18-24 years	OVC2_1	Independent var
Overarching Conditions	Current lack of health insurance among adults aged 18-64 years	OVC1_1	Independent var
Tobacco	Current smoking among adults aged >= 18 years	TOB1_2	Independent var
Tobacco	Current smokeless tobacco use among adults aged >= 18 years	TOB2_2	Independent var
Chronic Kidney Disease	Prevalence of chronic kidney disease among adults aged >= 18 years	CKD3_0	Independent var

```
# filter down to just the 12 selected questions
disease_with_walkability_filtered <- filter(disease_with_walkability, DataValueType == "Crude Prevalenc
  StratificationCategory1 == "Overall" & !is.na(NatWalkInd))
filtered_qids <- c("NPAW2_1", "ALC5_1", "ART1_1", "AST1_1", "COPD2_0", "ORH1_1", "OVC2_1", "OVC1_1", "T
disease_with_walkability_filtered <- disease_with_walkability_filtered[disease_with_walkability_filtere

# reshape the data by grouping by all the unique properties per block group, and expanding
# columns with the result of each of the relevant 12 questions for that observation.
# for now we will also keep some descriptive variables (like LocationAbbr)
# which may come in handy for visual data exploration
collapsed_cols = c("YearStart", "LocationAbbr", "LocationDesc", "STATEFP", "COUNTYFP", "TRACTCE", "BLKG

disease_with_walkability_collapsed <- disease_with_walkability_filtered %>%
  pivot_wider(id_cols = collapsed_cols, names_from = c("QuestionID"), values_from = DataValueAlt)
```

```
## Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
## i Please use `all_of()` or `any_of()` instead.
##   # Was:
##   data %>% select(collapsed_cols)
##
##   # Now:
##   data %>% select(all_of(collapsed_cols))
##
## See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```

```
# rename the question columns to something easier to read
disease_with_walkability_collapsed <- disease_with_walkability_collapsed %>%
  rename("Overweight" = "NPAW2_1") %>%
  rename("Alcohol" = "ALC5_1") %>%
  rename("Arthritis" = "ART1_1") %>%
  rename("Asthma" = "AST1_1") %>%
  rename("COPD" = "COPD2_0") %>%
  rename("Dentist" = "ORH1_1") %>%
```

```
  rename("HighSchool" = "OVC2_1") %>%
  rename("NoInsurance" = "OVC1_1") %>%
  rename("Smoking" = "TOB1_2") %>%
  rename("Vapeing" = "TOB2_2") %>%
  rename("KidneyDisease" = "CKD3_0")
```

Next we will convert our treatment variable, NatWalkInd to binary. Our strategy will be to select the bottom 40% least walkable block groups and consider these "not walkable" and the top 40% most walkable block groups will be considered "walkable." The middle 20% will not be used

```
nwi40 <- quantile(disease_with_walkability_collapsed$NatWalkInd, .40)
nwi60 <- quantile(disease_with_walkability_collapsed$NatWalkInd, .60)
disease_with_walkability_collapsed = filter(disease_with_walkability_collapsed, NatWalkInd < nwi40 | Na
disease_with_walkability_collapsed$Walkable <- ifelse(disease_with_walkability_collapsed$NatWalkInd >= 
# rename the dataframe to a shorter name
dww <- disease_with_walkability_collapsed
```

Understand the variables in the dataset

The table below describes the variables that are used in this analysis, including the variables that contain the values of the questions we selected. The "Usage In This Analysis" column categorizes how these will be used in analysis.

Variable Name	Variable Description	Usage In This Analysis
Walkable	Binary variable, 1 if the block is walkable, 0 otherwise	Treatment Variable
Overweight	Percentage of the population overweight or obese	Dependent Variable
R_PCTLOWWAGE	Percentage of the population that makes less than $1250/month	Independent Variable
HighSchool	Percentage of the population ages 18-24 who have completed high school	Independent Variable
Alcohol	Percentage of the population with high alcohol use	Independent Variable
Arthritis	Percentage of the population with arthritis	Independent Variable
Asthma	Percentage of the population with asthma	Independent Variable
KidneyDisease	Percentage of the population with chronic kidney disease	Independent Variable
Vapeing	Percentage of the population using smokeless tobacco	Independent Variable
Smoking	Percentage of the population that smokes	Independent Variable
COPD	Percentage of the population with Chronic Obstructive Pulmonary Disease	Independent Variable
NoInsurance	Percentage of the population without health care coverage	Independent Variable
Dentist	Percentage of the population with dentist visits	Independent Variable

Get familiar with the data using summary statistics

The first thing we will note is that what started as a very large data set with almost a million observations has diminished down to a relatively small sample size (318 observations) with all of the necessary filtering and aggregation that was done. We will need to do the best we can with it. Due to the small sample size, propensity score matching may not be a viable option, and generally we need to take care with the conclusions we draw from the analysis.

```
str(dww)
```

```
## tibble [318 x 22] (S3: tbl_df/tbl/data.frame)
##  $ YearStart    : int [1:318] 2012 2013 2010 2014 2012 2011 2017 2015 2014 2011 ...
##  $ LocationAbbr : chr [1:318] "MI" "WI" "NJ" "KS" ...
##  $ LocationDesc : chr [1:318] "Michigan" "Wisconsin" "New Jersey" "Kansas" ...
##  $ STATEFP      : int [1:318] 26 55 34 20 41 5 34 44 66 9 ...
##  $ COUNTYFP     : int [1:318] 39 141 29 159 69 119 29 3 10 7 ...
##  $ TRACTCE      : int [1:318] 960200 11000 717101 967200 960100 4400 717101 20300 952900 541600 ...
##  $ BLKGRPCE     : int [1:318] 2 5 1 1 2 1 1 3 6 1 ...
##  $ GEOID        : num [1:318] 2.60e+11 5.51e+11 3.40e+11 2.02e+11 4.11e+11 ...
##  $ R_PCTLOWWAGE : num [1:318] 0.258 0.253 0.22 0.337 0.384 ...
##  $ NatWalkInd   : num [1:318] 8.17 11.33 6.5 12.33 5.5 ...
##  $ HighSchool   : num [1:318] 86.1 87.5 86.7 87.3 87.4 84.5 89.2 90.2 NA 85.6 ...
##  $ Alcohol      : num [1:318] 6.1 NA NA 5.1 NA 6.1 NA NA 8.7 6.6 ...
##  $ Arthritis    : num [1:318] 31.8 NA NA 25.4 NA 28.7 NA NA 15.7 22.5 ...
##  $ Asthma       : num [1:318] 10.5 NA NA 8.7 NA 9.5 NA NA 5.6 9.9 ...
##  $ KidneyDisease: num [1:318] 3.4 NA NA 2.6 NA 3 NA NA 4.2 1.9 ...
##  $ Overweight   : num [1:318] 65.6 NA NA 66 NA 64.9 NA NA 63.4 59.6 ...
##  $ Vapeing      : num [1:318] 3.9 NA NA 5.7 NA 7.1 NA NA 6.6 1.5 ...
##  $ Smoking      : num [1:318] 23.3 NA NA 18.1 NA 27 NA NA 29.2 17.1 ...
##  $ COPD         : num [1:318] 7.4 NA NA 6.6 NA 8 NA NA 3.2 6.1 ...
##  $ NoInsurance  : num [1:318] 16.5 NA NA 17.5 NA 29.1 NA NA 26.6 14.7 ...
##  $ Dentist      : num [1:318] 68 NA NA 67.3 NA NA NA NA 54 NA ...
##  $ Walkable     : num [1:318] 1 0 1 0 1 0 1 0 1 0 ...
```

Another observation from the summary statistics is that there are a lot of NA values for the disease indicators (such as Alcohol, Arthritis, Asthma, etc.). This will also impact the confidence of any estimated causality conclusions drawn from this data.

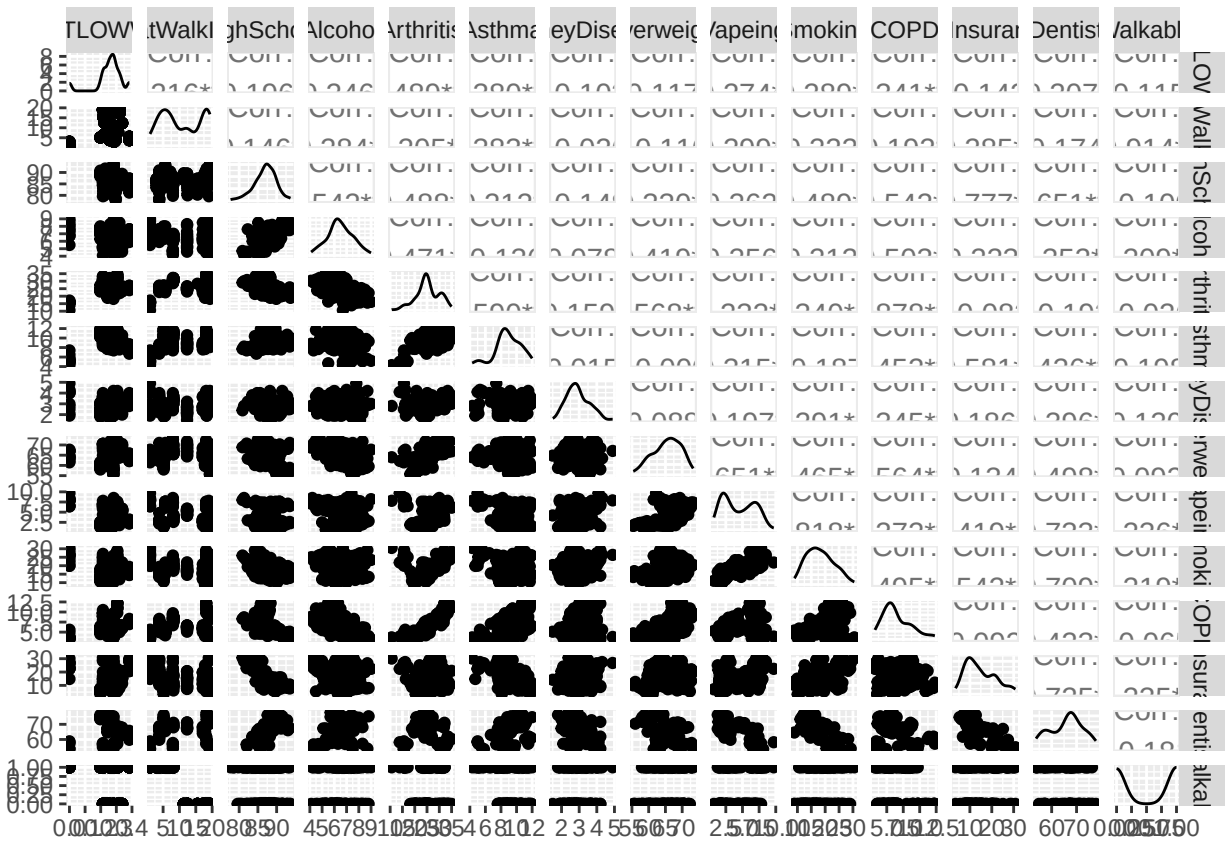
```
summary(dww)
```

```
##    YearStart    LocationAbbr       LocationDesc          STATEFP
##  Min.   :2010   Length:318         Length:318         Min.   : 5.00
##  1st Qu.:2012   Class :character   Class :character   1st Qu.:19.00
##  Median :2015   Mode  :character   Mode  :character   Median :34.00
##  Mean   :2015                                         Mean   :32.86
##  3rd Qu.:2018                                         3rd Qu.:44.75
##  Max.   :2021                                         Max.   :78.00
##
##     COUNTYFP         TRACTCE          BLKGRPCE         GEOID
##  Min.   :  3.00   Min.   :  1303   Min.   :1.000   Min.   :5.119e+10
##  1st Qu.: 27.00   1st Qu.: 20300   1st Qu.:1.000   1st Qu.:1.908e+11
##  Median : 53.00   Median :759000   Median :2.000   Median :3.403e+11
##  Mean   : 57.69   Mean   :557102   Mean   :2.069   Mean   :3.292e+11
##  3rd Qu.: 83.00   3rd Qu.:960200   3rd Qu.:3.000   3rd Qu.:4.481e+11
##  Max.   :159.00   Max.   :971700   Max.   :6.000   Max.   :7.803e+11
##
##   R_PCTLOWWAGE      NatWalkInd       HighSchool       Alcohol
##  Min.   :0.0000   Min.   : 1.000   Min.   :77.70   Min.   :4.10
##  1st Qu.:0.2287   1st Qu.: 5.500   1st Qu.:85.33   1st Qu.:5.90
##  Median :0.2700   Median : 8.167   Median :87.20   Median :6.40
```

```
##  Mean   :0.2582   Mean   :10.747   Mean   :86.95   Mean   :6.48
##  3rd Qu.:0.2958   3rd Qu.:17.833   3rd Qu.:88.97   3rd Qu.:7.20
##  Max.   :0.3835   Max.   :19.333   Max.   :93.70   Max.   :9.00
##                                    NA's   :48      NA's   :161
##    Arthritis         Asthma       KidneyDisease     Overweight
##  Min.   :10.60   Min.   : 4.300   Min.   :1.500   Min.   :55.40
##  1st Qu.:23.10   1st Qu.: 8.100   1st Qu.:2.500   1st Qu.:62.40
##  Median :24.80   Median : 8.900   Median :2.850   Median :65.80
##  Mean   :25.26   Mean   : 8.995   Mean   :2.908   Mean   :65.45
##  3rd Qu.:29.18   3rd Qu.:10.150   3rd Qu.:3.300   3rd Qu.:68.70
##  Max.   :35.00   Max.   :12.000   Max.   :4.900   Max.   :73.30
##  NA's   :160     NA's   :163      NA's   :164     NA's   :163
##     Vapeing         Smoking          COPD         NoInsurance
##  Min.   :0.700   Min.   : 8.90   Min.   : 3.200   Min.   : 3.40
##  1st Qu.:2.300   1st Qu.:14.50   1st Qu.: 5.000   1st Qu.: 9.45
##  Median :4.200   Median :17.70   Median : 5.900   Median :12.90
##  Mean   :4.417   Mean   :18.15   Mean   : 6.516   Mean   :14.73
##  3rd Qu.:6.600   3rd Qu.:21.30   3rd Qu.: 7.850   3rd Qu.:20.20
##  Max.   :9.700   Max.   :30.50   Max.   :12.300   Max.   :30.80
##  NA's   :165     NA's   :163     NA's   :159      NA's   :163
##     Dentist         Walkable
##  Min.   :53.70   Min.   :0.0000
##  1st Qu.:61.05   1st Qu.:0.0000
##  Median :67.25   Median :1.0000
##  Mean   :65.92   Mean   :0.5094
##  3rd Qu.:70.33   3rd Qu.:1.0000
##  Max.   :77.80   Max.   :1.0000
##  NA's   :244
```

Here we take a look at the correlation between the variables. The rendering in markdown is a bit small to see so you can also un-comment the "ggsave" command to export a larger rendering. Key observations: 1. Walkable (the treatment variable) is correlated with higher levels of alcohol use, smoking, vapeing, and noinsurance. 2. Overweight (the dependent variable) is correlated with lower likeiyhood of high school graduation, lower levels of regular dental visits, and higher levels of arthritis, COPD, smoking and vapeing.

```
dww_numeric = dww %>% select(-YearStart, -LocationAbbr, -LocationDesc, -STATEFP, -COUNTYFP,
       -TRACTCE, -BLKGRPCE, -GEOID)
#pairs(dww_numeric)
dww_ggpairs = GGally::ggpairs(dww_numeric)
#ggsave("dww_ggpairs.png", plot=dww_ggpairs, width=20, height=20)
dww_ggpairs
```



Fit an appropriate forecasting model to predict Overweight

Regressing Overweight on Walkable indicates that Walkable predicts about a 0.76% increase in the population that is Overweight. However the Walkable coefficient is not statistically significant and the Adjusted R-squared is very low, which indicates that a lot of variance in Overweight is not accurately predicted by this model.

```
lm_fit <- lm(Overweight ~ Walkable, dww)
summary(lm_fit)
```

```
##
## Call:
## lm(formula = Overweight ~ Walkable, data = dww)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7205 -2.8705  0.4795  3.2795  7.4224
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  65.1205     0.4380 148.682   <2e-16 ***
## Walkable      0.7572     0.6662   1.137    0.257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.109 on 153 degrees of freedom
##   (163 observations deleted due to missingness)
```

```
## Multiple R-squared:  0.008373,   Adjusted R-squared:  0.001891
## F-statistic: 1.292 on 1 and 153 DF,  p-value: 0.2575
```

Adding in our other variables reduces the coefficient of Walkable by about 10x and it is still not statistically significant. Arthritis, Asthma, KidneyDisease, Vapeing and Smoking are much better predictors of the precentage of the population that is overweight. This model has a much higher adjusted R-squared, so it is much better at predicting the variance in Overweight

```
lm_fit <- lm(Overweight ~ R_PCTLOWWAGE+HighSchool+Alcohol+Arthritis+Asthma+KidneyDisease+Vapeing+Smokin
summary(lm_fit)
```

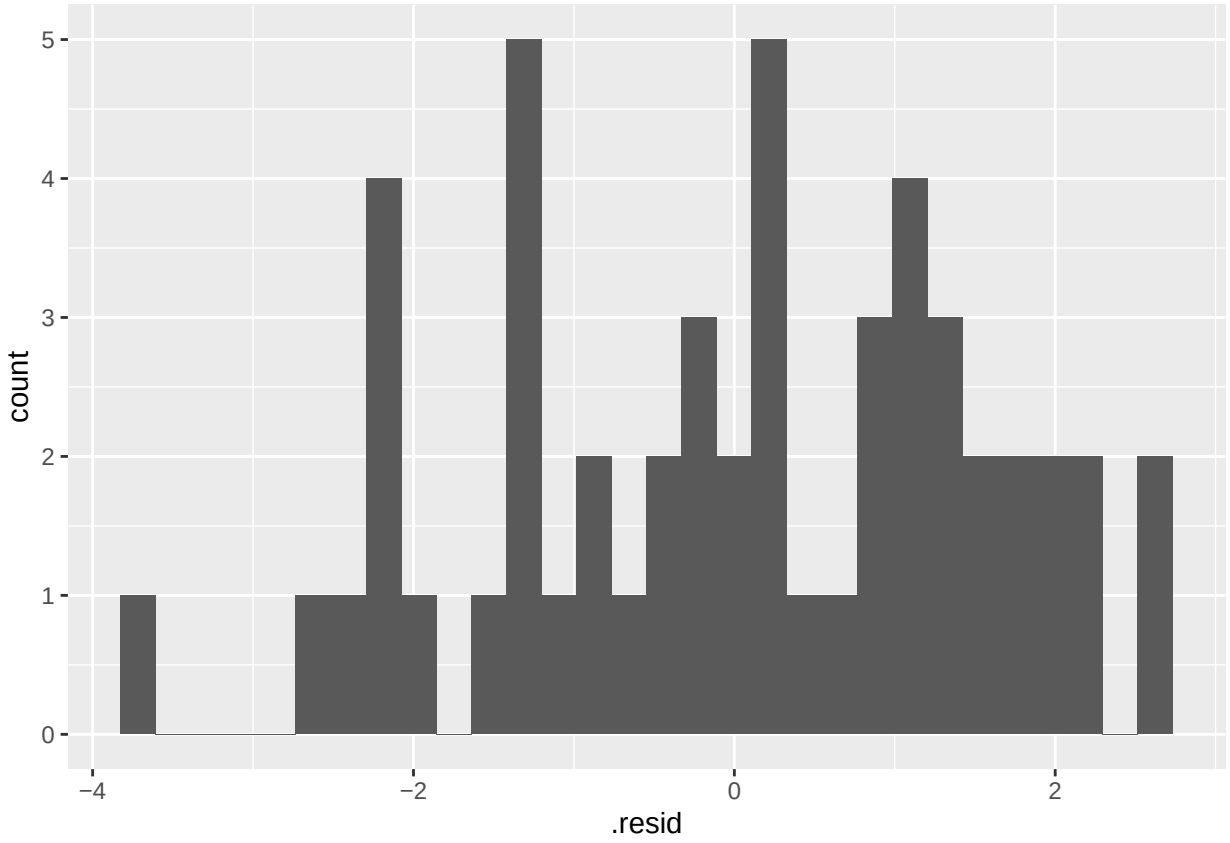
```
##
## Call:
## lm(formula = Overweight ~ R_PCTLOWWAGE + HighSchool + Alcohol +
##     Arthritis + Asthma + KidneyDisease + Vapeing + Smoking +
##     COPD + NoInsurance + Dentist + Walkable, data = dww)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7483 -1.3127  0.1808  1.1772  2.5920
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   35.922048  20.453922   1.756 0.086898 .
## R_PCTLOWWAGE   9.274197   7.268772   1.276 0.209539
## HighSchool     0.336075   0.219580   1.531 0.133957
## Alcohol       -0.292404   0.436230  -0.670 0.506618
## Arthritis      1.006637   0.244157   4.123 0.000189 ***
## Asthma        -1.168938   0.393499  -2.971 0.005066 **
## KidneyDisease -1.703910   0.624412  -2.729 0.009484 **
## Vapeing        0.928298   0.334727   2.773 0.008468 **
## Smoking       -0.446216   0.164611  -2.711 0.009929 **
## COPD          -0.004116   0.404486  -0.010 0.991932
## NoInsurance   -0.081151   0.091927  -0.883 0.382765
## Dentist       -0.100848   0.119477  -0.844 0.403771
## Walkable       0.074598   0.639220   0.117 0.907695
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.751 on 39 degrees of freedom
##   (266 observations deleted due to missingness)
## Multiple R-squared:  0.8623, Adjusted R-squared:  0.8199
## F-statistic: 20.35 on 12 and 39 DF,  p-value: 3.84e-13
```

Diagnostic plots to judge the model's fit

Let's look at some diagnostics to further understand how good the predictive model is. This will provide some clues and baseline when we move on to estimating causality in the next section.

Residuals Histogram First consider the histogram plot of the residuals. This is to help judge if the errors are normally distributed. The residuals are not very normal looking here, likely due to issues mentioned before - small sample size and a lot of NA values.

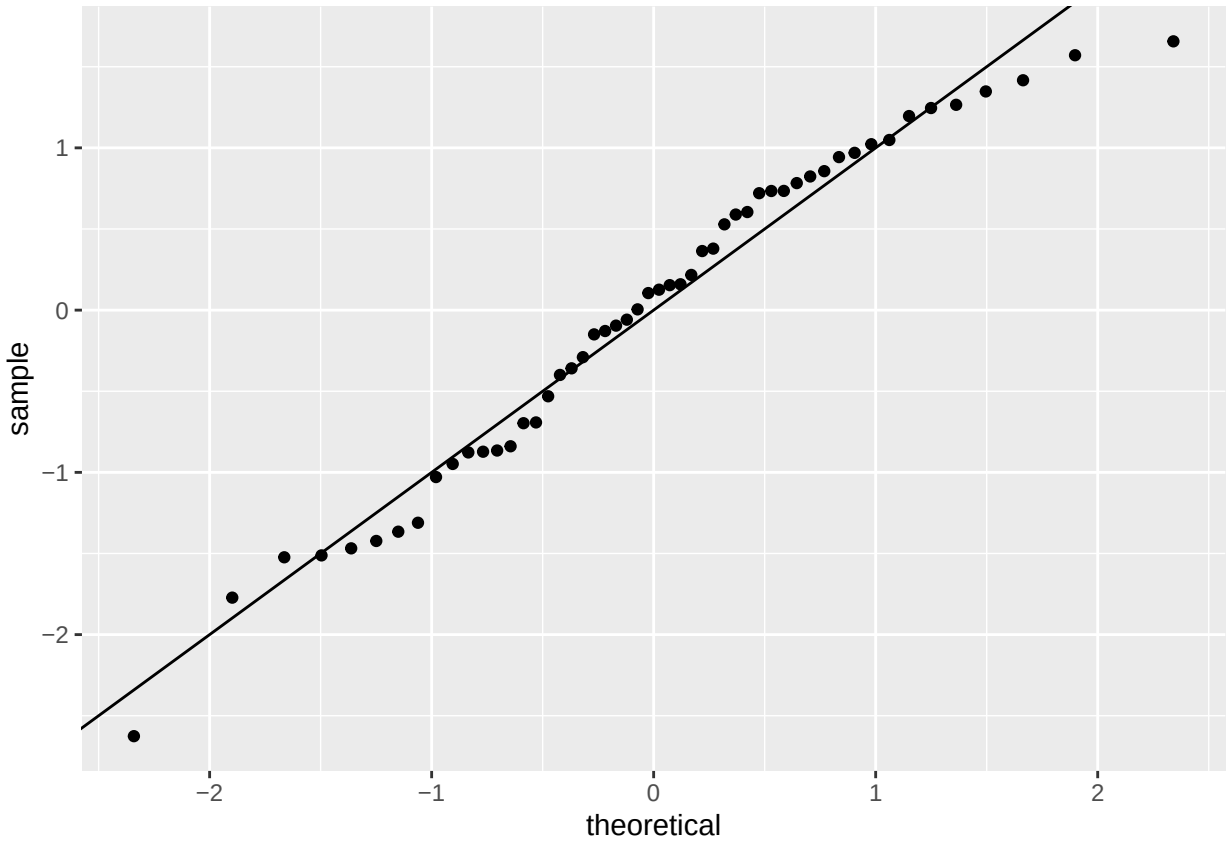
```
resid_hist <- ggplot(data = lm_fit, aes(x=.resid)) + geom_histogram()
resid_hist
```



Residuals qq-plot

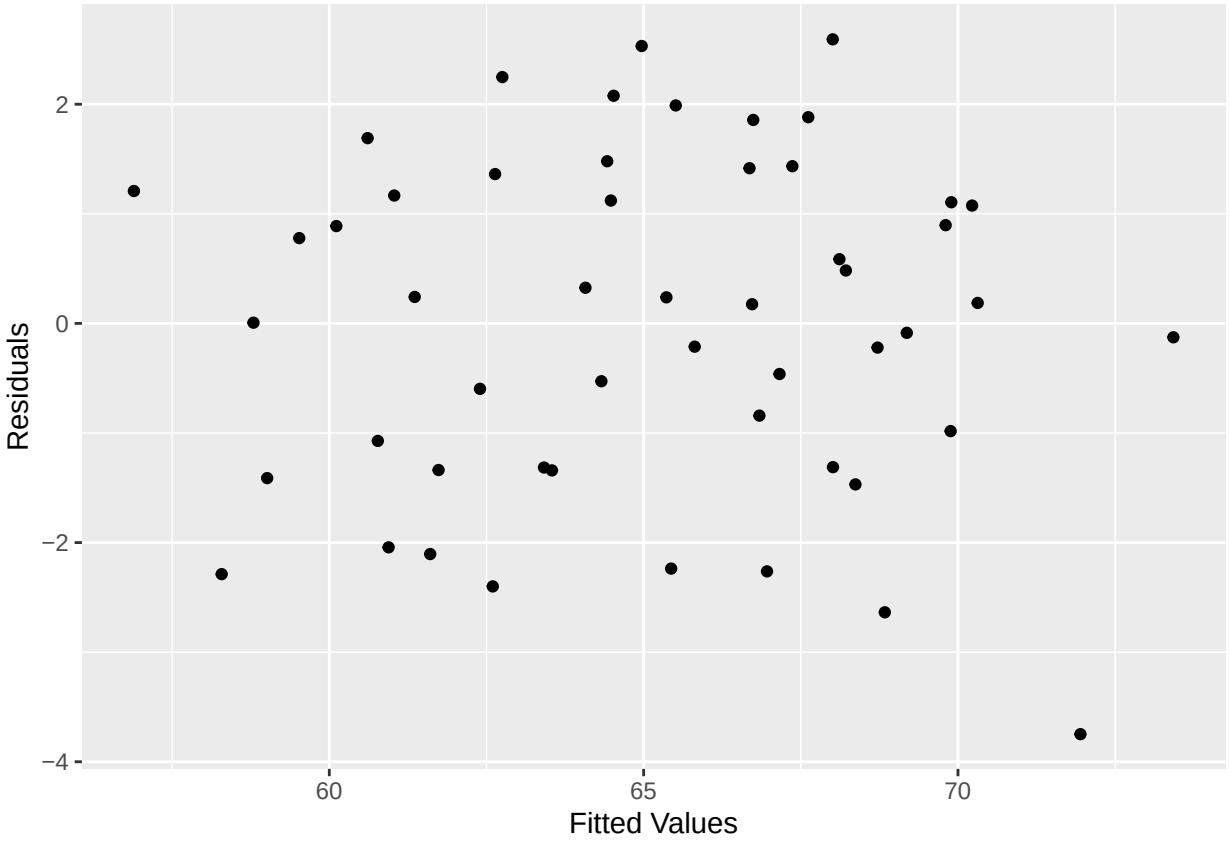
Next consider the qq-plot to again judge the normality of the errors. The qq-plot does not look as bad.

```
resid_qqplot <- ggplot(data = lm_fit, aes(sample=.stdresid)) + stat_qq() + geom_abline()
resid_qqplot
```



Residuals versus Fitted scatterplot Finally, plot the residuals against the fitted values to see if there is additional structure not captured by the model. It looks pretty unstructured - no problems here.

```
resid_fitted <- ggplot(data = lm_fit, aes(x=.fitted, y=.resid)) + geom_point() +
              labs(x="Fitted Values", y="Residuals")
resid_fitted
```



Conclusions from the predictive model

Overweight can be predicted reasonably well from the variables in our dataset. Interestingly Walkable is not statistically significant in the model. We will do some further analysis in the next sections to see if confounders may be impacting this model and there may actually be hidden causality between Walkable and Overweight.

TODO: Insert rest of paper here

Appendix

Original data pre-processing methodology

As described in the objective section, the original data came from two sources. The disease indicators data contains location information in the form of latitude and longitude. The walkability data contains location information in the form of Federal census location codes (FIPS codes). The pre-processing technique below was used to convert the latitude and longitude to FIPS codes, and then perform a join operation utilizing the FIPS codes. The resulting data is the original disease indicators data, augmented with the walkability information for the location corresponding to the original latitude and longitude.

In other words, for every row in the disease indicators data set, the corresponding walkability information for the region was added to that row. All of the commands are commented out to prevent them from being executed on knit since they take a long time to run.

```
#download.file("https://edg.epa.gov/EPADataCommons/public/OA/EPA_SmartLocationDatabase_V3_Jan_2021_Fina
#download.file("https://data.cdc.gov/api/views/g4ie-h725/rows.csv?accessType=DOWNLOAD", destfile="disea
```

Download the raw data

```
#walkability <- read.csv("walkability.csv")
## some of the disease data has no GeoLocation, which we cannot use for our analysis, so filter those o
#disease <- filter(read.csv("diseaseindicators.csv"), GeoLocation != "")
```

Load the data into R

```
## Extract the latitude and longitude values from the GeoLocation column using str_extract_all()
#geo_df <- str_extract_all(disease$GeoLocation, "-?[0-9]+\\.[0-9]+")

## Convert the extracted values to numeric and assign them to the corresponding latitude and longitude
#disease$lat <- as.numeric(sapply(geo_df, function(x) x[2]))
#disease$long <- as.numeric(sapply(geo_df, function(x) x[1]))
```

Extract the latitude and longitude into separate columns

Fetch the geographic information required to map latitude and logitude to FIPS blocks The tigris library provides a function "block_groups" which returns geographic information about every FIPS block. This geographic information can be used to convert latitude and longitude to FIPS block. The following code downloads all of the block_groups for every block in the walkability data set.

```
## create data frame for block_groups data
#allblockgroups <- data.frame(matrix(ncol=6, nrow=0))
#colnames(allblockgroups) <- c('STATEFP', 'COUNTYFP', 'TRACTCE', 'BLKGRPCE', 'GEOID', 'geometry')

## get block geography data for each state in the walkability dataset
#stateCodes <- data.frame(unique(walkability$STATEFP))
#for (i in 1:nrow(stateCodes)) {
#  stateCode=stateCodes[[1]][i]
#  counties = distinct(filter(walkability, STATEFP == stateCode), COUNTYFP)$COUNTYFP
#  new_blocks <- block_groups(state=stateCodes[[1]][i], counties) %>%
#    select(STATEFP, COUNTYFP, TRACTCE, BLKGRPCE, GEOID, geometry)
#  allblockgroups <- rbind(allblockgroups, new_blocks)
#}
```

```
#my_points <- data.frame(
#  x = disease$lat,
```

```
#  y = disease$long
#) %>%
#  st_as_sf(coords = c("y", "x"),
#     crs = st_crs(allblockgroups))

#my_points_blocks <- st_join(my_points, allblockgroups)
#disease$STATEFP = as.integer(my_points_blocks$STATEFP)
#disease$COUNTYFP = as.integer(my_points_blocks$COUNTYFP)
#disease$TRACTCE = as.integer(my_points_blocks$TRACTCE)
#disease$BLKGRPCE = as.integer(my_points_blocks$BLKGRPCE)
#disease$GEOID = as.numeric(my_points_blocks$GEOID)
```

Use block geographies to convert longitude and latitude to FIPS blocks

```
# Join the disease data with the walkability data
#disease_with_walkability <- left_join(disease, walkability,
#                                      by = c("STATEFP", "COUNTYFP", "TRACTCE", "BLKGRPCE"))
```

Join the disease indicators and walkability data sets based on FIPS blocks

```
#write.csv(disease_with_walkability, file = "disease_with_walkability.csv")
```

Export the joined data to be used for further processing later.