Peer Review

After presenting our project, we received several constructive feedbacks and made some corresponding adjustments. First, we revised the data of X (the percentage of long-time commuters whose commuting times are longer than 30 minutes). Our original data of X only included car, truck, and van driving alone as the commuting transportation. To make the transportation more inclusive, we enlarged our dataset of X by including all the public transportation. Also, to reduce the noise in X and to eliminate the attenuation bias, we smoothed the X data using natural splines function and random forests and then built the second multiple linear regression model with the smoothed X data. Moreover, a third confounder, the median household income of each county, was added to our multiple linear regression model. In the end, we explained why our X and Y (the adult obesity rate) were selected and how they are correlated more specifically.

Regression analysis of the effect of the percentage of long-time commuters on the adult obesity rate

STAT 333

Spring 2018

Group Name: MADATA

Group Members: Jingqi Duan, Elaine Lin, Wengie Wang, Hengjiali Xu, Yumin Zhang

Introduction

As the rate of obesity significantly increases over the time, obesity has now become one of the leading public issues in the United States. According to the National Center for Health Statistics, the rate of obesity has peaked in all adults, as nearly 40% of adults and 19% of youth are obese (Larned, 2017). The impacts of obesity can be both extensive and destructive. Not only would obesity and its associated health problems undermine the lengths and quality of lives of individuals, but they also increase the healthcare costs nationwide. Indeed, on the individual level, obesity can have adverse effects on individuals' health, such as heart problems, diabetes and some cancers, while on the societal level, obesity may directly and indirectly increase the medical costs of a society as a whole. Specifically, direct medical costs include preventive, diagnostic, and treatment services related to obesity, and indirect costs involve morbidity and mortality costs including productivity ("Adult Obesity Causes & Consequences", 2018).

Considering these severe consequences of obesity, our group was interested in identifying a potential factor behind the high obesity rate in the United States. We assumed that the commuting time spent in private motor vehicles and public transportations might be correlated with the high adult obesity rate, because people spending a large amount of time in those vehicles tend to be more sedentary and physically inactive and therefore tend to become obese. Specifically, they commute for such a long time that they might have less time to exercise and thus are more likely to be obese. Therefore, we thought counties where more people commute for long time (commuting time > 30 min) might be more likely to have higher adult obesity rate.

As we examined the relationship between the percentage of long-time commuters (commuting time > 30 min) (X) and the adult obesity rate (Y), we have found several potential

confounders that might influence the relationship between X and Y, including disability rate (Z_1), poverty rate (Z_2) and median household income (Z_3), which we thought can both affect the percentage of long-time commuters and the obesity rate. Our group then proposed that **counties where more people commuting for a long time tend to have higher obesity rate, after smoothing the percentage of long-time commuters and controlling the disability rate, poverty rate and median household income.**

Methods

Part 1: Data Definitions, Descriptions, & Collections

(i) Dependent and Independent Variables

Our independent variable X is the percentage of commuters whose commuting times in private motor vehicles and public transportations are larger than 30 minutes over the total number of people commuting by private motor vehicles and public transportations. According to the American Community Survey (ACS), the data are collected by self-reported surveys and involve cars, trucks, vans, buses, streetcars, subways, railroads and ferryboats. For each category, the data include the total numbers of workers who use that kind of transit method for 0 to 29 minutes, 30 to 34 minutes, 35 to 44 minutes, 45 to 59 minutes, and 60 or more minutes. Because the populations of different regions vary, in order to normalize the long-time commuters, we have calculated the percentage of the number of long-time commuters (commuting time > 30 min) over that of total commuters. As for the dependent variable Y, the data of which are collected by self-reported surveys, the adult obesity rate is defined as the percentage of the adult population (age 20 and older) who report a body mass index (BMI) greater than or equal to 30 kg/m^2. Released by CDC's Behavioral Risk Factor Surveillance System (BRFSS), the data have

information from 2004 to 2013 and include the number of obese adults, the percent of obese adults, lower confidence limit, upper confidence limit, age adjusted percent of obese adults, age-adjusted lower confidence limit, and age-adjusted upper confidence limit.

(ii) Confounders

Considering there might be some potential factors that influence the relationship between the percentage of long-time commuters (X) and the adult obesity rate (Y), we have identified three confounders--the disability rate (Z_1), the poverty rate (Z_2) and household median income (Z_3). First, if the disability rate of a county is higher, its percentage of people commuting for over 30 minutes would be higher because physically disabled people might move with difficulty and thus have longer commuting time every day. Also, higher disability rate may cause higher obesity rate because disable people tend to be physical inactive, which might result in obesity. In other words, higher disability rate can lead to both higher percentage of long-time commuters (X) and higher adult obesity rate (Y). Second, higher poverty rate of counties can cause higher percentage of long-time commuters because people in poverty tend to have longer time commuting due to limited budget to commute fast. Also, higher poverty rate may lead to higher obesity rate, since poor people have limited access to healthier food and lifestyle, thus resulting in a higher obesity rate. That is, higher poverty rate can also cause both higher percent of long-time commuters (X) and higher adult obesity rate (Y). Third, higher median household income can lead to lower percentage of long-time commuters because households with higher income have enough budget to commute more efficiently. Moreover, higher median income can bring about lower adult obesity rate because households with higher income tend to choose healthier lifestyle. For instance, they budget for a gym membership and dedicate time to fitness, resulting in a lower adult obesity rate. That is to say, higher median household income can lead to lower

percent of long-time commuters (X) and lower adult obesity rate (Y). Thus, since our three confounders--the disability rate, the poverty rate and the median household income can all cause changes in the percentage of long-time commuters (X) and the adult obesity rate (Y), it is reasonable for us to include them as confounders.

According to American Community Survey (ACS), the data of disability rate (Z_1) are collected via surveys and include the total population, the disability population, and the percentage of different kinds of disability for different people from 2011 to 2015. Additionally, collected by Census Bureau in 2013, the data of the poverty rate (Z_2) include the number of individuals who are in poverty and the population size of each county. Lastly, according to the Census Bureau, the data of median household income (Z_3) include all sources of a household's income such as wage, investments, proprietors' income, etc. in 2013.

Part 2: Data Cleaning

The unit of observations in our dataset was county sorted by ascending order of Federal Information Processing Standards county code. While the data of the percentage of long-time commuters (X), the adult obesity rate (Y), and the median income (Z_3) included 50 states and District of Columbia, the data of the disability rate (Z_1) and the poverty rate (Z_2) also included Puerto Rico besides 50 states. To maintain the same number of observations, we deleted the data collected in Puerto Rico from Z_1 and Z_2 and performed the regression analysis with data obtained in 50 states and District of Columbia. This adjusted dataset gave us in total 3143 observations of 5 variables. Due to the limitation in data collection, there were 3 missing values in X, 2 missing values in Y, 1 missing value in Z_1, and 1 missing value in Z_3. Using the function in R to fit linear models, we handled the missing values with the default settings of linear model function, which was the case wise deletion. Specifically, R deleted the observations with missing values in any

variable when fitting the model and then calculated the fitted values and residuals of the observations. Since only 6 observations out of 3143 observations had missing data of 5 variables and these missing observations had no specific pattern, the list wise deletion was suitable to handle the missing values.

Part 3: Description of Statistical Techniques

In order to find the true relationship between the adult obesity rate (Y) and the percentage of long-time commuters (X), we built a multiple linear regression model by regressing Y on X, the disability rate (Z_1), the poverty rate (Z_2), and the median household income (Z_3).

To eliminate the noise in the original data of X, which could probably cause the attenuation bias and deviate the estimated β_1 (the coefficient of X) from the true β_1, we applied the natural splines function in R to smooth the data of X. Using the algorithm of random forests, we then classified each county based on its geographical location, which was referred from the dataset of the county's latitude and longitude. The smoothed dataset of X was denoted as X_s and was also included in our dataset. The visualization and comparison of X and X_s are shown in *Figure 1.1* and *Figure 1.2*. In the end, we built another multiple linear regression model with the smoothed data of X by regressing Y on X_s, Z_1, Z_2, and Z_3.

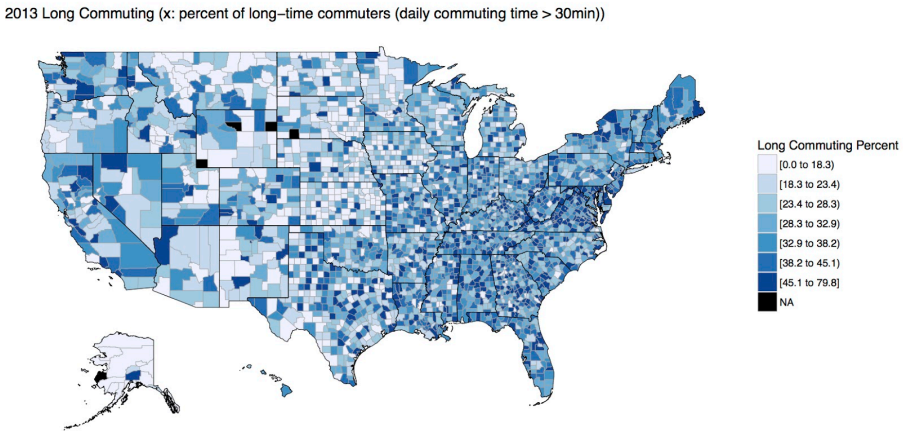


Figure 1.1. The Distribution of Long-Time Commuters in the United States Before Data Processing

This map shows little pattern among counties, which means our data are quite noisy. To reduce the noise, natural splines is applied to smooth the data of X.

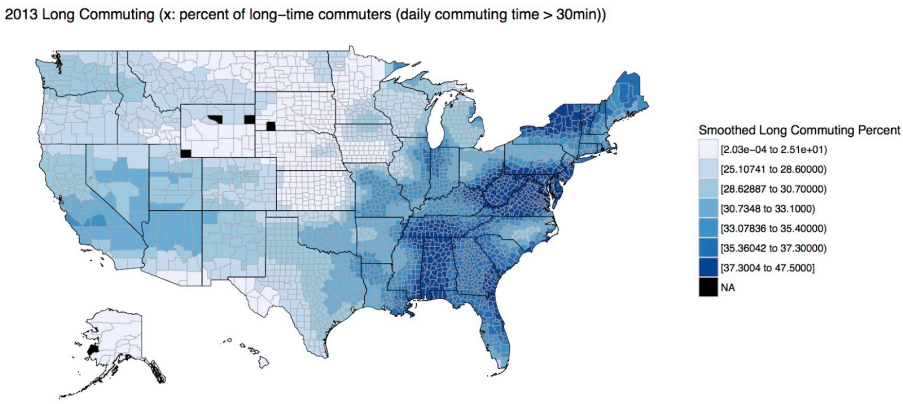


Figure 1.2. The Distribution of Long-Time Commuters in the United States After Data Processing

This map demonstrates the distribution of percent of long-time commuters in the United States after applying random forest with degree of freedom of 10 based on the latitude and longitude of each county.

Results

Figure 1.1, the map of the long-time commuters, demonstrated that our original X data was quite noisy. In *Table 1*, the p-values of disability rate, poverty rate and median income were all statistically significant, while the p-value of long-time commuters was 0.596, which was not statistically significant.

After smoothing the original commuting data, we found some geographic patterns, as shown in *Figure 1.2* and *Figure 1.3*. The scatterplot of the adult obesity rate (Y) against the smoothed percentage of long-time commuters (X_s) implied that they might be positively correlated. In order to find the relationship between X_s and Y with the presence of three confounders, we built a multiple linear regression model on those variables. In *Table 2*, the p-value of the F-statistic was extremely small, which meant in this model, at least one coefficient was significant. More specifically, four p-values regarding to the percentage of long-time commuters, disability rate, poverty rate and median income, as shown in *Table 2*, were all statistically significant. *Table 2* told that keeping all other variables constant, a unit increase in percentage of long-time commuters, disability rate and poverty rate could cause obesity rate to increase 5.338%, 9.871%, 5.404% respectively. A dollar increase in median income could lead to 0.01233% decrease in obesity rate.

To further investigate our model, we used R to plot the regression model, and the group of four plots, *Figure 1.4* and *Figure 1.5,* for each model was shown as following. The plots indicate that both of our model were appropriately fitted.

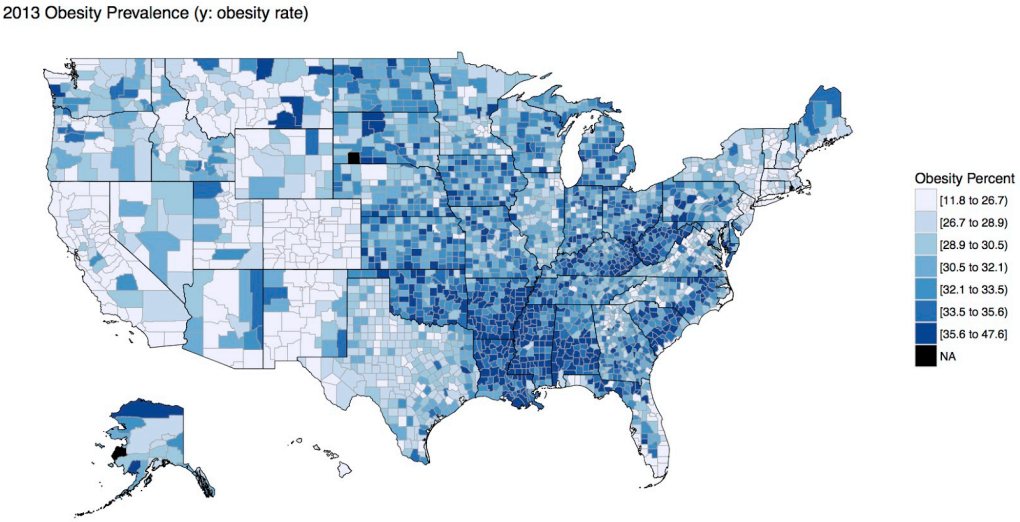


Figure 1.3. The Prevalence of Obesity for Counties in the United States

This map illustrates that horizontally the obesity rate on the west coast is much lower than that in the southeastern area. In the Midwest and east coast area, vertically the southern counties tend to have higher obesity rate than the north. This pattern can also be seen in *Figure 1.2* (map of smoothed X).

```
##
## Call:
## lm(formula = Obesity ~ Commute + Disability + Poverty + Income,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3497  -2.3449   0.2543   2.7036  13.3753
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.326e+01  9.245e-01  35.976  < 2e-16 ***
## Commute      3.181e-03  5.994e-03   0.531    0.596
## Disability   1.188e-01  2.277e-02   5.215 1.96e-07 ***
## Poverty      6.738e-02  1.665e-02   4.046 5.34e-05 ***
## Income      -1.157e-04  1.114e-05 -10.388  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.011 on 3132 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.2143, Adjusted R-squared:  0.2133
## F-statistic: 213.6 on 4 and 3132 DF,  p-value: < 2.2e-16
```

Table 1. Summary table: The linear model on the obesity rate with original percentage of long-time commuters, disability, poverty and median income

```
##
## Call:
## lm(formula = Obesity ~ Smoothed_Commute + Disability + Poverty +
##     Income, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.271  -2.355   0.234   2.741  14.831
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.257e+01  9.396e-01  34.659  < 2e-16 ***
## Smoothed_Commute  5.338e-02  1.383e-02   3.861 0.000115 ***
## Disability        9.871e-02  2.318e-02   4.258 2.12e-05 ***
## Poverty           5.404e-02  1.695e-02   3.188 0.001445 **
## Income           -1.233e-04  1.125e-05 -10.960  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.002 on 3132 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.218,  Adjusted R-squared:  0.217
## F-statistic: 218.2 on 4 and 3132 DF,  p-value: < 2.2e-16
```

Table 2. Summary Table: The linear regression model of the obesity rate on the smoothed percentage of long-time commuters, the disability rate, the poverty rate and the median household income.

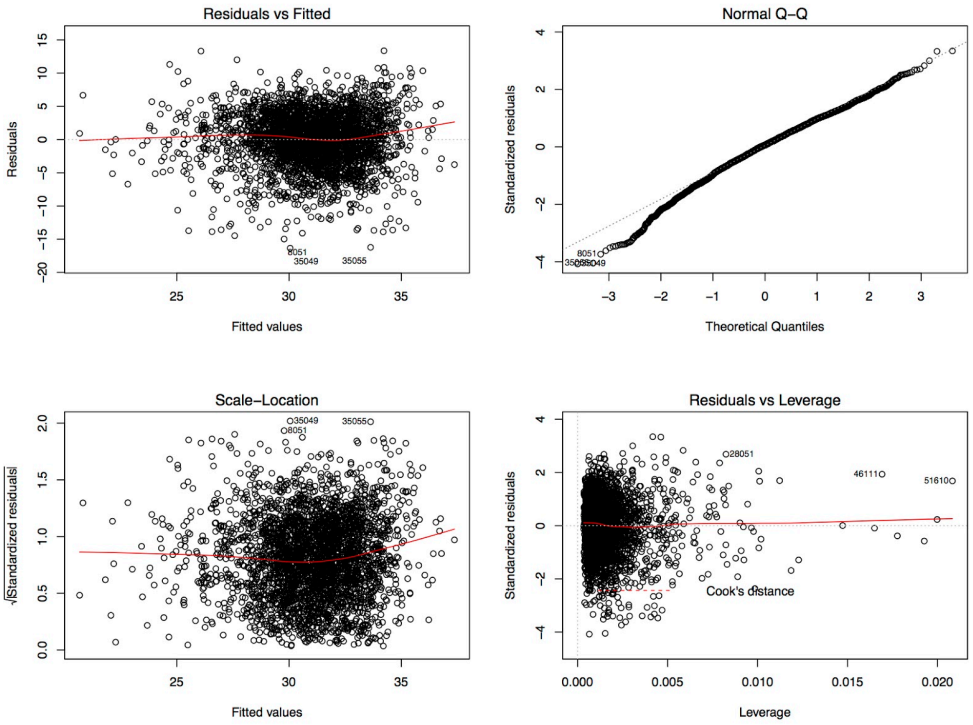


Figure 1.4. Diagnostic plots of multiple regression model (before smoothing): a plot of residuals against fitted values, a scale-location plot against fitted values, a Normal Q-Q plot, a plot of residuals against leverages.

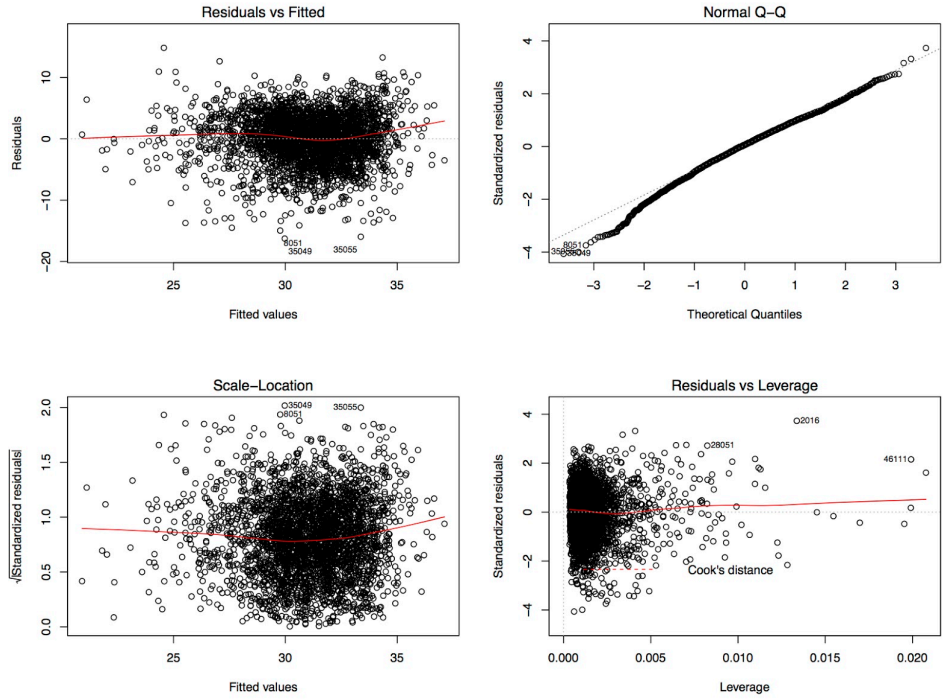


Figure 1.5. Diagnostic plots of multiple regression model (after smoothing): a plot of residuals against fitted values, a scale-location plot against fitted values, a Normal Q-Q plot, a plot of residuals against leverages.

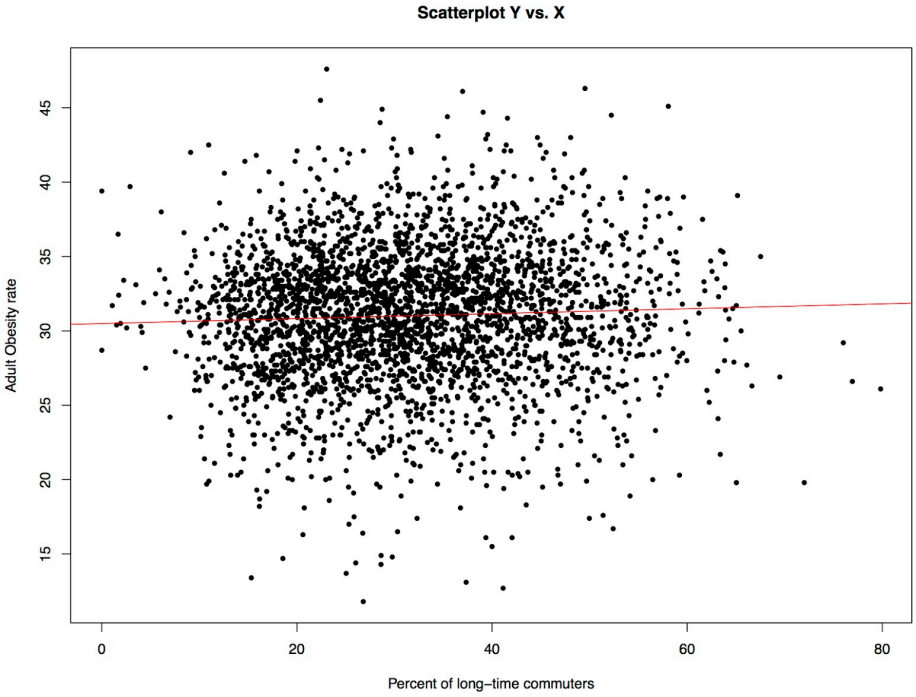


Figure 1.6. Scatter plot of the adult obesity rate against the percent of long-time commuters

The regression line (red line) has a slightly positive slope, indicating that there is a positive correlation between the adult obesity rate and the percentage of long-time commuters (before smoothing).

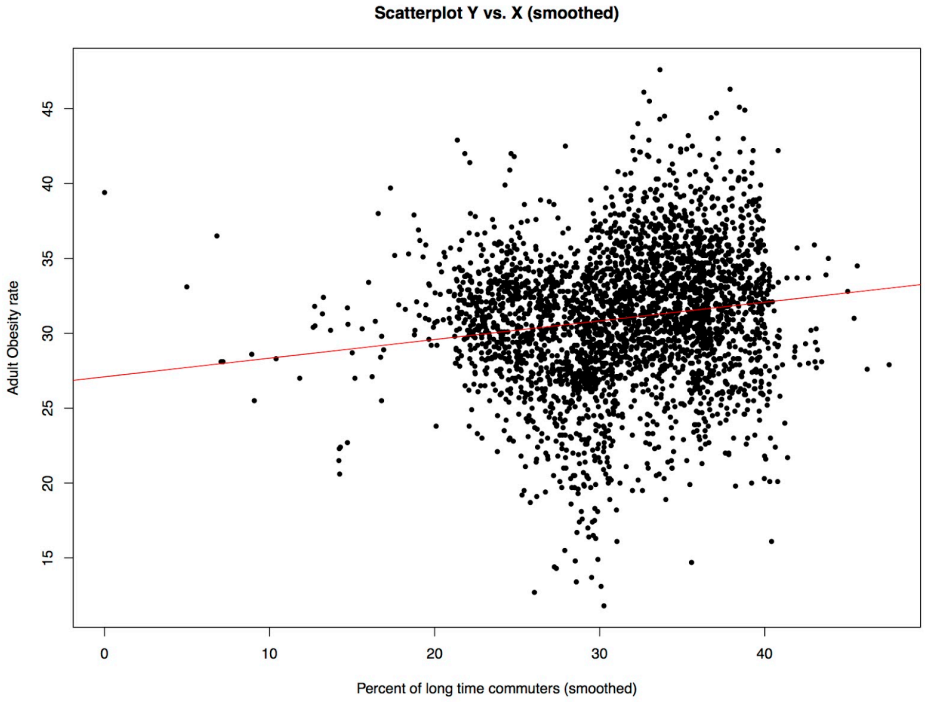


Figure 1.7. Scatter plot of the adult obesity rate against the smoothed percent of long-time commuters

The regression line (red line) has a slightly positive slope, indicating that there is a positive correlation between the adult obesity rate and the percentage of long-time commuters (after smoothing).

Conclusion

Part 1: Discussion & Uncertainty

As mentioned in the result section, two multiple linear regression models have been built and analyzed. At first, we created two maps, each based on the percentage of long-time commuters whose commuting times are longer than 30 minutes (X) and the adult obesity rate (Y). Even though we noticed the noise of the map of X, we built a linear model using the original noisy data of X. According to *Table 1*, the p-value of the coefficient of X (β_1) was insignificant and the estimated slope coefficient was close to zero. Accordingly, our first model showed no correlation between the percentage of long time commuters (X) and the adult obesity rate (Y). This null result made us reconsider the noise in X. This noise might produce the attenuation bias

and decrease the power of β_1, making the estimated β_1 much lower than the true value. To eliminate this attenuation bias, we reconsidered the measurement of our X. Because residents in adjacent counties might have similar commuting time, we smoothed the data of X by averaging the percentage of long-time commuters of the adjacent counties. Then we built the second model using the smoothed data X_s. According to *Table 2*, the p-value of the coefficient of X_s was significant, suggesting a positive correlation between X_s and Y. As these two models led to two different results, we had to choose a better one. After we mapped the smoothed X, it (*Figure 1.2*) showed much less noise. The previous null result caused by the attenuation bias has changed after we replaced the original data of X with the smoothed data. With the sufficient reasons mentioned previously to smooth X and the appropriate results after smoothing, the smoothed data X_s were more reliable than the original ones. We thus drew the conclusion that counties with higher smoothed percentage of long-time commuters tend to have higher adult obesity rate, holding the disability rate, the poverty rate and the median household income fixed.

Although we believed the reasons to choose the second model were sufficient, the uncertainty still exists. The main source of this uncertainty came from the smoothing process, because we were unsure whether we chose the correct model. That is, we were not sure whether the noise actually revealed the variation among counties. If this variation among counties was a true feature of the data of X, the first model might be better.

Part 2: Limitations

There were several sources of limitations in our project. First, the data of the disability rate was collected in a different time interval from other variables. While the datasets of other variables were from 2013, the dataset of the disability rate contained 5-year estimates. This inconsistency might decrease the accuracy of our model. Second, the reliability of the datasets of

the adult obesity rate (Y) and the disability rate (Z_1) remained uncertain since these data were collected by self-reported surveys, which may produce subjective biases to the data.

Apart from the confounders we included in the project, we have also considered other potential variables such as state effects. In particular, we thought that population density might be a factor causing the variations of commuting time between different counties/states. Specifically, people in counties with lower population density might have longer commuting time due to limited transportation system to commute fast, thus increasing percentage of long time commuters. In contrast, counties with higher population density tend to have better transportation system, which makes commuting more efficient and convenient, leading to a lower percentage of long time commuters. However, even though we thought population density might bring about different results to our model, due to limited time and resources, we did not include state effects in our model, which could also limit the accuracy of our model.

References

Larned, V. (2017, October 13). Obesity among all US adults reaches all-time high. Retrieved April 15, 2018, from https://www.cnn.com/2017/10/13/health/adult-obesity-increase-study/index.html

"Adult Obesity Causes & Consequences". (2018, March 05). Retrieved April 15, 2018, from https://www.cdc.gov/obesity/adult/causes.html

White, G. B. (2015, June 10). Long Commutes Are Awful, Especially for the Poor. Retrieved April 15, 2018, from https://www.theatlantic.com/business/archive/2015/06/long-commutes-are-awful-especially-for-the-poor/395519/

Data Sources:

http://www.countyhealthrankings.org/app/wisconsin/2017/measure/factors/137/description
https://www.cdc.gov/diabetes/data/countydata/countydataindicators.html
http://rtc.ruralinstitute.umt.edu/geography/
https://datausa.io/map/?level=county&key=income_below_poverty:pop_poverty_status,income_below_poverty,income_below_poverty_moe,pop_poverty_status,pop_poverty_status_moe
https://www.census.gov/data/datasets/2013/demo/saipe/2013-state-and-county.html

Appendix

Jingqi Duan, Elaine Lin, Wengie Wang, Hengjiali Xu, Yumin Zhang

4/22/2018

```
rm(list=ls())
library(choroplethr)
```

```
## Loading required package: acs
```

```
## Loading required package: stringr
```

```
## Loading required package: XML
```

```
##
## Attaching package: 'acs'
```

```
## The following object is masked from 'package:base':
##
##     apply
```

```
library(blscrapeR)
library(splines)
x <- get_bls_county()

data <- read.csv("RAW data.csv", header=T, colClasses="character")
data$Fips <- 0
data[,c(3,5)] <- data[,c(5,3)]
colnames(data) <- c("State","County","Fips","Commute","Obesity")

##### Fips #####
x <- x[-grep(pattern="PR", x=x$area_title),c(4,10)]
x[550:(nrow(x)+1),] <- x[549:nrow(x),]
x$area_title[549] <- "Kalawao County"; x$fips[549] <- "15005"
x[2918:(nrow(x)+1),] <- x[2917:nrow(x),]
x$area_title[2917] <- "Bedford City"; x$fips[2917] <- "51515"

data[83:(nrow(data)+1),] <- data[82:nrow(data),]
data[82,] <- c("Alaska", "Kusilvak Census Area", 0, NA, NA)
# remove 2004 to 2008 Census
data <- data[-grep(pattern="2004 to 2008", x=data$County),]
# reomve Wade Hampton Census Area AK (row 94)
data <- data[-grep(pattern="Wade Hampton", x=data$County),]

data$Fips <- x$fips
rownames(data) <- 1:nrow(data)

data$Obesity <- as.numeric(data$Obesity)
data$Commute <- as.numeric(data$Commute) * 100

##### Poverty #####
p <- read.csv("RAW Poverty.csv", header=T)
p <- p[which(p$year == "2013"),]
p$poverty_rate <- p[,4] / p[,5] * 100
p <- p[,c(2,3,6)]
```

```
p$geo <- gsub(pattern="05000US", replacement="", x=p$geo)
p <- p[-grep(pattern="^72", x=p$geo),]
p <- p[order(p$geo),]

data$Poverty <- p$poverty_rate

##### Disability #####
d <- read.csv("RAW disability.csv", header=T)
d <- d[-which(d$State == "Puerto Rico"),]
d[2918:(nrow(d)+1),] <- d[2917:nrow(d),]
d$County.Name[2917] <- "Bedford city"; d$Percent.with.a.Disability[2917] <- NA

data$Disability <- d$Percent.with.a.Disability

##### Median Income #####
i <- read.csv("RAW Income.csv", header=T)
i <- i[-which(i$County.FIPS.Code==0),]; rownames(i) <- 1:nrow(i)

data$Income <- as.numeric(as.character(gsub(",", "", i$Median.Household.Income)))

##### Latitude and Longitude #####
l <- read.csv("RAW Lat_Long.csv", header=T)
l <- l[-which(l$USPS=="PR"),]
l$NAME <- as.character(l$NAME)
l[2918:(nrow(l)+1),] <- l[2917:nrow(l),]
l$NAME[2917] <- "Bedford City"; l$INTPTLAT[2917] <- NA; l$INTPTLONG[2917] <- NA

data$Latitude <- l$INTPTLAT
data$Longitude <- l$INTPTLONG

##### Random Forests #####
#lm.5 <- lm(Commute ~ ns(Latitude,5) : ns(Longitude, 5), data=data)
#Smoothed_Commute <- unname(lm.5$fitted.values)
lm.10 <- lm(Commute ~ ns(Latitude,10) : ns(Longitude, 10), data=data)
Smoothed_Commute <- unname(lm.10$fitted.values)
#lm.20 <- lm(Commute ~ ns(Latitude,20) : ns(Longitude, 20), data=data)
#Smoothed_Commute <- unname(lm.20$fitted.values)
which(is.na(data$Commute)); which(is.na(data$Latitude))
```

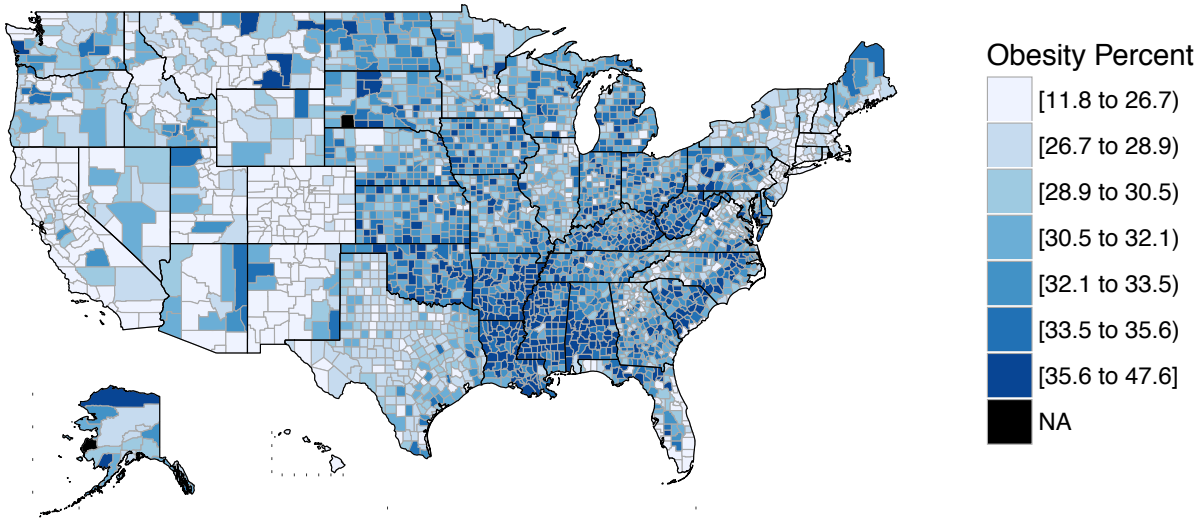
```
## [1]   82 3141 3142 3143
```

```
## [1] 2917
```

```
Smoothed_Commute[83:(length(Smoothed_Commute)+1)] <- Smoothed_Commute[82:(length(Smoothed_Commute))]
Smoothed_Commute[82] <- NA
Smoothed_Commute[2918:(length(Smoothed_Commute)+1)] <- Smoothed_Commute[2917:(length(Smoothed_Commute))]
Smoothed_Commute[2917] <- NA
Smoothed_Commute[3141:3143] <- NA
data$Smoothed_Commute <- Smoothed_Commute

# Save data
rownames(data) <- data$Fips

write.csv(data, file="data.csv", row.names=F)
```

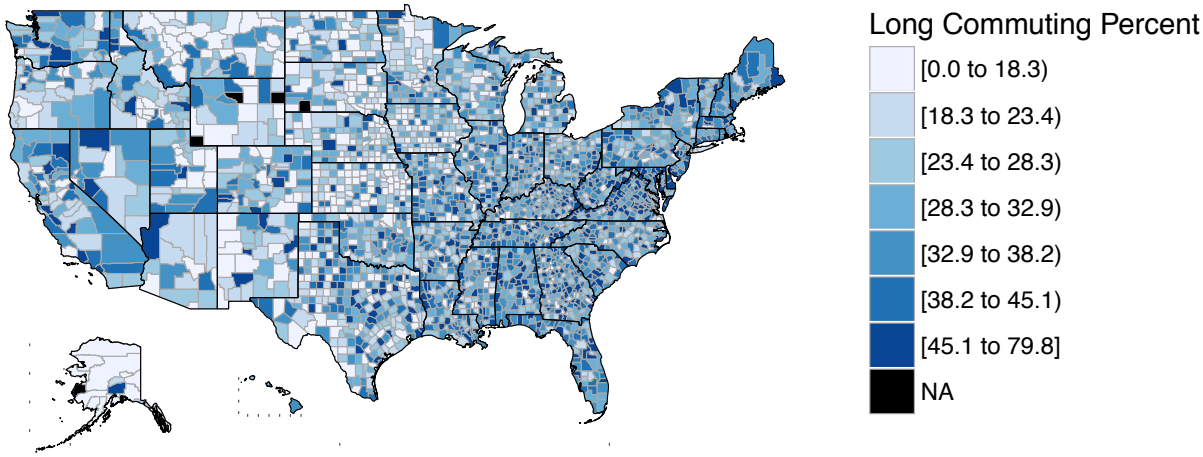
```
data <- read.csv("data.csv", header=T)

rownames(data) <- data$Fips

df <- data.frame(value=data$Obesity, region=as.numeric(data$Fips))
df2 <- data.frame(value=data$Commute, region=as.numeric(data$Fips))
df3 <-data.frame(value=data$Smoothed_Commute, region=as.numeric(data$Fips))

county_choropleth(df, title="2013 Obesity Prevalence (y: obesity rate)", legend="Obesity Percent")
```

2013 Obesity Prevalence (y: obesity rate)



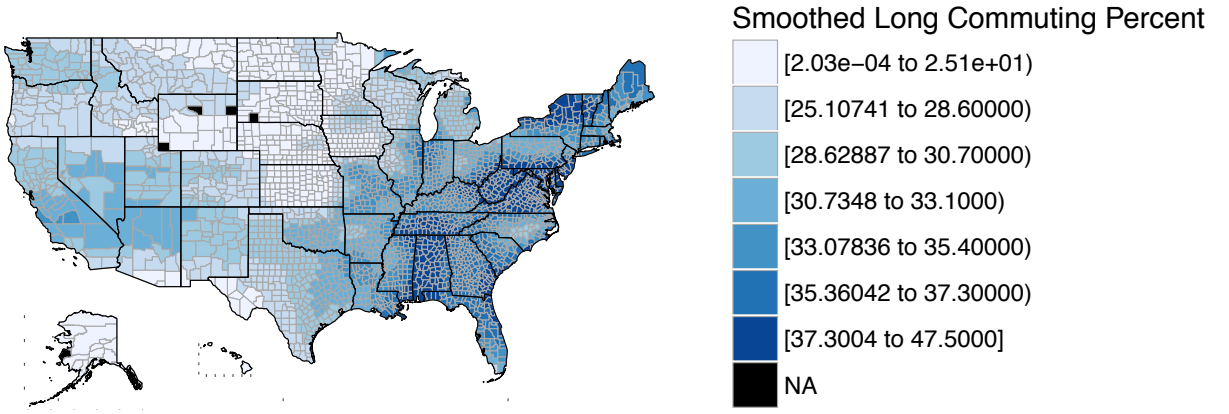

```
county_choropleth(df2, title="2013 Long Commuting (x: percent of long-time commuters (daily commuting t
```

2013 Long Commuting (x: percent of long−time commuters (daily commuting time




```
county_choropleth(df3, title="2013 Long Commuting (x: percent of long-time commuters (daily commuting t
```

2013 Long Commuting (x: percent of long−time commuters (daily commuting time




```
# x: Long Commute Percent
# y: Obesity Percent
# z1: Disability Percent
# z2: Poverty Percent
# z3: Median Income

lm1 <- lm(Obesity~Commute+Disability+Poverty+Income, data=data)
summary(lm1)
```

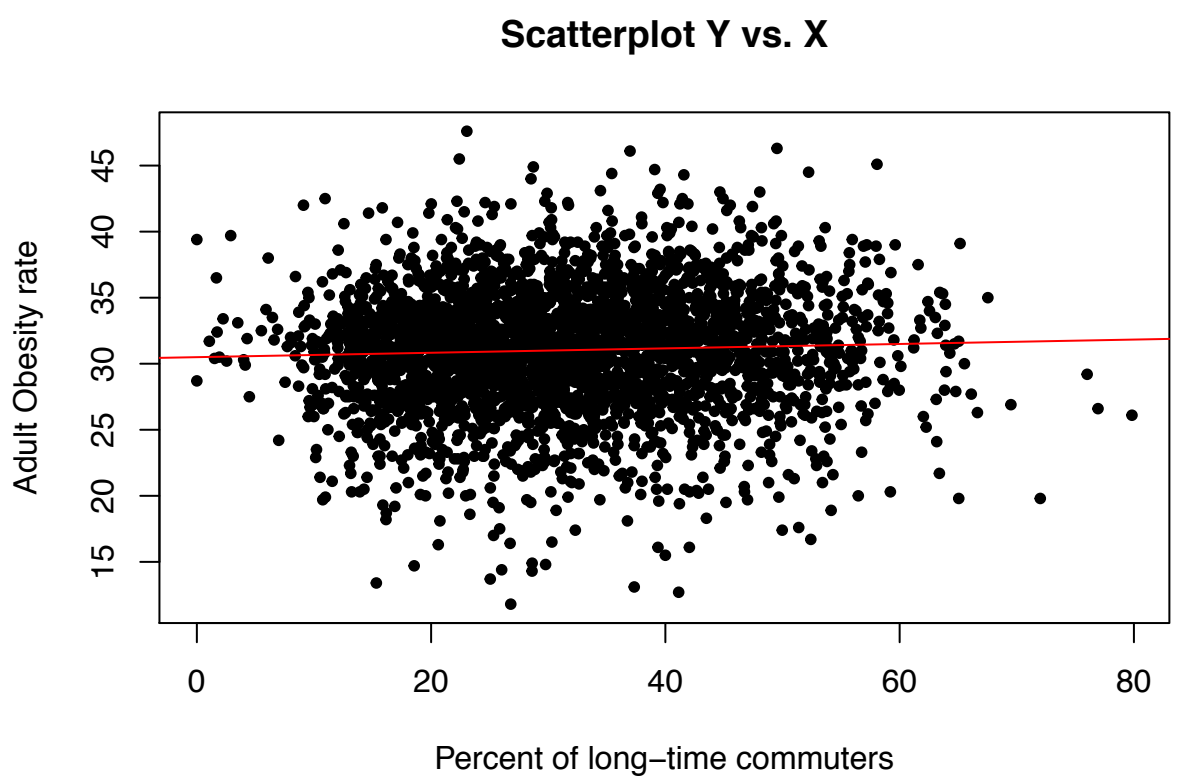
```
##
## Call:
## lm(formula = Obesity ~ Commute + Disability + Poverty + Income,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3497  -2.3449   0.2543   2.7036  13.3753
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.326e+01  9.245e-01  35.976  < 2e-16 ***
## Commute      3.181e-03  5.994e-03   0.531    0.596
## Disability   1.188e-01  2.277e-02   5.215 1.96e-07 ***
## Poverty      6.738e-02  1.665e-02   4.046 5.34e-05 ***
## Income      -1.157e-04  1.114e-05 -10.388  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.011 on 3132 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.2143, Adjusted R-squared:  0.2133
## F-statistic: 213.6 on 4 and 3132 DF,  p-value: < 2.2e-16
```

```
lm2 <- lm(Obesity~Smoothed_Commute+Disability+Poverty+Income, data=data)
summary(lm2)
```

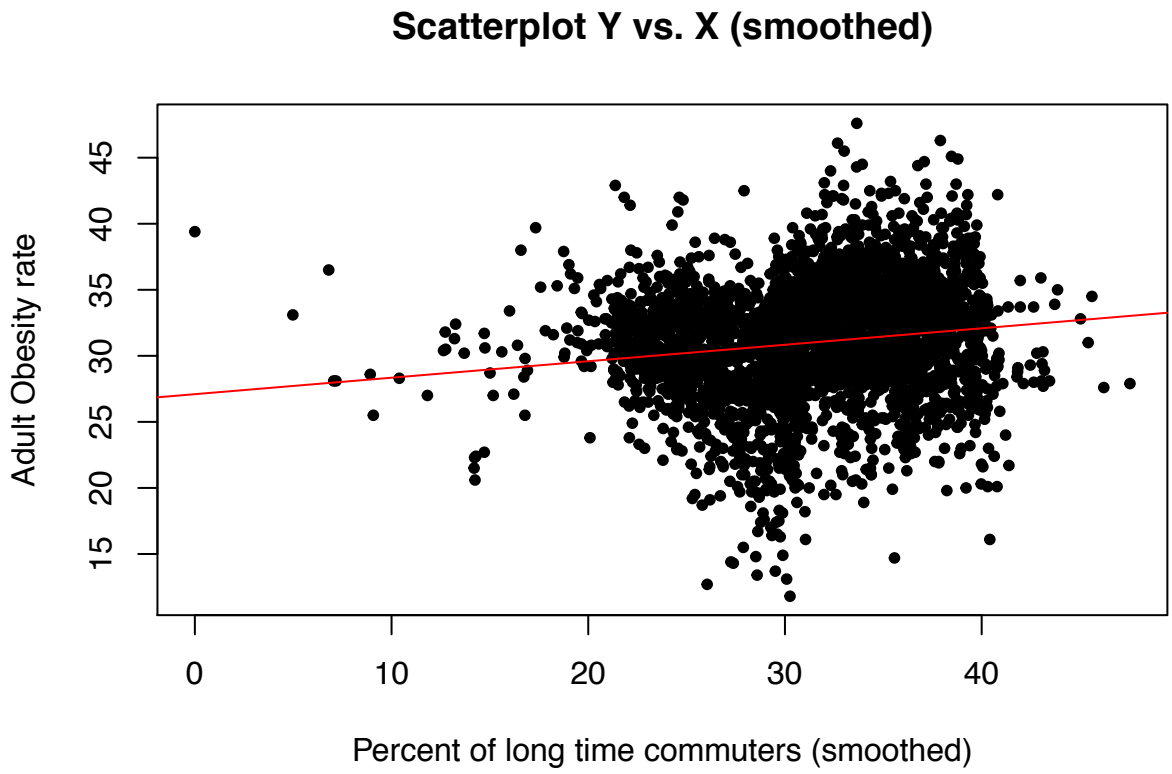
```
##
## Call:
## lm(formula = Obesity ~ Smoothed_Commute + Disability + Poverty +
##     Income, data = data)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.271  -2.355   0.234   2.741  14.831
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.257e+01  9.396e-01  34.659  < 2e-16 ***
## Smoothed_Commute  5.338e-02  1.383e-02   3.861 0.000115 ***
## Disability        9.871e-02  2.318e-02   4.258 2.12e-05 ***
## Poverty           5.404e-02  1.695e-02   3.188 0.001445 **
## Income           -1.233e-04  1.125e-05 -10.960  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.002 on 3132 degrees of freedom
##   (6 observations deleted due to missingness)
## Multiple R-squared:  0.218,  Adjusted R-squared:  0.217
## F-statistic: 218.2 on 4 and 3132 DF,  p-value: < 2.2e-16
```

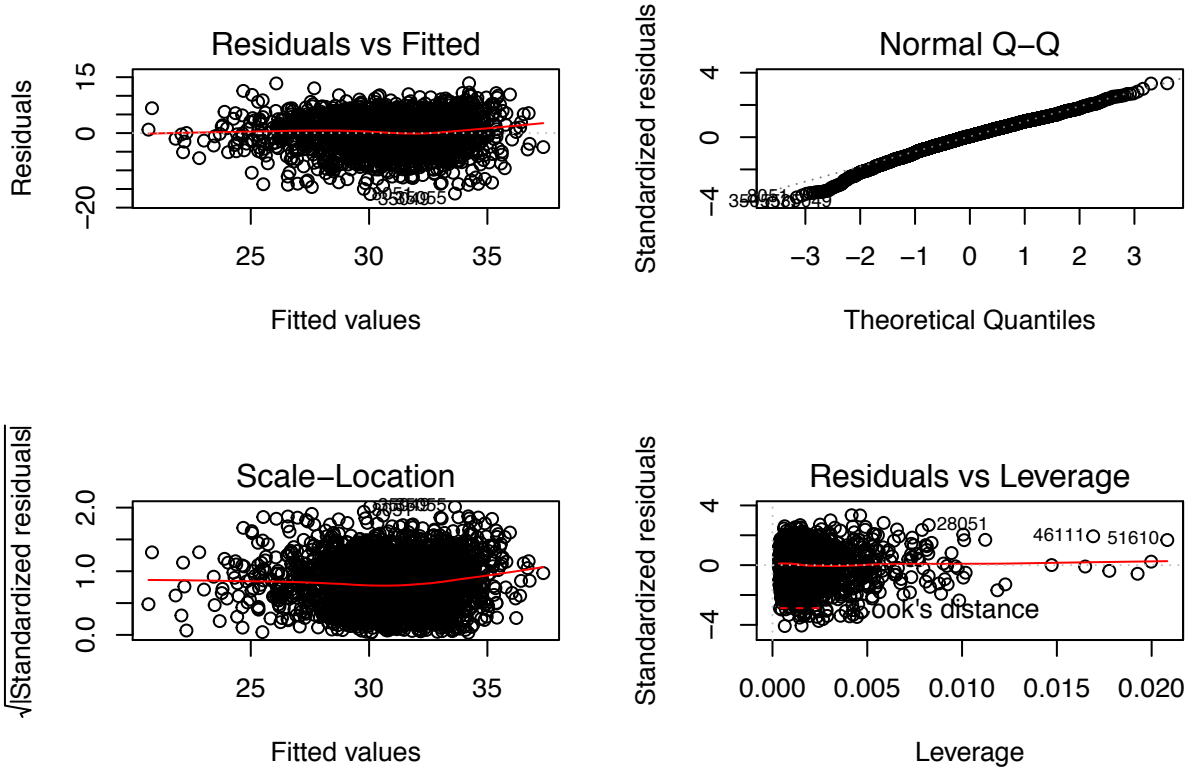
```r
par(mfrow=c(1,1))
plot(x=data$Commute, y=data$Obesity, xlab="Percent of long-time commuters",
     ylab="Adult Obesity rate", main="Scatterplot Y vs. X", pch=20)
abline(lm(Obesity~Commute, data=data), col="red")
```



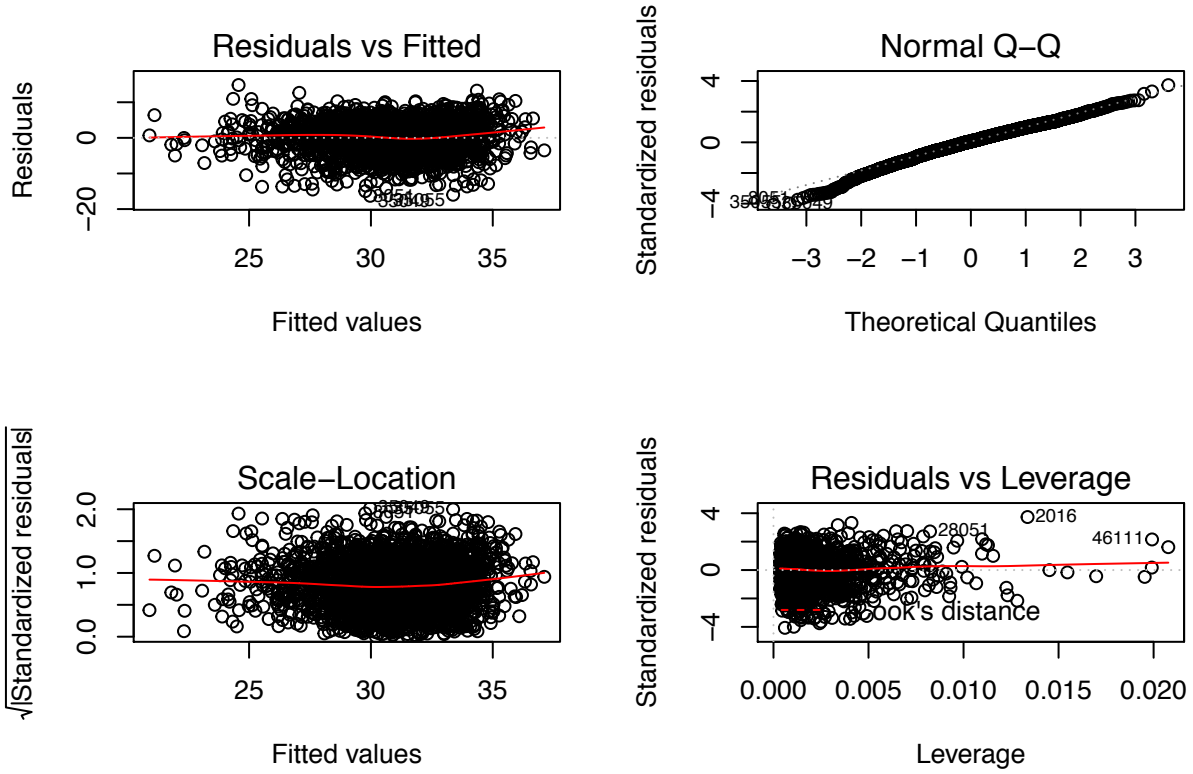

```r
plot(x=data$Smoothed_Commute, y=data$Obesity, xlab="Percent of long time commuters (smoothed)",
     ylab="Adult Obesity rate", main="Scatterplot Y vs. X (smoothed)", pch=20)
abline(lm(Obesity~Smoothed_Commute, data=data), col="red")
```




```
par(mfrow=c(2,2))
plot(lm1)
```



```
plot(lm2)
```


Residuals vs Fitted
Residuals
Fitted values
Normal Q–Q
Standardized residuals
Theoretical Quantiles
Scale–Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
2016
28051
46111