HM3

Zikang Zheng

7/16/2020

Load necessary packages.

```
library(tidyverse)
library(stargazer)
library(lubridate)
library(sf)
library(tidycensus)
library(stargazer,quietly = TRUE)
library(reshape)
library(broom)
library(sandwich)
```

SECTION 1

Question 1.1-1.2

Load data.

```
thefts <- read_csv("thefts.csv") %>%
  select(-X1)
```

Question 1.3

Extract the year, month, day, week, and hour columns. Then, drop any row that has NA in either latitude or longtitude.

```
thefts <- thefts %>%
  mutate(date=ymd_hms(date)) %>%
  mutate(year=year(date)) %>%
  mutate(month=month(date)) %>%
  mutate(day=day(date)) %>%
  mutate(hour=hour(date))

thefts <- thefts[!is.na(thefts$latitude),]
thefts <- thefts[!is.na(thefts$longitude),]
```

Question 1.4

Create a new column called classification.

```
thefts <- thefts %>%
  mutate(classification=ifelse(description=='$500 AND UNDER'|description=='POCKET-PICKING'|description=
```

SECTION 2

Question 2.1

First, convert latitude and longitude columns into one column called geometry. Then, only keep data in the first two month in 2019.

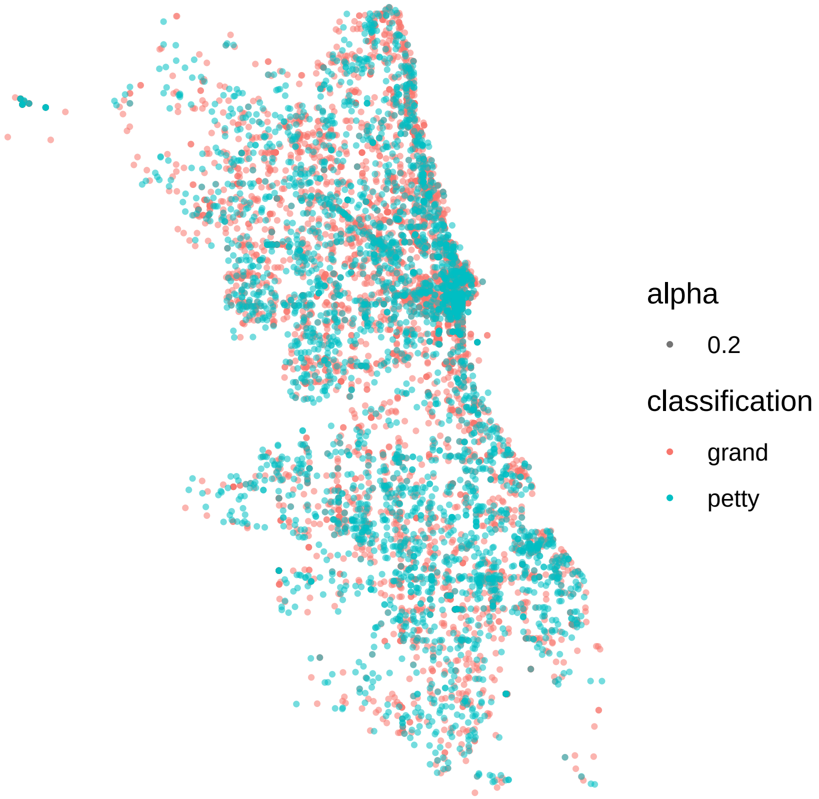
```
thefts <- st_as_sf(thefts,coords = c('longitude','latitude'),crs=4326,remove = FALSE)

thefts_fil <- thefts %>%
  filter(year==2019&(month==1|month==2))
```

Question 2.2

Try to plot the data in a spatial way.

```
thefts_fil %>%
  ggplot()+
  geom_sf(aes(color=classification,alpha=0.2),size=0.5)+
  theme_void()
```



## Question 2.3 Install the API key.

```
census_api_key("81335e4d249d74381a006496b175d659dd0188b4", install=TRUE, overwrite=TRUE)
```

```
## [1] "81335e4d249d74381a006496b175d659dd0188b4"
```

Use read_csv to read data since the connection problem still exsists.

```
cook <- st_read('Chicago_population_tracts.shp')
```

```
## Reading layer `Chicago_population_tracts' from data source `/Users/cyan/R/HM3/Chicago_population_trac
## replacing null geometries with empty geometries
## Simple feature collection with 1319 features and 5 fields (with 1 geometry empty)
## geometry type:  POLYGON
## dimension:      XY
## bbox:           xmin: -88.26364 ymin: 41.46971 xmax: -87.52416 ymax: 42.15429
## CRS:            4326
```

Merge thefts and cook database.

```
thefts_merged <- cook %>%
  st_join(thefts,by='geometry')
```

Question 2.4

Set geometry to NULL.

```
st_geometry(thefts_merged) <- NULL
```

Group by and summarize. Then assign it to a new dataframe called thefts_agg.

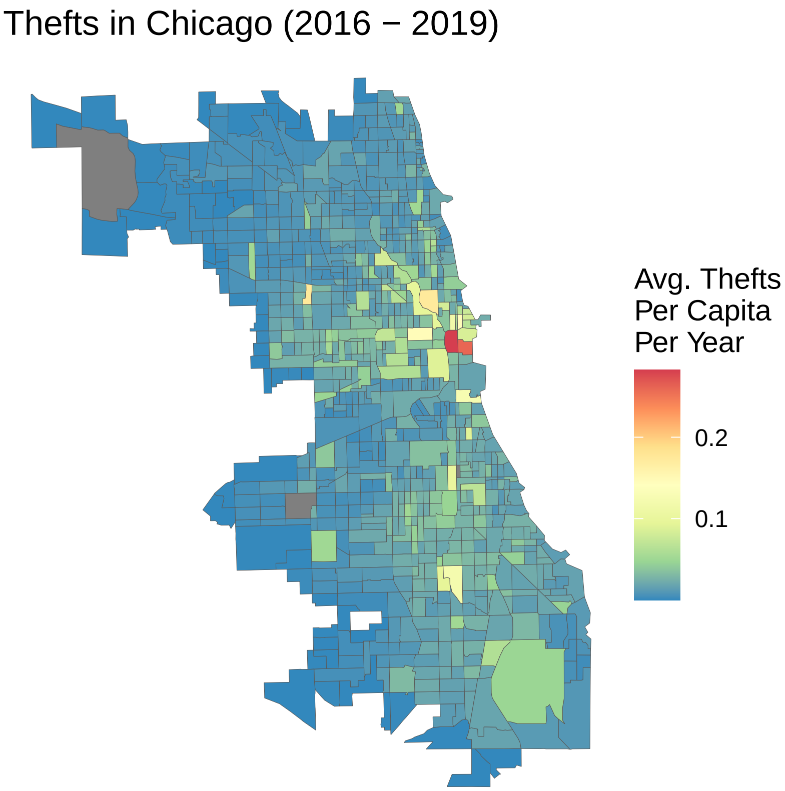
```
thefts_agg <- thefts_merged %>%
  filter(!is.na(year)) %>%
  group_by(GEOID,year) %>%
  summarise(count=n()) %>%
  group_by(GEOID) %>%
  summarise(AVG_TFTS_YEAR=mean(count))
```

Join thefts_agg to cook.

```
cook <- cook %>%
  left_join(thefts_agg,by='GEOID') %>%
  na.omit() %>%
  mutate(thefts_pc=AVG_TFTS_YEAR/estimate)
```

Replicate the plot.

```r
cook %>% ggplot()+
  geom_sf(aes(fill=thefts_pc),lwd=0)+
  theme_void()+
  scale_fill_distiller(palette = "Spectral")+
  labs(caption = 'Source: City of Chicago Data Portal',
       title = 'Thefts in Chicago (2016 - 2019)',
       fill='Avg. Thefts\nPer Capita\nPer Year')
```



Thefts in Chicago (2016 − 2019)

Source: City of Chicago Data Portal

It might because communities there have a higher proverty rate. Thus, less people are able to afford higher education, which further hurt their opportunities to overcome the dilemma. To make our plot even clearer, firstly, we can add a grid line around each tract. Secondly, we can make our plot interactive. For instance, when we click one specific area, the name of that area could be shown on the screen.

# SECTION 3
## Question 3.1 Read csv data.

```
cook_q3 <- read_csv('cook_q3.csv') %>%
  select(-X1)
```

Convert the long data into wide data. Then, rename some variables (using reshape library).

```
cook_1 <- cook_q3 %>%
  select(-moe) %>%
  spread(key=variable,value=estimate) %>%
  mutate(pct_wht=B02001_002/B01001_001*100) %>%
  mutate(pct_pvt=B17007_002/B01001_001*100) %>%
  mutate(pct_bcl=B23006_023/B01001_001*100) %>%
  rename(c(B01001_001='Population')) %>%
  rename(c(B19013_001='Median_Household_Income'))
```

Merge cook_1 and cook to get the prepared regression dataset.

```
cook_regress <- full_join(cook_1,cook,by="NAME") %>%
  mutate(thefts_pc=thefts_pc*1000)
```

Use linear model to regress.

```
linear_model <- lm(thefts_pc~Population+Median_Household_Income+pct_wht+pct_pvt+pct_bcl, data = cook_reg
```

To make our output looks good, use stargazer library to create a table.

```
stargazer::stargazer(linear_model, type = "latex",
 title = "Regression Table 1",
 header = FALSE,
 dep.var.labels = 'Average Thefts Per 1000 Per Year',
 covariate.labels = c('Population','Median Household Income','Pct. White','Pct. Poverty','Pct. Bachelor
```

From this table, we can see that, population and percentage of white has a negative relationship with thefts. Whereas, Median Household Income, percentage of poverty and percentage of bachelor degree have a positive relationship with thefts. However, in my opinion, the Pct. Bachelor term's coefficient is likely to be incorrect since it is disobeying our common sense. If there are more people who have received higher education in one community, cases of thefts will definately go down.

Question 3.2

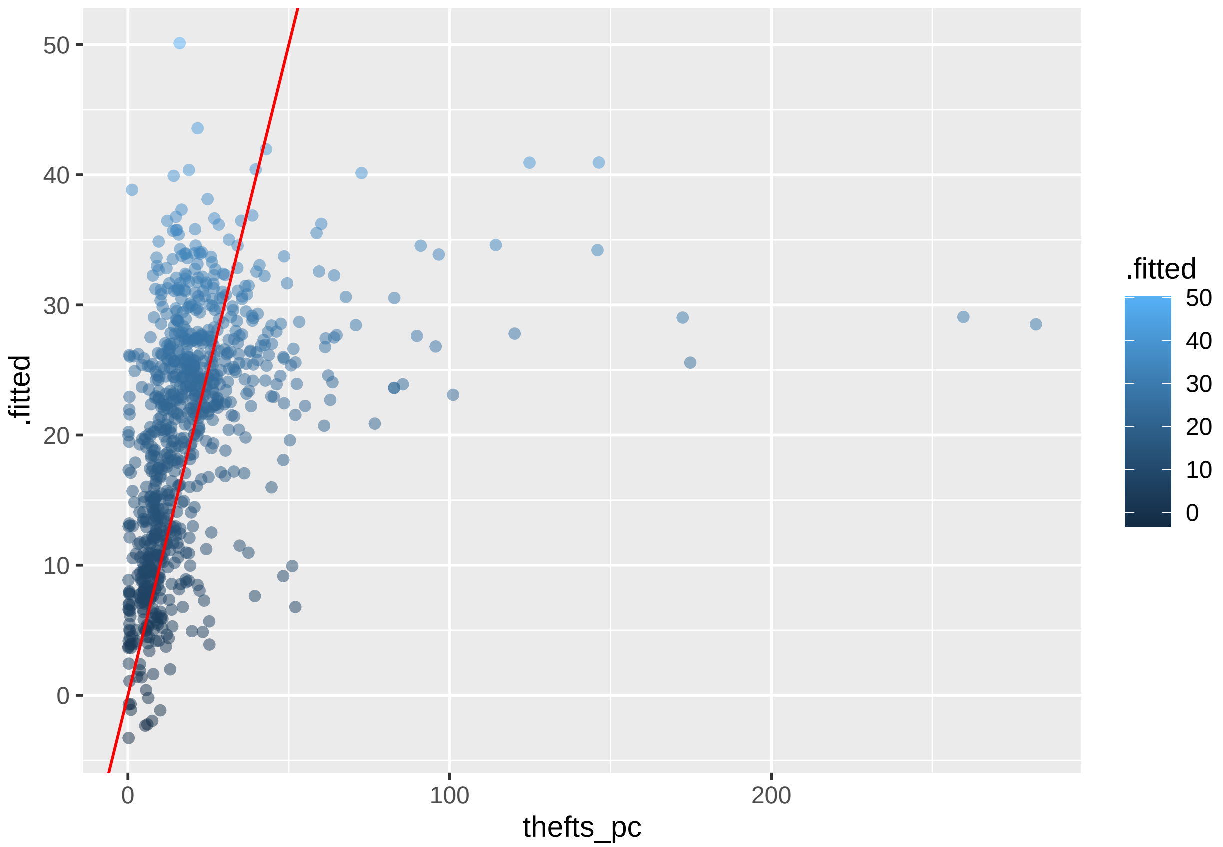
Extract the observed value and pridicted value to plot. Add a 45 degree line to get a reference standard.

```
linear_model_1 <- augment(linear_model)

linear_model_1 %>% ggplot()+
  geom_point(aes(thefts_pc,.fitted,color=.fitted),alpha=0.5)+
  geom_abline(slope = 1,intercept = 0,color='red')
```

Table 1: Regression Table 1

	Dependent variable:
	Average Thefts Per 1000 Per Year
Population	-0.001^{**}
	(0.0004)
Median Household Income	0.0001^{***}
	(0.00004)
Pct. White	-0.279^{***}
	(0.032)
Pct. Poverty	0.517^{***}
	(0.178)
Pct. Bachelor	0.317^{***}
	(0.072)
Constant	17.852^{***}
	(2.792)
Observations	849
R^2	0.168
Adjusted R^2	0.163
Residual Std. Error	20.481 (df = 843)
F Statistic	34.046^{***} (df = 5; 843)
Note:	*p<0.1; **p<0.05; ***p<0.01



From the R square, we can know that the model does not fit well. However, all of the covariables are significant, this is a good sign. The plot suggests that predicted values and observed values fit each other pretty well when thefts_pc are not high. Meanwhile, when thefts_pc are high, the model is not so reliable.

Question 3.3

Extract the standard error.

```
matrix_cov <- vcov(linear_model)
se <- sqrt(diag(matrix_cov))
```

Get robust SE.

```
se_robust <- sqrt(diag(vcovHC(linear_model,type='HC1')))
```

Use stargazer to illustrate.

```
stargazer::stargazer(linear_model,
                     linear_model,
                     type='latex',
                     title='Regression Table 2',
                     header=FALSE,
                     dep.var.labels = 'Average Thefts Per 1000 Per Year',
                     column.labels = c('Robust SE','Normal SE'),
                     covariate.labels = c('Population','Median Household Income','Pct. White','Pct. Pove
                     keep.stat = c('n','rsq'))
```

Table 2: Regression Table 2

	Dependent variable:	
	Average Thefts Per 1000 Per Year	
	Robust SE	Normal SE
	(1)	(2)
Population	-0.001^{**}	-0.001^{**}
	(0.0004)	(0.0004)
Median Household Income	0.0001^{***}	0.0001^{***}
	(0.00004)	(0.00004)
Pct. White	-0.279^{***}	-0.279^{***}
	(0.032)	(0.032)
Pct. Poverty	0.517^{***}	0.517^{***}
	(0.178)	(0.178)
Pct. Bachelor	0.317^{***}	0.317^{***}
	(0.072)	(0.072)
Constant	17.852^{***}	17.852^{***}
	(2.792)	(2.792)
Observations	849	849
R^2	0.168	0.168
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$