

# Homework 2

2025-09-09

Priya Mantraratum Linear Regression Analysis and Visualization

Description of dataset: the Food Environment Atlas from the Economic Research Service of the US Department of Agriculture. Data range from 2012 to 2013 but do not include every year and are sometimes grouped into three-year averages, percent changes over an interval of a few years, etc. County-level data and state-level data are both included (as with the years, not consistently). I choose a target variable measured at the county level, so I restrict my predictor variables to the county level and to years preceding the time range of the target.

Continuous target variable: Population, low access to store (% change), 2015-19 [PCH\_LACCESS\_POP\_15\_19] [ACCESS]

Continuous predictor variables: nine variables with per capita alternative for each to consider later on  
1. Grocery stores (% change), 2016-20 [PCH\_GROC\_16\_20] [STORES] Grocery stores/1,000 pop (% change), 2016-20 [PCH\_GROCPHT\_16\_20] 2. Supercenters & club stores (% change), 2016-20 [PCH\_SUPERC\_16\_20] Supercenters & club stores/1,000 pop (% change), 2016-20 [PCH\_SUPERCPHT\_16\_20] 3. Convenience stores (% change), 2016-20 [PCH\_CONVS\_16\_20] Convenience stores/1,000 pop (% change), 2016-20 [PCH\_CONVSPHT\_16\_20] 4. Specialized food stores (% change), 2016-20 [PCH\_SPECS\_16\_20] Specialized food stores/1,000 pop (% change), 2016-20 [PCH\_SPECSPHT\_16\_20] 5. SNAP-authorized stores (% change), 2017-23 [PCH\_SNAPS\_17\_23] SNAP-authorized stores/1,000 pop (% change), 2017-23 [PCH\_SNAPSPTH\_17\_23] 6. WIC-authorized stores (% change), 2016-22 [PCH\_WICS\_16\_22] WIC-authorized stores/1,000 pop (% change), 2016-22 [PCH\_WICSPHT\_16\_22] 7. Fast-food restaurants (% change), 2016-20 [PCH\_FFR\_16\_20] Fast-food restaurants/1,000 pop (% change), 2016-20 [PCH\_FFRPTH\_16\_20] 8. Full-service restaurants (% change), 2016-20 [PCH\_FSR\_16\_20] Full-service restaurants/1,000 pop (% change), 2016-20 [PCH\_FSRPTH\_16\_20] 9. Direct farm sales (% change), 2012 - 17 [PCH\_DIRSALES\_12\_17] [LOCAL] Direct farm sales per capita (% change), 2012 - 17 [PCH\_PC\_DIRSALES\_12\_17]

Binary variable to add later on: Persistent-poverty counties, 2017-21 [PERPOV17\_21] [SOCIOECONOMIC]  
`install.packages("ggfortify", repos="http://cran.us.r-project.org")`

```
## Installing package into 'C:/Users/harip/AppData/Local/R/win-library/4.5'  
## (as 'lib' is unspecified)  
  
## package 'ggfortify' successfully unpacked and MD5 sums checked  
##  
## The downloaded binary packages are in  
##   C:\Users\harip\AppData\Local\Temp\Rtmpwlv47M\downloaded_packages  
install.packages("mvnormtest", repos="http://cran.us.r-project.org")  
  
## Installing package into 'C:/Users/harip/AppData/Local/R/win-library/4.5'  
## (as 'lib' is unspecified)  
  
## package 'mvnormtest' successfully unpacked and MD5 sums checked  
##  
## The downloaded binary packages are in  
##   C:\Users\harip\AppData\Local\Temp\Rtmpwlv47M\downloaded_packages
```

```

install.packages("datarium", repos="http://cran.us.r-project.org")

## Installing package into 'C:/Users/harip/AppData/Local/R/win-library/4.5'
## (as 'lib' is unspecified)

## package 'datarium' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##   C:\Users\harip\AppData\Local\Temp\Rtmpwlv47M\downloaded_packages
install.packages("ggplot2", repos="http://cran.us.r-project.org")

## Installing package into 'C:/Users/harip/AppData/Local/R/win-library/4.5'
## (as 'lib' is unspecified)

## package 'ggplot2' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##   C:\Users\harip\AppData\Local\Temp\Rtmpwlv47M\downloaded_packages
install.packages("car", repos="http://cran.us.r-project.org")

## Installing package into 'C:/Users/harip/AppData/Local/R/win-library/4.5'
## (as 'lib' is unspecified)

## package 'car' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##   C:\Users\harip\AppData\Local\Temp\Rtmpwlv47M\downloaded_packages
library(MASS)
library(car)

## Loading required package: carData
library(datarium)
library(ggplot2)
library(broom)
library(ggfortify)
library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr     1.1.4    v readr     2.1.5
## vforcats   1.0.0    v stringr   1.5.1
## v lubridate 1.9.4    v tibble    3.3.0
## v purrr    1.1.0    v tidyr    1.3.1

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()
## x dplyr::recode() masks car::recode()
## x dplyr::select() masks MASS::select()
## x purrr::some()  masks car::some()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
library(mvnormtest)

```

Scatter plots and regression:

```

library(readxl)
X2025_food_environment_atlas_data <- read_excel("2025-food-environment-atlas-data.xlsx",
  sheet = "ACCESS", skip = 1)
access <- read_excel("2025-food-environment-atlas-data.xlsx",
  sheet = "ACCESS", skip = 1)
access

## # A tibble: 3,144 x 68
##   FIPS State County LACCESS_POP15 LACCESS_POP19 PCH_LACCESS_POP_15_19
##   <chr> <chr> <chr>     <dbl>      <dbl>          <dbl>
## 1 01001 AL Autauga    18093.    18503.         2.27
## 2 01003 AL Baldwin    46400.    45789.        -1.32
## 3 01005 AL Barbour    6684.     5634.        -15.7
## 4 01007 AL Bibb       296.      365.         23.5
## 5 01009 AL Blount    5856.     3902.        -33.4
## 6 01011 AL Bullock   6100.     7480.         22.6
## 7 01013 AL Butler    2478.     2508.         1.23
## 8 01015 AL Calhoun   34221.    42575.        24.4
## 9 01017 AL Chambers   6794.     6745.        -0.720
## 10 01019 AL Cherokee  3519.     3506.        -0.358
## # i 3,134 more rows
## # i 62 more variables: PCT_LACCESS_POP15 <dbl>, PCT_LACCESS_POP19 <dbl>,
## #   LACCESS_LOWI15 <dbl>, LACCESS_LOWI19 <dbl>, PCH_LACCESS_LOWI_15_19 <dbl>,
## #   PCT_LACCESS_LOWI15 <dbl>, PCT_LACCESS_LOWI19 <dbl>, LACCESS_HHNV15 <dbl>,
## #   LACCESS_HHNV19 <dbl>, PCH_LACCESS_HHNV_15_19 <dbl>,
## #   PCT_LACCESS_HHNV15 <dbl>, PCT_LACCESS_HHNV19 <dbl>, LACCESS_SNAP15 <dbl>,
## #   LACCESS_SNAP19 <dbl>, PCH_LACCESS_SNAP_15_19 <dbl>, ...

library(readxl)
X2025_food_environment_atlas_data <- read_excel("2025-food-environment-atlas-data.xlsx",
  sheet = "STORES", skip = 1)
stores <- read_excel("2025-food-environment-atlas-data.xlsx",
  sheet = "STORES", skip = 1)
stores

## # A tibble: 3,144 x 41
##   FIPS State County GROC16 GROC20 PCH_GROC_16_20 GROCPTH16 GROCPTH20
##   <chr> <chr> <chr>   <dbl>  <dbl>        <dbl>      <dbl>          <dbl>
## 1 01001 AL Autauga    3      4        33.3    0.0542        0.0712
## 2 01003 AL Baldwin    29     29        0       0.140        0.126
## 3 01005 AL Barbour    4      5        25      0.155        0.203
## 4 01007 AL Bibb       5      4       -20      0.221        0.181
## 5 01009 AL Blount    5      4       -20      0.0870       0.0691
## 6 01011 AL Bullock   3     -9999     -9999    0.289       -9999
## 7 01013 AL Butler    3      3        0       0.150        0.154
## 8 01015 AL Calhoun   27     21       -22.2    0.235        0.185
## 9 01017 AL Chambers   7      5       -28.6    0.207        0.152
## 10 01019 AL Cherokee  5     -9999     -9999    0.194       -9999
## # i 3,134 more rows
## # i 33 more variables: PCH_GROCPTH_16_20 <dbl>, SUPERC16 <dbl>, SUPERC20 <dbl>,
## #   PCH_SUPERC_16_20 <dbl>, SUPERCPTH16 <dbl>, SUPERCPTH20 <dbl>,
## #   PCH_SUPERCPTH_16_20 <dbl>, CONVS16 <dbl>, CONVS20 <dbl>,
## #   PCH_CONVS_16_20 <dbl>, CONVSPTH16 <dbl>, CONVSPTH20 <dbl>,
## #   PCH_CONVSPTH_16_20 <dbl>, SPECS16 <dbl>, SPECS20 <dbl>,
## #   PCH_SPECS_16_20 <dbl>, SPECSPTH16 <dbl>, SPECSPTH20 <dbl>, ...

```

```

library(readxl)
X2025_food_environment_atlas_data <- read_excel("2025-food-environment-atlas-data.xlsx",
  sheet = "RESTAURANTS", skip = 1)
restaurants <- read_excel("2025-food-environment-atlas-data.xlsx",
  sheet = "RESTAURANTS", skip = 1)
restaurants

## # A tibble: 3,144 x 15
##   FIPS State County   FFR16 FFR20 PCH_FFR_16_20 FFRPTH16 FFRPTH20
##   <chr> <chr> <chr>   <dbl> <dbl>        <dbl>      <dbl>      <dbl>
## 1 01001 AL Autauga     44     45        2.27    0.796    0.801
## 2 01003 AL Baldwin    156    172       10.3    0.751    0.750
## 3 01005 AL Barbour    23     24        4.35    0.891    0.976
## 4 01007 AL Bibb        7      7         0        0.310    0.316
## 5 01009 AL Blount     23     24        4.35    0.400    0.415
## 6 01011 AL Bullock    3      3         0        0.289    0.301
## 7 01013 AL Butler     18     21       16.7    0.898    1.08
## 8 01015 AL Calhoun    95    104       9.47    0.826    0.917
## 9 01017 AL Chambers   29     32       10.3    0.859    0.974
## 10 01019 AL Cherokee   15     18       20      0.582    0.685
## # i 3,134 more rows
## # i 7 more variables: PCH_FFRPTH_16_20 <dbl>, FSR16 <dbl>, FSR20 <dbl>,
## #   PCH_FSR_16_20 <dbl>, FSRPTH16 <dbl>, FSRPTH20 <dbl>, PCH_FSRPTH_16_20 <dbl>

library(readxl)
X2025_food_environment_atlas_data <- read_excel("2025-food-environment-atlas-data.xlsx",
  sheet = "LOCAL", skip = 1)
local <- read_excel("2025-food-environment-atlas-data.xlsx",
  sheet = "LOCAL", skip = 1)
local

## # A tibble: 3,161 x 98
##   FIPS State County DIRSALES_FARMS12 DIRSALES_FARMS17 PCH_DIRSALES_FARMS_1~1
##   <chr> <chr> <chr>   <dbl>          <dbl>          <dbl>
## 1 01001 AL Autauga      51            16           -68.6
## 2 01003 AL Baldwin     103            78           -24.3
## 3 01005 AL Barbour     13             9           -30.8
## 4 01007 AL Bibb         13            11           -15.4
## 5 01009 AL Blount      88            40           -54.5
## 6 01011 AL Bullock     12             2           -83.3
## 7 01013 AL Butler      31            20           -35.5
## 8 01015 AL Calhoun     50            52              4
## 9 01017 AL Chambers    22            13           -40.9
## 10 01019 AL Cherokee    14            14              0
## # i 3,151 more rows
## # i abbreviated name: 1: PCH_DIRSALES_FARMS_12_17
## # i 92 more variables: PCT_LOCLFARM12 <dbl>, PCT_LOCLFARM17 <dbl>,
## #   PCT_LOCLSALE12 <dbl>, PCT_LOCLSALE17 <dbl>, DIRSALES12 <dbl>,
## #   DIRSALES17 <dbl>, PCH_DIRSALES_12_17 <dbl>, PC_DIRSALES12 <dbl>,
## #   PC_DIRSALES17 <dbl>, PCH_PC_DIRSALES_12_17 <dbl>, FMRKT13 <dbl>,
## #   FMRKT18 <dbl>, PCH_FMRKT_13_18 <dbl>, FMRKTPTH13 <dbl>, ...
atlas = merge(access, stores, by.x = "FIPS", by.y = "FIPS")
atlas = merge(atlas, restaurants, by.x = "FIPS", by.y = "FIPS")
atlas = merge(atlas, local, by.x = "FIPS", by.y = "FIPS")

```

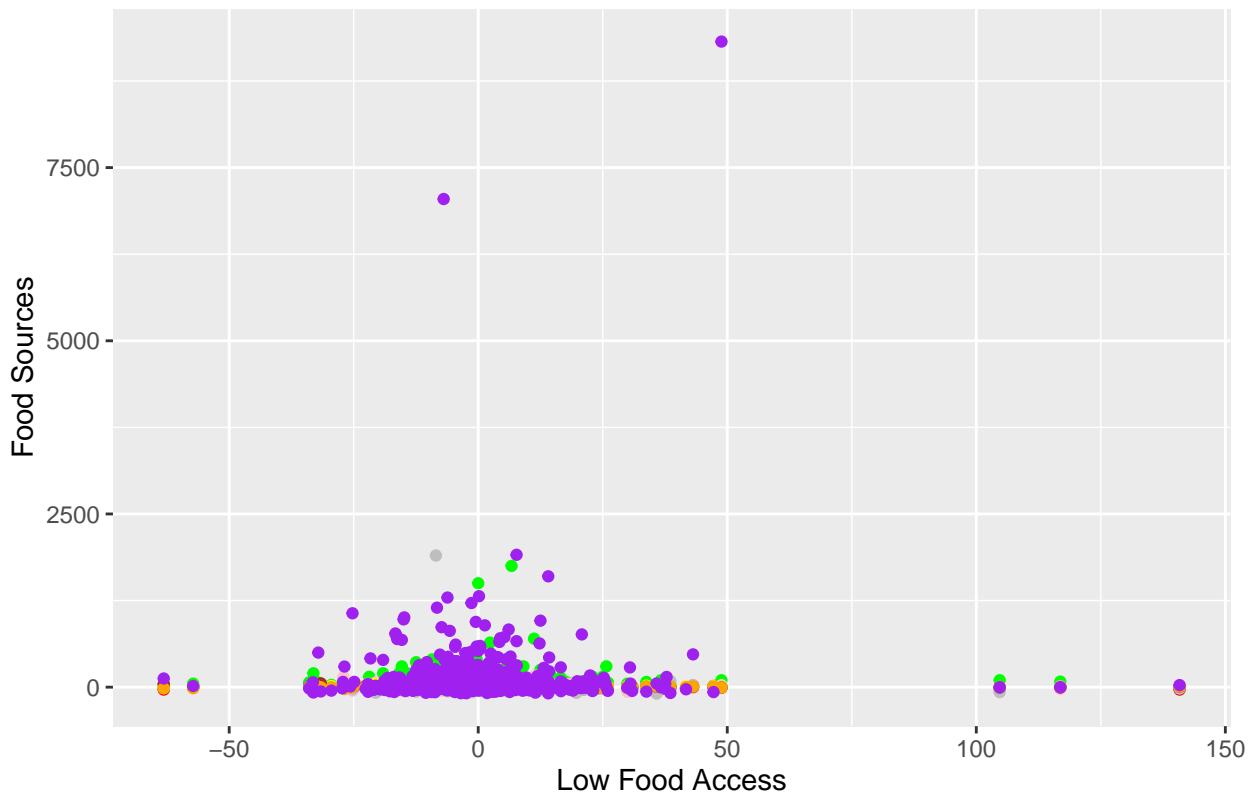
```

## Warning in merge.data.frame(atlas, local, by.x = "FIPS", by.y = "FIPS"): column
## names 'State.x', 'County.x', 'State.y', 'County.y' are duplicated in the result
atlas = select(atlas, PCH_LACCESS_POP_15_19, PCH_GROC_16_20, PCH_SUPERC_16_20, PCH_CONVS_16_20, PCH_SPE
atlas = filter(atlas, PCH_LACCESS_POP_15_19 != -9999, PCH_GROC_16_20 != -9999, PCH_SUPERC_16_20 != -9999
atlas = filter(atlas, PCH_LACCESS_POP_15_19 != -8888, PCH_GROC_16_20 != -8888, PCH_SUPERC_16_20 != -8888

options(repr.plot.width=4, repr.plot.height=4) # Adjust Jupyter plot size
ggplot(atlas) +
  geom_point(aes(y = PCH_GROC_16_20, x = PCH_LACCESS_POP_15_19), color="red")+
  geom_point(aes(y = PCH_SUPERC_16_20, x = PCH_LACCESS_POP_15_19), color="green")+
  geom_point(aes(y = PCH_CONVS_16_20, x = PCH_LACCESS_POP_15_19), color="blue")+
  geom_point(aes(y = PCH_SPECS_16_20, x = PCH_LACCESS_POP_15_19), color="pink")+
  geom_point(aes(y = PCH_SNAPS_17_23, x = PCH_LACCESS_POP_15_19), color="yellow")+
  geom_point(aes(y = PCH_WICS_16_22, x = PCH_LACCESS_POP_15_19), color="gray")+
  geom_point(aes(y = PCH_FFR_16_20, x = PCH_LACCESS_POP_15_19), color="brown")+
  geom_point(aes(y = PCH_FSR_16_20, x = PCH_LACCESS_POP_15_19), color="orange")+
  geom_point(aes(y = PCH_DIRSALES_12_17, x = PCH_LACCESS_POP_15_19), color="purple")+
  labs(title = "Low Food Access vs. Food Sources",
       x = "Low Food Access",
       y = "Food Sources")

```

## Low Food Access vs. Food Sources



```

#There are two points that are clearly outliers at thousands of percents higher than the others, so I am
atlas2 = filter(atlas, PCH_LACCESS_POP_15_19 < 5000, PCH_GROC_16_20 < 5000, PCH_SUPERC_16_20 < 5000, PCH_C
atlas2_long <- atlas2 |>
  pivot_longer(cols = c("PCH_GROC_16_20", "PCH_SUPERC_16_20", "PCH_CONVS_16_20", "PCH_SPECS_16_20", "PCH_S

```

```

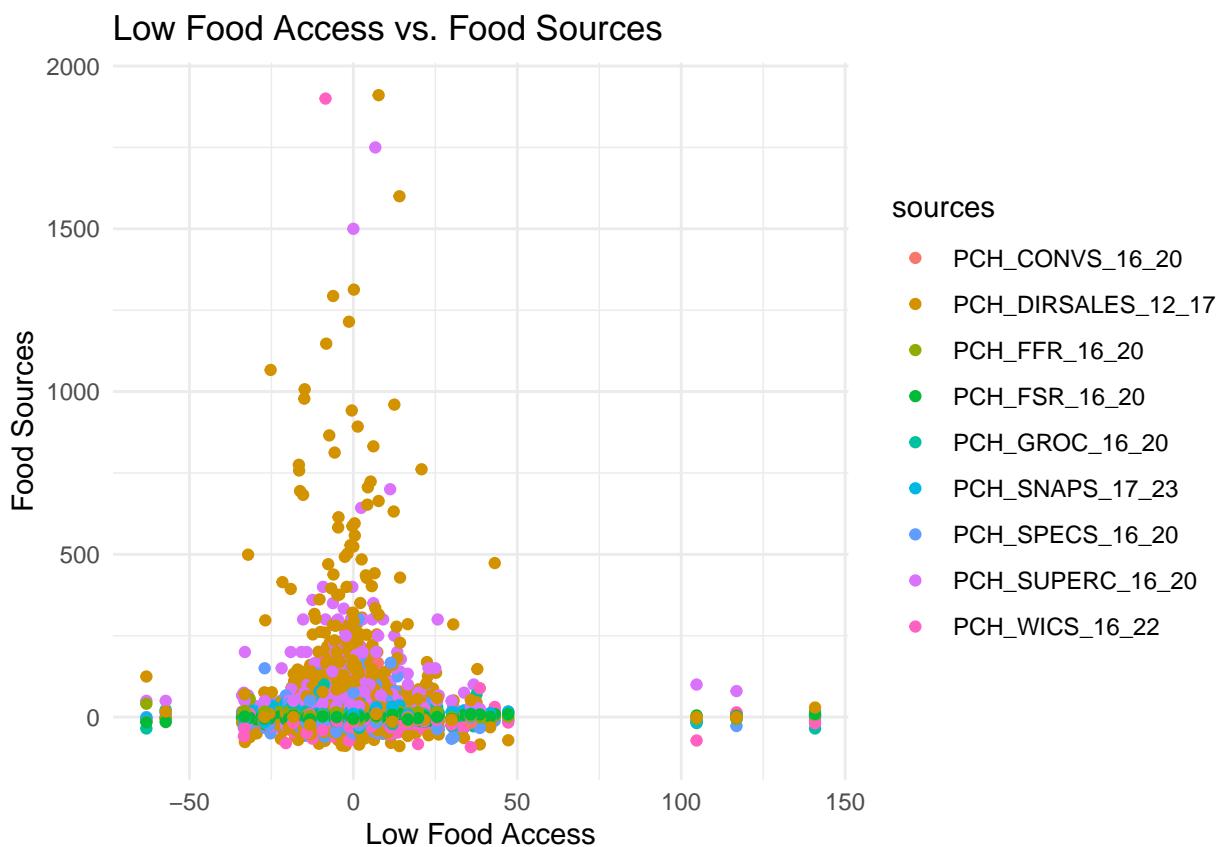
names_to = "sources",
values_to = "food")

atlas2_long

## # A tibble: 5,400 x 3
##   PCH_LACCESS_POP_15_19 sources      food
##   <dbl> <chr>          <dbl>
## 1 -1.32 PCH_GROC_16_20    0
## 2 -1.32 PCH_SUPERC_16_20  28.6
## 3 -1.32 PCH_CONVS_16_20   0
## 4 -1.32 PCH_SPECS_16_20  7.41
## 5 -1.32 PCH_SNAPS_17_23  13.5
## 6 -1.32 PCH_WICS_16_22  -6.90
## 7 -1.32 PCH_FFR_16_20   10.3
## 8 -1.32 PCH_FSR_16_20   4.24
## 9 -1.32 PCH_DIRSALES_12_17 79.0
## 10 -8.19 PCH_GROC_16_20    0
## # i 5,390 more rows

options(repr.plot.width = NULL, repr.plot.height = NULL)
ggplot(atlas2_long, aes(x = PCH_LACCESS_POP_15_19, y = food, color = sources)) +
  geom_point() +
  labs(title = "Low Food Access vs. Food Sources",
       x = "Low Food Access",
       y = "Food Sources") +
  theme_minimal()

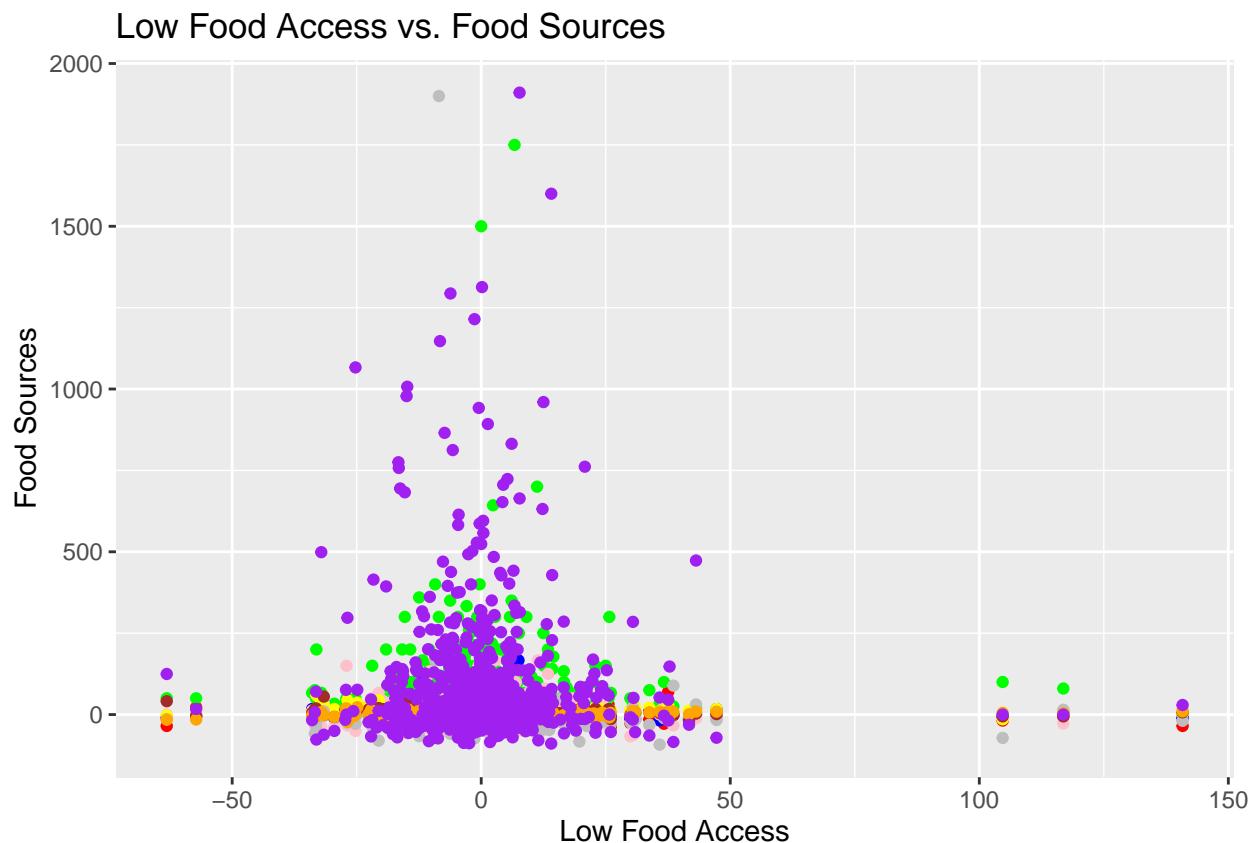
```



```

#trying the other way again just to practice the code
options(repr.plot.width=4, repr.plot.height=4) # Adjust Jupyter plot size
ggplot(atlas2) +
  geom_point(aes(y = PCH_GROC_16_20, x = PCH_LACCESS_POP_15_19), color="red")+
  geom_point(aes(y = PCH_SUPERC_16_20, x = PCH_LACCESS_POP_15_19), color="green")+
  geom_point(aes(y = PCH_CONVS_16_20, x = PCH_LACCESS_POP_15_19), color="blue")+
  geom_point(aes(y = PCH_SPECS_16_20, x = PCH_LACCESS_POP_15_19), color="pink")+
  geom_point(aes(y = PCH_SNAPS_17_23, x = PCH_LACCESS_POP_15_19), color="yellow")+
  geom_point(aes(y = PCH_WICS_16_22, x = PCH_LACCESS_POP_15_19), color="gray")+
  geom_point(aes(y = PCH_FFR_16_20, x = PCH_LACCESS_POP_15_19), color="brown")+
  geom_point(aes(y = PCH_FSR_16_20, x = PCH_LACCESS_POP_15_19), color="orange")+
  geom_point(aes(y = PCH_DIRSALES_12_17, x = PCH_LACCESS_POP_15_19), color="purple")+
  labs(title = "Low Food Access vs. Food Sources",
       x = "Low Food Access",
       y = "Food Sources")

```



```
colMeans(atlas2)
```

## PCH_LACCESS_POP_15_19	PCH_GROC_16_20	PCH_SUPERC_16_20
## 0.2320956	-0.7079341	68.9385445
## PCH_CONVS_16_20	PCH_SPECS_16_20	PCH_SNAPS_17_23
## 2.6007777	-0.6579917	11.6952787
## PCH_WICS_16_22	PCH_FFR_16_20	PCH_FSR_16_20
## -5.6402305	6.6306411	1.8834212
## PCH_DIRSALES_12_17		
## 106.6679969		

```

cor(atlas2)

##                                     PCH_LACCESS_POP_15_19 PCH_GROC_16_20 PCH_SUPER_C_16_20
## PCH_LACCESS_POP_15_19             1.00000000 -0.12825303  0.01054011
## PCH_GROC_16_20                  -0.12825303  1.00000000 -0.03476868
## PCH_SUPER_C_16_20                0.01054011 -0.03476868  1.00000000
## PCH_CONVS_16_20                 -0.04837386 -0.10194873  0.08313836
## PCH_SPECS_16_20                 -0.01327074 -0.03113158 -0.03446943
## PCH_SNAPS_17_23                 -0.07170763  0.14143289 -0.01624929
## PCH_WICS_16_22                  -0.04365784 -0.00269269 -0.03008571
## PCH_FFR_16_20                   -0.17928865  0.14941098 -0.09015193
## PCH_FSR_16_20                   0.04399629  0.08591949 -0.11639025
## PCH_DIRSALES_12_17              -0.04141853  0.05904253  0.06667660
##                                     PCH_CONVS_16_20 PCH_SPECS_16_20 PCH_SNAPS_17_23
## PCH_LACCESS_POP_15_19            -0.048373860 -0.01327074 -0.071707633
## PCH_GROC_16_20                  -0.101948734 -0.03113158  0.141432885
## PCH_SUPER_C_16_20                0.083138361 -0.03446943 -0.016249294
## PCH_CONVS_16_20                 1.000000000  0.03017534  0.001181517
## PCH_SPECS_16_20                 0.030175344  1.00000000  0.078258999
## PCH_SNAPS_17_23                 0.001181517  0.07825900  1.000000000
## PCH_WICS_16_22                  0.018046555 -0.03976880 -0.031334023
## PCH_FFR_16_20                   -0.012903865  0.06940831  0.095304847
## PCH_FSR_16_20                   0.095275731  0.06975211  0.090492833
## PCH_DIRSALES_12_17              0.026334746 -0.01929969 -0.036825560
##                                     PCH_WICS_16_22 PCH_FFR_16_20 PCH_FSR_16_20
## PCH_LACCESS_POP_15_19            -0.043657843 -0.17928865  0.043996289
## PCH_GROC_16_20                  -0.002692690  0.14941098  0.085919490
## PCH_SUPER_C_16_20                -0.030085712 -0.09015193 -0.116390245
## PCH_CONVS_16_20                 0.018046555 -0.01290386  0.095275731
## PCH_SPECS_16_20                 -0.039768796  0.06940831  0.069752112
## PCH_SNAPS_17_23                 -0.031334023  0.09530485  0.090492833
## PCH_WICS_16_22                  1.000000000 -0.01633849  0.005264114
## PCH_FFR_16_20                   -0.016338491  1.00000000  0.063304297
## PCH_FSR_16_20                   0.005264114  0.06330430  1.000000000
## PCH_DIRSALES_12_17              -0.029665757  0.06431184 -0.024182203
##                                     PCH_DIRSALES_12_17
## PCH_LACCESS_POP_15_19            -0.04141853
## PCH_GROC_16_20                  0.05904253
## PCH_SUPER_C_16_20                0.06667660
## PCH_CONVS_16_20                 0.02633475
## PCH_SPECS_16_20                 -0.01929969
## PCH_SNAPS_17_23                 -0.03682556
## PCH_WICS_16_22                  -0.02966576
## PCH_FFR_16_20                   0.06431184
## PCH_FSR_16_20                   -0.02418220
## PCH_DIRSALES_12_17              1.000000000

mshapiro.test(t(atlas2[,1:9]))

##
##  Shapiro-Wilk normality test
##
##  data: Z
##  W = 0.12859, p-value < 2.2e-16

```

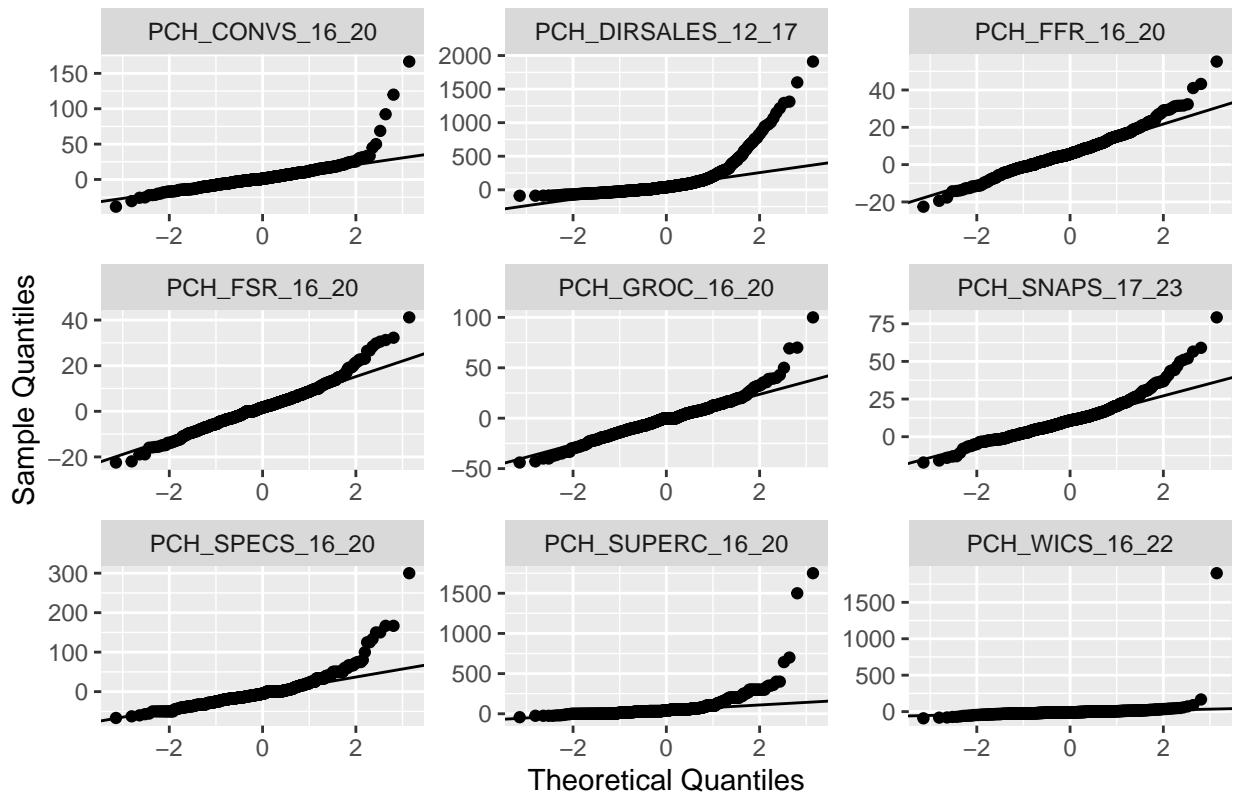
```

options(repr.plot.width = 8, repr.plot.height = 3)

ggplot(atlas2_long, aes(sample = food)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~sources, scales = "free") +
  labs(title = "Q-Q Plots for Food Sources",
       x = "Theoretical Quantiles", y = "Sample Quantiles")

```

## Q-Q Plots for Food Sources



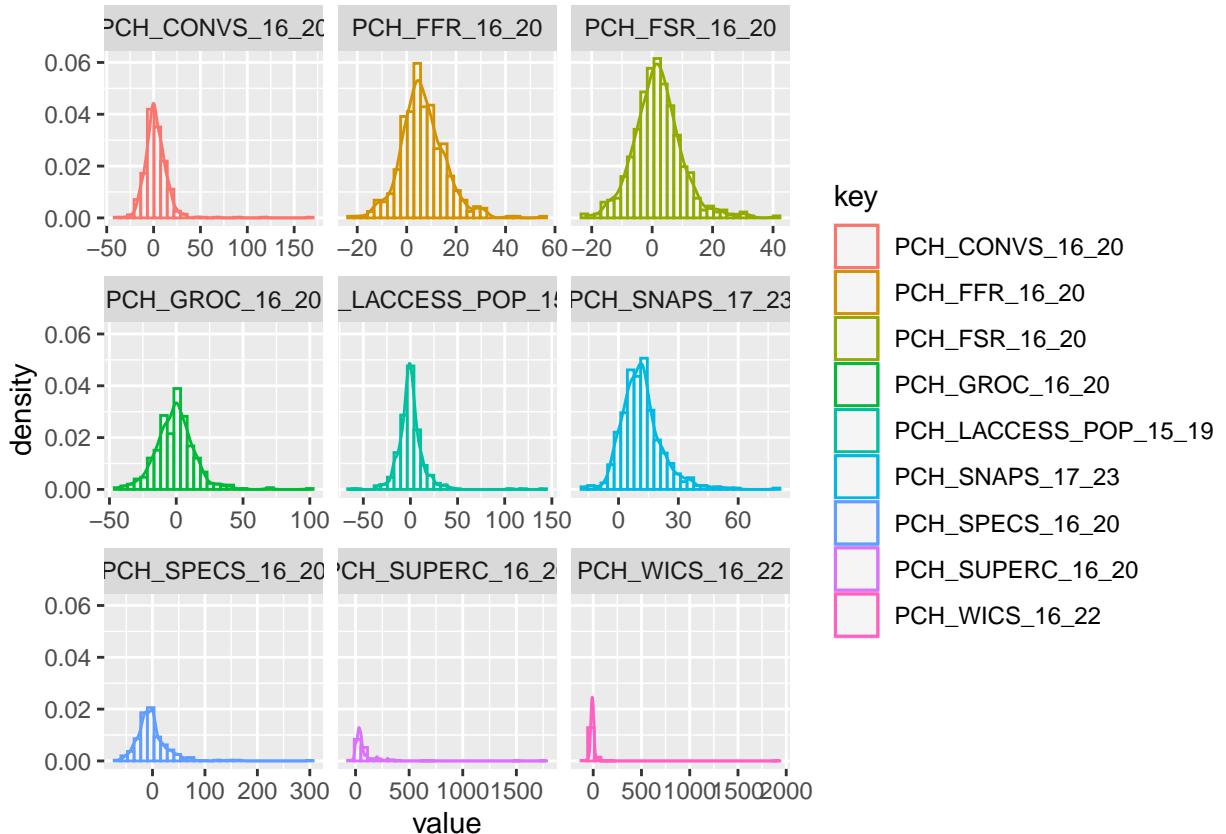
```

options(repr.plot.width = 8, repr.plot.height = 3)

d <- gather(atlas2[,c(1:9)])
ggplot(d,aes(x = value, color=key)) +
  facet_wrap(~key,scales = "free_x") +
  geom_histogram(aes(y=after_stat(density)), alpha=0.5,
                 position="identity",fill="white")+
  geom_density(alpha=.2)

## `stat_bin()` using `bins = 30` . Pick better value `binwidth` .

```



```

model <- lm(PCH_LACCESS_POP_15_19 ~ ., data = atlas)
model

##
## Call:
## lm(formula = PCH_LACCESS_POP_15_19 ~ ., data = atlas)
##
## Coefficients:
##             (Intercept)      PCH_GROC_16_20      PCH_SUPERC_16_20      PCH_CONVS_16_20
##             2.4421555       -0.1130423        0.0001464      -0.0720531
##     PCH_SPECS_16_20      PCH_SNAPS_17_23      PCH_WICS_16_22      PCH_FFR_16_20
##            -0.0015429       -0.0652893       -0.0084057      -0.2785555
##     PCH_FSR_16_20  PCH_DIRSALES_12_17
##            0.1342330        0.0019370

predict(model, newdata=data.frame(PCH_GROC_16_20=1, PCH_SUPERC_16_20=1, PCH_CONVS_16_20=1, PCH_SPECS_16_20=1))

##
##           1
## 2.039583

summary(model)

##
## Call:
## lm(formula = PCH_LACCESS_POP_15_19 ~ ., data = atlas)
##
## Residuals:
##    Min     1Q   Median     3Q    Max 
## -10.00  -4.00   0.00  10.00  20.00
## 
```

```

## -57.244 -6.961 -0.892 4.858 134.229
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.4421555 1.0803043  2.261  0.02415 *
## PCH_GROC_16_20      -0.1130423 0.0413562 -2.733  0.00646 **
## PCH_SUPERC_16_20     0.0001464 0.0050833  0.029  0.97704
## PCH_CONVS_16_20     -0.0720531 0.0434811 -1.657  0.09803 .
## PCH_SPECS_16_20     -0.0015429 0.0188463 -0.082  0.93478
## PCH_SNAPS_17_23     -0.0652893 0.0574588 -1.136  0.25630
## PCH_WICS_16_22     -0.0084057 0.0074930 -1.122  0.26240
## PCH_FFR_16_20       -0.2785555 0.0673914 -4.133 4.09e-05 ***
## PCH_FSR_16_20        0.1342330 0.0763641  1.758  0.07930 .
## PCH_DIRSALES_12_17   0.0019370 0.0011570  1.674  0.09461 .
##
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.72 on 592 degrees of freedom
## Multiple R-squared: 0.06106, Adjusted R-squared: 0.04679
## F-statistic: 4.278 on 9 and 592 DF, p-value: 2.128e-05

confint(model)

##                                2.5 %      97.5 %
## (Intercept)           0.3204602309 4.563850676
## PCH_GROC_16_20      -0.1942650806 -0.031819608
## PCH_SUPERC_16_20     -0.0098371134 0.010129860
## PCH_CONVS_16_20      -0.1574490691 0.013342955
## PCH_SPECS_16_20      -0.0385566098 0.035470893
## PCH_SNAPS_17_23      -0.1781371387 0.047558599
## PCH_WICS_16_22      -0.0231218207 0.006310358
## PCH_FFR_16_20        -0.4109109011 -0.146200082
## PCH_FSR_16_20        -0.0157444978 0.284210540
## PCH_DIRSALES_12_17   -0.0003352232 0.004209238

anova(model)

## Analysis of Variance Table
##
## Response: PCH_LACCESS_POP_15_19
##              Df Sum Sq Mean Sq F value    Pr(>F)
## PCH_GROC_16_20      1   2296   2296.2 10.5973 0.001197 **
## PCH_SUPERC_16_20     1      7      6.8  0.0313 0.859714
## PCH_CONVS_16_20      1    512    512.5  2.3651 0.124614
## PCH_SPECS_16_20      1     15     15.1  0.0696 0.792077
## PCH_SNAPS_17_23      1    358    357.8  1.6514 0.199267
## PCH_WICS_16_22       1    254    253.7  1.1706 0.279710
## PCH_FFR_16_20        1   3662   3661.5 16.8984 4.5e-05 ***
## PCH_FSR_16_20        1    631    630.8  2.9113 0.088486 .
## PCH_DIRSALES_12_17   1    607    607.4  2.8031 0.094613 .
## Residuals            592 128273   216.7
##
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#only the fast food restaurants and grocery stores appear significant, so I will rerun the regression u

```

```

atlas3 = select(atlas, PCH_LACCESS_POP_15_19, PCH_GROC_16_20, PCH_FFR_16_20)

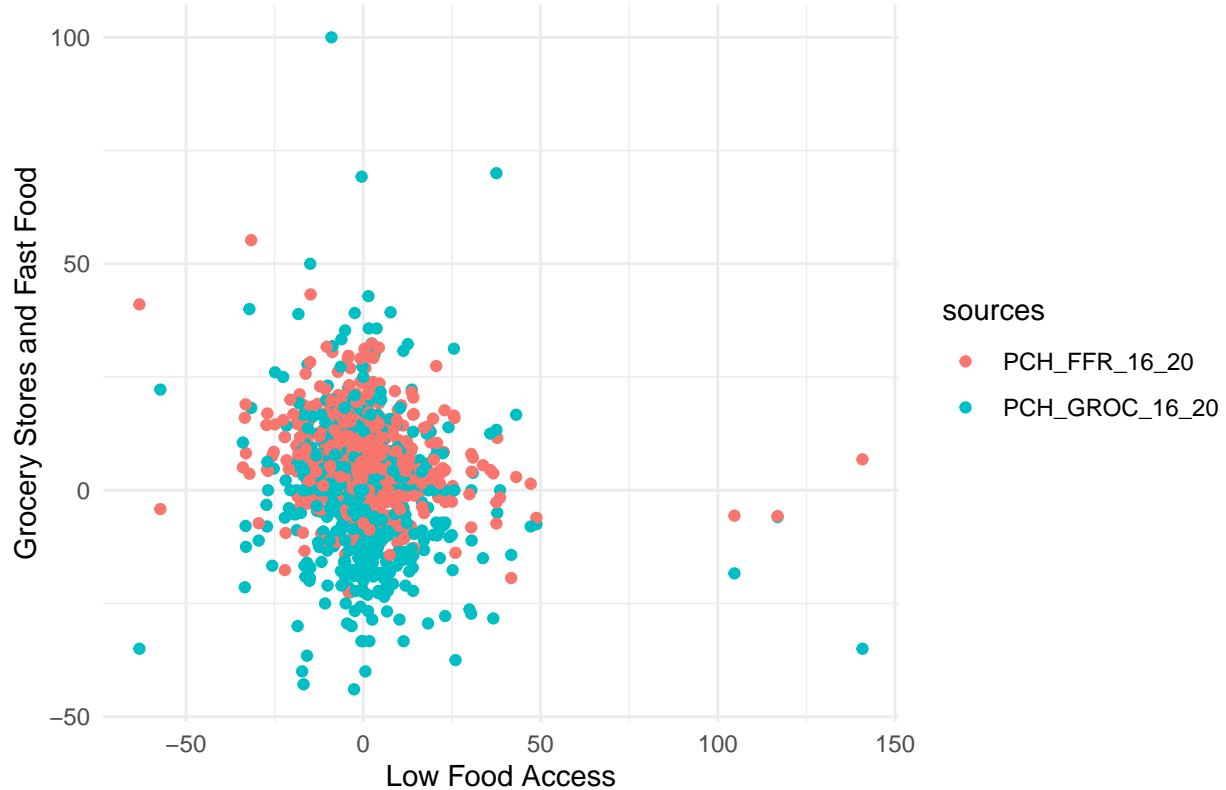
atlas3_long <- atlas3 |>
  pivot_longer(cols = c("PCH_GROC_16_20", "PCH_FFR_16_20"),
               names_to = "sources",
               values_to = "food")
atlas3_long

## # A tibble: 1,204 x 3
##   PCH_LACCESS_POP_15_19 sources      food
##   <dbl>     <chr>      <dbl>
## 1 -1.32    PCH_GROC_16_20  0
## 2 -1.32    PCH_FFR_16_20  10.3
## 3 -8.19    PCH_GROC_16_20  0
## 4 -8.19    PCH_FFR_16_20  4.31
## 5 -0.0797  PCH_GROC_16_20  5.61
## 6 -0.0797  PCH_FFR_16_20 -3.81
## 7 10.3     PCH_GROC_16_20  0
## 8 10.3     PCH_FFR_16_20  8.45
## 9 -3.73    PCH_GROC_16_20 -2.17
## 10 -3.73   PCH_FFR_16_20  9.01
## # i 1,194 more rows

options(repr.plot.width = NULL, repr.plot.height = NULL)
ggplot(atlas3_long, aes(x = PCH_LACCESS_POP_15_19, y = food, color = sources)) +
  geom_point() +
  labs(title = "Low Food Access vs. Grocery Stores and Fast Food",
       x = "Low Food Access",
       y = "Grocery Stores and Fast Food") +
  theme_minimal()

```

## Low Food Access vs. Grocery Stores and Fast Food



```
colMeans(atlas3)
```

```
## PCH_LACCESS_POP_15_19          PCH_GROC_16_20          PCH_FFR_16_20
##      0.3009250       -0.7154586        6.6050853
```

```
cor(atlas3)
```

```
##                                     PCH_LACCESS_POP_15_19 PCH_GROC_16_20 PCH_FFR_16_20
## PCH_LACCESS_POP_15_19           1.0000000   -0.1296448  -0.1846587
## PCH_GROC_16_20                  -0.1296448    1.0000000   0.1501111
## PCH_FFR_16_20                  -0.1846587    0.1501111    1.0000000
```

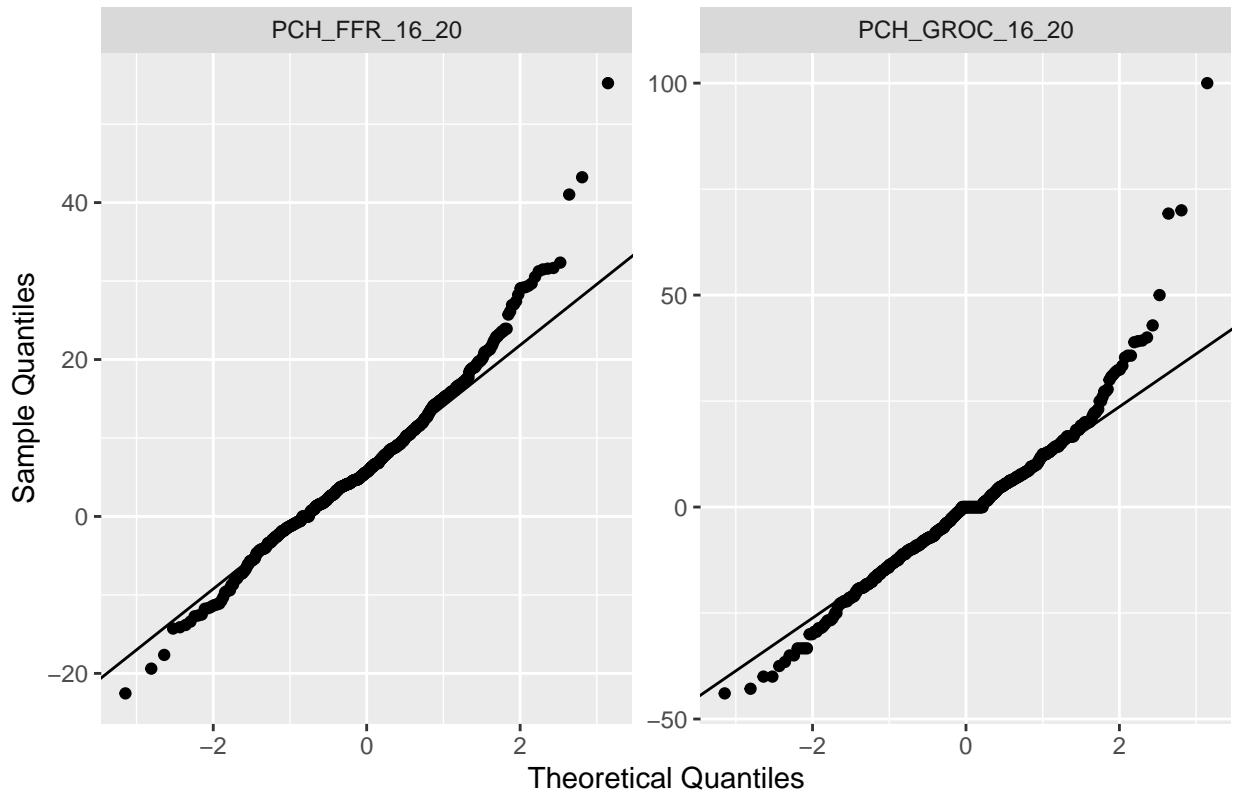
```
mshapiro.test(t(atlas3[,1:3]))
```

```
##
##  Shapiro-Wilk normality test
##
## data: Z
## W = 0.82337, p-value < 2.2e-16
```

```
options(repr.plot.width = 8, repr.plot.height = 3)
```

```
ggplot(atlas3_long, aes(sample = food)) +
  stat_qq() +
  stat_qq_line() +
  facet_wrap(~sources, scales = "free") +
  labs(title = "Q-Q Plots for Food Sources",
       x = "Theoretical Quantiles", y = "Sample Quantiles")
```

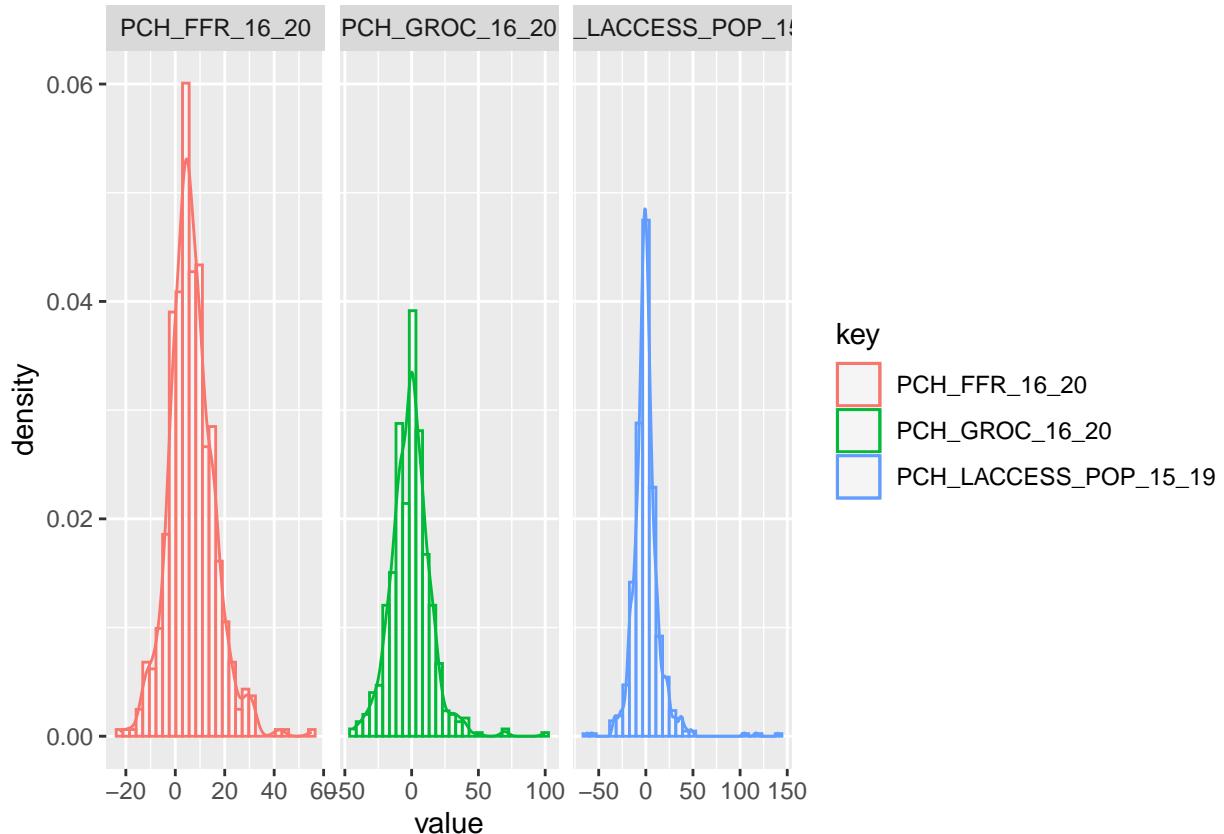
## Q–Q Plots for Food Sources



```
options(repr.plot.width = 8, repr.plot.height = 3)

d <- gather(atlas3[,c(1:3)])
ggplot(d,aes(x = value, color=key)) +
  facet_wrap(~key,scales = "free_x") +
  geom_histogram(aes(y=after_stat(density)), alpha=0.5,
                 position="identity",fill="white")+
  geom_density(alpha=.2)

## `stat_bin()` using `bins = 30`. Pick better value `binwidth`.
```



```

model2 <- lm(PCH_LACCESS_POP_15_19 ~ ., data = atlas3)
model2

##
## Call:
## lm(formula = PCH_LACCESS_POP_15_19 ~ ., data = atlas3)
##
## Coefficients:
##   (Intercept)  PCH_GROC_16_20  PCH_FFR_16_20
##         2.0757        -0.1051        -0.2801
predict(model2, newdata=data.frame(PCH_GROC_16_20=1, PCH_FFR_16_20=1))

##           1
## 1.690526
summary(model2)

##
## Call:
## lm(formula = PCH_LACCESS_POP_15_19 ~ ., data = atlas3)
##
## Residuals:
##    Min     1Q Median     3Q    Max 
## -58.144 -6.782 -1.045  4.968 136.970 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  2.0757    0.0520  40.28  <2e-16 ***
## PCH_GROC_16_20 -0.1051    0.0140 -7.57  1.1e-12 ***
## PCH_FFR_16_20 -0.2801    0.0140 -20.00  <2e-16 ***
##   
```

```

## (Intercept) 2.07570 0.74974 2.769 0.0058 **
## PCH_GROC_16_20 -0.10509 0.04071 -2.582 0.0101 *
## PCH_FFR_16_20 -0.28008 0.06694 -4.184 3.29e-05 ***
##
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.76 on 599 degrees of freedom
## Multiple R-squared: 0.04473, Adjusted R-squared: 0.04154
## F-statistic: 14.02 on 2 and 599 DF, p-value: 1.117e-06
confint(model2)

## 2.5 % 97.5 %
## (Intercept) 0.6032542 3.5481361
## PCH_GROC_16_20 -0.1850357 -0.0251419
## PCH_FFR_16_20 -0.4115448 -0.1486166
anova(model2)

## Analysis of Variance Table
##
## Response: PCH_LACCESS_POP_15_19
## Df Sum Sq Mean Sq F value Pr(>F)
## PCH_GROC_16_20 1 2296 2296.2 10.539 0.001234 **
## PCH_FFR_16_20 1 3814 3814.2 17.507 3.292e-05 ***
## Residuals 599 130504 217.9
##
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

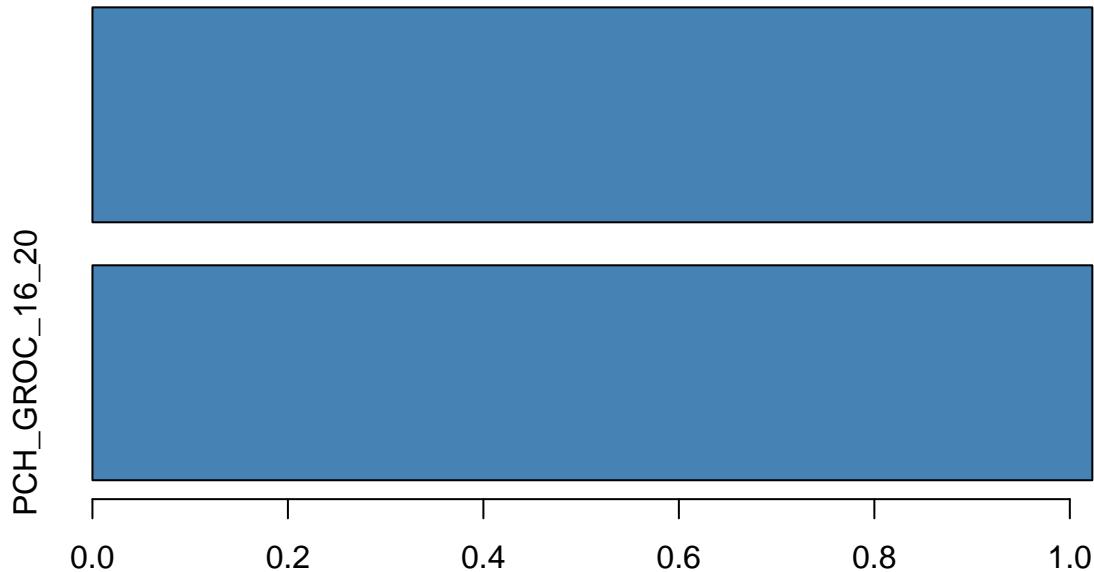
Assumptions: testing linearity, independence, homoskedasticity, etc.

```

vif_values <- vif(model2)           #create vector of VIF values
barplot(vif_values, main = "VIF Values", horiz = TRUE, col = "steelblue") #create horizontal bar chart
abline(v = 5, lwd = 3, lty = 2)      #add vertical line at 5 as after 5 there is severe correlation

```

## VIF Values



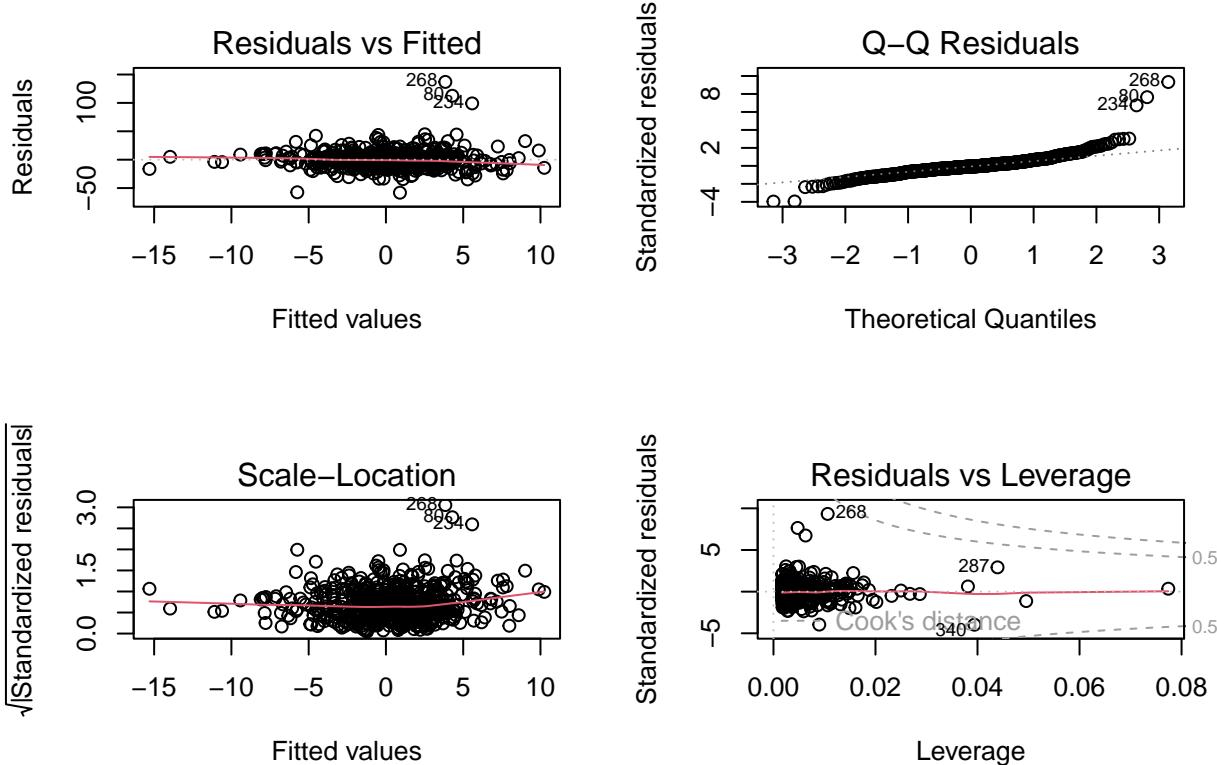
```
data_x <- atlas3[,1:3]                                # independent variables
var <- cor(data_x)                                     # independent variables correlation matrix
var_inv <- ginv(var)         # or solve                # independent variables inverse c
colnames(var_inv) <- colnames(data_x)                  # rename the row names and column names
rownames(var_inv) <- colnames(data_x)
var_inv

##                               PCH_LACCESS_POP_15_19  PCH_GROC_16_20  PCH_FFR_16_20
## PCH_LACCESS_POP_15_19           1.0468213     0.1091574    0.1769190
## PCH_GROC_16_20                  0.1091574     1.0344352   -0.1351233
## PCH_FFR_16_20                  0.1769190    -0.1351233    1.0529531

model2.arg <- augment(model2, se_fit = TRUE)
head(model2.arg)

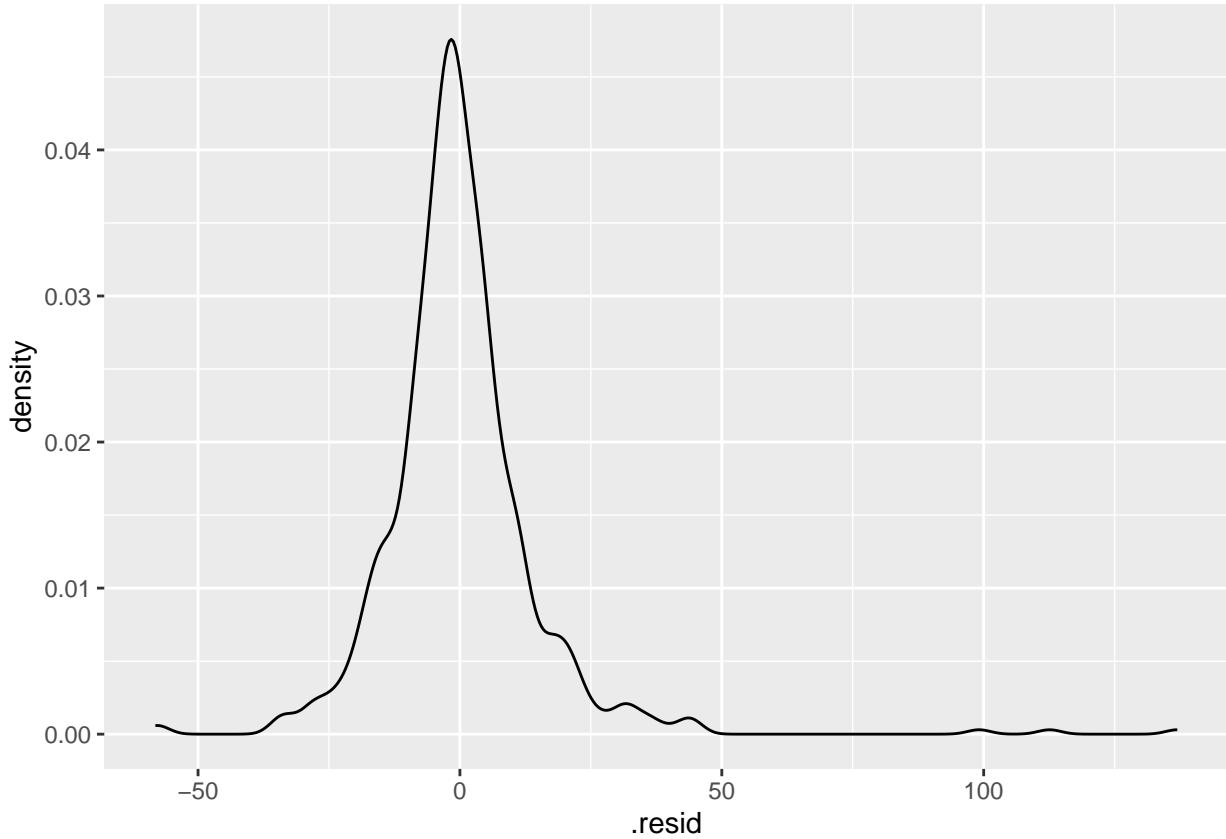
## # A tibble: 6 x 10
##   PCH_LACCESS_POP_15_19  PCH_GROC_16_20  PCH_FFR_16_20 .fitted .se.fit .resid
##   <dbl>            <dbl>            <dbl>        <dbl>    <dbl>    <dbl>
## 1      -1.32             0             10.3       -0.797    0.648   -0.521
## 2      -8.19             0              4.31       0.868    0.623   -9.06
## 3     -0.0797            5.61          -3.81       2.55     0.984   -2.63
## 4      10.3              0              8.45       -0.291    0.614   10.6
## 5     -3.73            -2.17          9.01       -0.219    0.628  -3.51
## 6     -4.41             4.55          17.5       -3.30     0.945  -1.10
## # i 4 more variables: .hat <dbl>, .sigma <dbl>, .cooksdi <dbl>, .std.resid <dbl>
```

```
par(mfrow = c(2, 2))
plot(model2)
```



```
mshapiro.test(t(model2.args$resid))
```

```
##
##  Shapiro-Wilk normality test
##
## data: Z
## W = 0.81791, p-value < 2.2e-16
ggplot(model2.args) +
  geom_density(aes(x=.resid))
```



```
# Add observations indices and
# drop some columns (.se.fit, .sigma) for simplification
model2.arg %>%
  mutate(index = 1:nrow(model2.arg)) %>%
  filter(index %in% c(6,76,131))

## # A tibble: 3 x 11
##   PCH_LACCESS_POP_15_19 PCH_GROC_16_20 PCH_FFR_16_20 .fitted .se.fit .resid
##   <dbl>           <dbl>           <dbl>       <dbl>     <dbl>    <dbl>
## 1      -4.41            4.55          17.5      -3.30    0.945   -1.10
## 2      -0.491           -33.3          7.81      3.39     1.47    -3.88
## 3      14.1             -22.2            0        4.41     1.10    9.65
## # i 5 more variables: .hat <dbl>, .sigma <dbl>, .cooksdi <dbl>,
## # .std.resid <dbl>, index <int>
```

The VIF looks to be slightly higher than 1, suggesting moderate multicollinearity. The Residual v. Fitted Plot suggests linearity because there is a horizontal line without any distinct patterns. The Normal Q-Q Plot suggests that the residuals are normally distributed because they follow the red line. The Scale-Location suggests that we have homoscedasticity because there is a horizontal line with equally spread points.

Results, interpretation, and insights:

Comparing the first and second models, the first model's Shapiro-Wilk test had a W of 0.12859 with a significant p-value, so we cannot say that the data are multivariate normal. The second model's W is much closer to 1 at 0.82337 with a significant p-value, so we can say that the data are multivariate normal.

In the first model, the intercept for grocery stores was -.113 with a p-value of .00646, and the intercept for fast food was -.279 with a p-value of 4.09e-05. In the second model, the intercept for grocery stores was -.105

with a p-value of .0101 (now slightly less significant than before), and the intercept for fast food was -.28008 with a p-value of 3.29e-05.

The first model had a residual standard error of 14.72 on 592 df, an adjusted R<sup>2</sup> of .04679—very low. Its F-statistic was 4.278 with a p-value of 2.128e-05. The second model had a residual standard error of 14.76 on 599 df and an adjusted R<sup>2</sup> of .04154—even lower. Its F-statistic was 14.02 with a p-value of 1.117e-06.

Overall, I need to introduce other variables that are responsible for more of the variation in the response variable in order to bring up the R<sup>2</sup>. I may need to change the dataset that I end up using.

- The Gaussian distribution assumption leads to MLE estimates that are similar in spirit to minimizing the absolute error. Along with the Gauss-Markov assumptions, we also now assume that the error term is normally distributed around the mean at zero.  $E(Y|X_1, \dots, X_p) + \varepsilon \sim N(0, \sigma^2)$
- The MLE method uses the sample mean  $\bar{x}$  and standard deviation  $s$  as inputs for the Gaussian distribution's likelihood function  $L(\mu, \sigma)$ . Then we find the values of  $\theta$  that are the most probable from  $L(\theta) = \prod_{i=1}^n \log P_{\theta}(y_i)$ .
- The concept of absolute error is to measure loss as the positive distance between  $Y$  and  $f(x)$ :  $L(Y, f(x)) = |Y - f(x)|$ . OLS uses squared error,  $L(Y, f(x)) = (Y - f(x))^2$ , which is a similar method that places more weight on differences larger than 1. Using  $P$  and  $\sigma^2$  as inputs for the MLE, the output includes the residual sum of squares. This means that MLE and OLS have outputs in common, so MLE must also be similar to absolute error.

Details of Gauss-Markov Proof, p. 13 on slides:

$$\begin{aligned}
 \text{Var}(\hat{\beta}) &= A \text{Var}(\varepsilon) A^T \\
 &= (X^+ + H^T) \text{Var}(\varepsilon) (X^+ + H^T)^T \\
 &= (X^+ + H^T) \sigma^2 (X^+ + H^T)^T \\
 &= \sigma^2 (X^+ + H^T) (X^+ + H^T)^T = \sigma^2 (X^+ + H^T) ((X^+)^T + H) \\
 &= \sigma^2 (X^+ (X^+)^T + X^+ H + H^T (X^+)^T + H^T H) \\
 &= \sigma^2 [(X^T X)^{-1} X^T ((X^T X)^{-1} X^T)^T + (X^T X)^{-1} X^T H + H^T ((X^T X)^{-1} X^T)^T + H^T H] \\
 &= \sigma^2 [(X^T X)^{-1} X^T X (X^T X)^{-1} + (X^T X)^{-1} X^T H + H^T X (X^T X)^{-1} + H^T H] \\
 &\quad \text{substitute } X^T X (X^T X)^{-1} = I \\
 &\quad \text{substitute } H^T X = 0 \text{ and } (H^T X)^T = 0^T \\
 &\quad \text{substitute } \text{Var}[\hat{\beta}^{OLS}] = \sigma^2 (X^T X)^{-1} \\
 &= \sigma^2 (X^T X)^{-1} + \sigma^2 H^T H \\
 &= \text{Var}[\hat{\beta}^{OLS}] + \sigma^2 H^T H.
 \end{aligned}$$

$$\begin{aligned}
 \sigma^2 H^T H \geq 0 \Rightarrow \text{Var}[\hat{\beta}] &\geq \text{Var}[\hat{\beta}^{OLS}] \\
 \Rightarrow \text{Var}[C^T \hat{\beta}] &\geq \text{Var}[C^T \hat{\beta}^{OLS}].
 \end{aligned}$$