Model

Ben Wallace, Lily Zhu

10/31/2020

```
unemployment_change <- fread(here("data", "unemployment-change.csv"))

mobility_avg <- fread(here("data", "mobility-avg.csv")) %>%
  mutate(county = sub_region_2) %>%
  select(-sub_region_2)

socioeconomic <- fread(here("data", "socioeconomic.csv")) %>%
  mutate(county = area_name) %>%
  select(-area_name)
```

Merge data

```
merged_data <- unemployment_change %>%
  inner_join(mobility_avg, by = "county") %>%
  inner_join(socioeconomic, by = "county") %>%
  select(-parks_avg, -transit_avg) %>%
  na.omit

set.seed(0)
sample_data <- sample_n(merged_data, 500) %>%
  rename(pct_some_college = percent_of_adults_completing_some_college_or_associate_s_degree_2014_18) %>%
  rename(pct_bachelor_or_higher = percent_of_adults_with_a_bachelor_s_degree_or_higher_2014_18) %>%
  rename(pct_only_hs = percent_of_adults_with_a_high_school_diploma_only_2014_18) %>%
  rename(pct_less_hs = percent_of_adults_with_less_than_a_high_school_diploma_2014_18) %>%
  select(county, unemployment_change, retail_avg, grocery_avg, workplaces_avg, residential_avg,
         medhhinc_2018, pct_bachelor_or_higher, pctpovall_2018, medhhinc_2018) %>%
  mutate(medhhinc_2018 = as.numeric(gsub(",", "", medhhinc_2018)))
```

```
glimpse(sample_data)
```

```
## Rows: 500
## Columns: 9
## $ county                <chr> "Oconto County", "Cayuga County", "Napa Coun...
## $ unemployment_change   <dbl> -0.2, 3.9, 4.8, 3.8, 4.4, -0.4, 0.7, 4.3, 2....
## $ retail_avg            <dbl> -18.843137, -11.467391, -28.348485, -8.13756...
## $ grocery_avg           <dbl> 0.9607843, 5.9836957, -10.6894737, -1.891304...
## $ workplaces_avg        <dbl> -18.78804, -23.12887, -31.77835, -22.22165, ...
## $ residential_avg       <dbl> 5.4897959, 8.2546584, 10.3489583, 8.1250000,...
## $ medhhinc_2018         <dbl> 59983, 52945, 85624, 53559, 42909, 50941, 42...
## $ pct_bachelor_or_higher <dbl> 16.0, 22.3, 34.9, 14.9, 12.7, 20.7, 15.6, 40...
## $ pctpovall_2018        <dbl> 9.2, 13.0, 8.8, 17.3, 24.1, 12.0, 24.6, 6.4,...
```

```
sample_data %>%
  mutate()
```

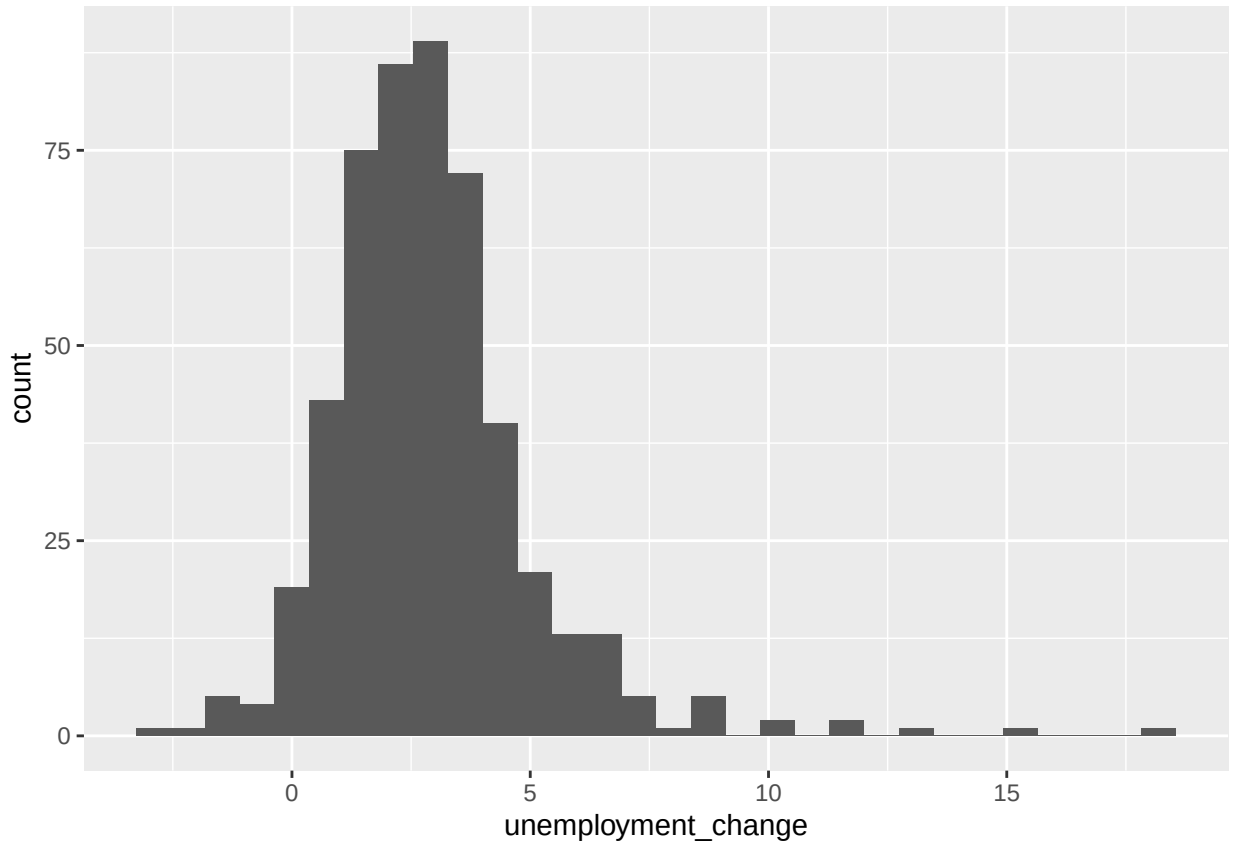
```
##                   county unemployment_change retail_avg grocery_avg
##   1:       Oconto County                -0.2 -18.843137   0.9607843
##   2:       Cayuga County                 3.9 -11.467391   5.9836957
##   3:         Napa County                 4.8 -28.348485 -10.6894737
##   4: San Patricio County                 3.8  -8.137566  -1.8913043
##   5:       Iberia Parish                 4.4  -3.005435   5.5706522
##  ---
## 496:         Polk County                 1.8  -6.327261   6.0765306
## 497: Queen Anne's County                 2.0  -6.157609   3.0706522
## 498:     Monterey County                -1.1 -26.949495   0.3181818
## 499:   Greenville County                 3.2 -15.909091   1.3888889
## 500:        Huron County                 2.0  -8.094862   8.5063830
##      workplaces_avg residential_avg medhhinc_2018 pct_bachelor_or_higher
##   1:      -18.78804        5.489796         59983                   16.0
##   2:      -23.12887        8.254658         52945                   22.3
##   3:      -31.77835       10.348958         85624                   34.9
##   4:      -22.22165        8.125000         53559                   14.9
##   5:      -20.11340        7.218310         42909                   12.7
##  ---
## 496:      -22.78673        7.400654         39048                   13.3
## 497:      -25.79688       12.197080         93751                   34.9
## 498:      -27.13636        8.080808         69665                   24.5
## 499:      -26.99495        8.398990         61162                   34.2
## 500:      -21.38095        5.213483         45817                   15.5
##      pctpovall_2018
##   1:            9.2
##   2:           13.0
##   3:            8.8
##   4:           17.3
##   5:           24.1
##  ---
## 496:           20.0
## 497:            6.5
## 498:           13.3
## 499:           11.1
## 500:           12.6
```

some college pct

Univariate analysis

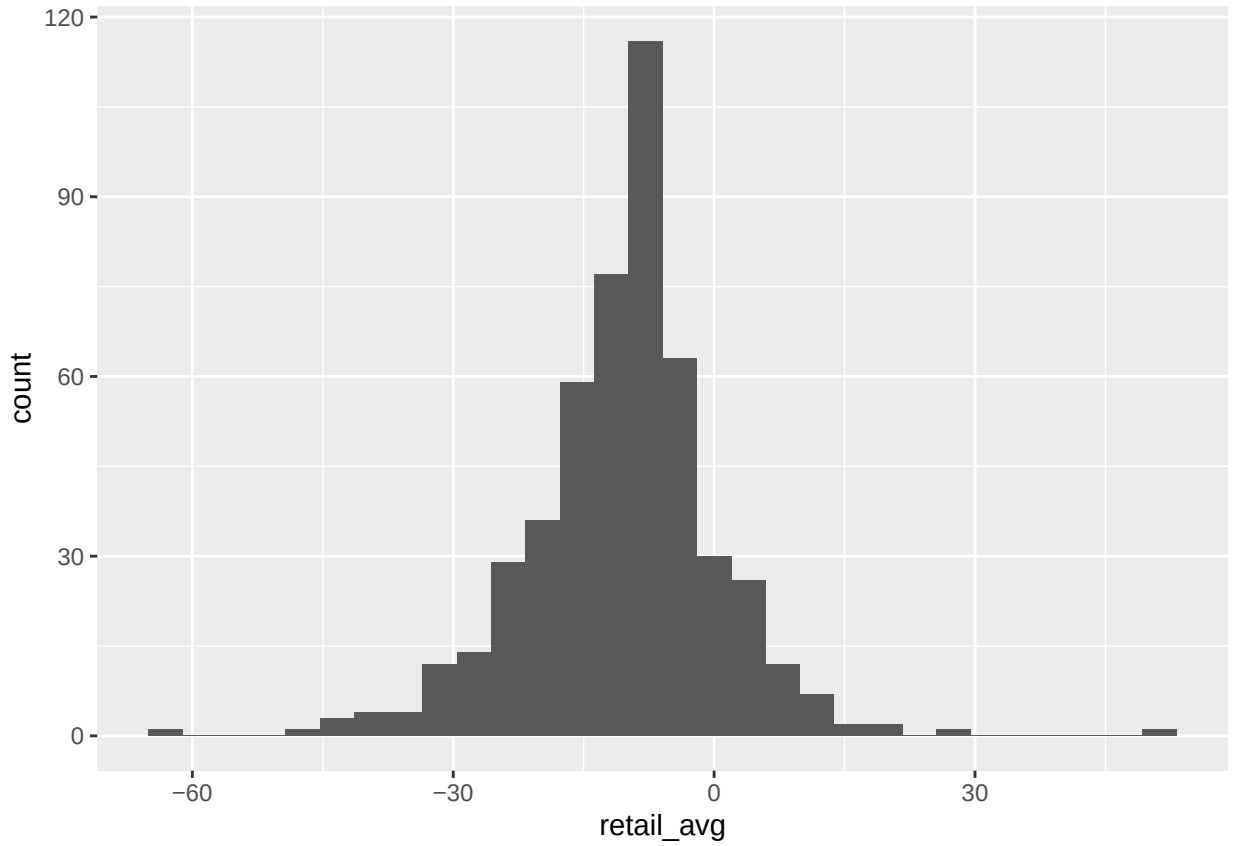
```
ggplot(data = sample_data, mapping = aes(x = unemployment_change)) +
  geom_histogram()
```

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



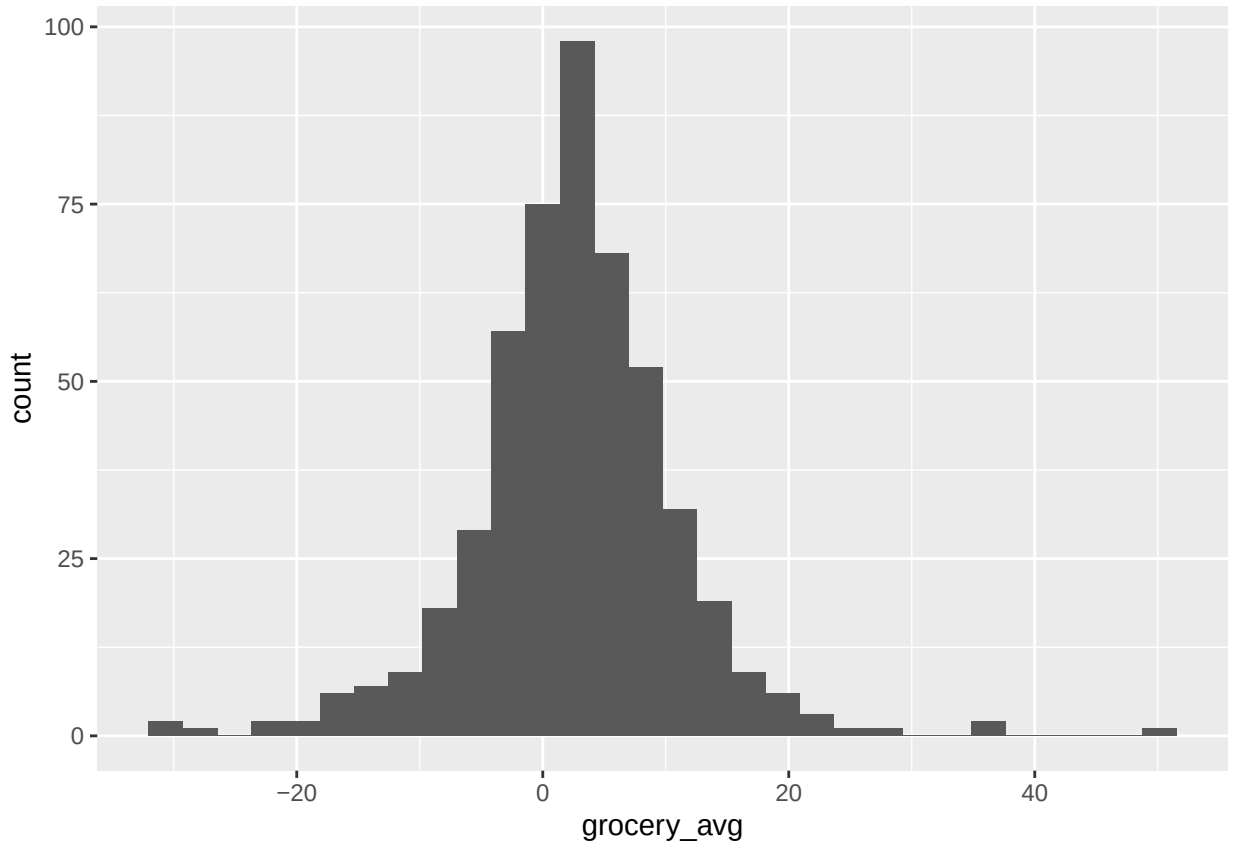

```
ggplot(data = sample_data, mapping = aes(x = retail_avg)) +
  geom_histogram()
```

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



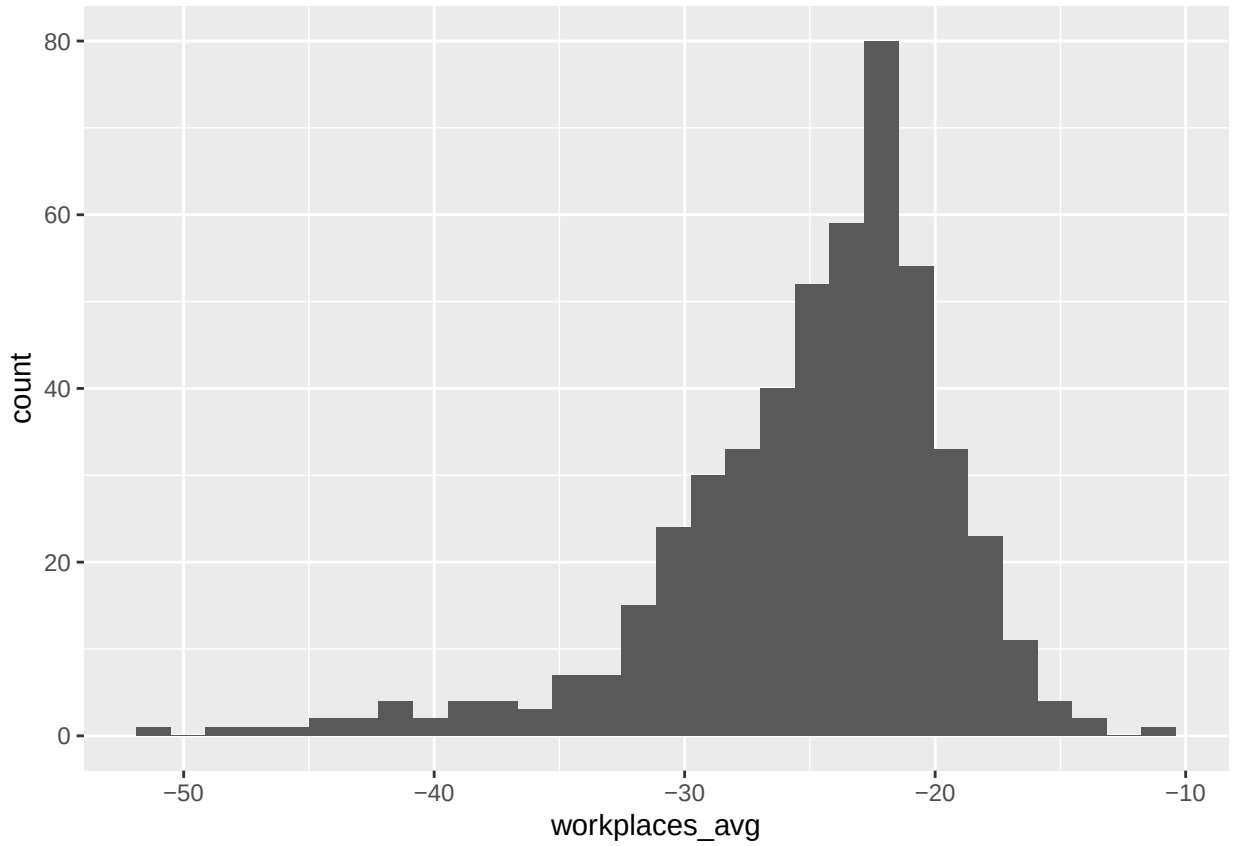

```
ggplot(data = sample_data, mapping = aes(x = grocery_avg)) +
  geom_histogram()
```

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



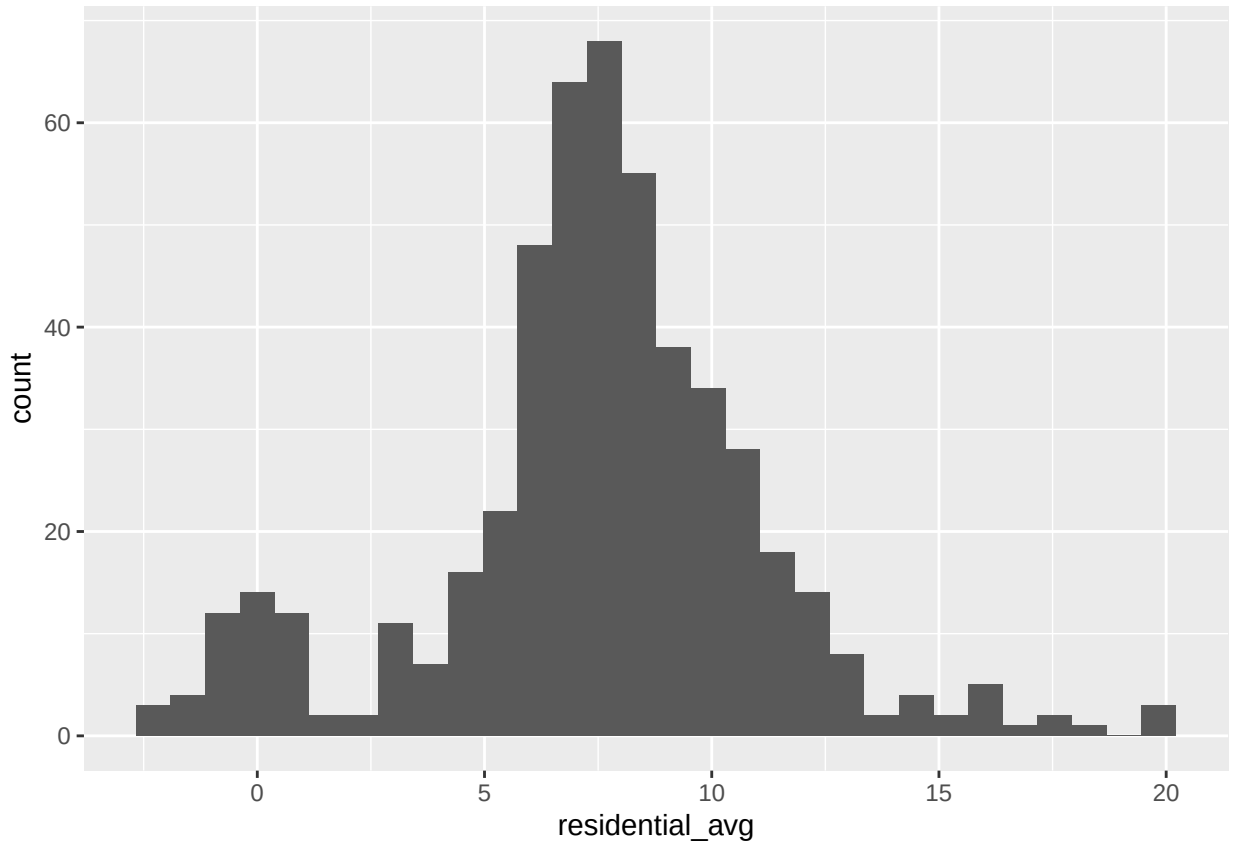

```
ggplot(data = sample_data, mapping = aes(x = workplaces_avg)) +
  geom_histogram()
```

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




```
ggplot(data = sample_data, mapping = aes(x = residential_avg)) +
  geom_histogram()
```

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




```
merged_data %>%
  summarise(mean = mean(unemployment_change),
            median = median(unemployment_change),
            sd = sd(unemployment_change),
            iqr = IQR(unemployment_change),
            min = min(unemployment_change),
            max = max(unemployment_change))
```

```
##       mean median       sd iqr  min  max
## 1 2.923181    2.7 2.044385 2.1 -2.7 18.4
```

```
merged_data %>%
  filter(!is.na(retail_avg)) %>%
  summarise(mean = mean(retail_avg),
            median = median(retail_avg),
            sd = sd(retail_avg),
            iqr = IQR(retail_avg),
            min = min(retail_avg),
            max = max(retail_avg))
```

```
##        mean    median       sd      iqr       min      max
## 1 -9.961328 -9.242424 10.65793 11.10077 -63.05556 56.82065
```

```
merged_data %>%
  filter(!is.na(grocery_avg)) %>%
  summarise(mean = mean(grocery_avg),
            median = median(grocery_avg),
            sd = sd(grocery_avg),
```

```
          iqr = IQR(grocery_avg),
          min = min(grocery_avg),
          max = max(grocery_avg))
```

```
##       mean   median       sd      iqr       min      max
## 1 3.108732 3.024055 8.372305 8.835333 -31.28804 49.50838
```

```
merged_data %>%
  filter(!is.na(workplaces_avg)) %>%
  summarise(mean = mean(workplaces_avg),
          median = median(workplaces_avg),
          sd = sd(workplaces_avg),
          iqr = IQR(workplaces_avg),
          min = min(workplaces_avg),
          max = max(workplaces_avg))
```

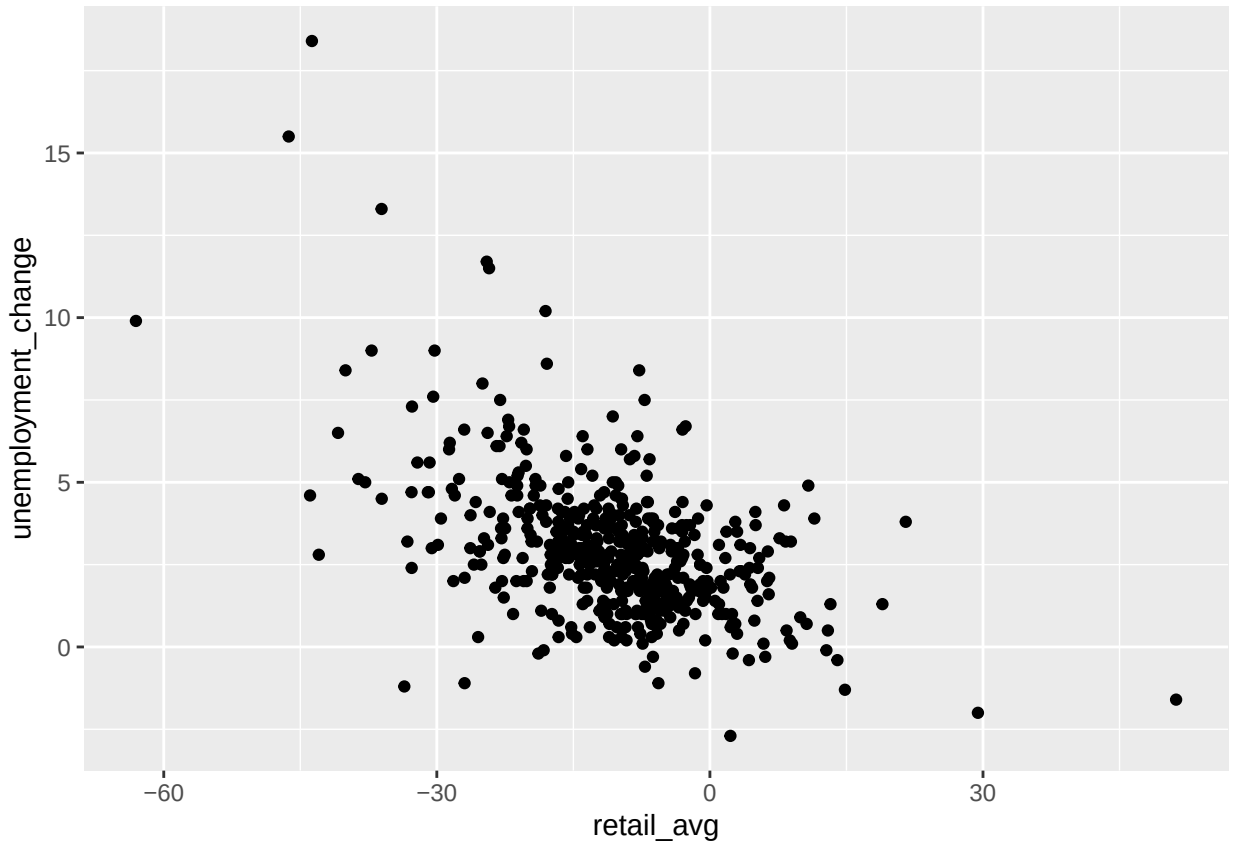
```
##        mean    median       sd      iqr       min       max
## 1 -24.82253 -23.84239 5.208127 5.977886 -50.95455 -10.81522
```

```
merged_data %>%
  filter(!is.na(residential_avg)) %>%
  summarise(mean = mean(residential_avg),
          median = median(residential_avg),
          sd = sd(residential_avg),
          iqr = IQR(residential_avg),
          min = min(residential_avg),
          max = max(residential_avg))
```

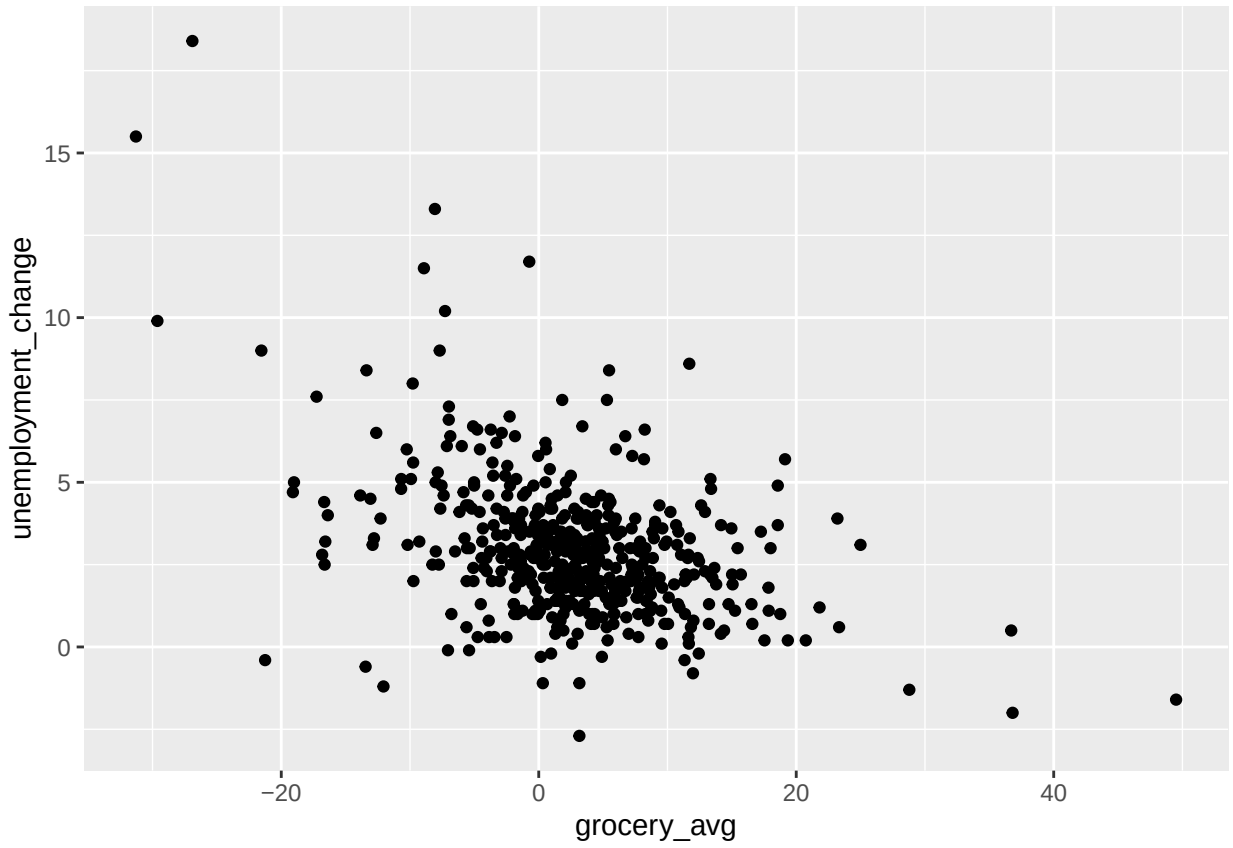
```
##       mean   median       sd      iqr min      max
## 1 7.499232 7.605405 3.451305 3.150776  -3 20.12626
```

Bivariate analysis

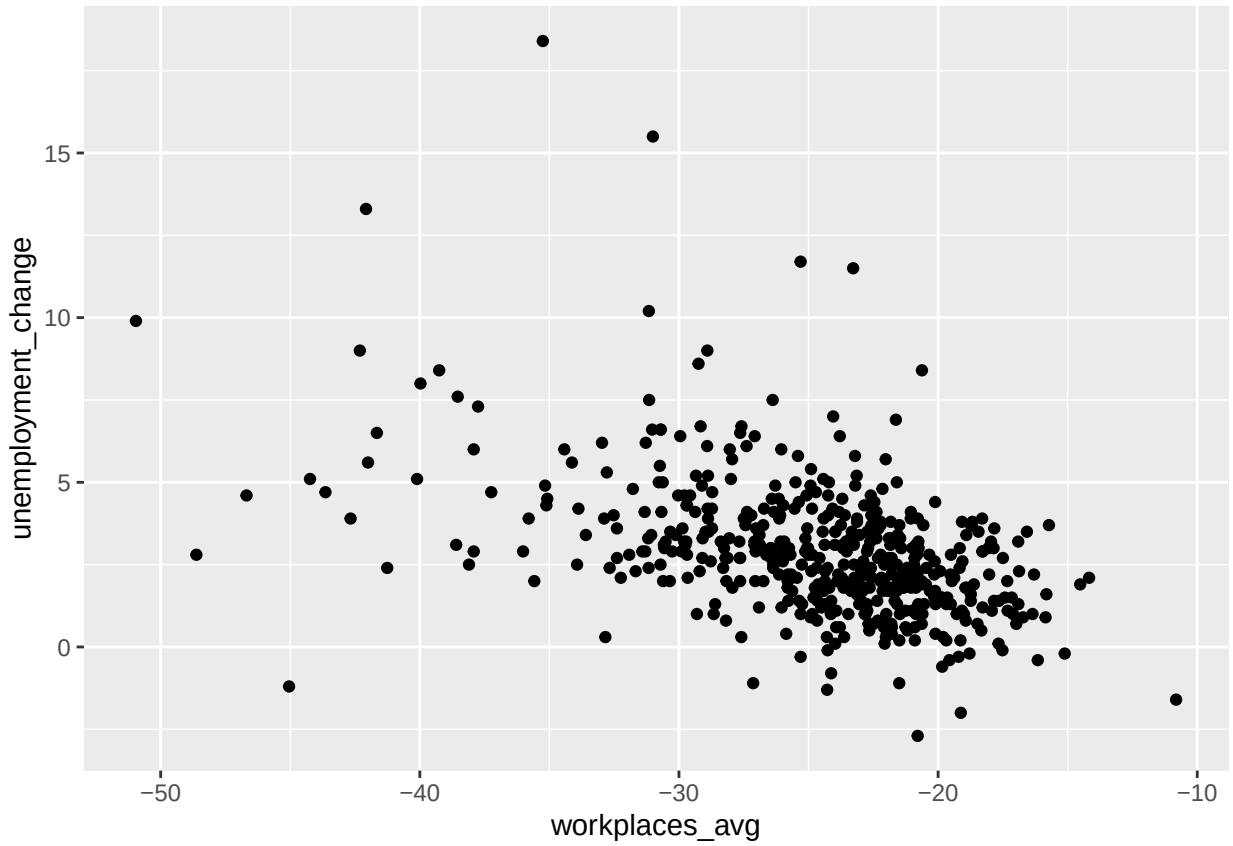
```
sample_data %>%
  ggplot(aes(y = unemployment_change, x = retail_avg)) +
  geom_point()
```



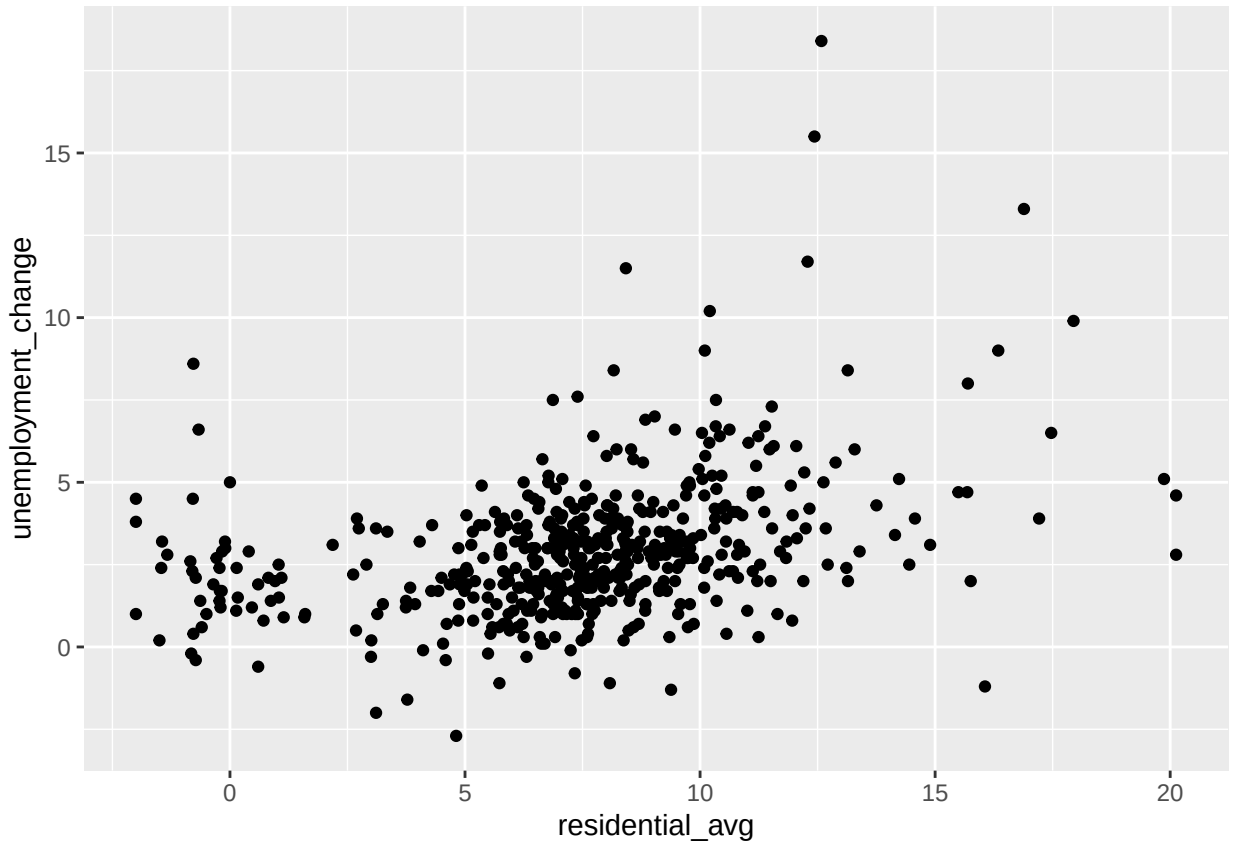

```
sample_data %>%
  ggplot(aes(y = unemployment_change, x = grocery_avg)) +
  geom_point()
```




```
sample_data %>%
  ggplot(aes(y = unemployment_change, x = workplaces_avg)) +
  geom_point()
```




```
sample_data %>%
  ggplot(aes(y = unemployment_change, x = residential_avg)) +
  geom_point()
```



Model

```
model <- lm(unemployment_change ~ retail_avg +
              grocery_avg +
              workplaces_avg +
              residential_avg +
              medhhinc_2018 +
              pct_bachelor_or_higher +
              pctpovall_2018 +
              medhhinc_2018,
            data = sample_data)

model %>%
  tidy(conf.int = TRUE) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.239	0.958	0.250	0.803	-1.643	2.122
retail_avg	-0.062	0.013	-4.667	0.000	-0.088	-0.036
grocery_avg	-0.025	0.014	-1.754	0.080	-0.054	0.003
workplaces_avg	-0.045	0.029	-1.551	0.122	-0.102	0.012
residential_avg	0.016	0.032	0.502	0.616	-0.047	0.080
medhhinc_2018	0.000	0.000	0.182	0.856	0.000	0.000
pct_bachelor_or_higher	0.012	0.013	0.938	0.349	-0.014	0.038
pctpovall_2018	0.030	0.027	1.102	0.271	-0.024	0.084

```r
int_only_model <- lm(unemployment_change ~ 1, data = sample_data)

final_model <- step(model, scope = formula(int_only_model), direction = "backward")
```

```
## Start:  AIC=608.59
## unemployment_change ~ retail_avg + grocery_avg + workplaces_avg +
##     residential_avg + medhhinc_2018 + pct_bachelor_or_higher +
##     pctpovall_2018 + medhhinc_2018
##
##                           Df Sum of Sq    RSS    AIC
## - medhhinc_2018            1     0.110 1635.7 606.62
## - residential_avg          1     0.838 1636.5 606.84
## - pct_bachelor_or_higher   1     2.923 1638.6 607.48
## - pctpovall_2018           1     4.038 1639.7 607.82
## <none>                                 1635.6 608.59
## - workplaces_avg           1     7.999 1643.6 609.03
## - grocery_avg              1    10.223 1645.9 609.70
## - retail_avg               1    72.394 1708.0 628.24
##
## Step:  AIC=606.62
## unemployment_change ~ retail_avg + grocery_avg + workplaces_avg +
##     residential_avg + pct_bachelor_or_higher + pctpovall_2018
##
##                           Df Sum of Sq    RSS    AIC
## - residential_avg          1     1.022 1636.8 604.93
## - pct_bachelor_or_higher   1     4.166 1639.9 605.89
## <none>                                 1635.7 606.62
## - pctpovall_2018           1     6.744 1642.5 606.68
## - workplaces_avg           1     8.462 1644.2 607.20
## - grocery_avg              1    10.434 1646.2 607.80
## - retail_avg               1    72.342 1708.1 626.26
##
## Step:  AIC=604.93
## unemployment_change ~ retail_avg + grocery_avg + workplaces_avg +
##     pct_bachelor_or_higher + pctpovall_2018
##
##                           Df Sum of Sq    RSS    AIC
## - pct_bachelor_or_higher   1     4.190 1641.0 604.21
## - pctpovall_2018           1     6.523 1643.3 604.92
## <none>                                 1636.8 604.93
## - grocery_avg              1    10.543 1647.3 606.14
## - workplaces_avg           1    11.916 1648.7 606.56
## - retail_avg               1    78.728 1715.5 626.42
##
## Step:  AIC=604.21
## unemployment_change ~ retail_avg + grocery_avg + workplaces_avg +
##     pctpovall_2018
##
##                   Df Sum of Sq    RSS    AIC
## - pctpovall_2018   1     3.183 1644.1 603.18
## <none>                         1641.0 604.21
## - grocery_avg      1    10.072 1651.0 605.27
## - workplaces_avg   1    26.528 1667.5 610.23
## - retail_avg       1    79.571 1720.5 625.89
```

```
##
## Step:  AIC=603.18
## unemployment_change ~ retail_avg + grocery_avg + workplaces_avg
##
##                   Df Sum of Sq    RSS    AIC
## <none>                         1644.1 603.18
## - grocery_avg      1    10.191 1654.3 604.27
## - workplaces_avg   1    23.383 1667.5 608.24
## - retail_avg       1    85.512 1729.7 626.53
```

```
final_model %>%
  tidy(conf.int = TRUE) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.838	0.509	1.645	0.101	-0.163	1.838
retail_avg	-0.066	0.013	-5.079	0.000	-0.091	-0.040
grocery_avg	-0.025	0.014	-1.753	0.080	-0.054	0.003
workplaces_avg	-0.059	0.022	-2.656	0.008	-0.103	-0.015

Interaction Term

```
reduced_model <- final_model
full_model <- lm(unemployment_change ~ retail_avg + grocery_avg + workplaces_avg + retail_avg*grocery_a

anova(reduced_model, full_model) %>%
  tidy() %>%
  kable(digits = 3)
```

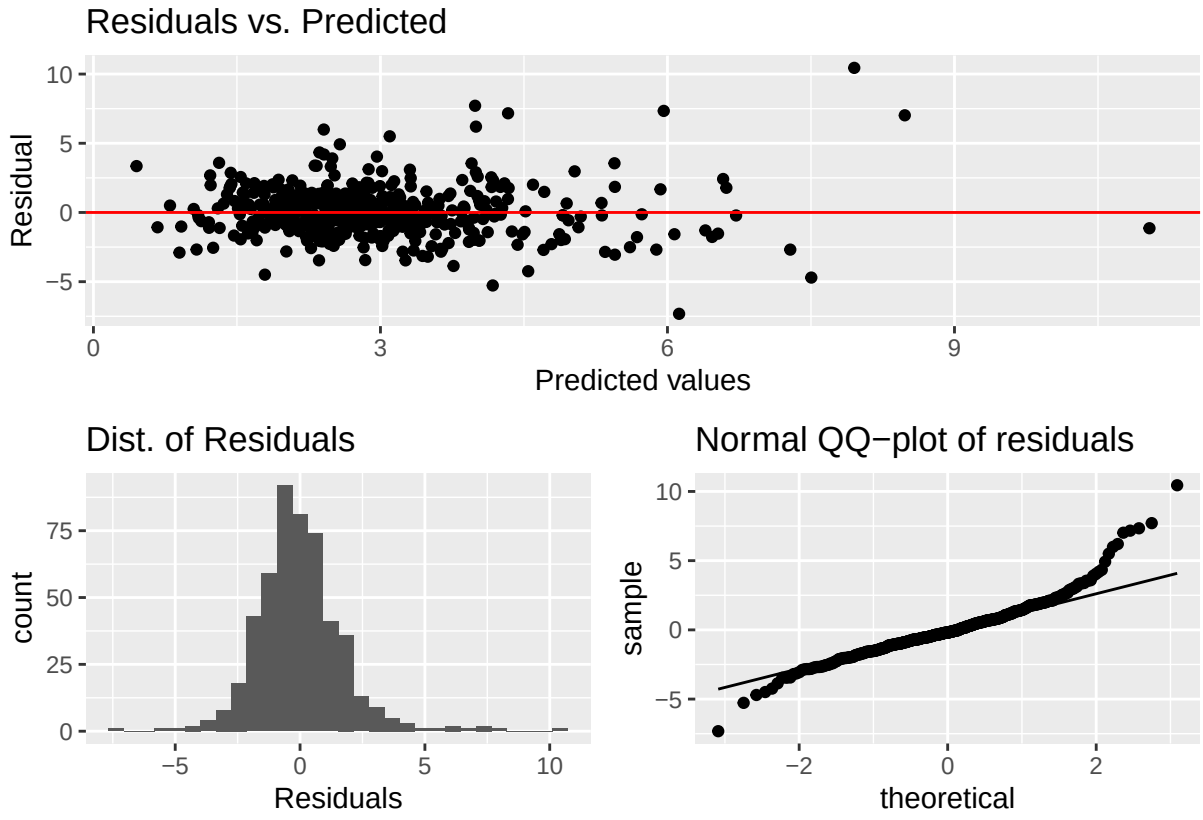
res.df	rss	df	sumsq	statistic	p.value
496	1644.137	NA	NA	NA	NA
495	1598.244	1	45.893	14.214	0

Model Conditions

```
model_aug <- augment(full_model) %>%
   mutate(obs_num = row_number()) #add row number to help with graphing

resid_fitted <- ggplot(data = model_aug, aes(x = .fitted, y = .resid)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  labs(x = "Predicted values",
    y = "Residual",
    title = "Residuals vs. Predicted")

resid_hist <- ggplot(data = model_aug, aes(x = .resid)) +
  geom_histogram() +
  labs(x = "Residuals", title = "Dist. of Residuals")

resid_qq <- ggplot(data = model_aug, aes(sample = .resid)) +
  stat_qq() +
```

```
  stat_qq_line() +
  labs(title = "Normal QQ-plot of residuals")

resid_fitted / (resid_hist + resid_qq)
```

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



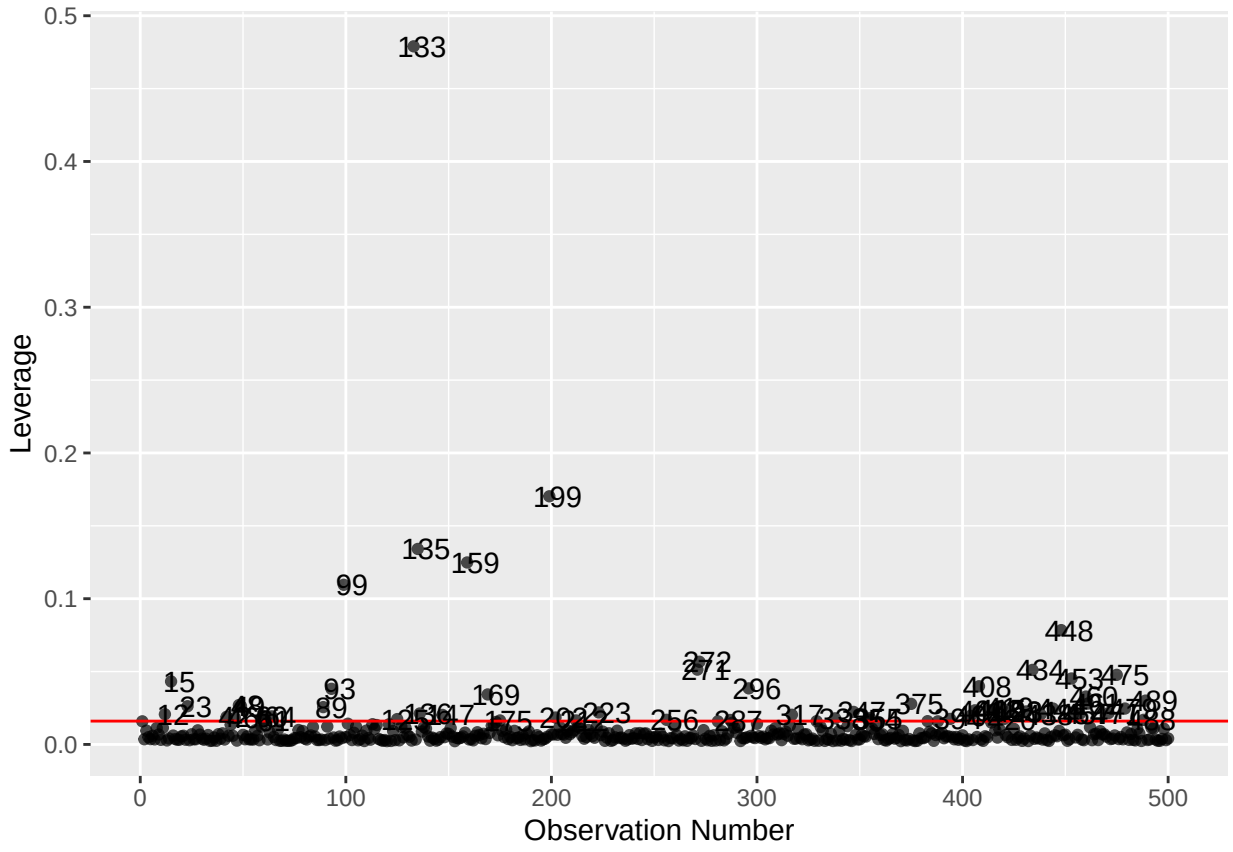
Model Diagnostics

Leverage

```
#calculate threshold
leverage_threshold <- 2*(3+1)/500
leverage_threshold
```

```
## [1] 0.016
```

```
ggplot(data = model_aug, aes(x = obs_num, y = .hat)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = leverage_threshold, color = "red") +
  labs(x = "Observation Number", y = "Leverage") +
  geom_text(aes(label=ifelse(.hat > leverage_threshold,
                             as.character(obs_num), "")), nudge_x = 4)
```



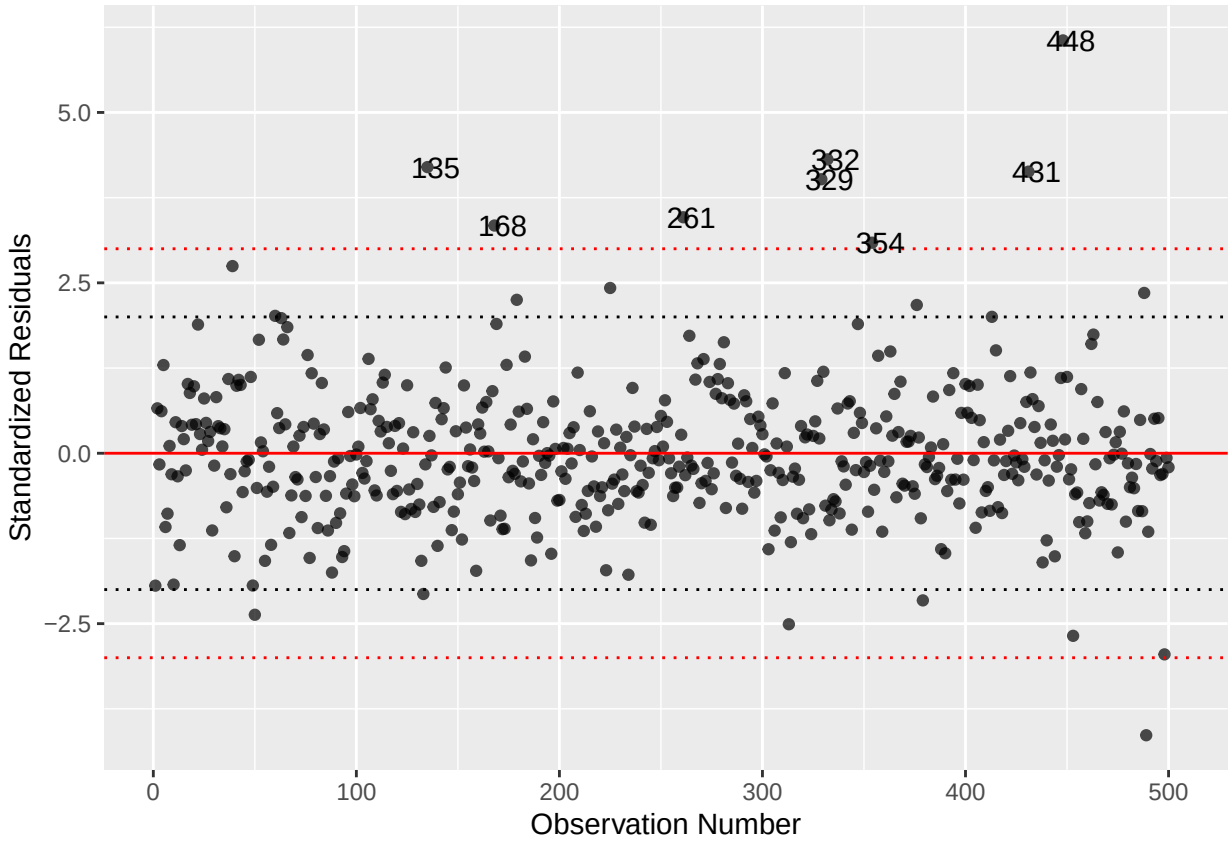

```
model_aug %>%
  filter(.hat > leverage_threshold)
```

```
## # A tibble: 60 x 11
##    unemployment_ch~ retail_avg grocery_avg workplaces_avg .fitted .resid
##               <dbl>      <dbl>       <dbl>          <dbl>   <dbl>  <dbl>
##  1             2.20     -17.3        15.0           -25.8   2.80  -0.602
##  2             2.4        5.27       -1.58          -32.7   2.04   0.363
##  3             1.30      19.0        16.5           -19.7   0.800  0.500
##  4             3.7       -5.69       18.6           -15.7   1.78   1.92
##  5             5.1      -27.6       -10.7           -40.1   5.31  -0.215
##  6             5.1      -22.8        13.3           -24.4   3.12   1.98
##  7            -0.6       -7.13      -13.4           -19.8   2.84  -3.44
##  8             4.9       10.8        18.6           -23.2   1.31   3.59
##  9             3         -7.47       18.0           -17.9   1.95   1.05
## 10             8.00     -25.0        -9.79          -40.0   5.03   2.97
## # ... with 50 more rows, and 5 more variables: .std.resid <dbl>, .hat <dbl>,
## #   .sigma <dbl>, .cooksd <dbl>, obs_num <int>
```

Standardized residuals

```
#scatterplot of std resid vs predicted
ggplot(data = model_aug, aes(x = obs_num, y = .std.resid)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = 0, color = "red") +
  geom_hline(linetype = "dotted", yintercept = c(-2,2)) +
  geom_hline(linetype = "dotted", yintercept = c(-3,3), color = "red") +
  labs(x = "Observation Number", y = "Standardized Residuals") +
```

```
  geom_text(aes(label=ifelse(.std.resid > 3,
                             as.character(obs_num), "")), nudge_x = 4)
```



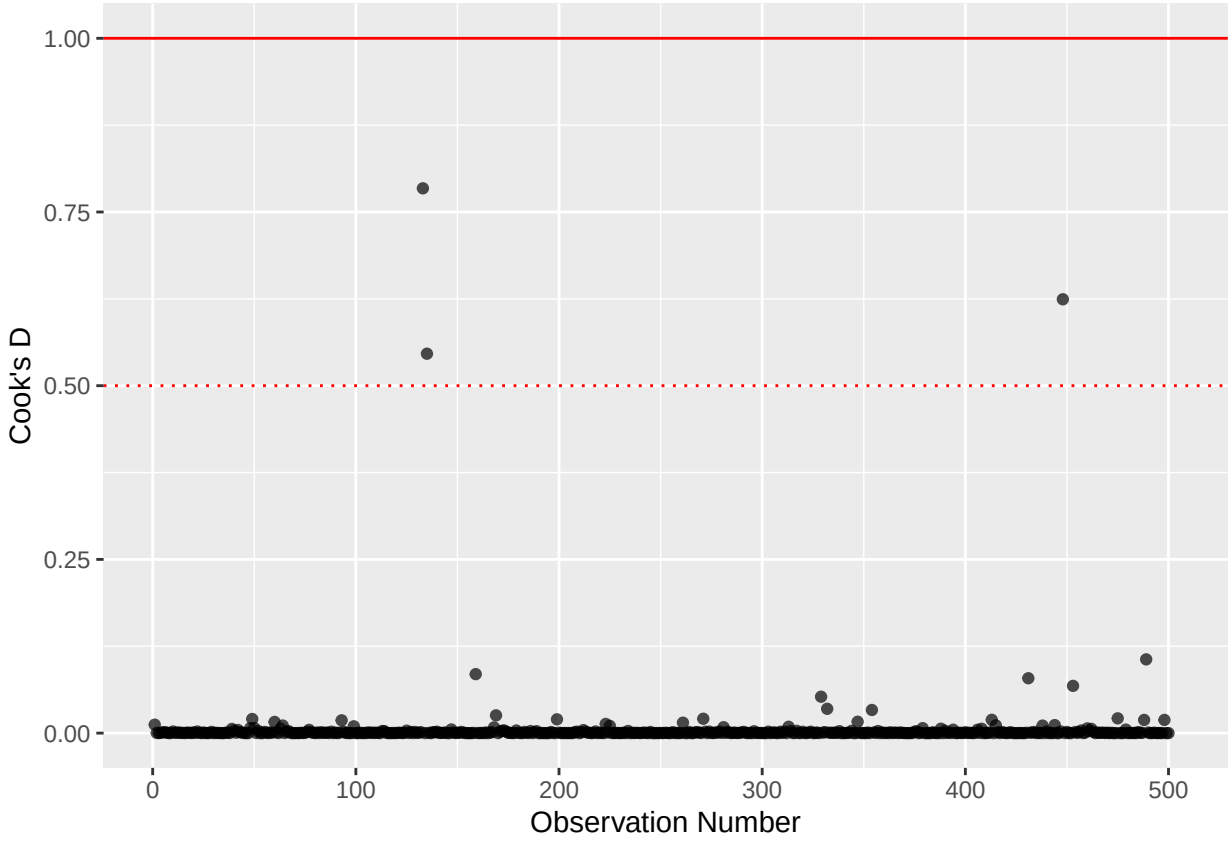

```
model_aug %>%
  filter(abs(.std.resid) > 3)
```

```
## # A tibble: 9 x 11
##   unemployment_ch~ retail_avg grocery_avg workplaces_avg .fitted .resid
##              <dbl>      <dbl>       <dbl>          <dbl>   <dbl>  <dbl>
## 1             15.5     -46.3      -31.3            -31.0    8.48   7.02
## 2              8.4      -7.77       5.46           -20.6    2.41   5.99
## 3             10.2     -18.1       -7.28           -31.2    4.00   6.20
## 4             11.5     -24.3       -8.92           -23.3    4.33   7.17
## 5             11.7     -24.5       -0.737          -25.3    3.99   7.71
## 6              8.6     -17.9       11.7            -29.2    3.10   5.50
## 7             13.3     -36.1       -8.07           -42.1    5.96   7.34
## 8             18.4     -43.7      -26.9            -35.3    7.95  10.4
## 9             -1.2     -33.6      -12.1            -45.0    6.12  -7.32
## # ... with 5 more variables: .std.resid <dbl>, .hat <dbl>, .sigma <dbl>,
## #   .cooksd <dbl>, obs_num <int>
```

Cook's distance

```
#scatterplot of cook's d vs obs num
ggplot(data = model_aug, aes(x = obs_num, y = .cooksd)) +
  geom_point(alpha = 0.7) +
  geom_hline(yintercept = 1, color = "red") +
  geom_hline(linetype = "dotted", yintercept = 0.5, color = "red") +
```

```
  labs(x = "Observation Number", y = "Cook's D") +
  geom_text(aes(label=ifelse(.cooksd > 1,
                             as.character(obs_num), "")), nudge_x = 4)
```



Multicollinearity

```
vif(full_model) %>%
  tidy() %>%
  kable(digits = 3)
```

```
## Warning: 'tidy.numeric' is deprecated.
## See help("Deprecated")
```

```
## Warning: `data_frame()` is deprecated as of tibble 1.1.0.
## Please use `tibble()` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_warnings()` to see where this warning was generated.
```

names	x
retail_avg	3.118
grocery_avg	2.207
workplaces_avg	2.510
retail_avg:grocery_avg	1.195

All of the predictor variables have a VIF less than 10. Thus, we can say that none of the predictor variables in our model are correlated.

Full model

$unemployment\_change = 1.266 - 0.73 \times retail\_avg - 0.015 \times grocery\_avg - 0.035 \times workplaces\_avg + 0.002 \times retail\_avg : groc$

```
tidy(full_model, conf.int = TRUE) %>%
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	1.266	0.515	2.458	0.014	0.254	2.278
retail_avg	-0.073	0.013	-5.680	0.000	-0.099	-0.048
grocery_avg	-0.015	0.014	-1.068	0.286	-0.044	0.013
workplaces_avg	-0.035	0.023	-1.537	0.125	-0.080	0.010
retail_avg:grocery_avg	0.002	0.000	3.770	0.000	0.001	0.002

Questions

- What is the relationship between _____ and _____? What does this say about the future of unemployment during the future of this pandemic, which experts believe will continue into the next years?
- Given the relationship between these variables, what is the expected unemployment rate of North Carolina counties?

Predictors:

- resilience:

Response:

- unemployment rate:

EDA:

See .rmd files in folder.

Model: We will use a multiple linear regression model with the following form:

unemployment rate = resilience + ___ + ___ + ___ + ___

Output: