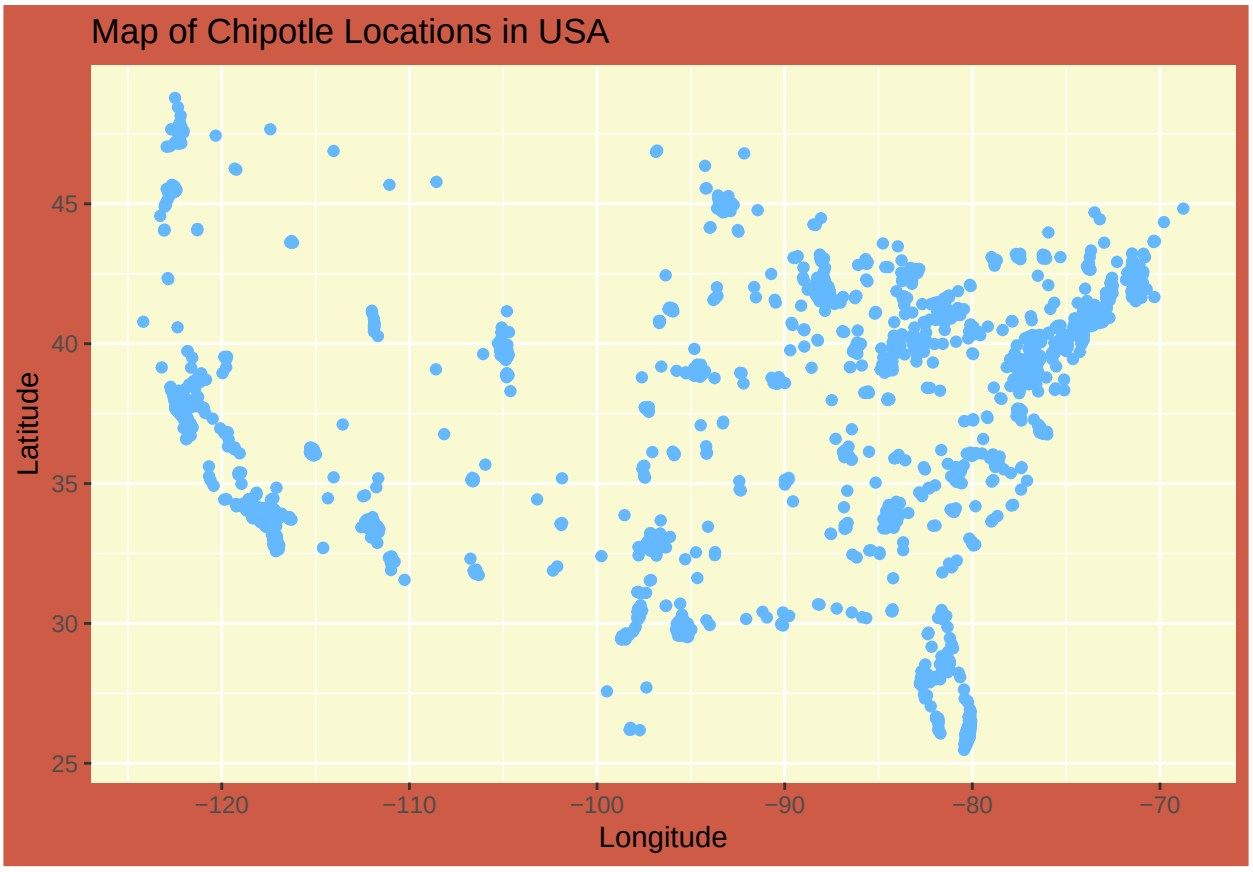
Chipotle

Razmig Zeitounian

9/8/2020

Links to data: https://www.kaggle.com/jeffreybraun/chipotle-locations https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html

```
#make a map for chipotle locations
ggplot(chipotle, aes(y=latitude, x=longitude)) +
  geom_point(color = "steelblue1") +
  labs(x = "Longitude", y = "Latitude", title = "Map of Chipotle Locations in USA") +
  theme(panel.background = element_rect(fill = "lightgoldenrodyellow"),
    plot.background = element_rect(fill = "coral3"))
```

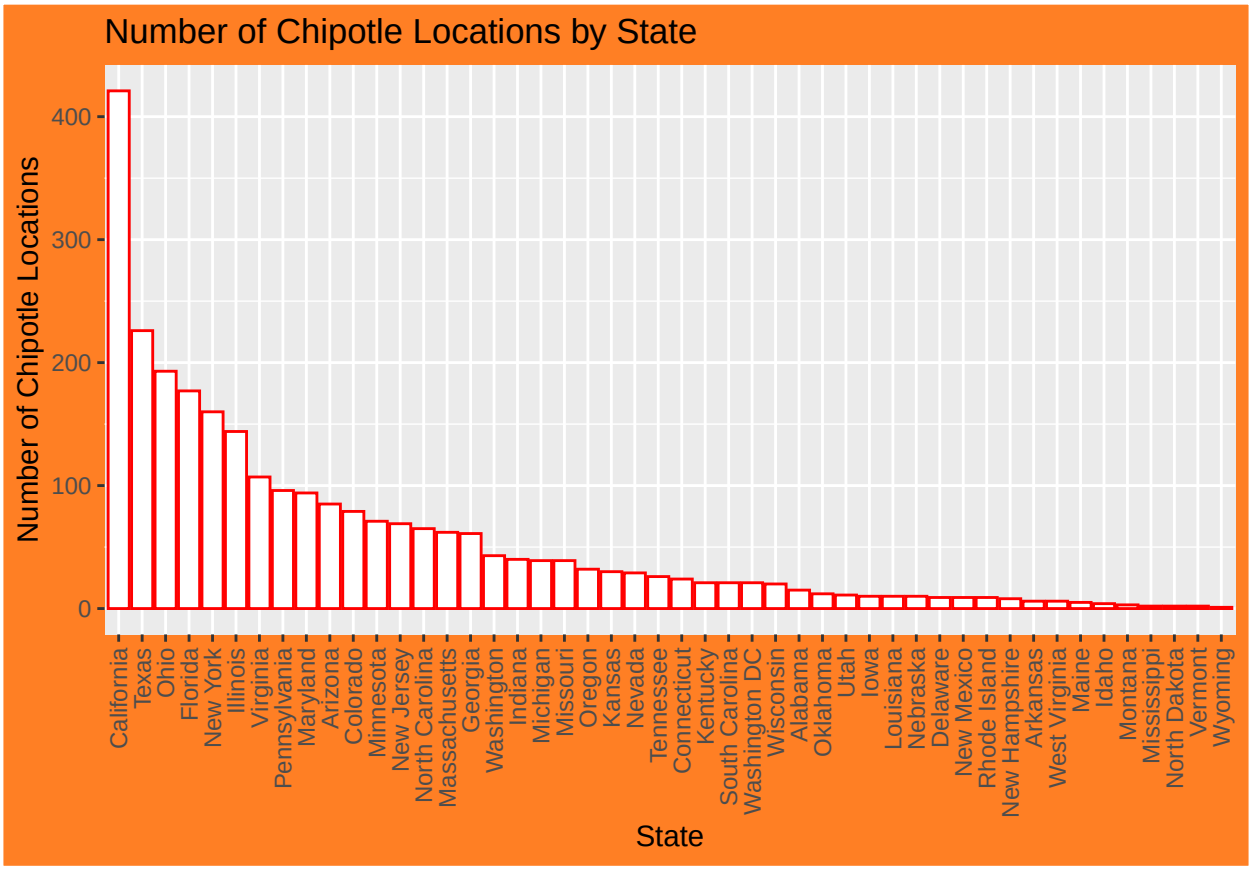



```
# See which states have the most Chipotle locations
chipotles.by.state <- chipotle %>%
  group_by(state) %>%
  count() %>%
  arrange(desc(n))
```

```
chipotles.by.state
```

```
## # A tibble: 48 x 2
## # Groups:   state [48]
##    state            n
##    <fct>        <int>
##  1 California     421
##  2 Texas          226
##  3 Ohio           193
##  4 Florida        177
##  5 New York       160
##  6 Illinois       144
##  7 Virginia       107
##  8 Pennsylvania    96
##  9 Maryland        94
## 10 Arizona         85
## # ... with 38 more rows
```

```
# visualize it in descending order
ggplot(chipotles.by.state, aes(reorder(state, -n), n)) +
  geom_bar(stat = "identity", fill = "white", color = "red") +
  labs(x = "State", y = "Number of Chipotle Locations", title = "Number of Chipotle Locations by State") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) +
  theme(plot.background = element_rect(fill = "chocolate1"))
```




```
# See which cities have the most locations
chipotles.by.city <- chipotle %>%
```

```
  group_by(location, state) %>%
  count() %>%
  arrange(desc(n))
chipotles.by.city
```

```
## # A tibble: 1,521 x 3
## # Groups:   location, state [1,521]
##    location      state             n
##    <fct>         <fct>         <int>
##  1 New York      New York         52
##  2 Chicago       Illinois         36
##  3 Houston       Texas            31
##  4 Washington DC Washington DC    21
##  5 Los Angeles   California       20
##  6 Columbus      Ohio             19
##  7 Dallas        Texas            19
##  8 Las Vegas     Nevada           19
##  9 Phoenix       Arizona          19
## 10 Cincinnati    Ohio             17
## # ... with 1,511 more rows
```

```
mod <- lm(n ~ cpp + CENSUS2010POP, data = chip)
mod.summary <- summary(mod)
mod.summary
```

```
##
## Call:
## lm(formula = n ~ cpp + CENSUS2010POP, data = chip)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -197.637  -12.022   -1.427    6.350  147.588
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.191e+01  9.605e+00  -1.240 0.221374
## cpp            9.763e-01  2.694e-01   3.624 0.000735 ***
## CENSUS2010POP  8.754e-06  9.574e-07   9.143 7.96e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45.13 on 45 degrees of freedom
## Multiple R-squared:  0.6695, Adjusted R-squared:  0.6548
## F-statistic: 45.57 on 2 and 45 DF,  p-value: 1.522e-11
```

```
cat("Our R^2 is", mod.summary$r.squared, ". This means", mod.summary$r.squared, "percent of variation i
```

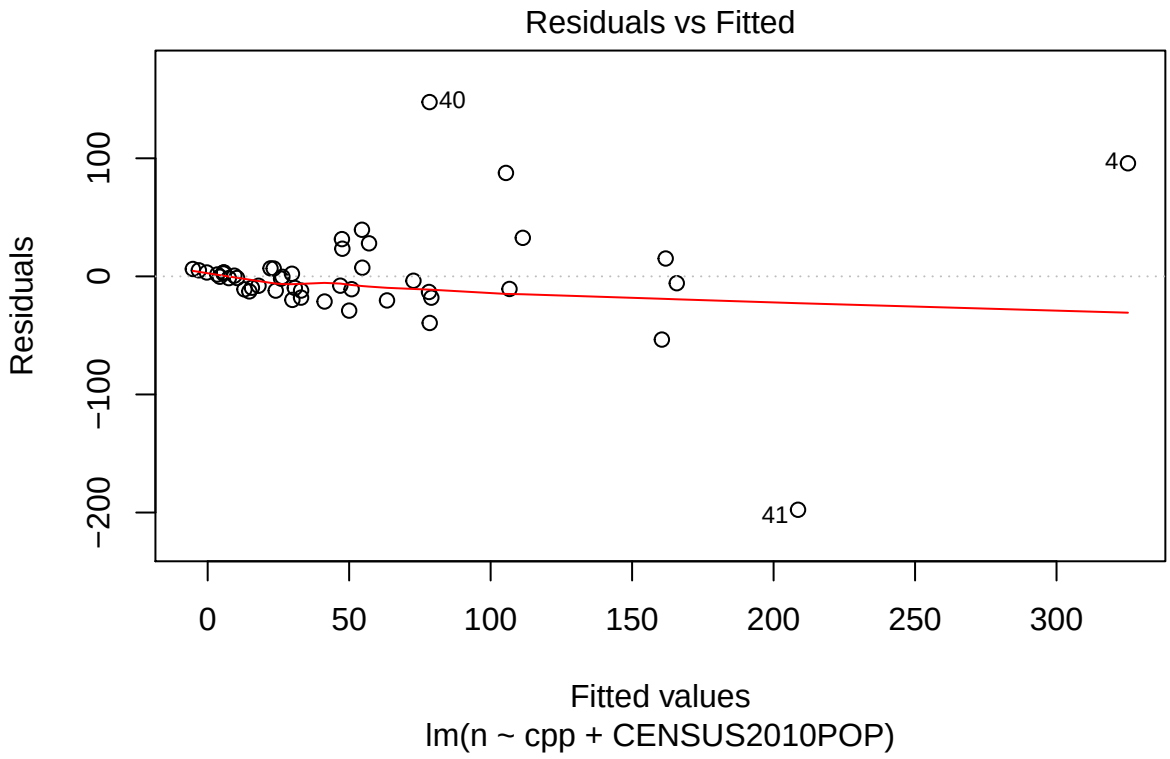
```
## Our R^2 is 0.669461 . This means 0.669461 percent of variation in the number of Chipotle locations p
```

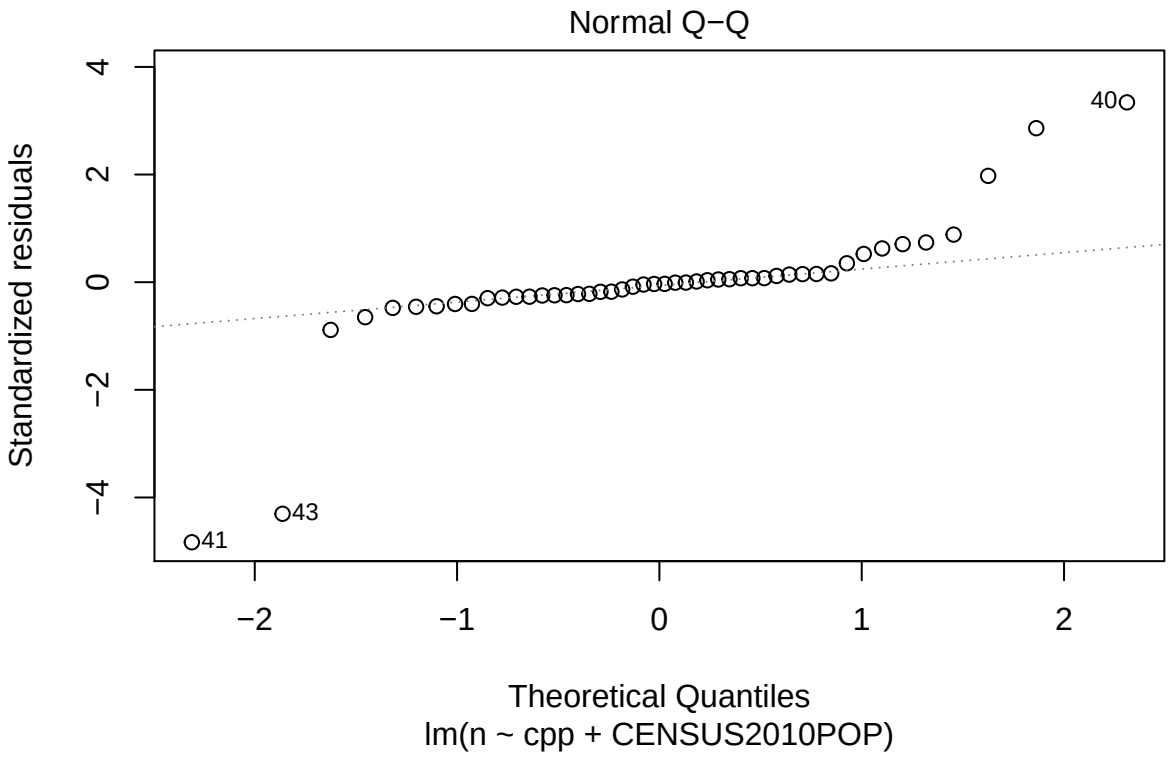
```
numeric.data <- chip %>%
  select(n, CENSUS2010POP, cpp)
#n and census2010 pop quite related
cor(numeric.data)
```

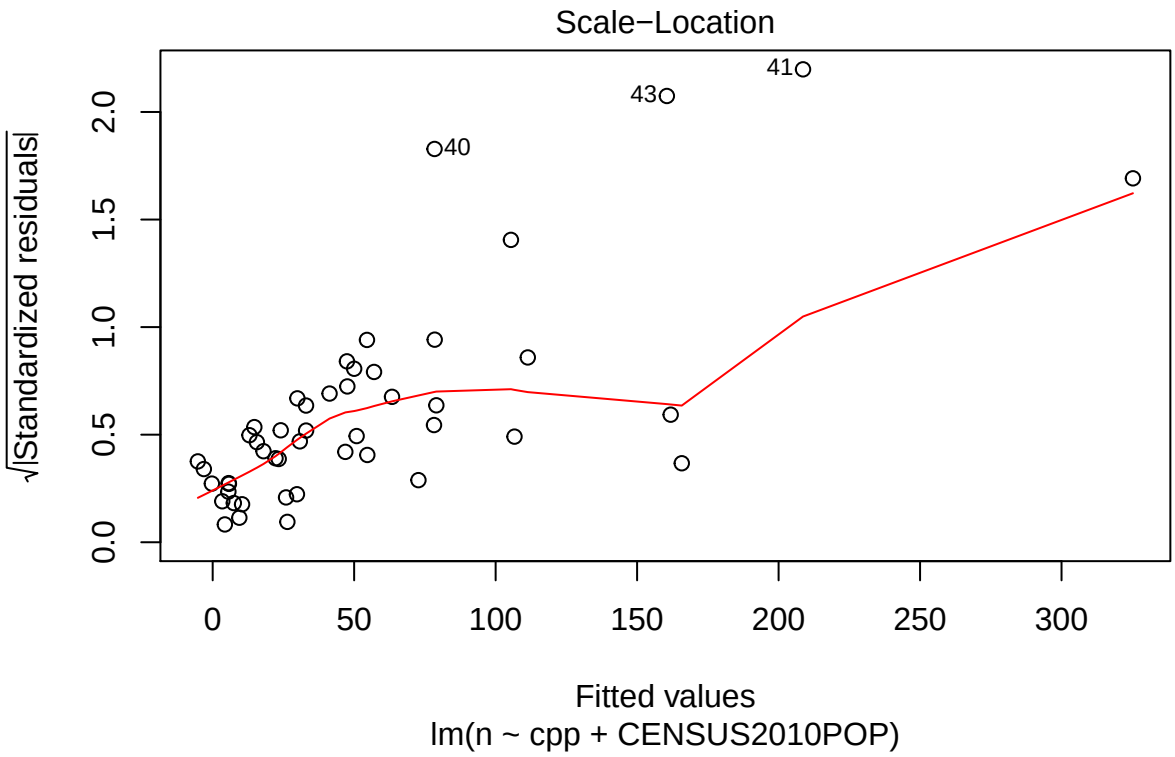
```
##                       n CENSUS2010POP         cpp
## n             1.0000000    0.75697051  0.23538451
```

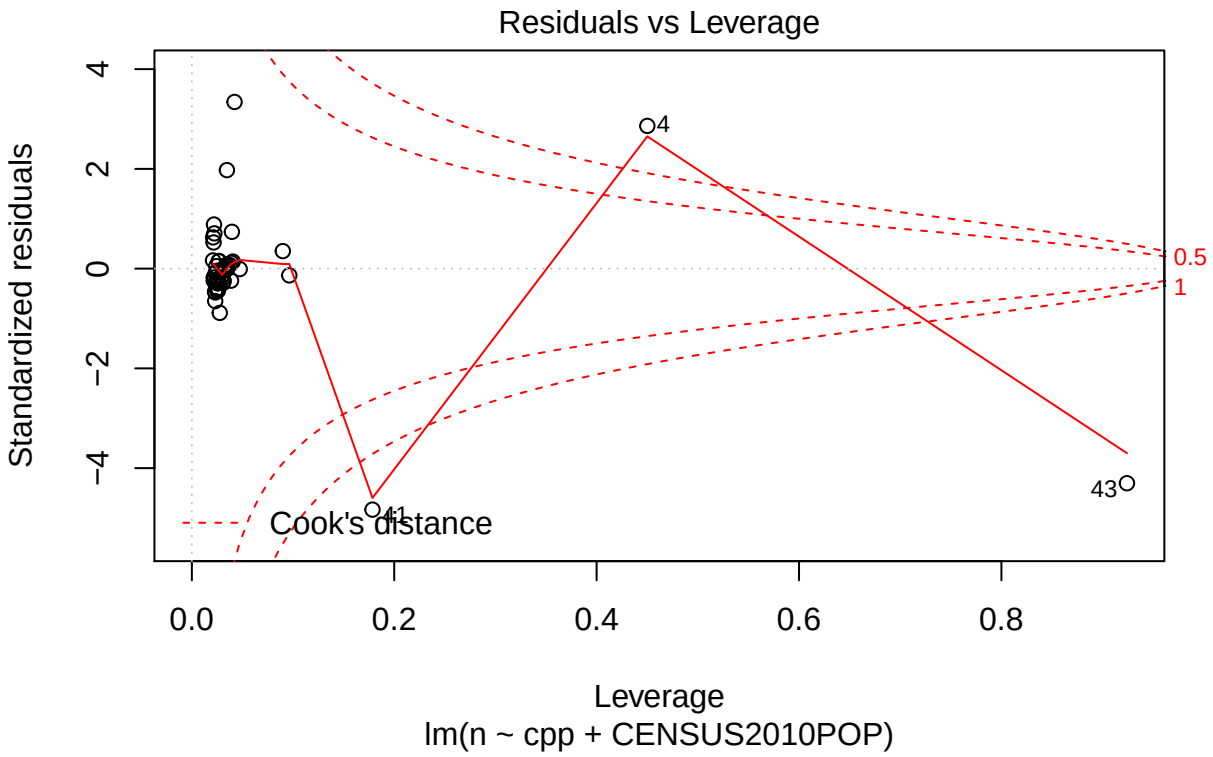
```
## CENSUS2010POP 0.7569705    1.00000000 -0.09738044
## cpp           0.2353845   -0.09738044  1.00000000
```

```
plot(mod)
```



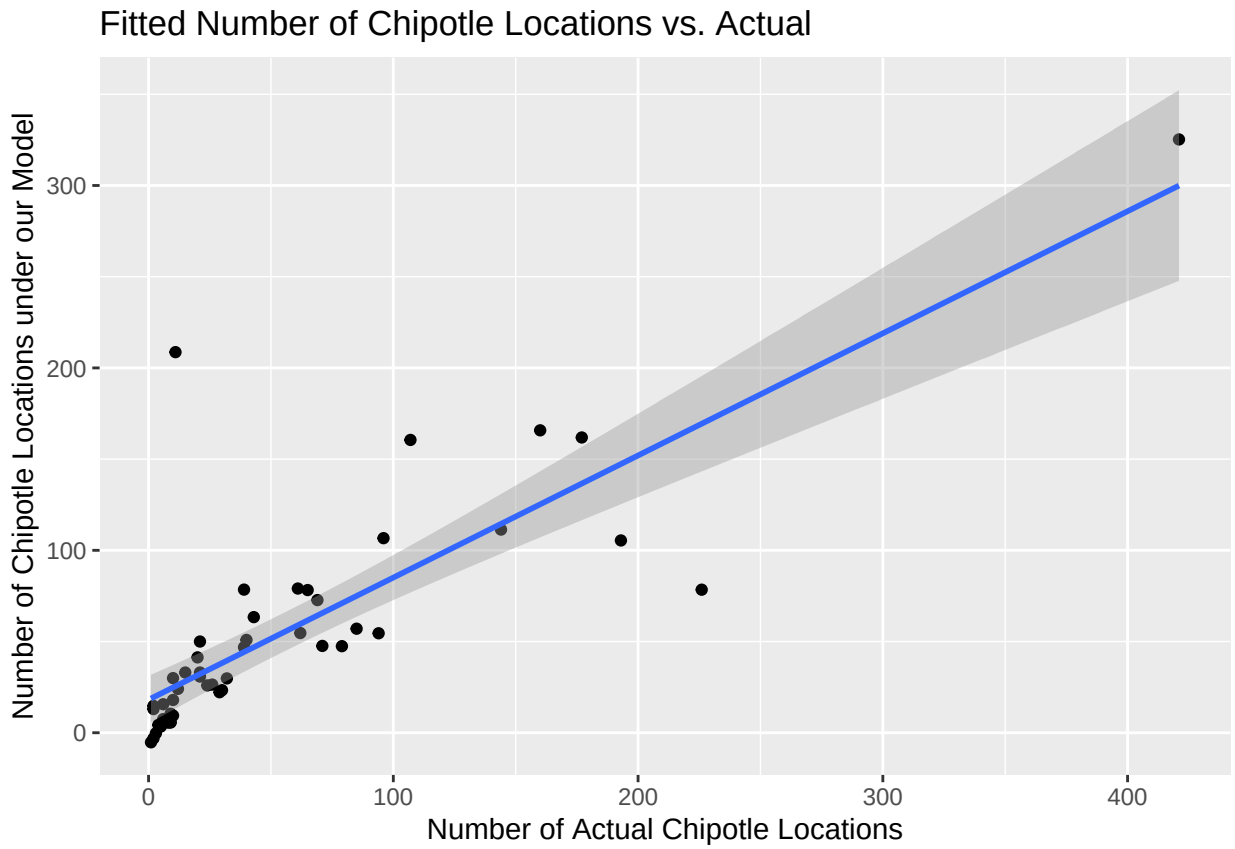

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(n ~ cpp + CENSUS2010POP)
40
43
41
−2
−1
0
1
2
−4
−2
0
2
4


Scale−Location
41
43
40
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0
50
100
150
200
250
300
Fitted values
lm(n ~ cpp + CENSUS2010POP)



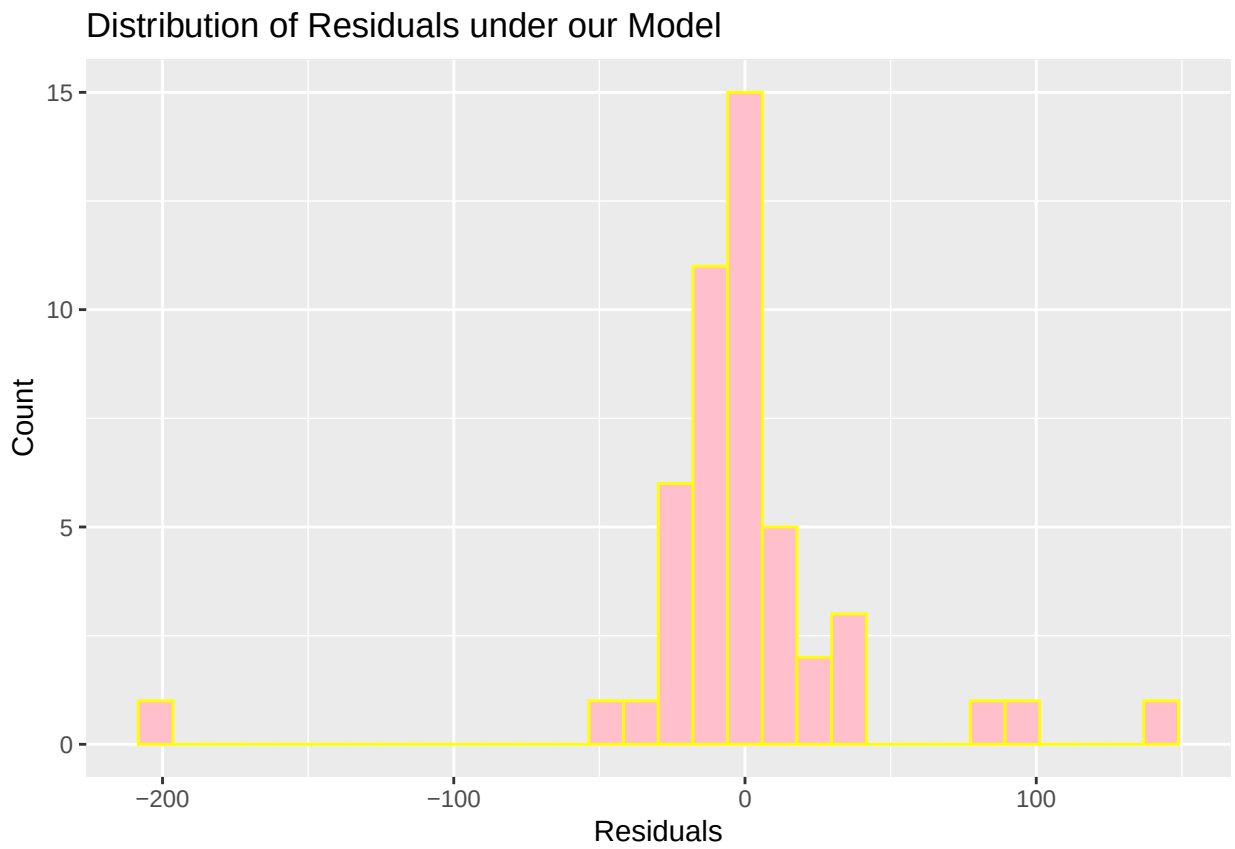

```
ggplot(chip, aes(y=mod$fitted.values, x=n)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(y = "Number of Chipotle Locations under our Model", x = "Number of Actual Chipotle Locations", ti
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
ggplot(chip, aes(mod$residuals)) +
  geom_histogram(fill = "pink", color = "yellow") +
  labs(y="Count", x="Residuals", title="Distribution of Residuals under our Model")
```

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



The model summary suggests that the two variables are significant in predicting the number of Chipotle Locations in a state.

It looks like the normality assumption of the residuals holds since the histogram is roughly bell shaped. Additionally, it looks to be centered around 0 with some constant variance σ.