NYC Childcare

Feng-Yi Liu

2/6/2020

NYC Childcare

ok.

```
library(rmarkdown)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.1
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)
library(janitor)
```

```
##
## Attaching package: 'janitor'
```

```
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(dplyr)
library(plyr)
```

```
## -------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## -------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following object is masked from 'package:purrr':
##
##     compact
```

```
library(raster)
```

```
## Loading required package: sp
```

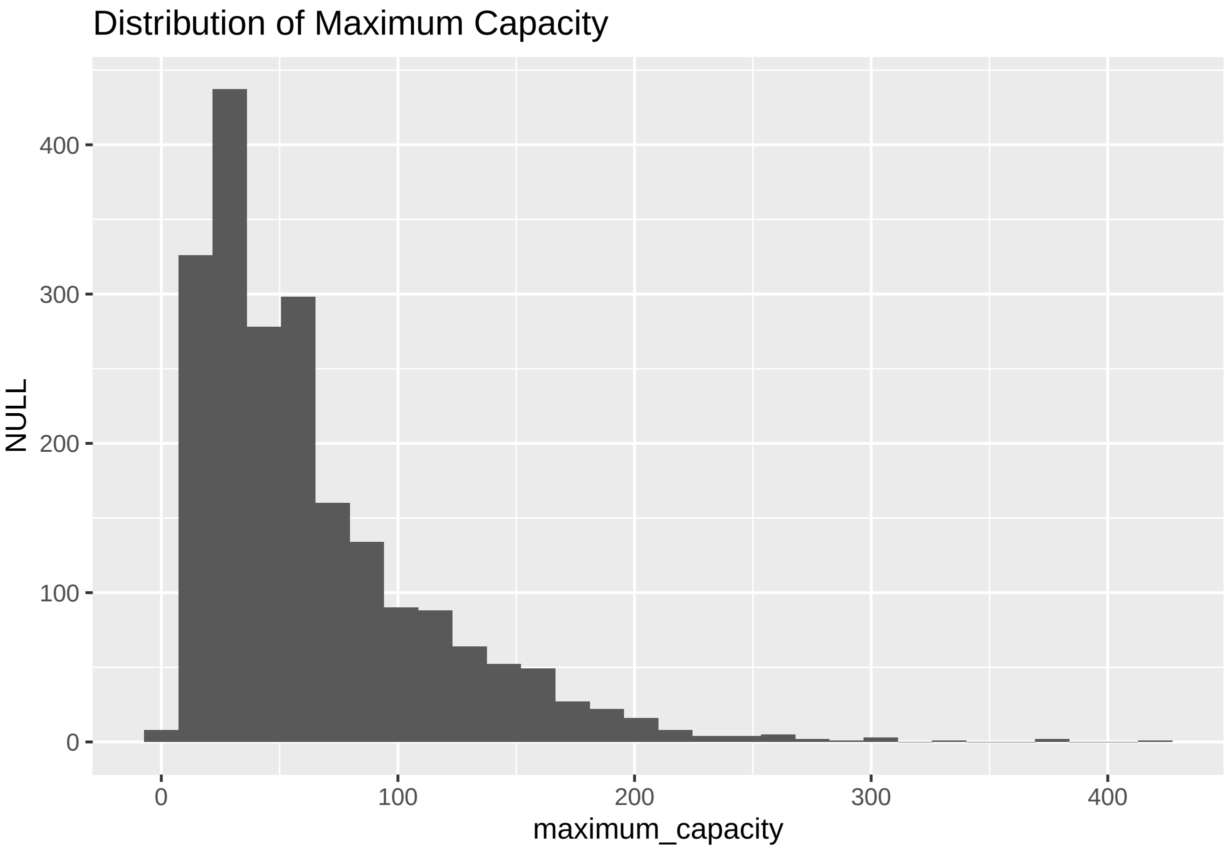
```
##
## Attaching package: 'raster'

## The following object is masked from 'package:janitor':
##
##     crosstab

## The following object is masked from 'package:dplyr':
##
##     select

## The following object is masked from 'package:tidyr':
##
##     extract
```

```r
data <- read.csv("NYC_CC_2020.csv", header=TRUE) %>%
  filter(Status == "Active" | Status == "Permitted") %>%
  dplyr::select(Borough,    ZipCode,Permit.Expiration,Date.Permitted,
Status,Age.Range,Maximum.Capacity,  Day.Care.ID,    Program.Type,   Facility.Type,
Child.Care.Type,Building.Identification.Number,Violation.Rate.Percent, Average.Violation.Rate.Percent,
Public.Health.Hazard.Violation.Rate,    Average.Public.Health.Hazard.Violation.Rate,    Critical.Violat
data<-data %>% distinct(Legal.Name, Day.Care.ID, .keep_all = TRUE)
```

```
## Warning: Trying to compute distinct() for variables not found in the data:
## - `Legal.Name`
## This is an error, but only a warning is raised for compatibility reasons.
## The following variables will be used:
## - Day.Care.ID
```

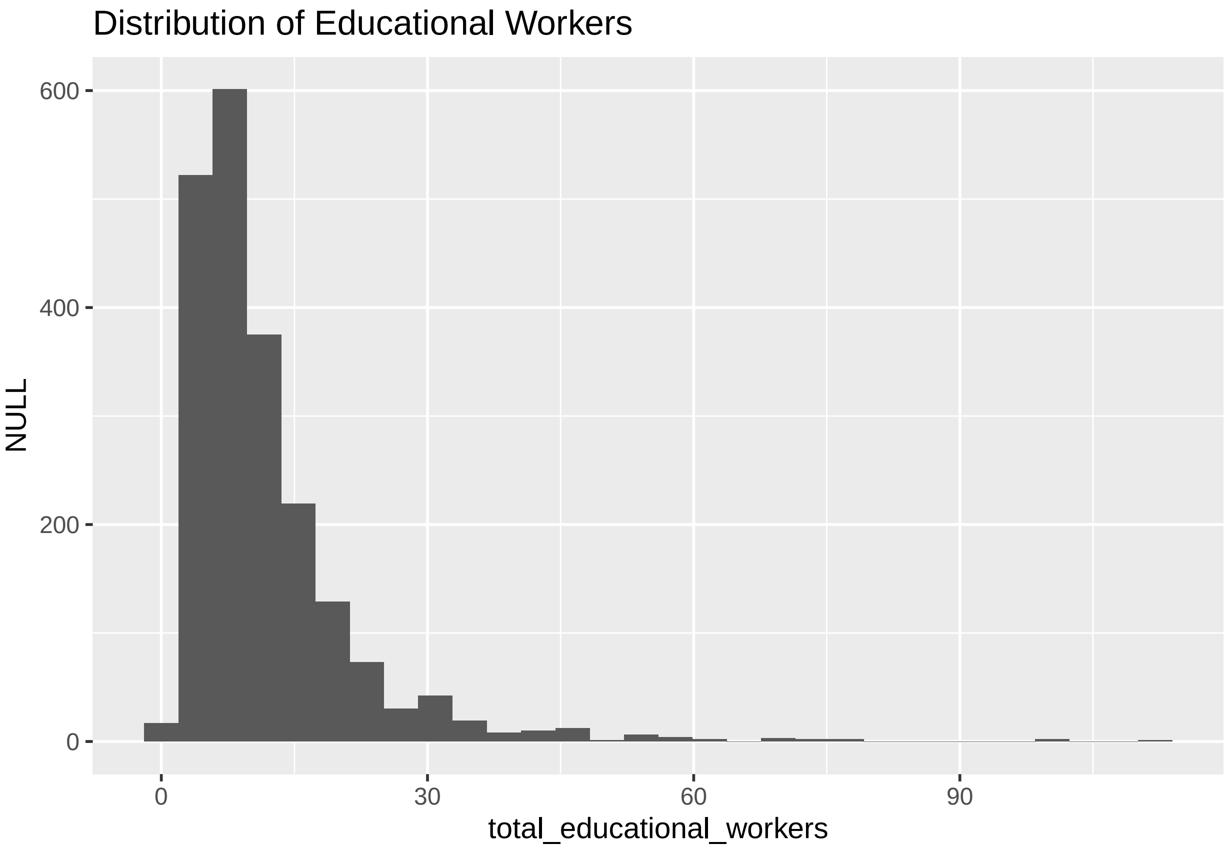
```r
# Remove rows with NA
data <- na.omit(data)
# Remove empty rows
data <- data %>%  filter(Date.Permitted != "")
# Clean Column names
data <- clean_names(data)
# Data
data <- data %>%  filter(maximum_capacity >0)
#CENTER_UNIT<-data %>% distinct(Building.Identification.Number, .keep_all = TRUE)
qplot(maximum_capacity, data = data, main = "Distribution of Maximum Capacity")
```

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



```
qplot(total_educational_workers, data = data, main = "Distribution of Educational Workers")
```

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



```r
#Look at the dataset
glimpse(data)
```

```
## Observations: 2,080
## Variables: 20
## $ borough                                   <fct> MANHATTAN, QUEENS,...
## $ zip_code                                  <int> 10016, 11415, 1138...
## $ permit_expiration                         <fct> 11/13/21, 9/5/20, ...
## $ date_permitted                            <fct> 6/28/04, 9/5/14, 1...
## $ status                                    <fct> Permitted, Permitt...
## $ age_range                                 <fct> 0 YEARS - 2 YEARS,...
## $ maximum_capacity                          <int> 44, 111, 61, 138, ...
## $ day_care_id                               <fct> DC2614, DC32009, D...
## $ program_type                              <fct> INFANT TODDLER, PR...
## $ facility_type                             <fct> GDC, GDC, GDC, GDC...
## $ child_care_type                           <fct> Child Care - Infan...
## $ building_identification_number            <int> 1087340, 4574091, ...
## $ violation_rate_percent                    <dbl> 12.5000, 25.0000, ...
## $ average_violation_rate_percent            <dbl> 28.0891, 30.5946, ...
## $ total_educational_workers                 <int> 17, 29, 9, 18, 16,...
## $ average_total_educational_workers         <dbl> 8.0442, 12.0664, 1...
## $ public_health_hazard_violation_rate       <dbl> 0.0000, 25.0000, 0...
## $ average_public_health_hazard_violation_rate <dbl> 10.6875, 12.5403, ...
## $ critical_violation_rate                   <dbl> 12.5000, 0.0000, 1...
## $ average_critical_violation_rate           <dbl> 24.9492, 27.1630, ...
```

```r
data$zip_code <- factor(data$zip_code)
data %>% group_by(zip_code) %>% tally( name="number.of.center")
```

```
## # A tibble: 175 x 2
##    zip_code number.of.center
##    <fct>               <int>
##  1 10001                  10
##  2 10002                  34
##  3 10003                   9
##  4 10004                   4
##  5 10005                   3
##  6 10006                   2
##  7 10007                   7
##  8 10009                  11
##  9 10010                  10
## 10 10011                  19
## # ... with 165 more rows
```

```r
summarise(data, mean_maximum_capacity =mean(maximum_capacity))
```

```
##   mean_maximum_capacity
## 1              64.06731
```

```r
summarise(data, mean_violation_rate =mean(violation_rate_percent))
```

```
##   mean_violation_rate
## 1            29.89969
```

```r
summarise(data, mean_workers =mean(total_educational_workers))
```

```
##   mean_workers
## 1     11.49087
```

```r
summarise(data, mean_health_hazard_violation =mean(public_health_hazard_violation_rate))
```

```
##   mean_health_hazard_violation
## 1                     12.01941
```

```r
#Now remove plyr and try again and you get the grouped summary.
detach(package:plyr)
zipcodeunite<-
  data %>%  group_by(zip_code) %>%
  summarise(total.count=n(),
            sum_capacity = sum(maximum_capacity),
            mean_maximum_capacity =mean(maximum_capacity),
            mean_workers =mean(total_educational_workers),
            mean_violation_rate =mean(violation_rate_percent),
            mean_health_hazard_violation =mean(public_health_hazard_violation_rate)
            )
zippoverty <- read.csv("~/ACS_16_5YR_B17001_EDDDD.csv", header=TRUE)
zippoverty$zip_code<-as.factor(zippoverty$zip_code)
test<-full_join(zippoverty,zipcodeunite, by = "zip_code" )
```

```
## Warning: Column `zip_code` joining factors with different levels, coercing
## to character vector
```

```r
test <- na.omit(test)
test <- clean_names(test)

f1 <- total_count ~ estimate_total+ below_poverty_level + sum_capacity + mean_workers
```

```r
f2<-  mean_violation_rate ~estimate_total+ below_poverty_level + sum_capacity + mean_workers
f3<-  mean_health_hazard_violation ~estimate_total+ below_poverty_level + sum_capacity + mean_workers
m1 <- lm(f1, data=test)
m2 <- lm(f2, data=test)
m3 <- lm(f3, data=test)

summary(m1)
```

```
## 
## Call:
## lm(formula = f1, data = test)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -8.0363 -1.3318 -0.1012  1.2107  6.4960 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          4.497e+00  6.053e-01   7.429 5.17e-12 ***
## estimate_total       5.666e-05  1.159e-05   4.889 2.34e-06 ***
## below_poverty_level -2.174e-04  3.317e-05  -6.555 6.47e-10 ***
## sum_capacity         1.456e-02  4.225e-04  34.456  < 2e-16 ***
## mean_workers        -3.685e-01  4.404e-02  -8.368 2.12e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 2.206 on 169 degrees of freedom
## Multiple R-squared:  0.9352, Adjusted R-squared:  0.9337 
## F-statistic: 609.8 on 4 and 169 DF,  p-value: < 2.2e-16
```

```r
summary(m2)
```

```
## 
## Call:
## lm(formula = f2, data = test)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -31.274  -8.921  -1.915   9.154  46.851 
## 
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)    
## (Intercept)          2.316e+01  3.729e+00   6.210 3.98e-09 ***
## estimate_total      -6.148e-05  7.141e-05  -0.861   0.3905    
## below_poverty_level  9.315e-04  2.044e-04   4.558 9.86e-06 ***
## sum_capacity        -6.606e-03  2.603e-03  -2.538   0.0121 *  
## mean_workers         5.355e-01  2.714e-01   1.973   0.0501 .  
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 13.59 on 169 degrees of freedom
## Multiple R-squared:  0.161,  Adjusted R-squared:  0.1412 
## F-statistic: 8.109 on 4 and 169 DF,  p-value: 5.265e-06
```

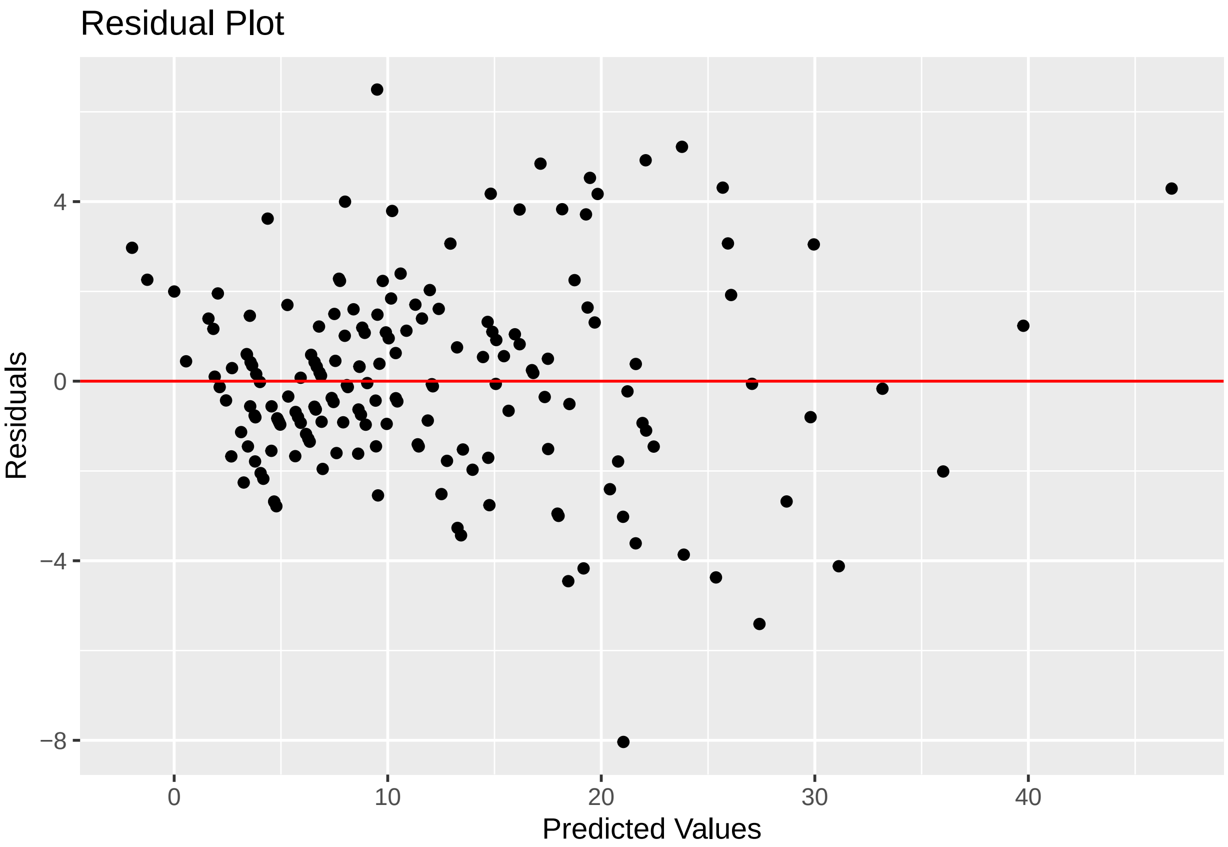
```r
summary(m3)
```

```
##
## Call:
## lm(formula = f3, data = test)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.493  -6.010  -0.627   3.895  45.994
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1.078e+01  2.455e+00   4.392 1.98e-05 ***
## estimate_total      -4.557e-05  4.701e-05  -0.969  0.33379
## below_poverty_level  4.120e-04  1.345e-04   3.062  0.00256 **
## sum_capacity        -3.747e-03  1.714e-03  -2.186  0.03018 *
## mean_workers         2.678e-01  1.787e-01   1.499  0.13581
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.947 on 169 degrees of freedom
## Multiple R-squared:  0.07665,    Adjusted R-squared:  0.05479
## F-statistic: 3.507 on 4 and 169 DF,  p-value: 0.008856
```

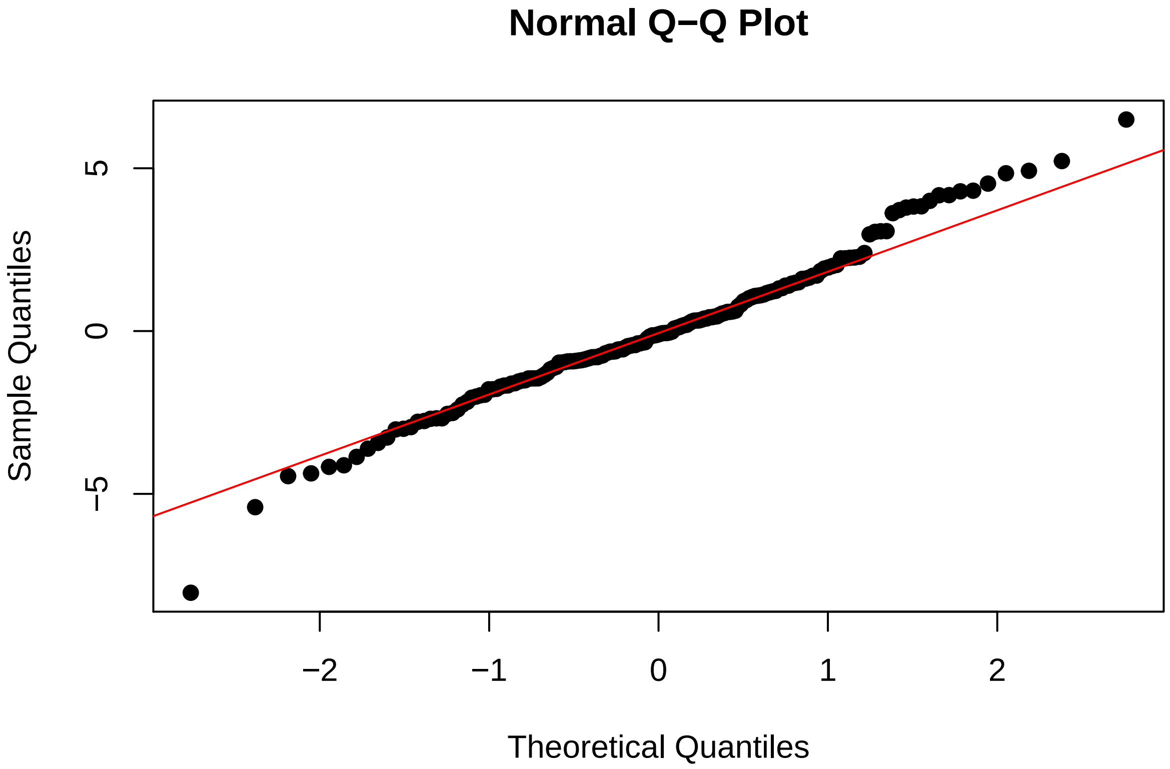
```r
anova(m1)
```

```
## Analysis of Variance Table
##
## Response: total_count
##                      Df Sum Sq Mean Sq   F value    Pr(>F)
## estimate_total        1 6024.5  6024.5 1238.5129 < 2.2e-16 ***
## below_poverty_level   1    0.5     0.5    0.0989    0.7536
## sum_capacity          1 5499.0  5499.0 1130.4733 < 2.2e-16 ***
## mean_workers          1  340.6   340.6   70.0193 2.123e-14 ***
## Residuals           169  822.1     4.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
predVal <- predict(m1)
residVal <- residuals(m1)
ggplot(mapping = aes(x = predVal, y = residVal)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residual Plot", x = "Predicted Values", y = "Residuals")
```



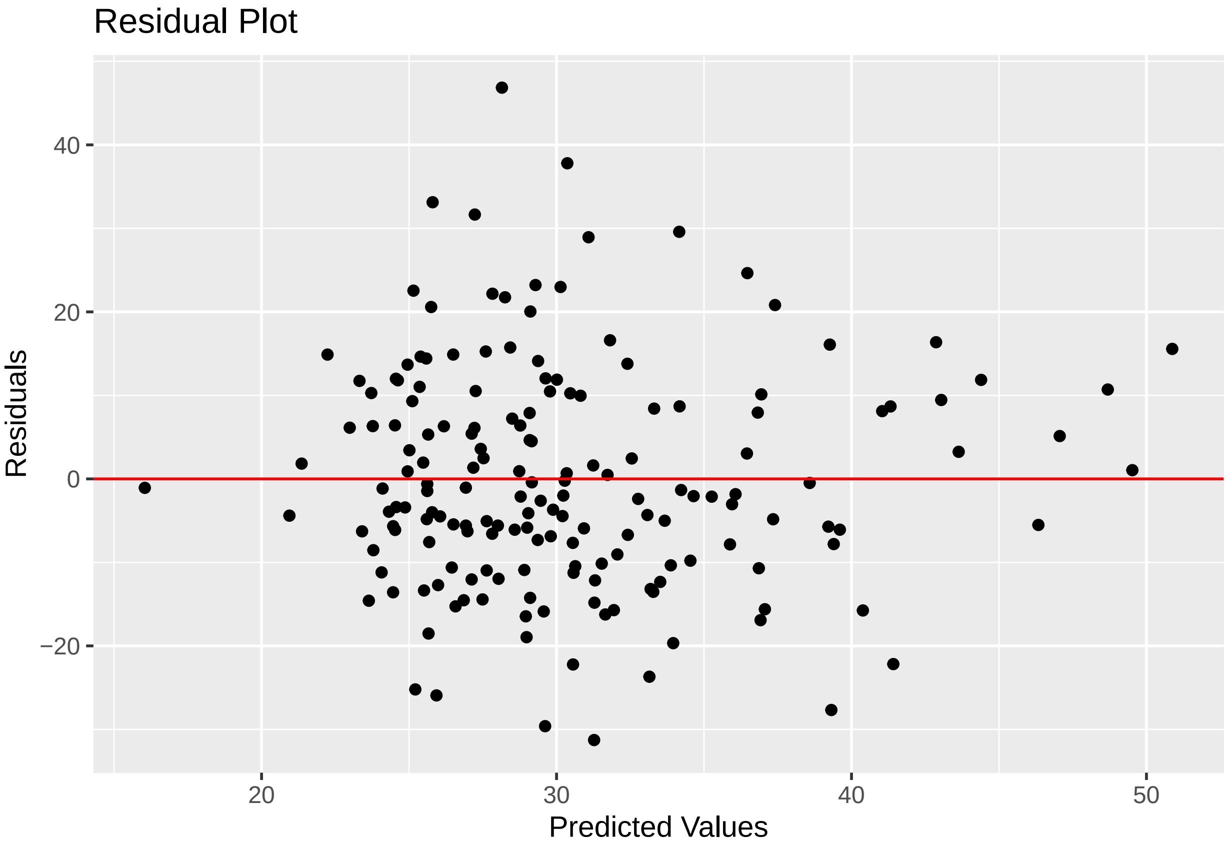
```r
qqnorm(residVal, pch=19)
qqline(residVal, col="red")
```



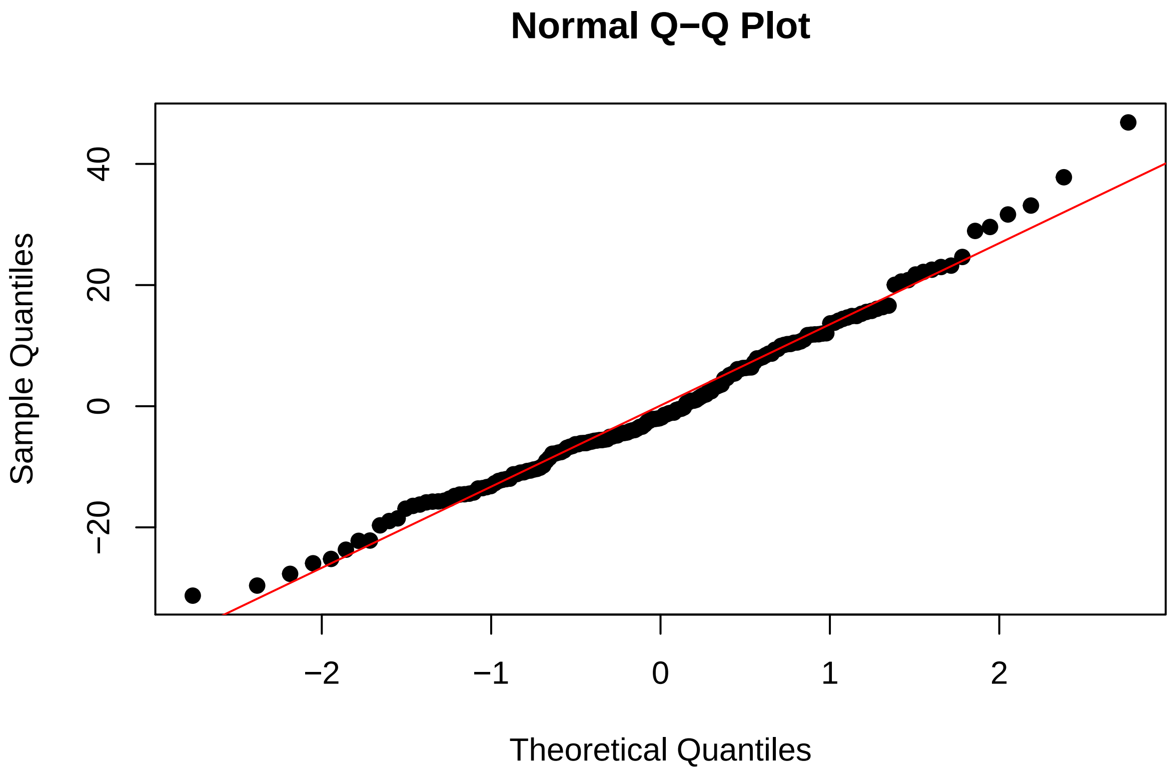
```r
anova(m2)
```

```
## Analysis of Variance Table
##
## Response: mean_violation_rate
##                      Df  Sum Sq Mean Sq F value   Pr(>F)
## estimate_total        1   940.0   940.0  5.0905  0.02534 *
## below_poverty_level   1  3364.8  3364.8 18.2216 3.27e-05 ***
## sum_capacity          1   965.7   965.7  5.2298  0.02344 *
## mean_workers          1   719.1   719.1  3.8941  0.05009 .
## Residuals           169 31207.5   184.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
predVal_2 <- predict(m2)
residVal_2 <- residuals(m2)
ggplot(mapping = aes(x = predVal_2, y = residVal_2)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residual Plot", x = "Predicted Values", y = "Residuals")
```



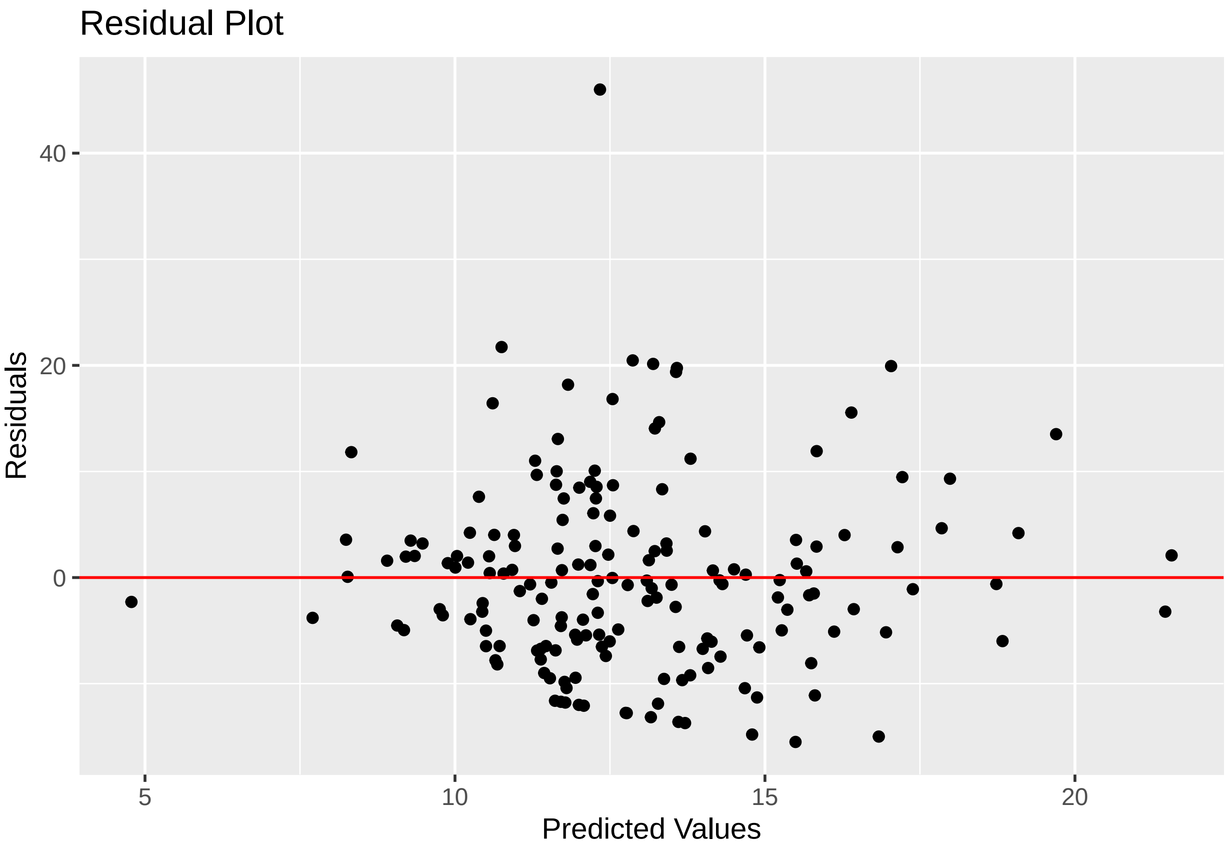
```r
qqnorm(residVal_2, pch=19)
qqline(residVal_2, col="red")
```



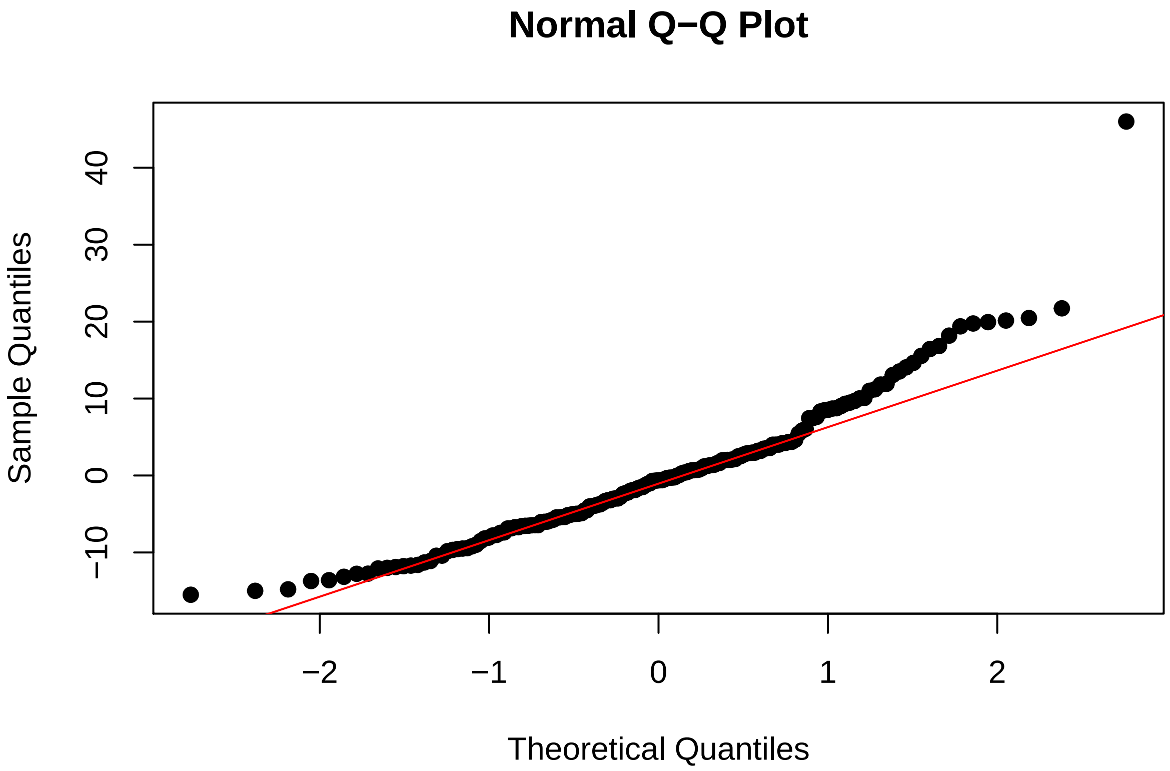
```r
anova(m3)
```

```
## Analysis of Variance Table
##
## Response: mean_health_hazard_violation
##                      Df  Sum Sq Mean Sq F value   Pr(>F)
## estimate_total        1     7.5    7.50  0.0937 0.759854
## below_poverty_level   1   616.0  615.99  7.6958 0.006158 **
## sum_capacity          1   319.6  319.63  3.9933 0.047286 *
## mean_workers          1   179.8  179.79  2.2462 0.135812
## Residuals           169 13527.3   80.04
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
predVal_3 <- predict(m3)
residVal_3 <- residuals(m3)
ggplot(mapping = aes(x = predVal_3, y = residVal_3)) +
  geom_point() +
  geom_hline(yintercept = 0, color = "red") +
  labs(title = "Residual Plot", x = "Predicted Values", y = "Residuals")
```



```
qqnorm(residVal_3, pch=19)
qqline(residVal_3, col="red")
```



```r
# Diagnostic
diag <- ls.diag(m3)
unusual_points <- test %>%
  mutate(h_i = diag$hat,
         stnd_res = diag$std.res,
         stud_res = diag$stud.res,
         cooks = diag$cooks)
# H_i
unusual_points %>%
  filter(h_i > 12/43538  | h_i > 18/43538) %>%
  head(5)
```

```
##   zip_code estimate_total below_poverty_level below_poverty_level_male
## 1    10001          22359                3922                     1874
## 2    10002          77429               21559                     9712
## 3    10003          47093                4655                     2301
## 4    10004           3044                 147                      119
## 5    10005           8710                1052                      420
##   below_poverty_level_male_under_5_years below_poverty_leve_male_5_years
## 1                                      0                             222
## 2                                    350                             183
## 3                                      7                              31
## 4                                      0                               0
## 5                                      0                               0
##   below_poverty_leve_female below_poverty_level_female_under_5_years
## 1                      2048                                        0
## 2                     11847                                      620
## 3                      2354                                       33
## 4                        28                                        0
## 5                       632                                        0
##   below_poverty_level_female_5_years total_count sum_capacity
## 1                                 23          10          608
## 2                                171          34         2551
## 3                                  0           9          429
## 4                                  0           4          122
## 5                                  0           3          205
##   mean_maximum_capacity mean_workers mean_violation_rate
## 1              60.80000     16.40000            25.74747
## 2              75.02941     14.44118            22.04831
## 3              47.66667     10.11111            28.51851
## 4              30.50000      6.50000            21.78570
## 5              68.33333     13.66667            13.69047
##   mean_health_hazard_violation        h_i    stnd_res    stud_res
## 1                    12.818180 0.02191748 -0.07634052 -0.07611563
## 2                     8.035712 0.07377812 -0.27999818 -0.27923333
## 3                    14.382711 0.01326848  0.30688743  0.30606342
## 4                    13.214275 0.03172542  0.13923476  0.13883018
## 5                     0.000000 0.02080977 -1.54874049 -1.55522764
##          cooks
## 1 2.611892e-05
## 2 1.248973e-03
## 3 2.532855e-04
## 4 1.270381e-04
## 5 1.019501e-02
```

```r
# Standardized residual
unusual_points %>%
  filter(abs(stnd_res) > 2 | abs(stnd_res) > 3) %>%
  head(5)
```

```
##   zip_code estimate_total below_poverty_level below_poverty_level_male
## 1    10018           9678                1492                      637
## 2    10454          38485               18060                     7819
## 3    10469          69058               10244                     4292
## 4    10475          43407                4648                     2016
## 5    11109           4964                 400                      135
##   below_poverty_level_male_under_5_years below_poverty_leve_male_5_years
## 1                                     24                               0
## 2                                   1062                             187
## 3                                    320                              61
## 4                                     91                             132
## 5                                      0                               0
##   below_poverty_leve_female below_poverty_level_female_under_5_years
## 1                       855                                       21
## 2                     10241                                      795
## 3                      5952                                      429
## 4                      2632                                      253
## 5                       265                                        0
##   below_poverty_level_female_5_years total_count sum_capacity
## 1                                  0           1           62
## 2                                261           9          588
## 3                                 93           8          598
## 4                                  0           2          104
## 5                                  0           2           90
##   mean_maximum_capacity mean_workers mean_violation_rate
## 1              62.00000      8.00000            50.00000
## 2              65.33333     10.33333            55.34391
## 3              74.75000     14.75000            46.19046
## 4              52.00000      7.50000            75.00000
## 5              45.00000     10.50000            50.00000
##   mean_health_hazard_violation        h_i stnd_res stud_res      cooks
## 1                     33.33330 0.02248379 2.313757 2.344331 0.02462698
## 2                     36.96649 0.03381068 2.266477 2.294909 0.03595211
## 3                     32.94642 0.02707901 2.196123 2.221544 0.02684716
## 4                     58.33335 0.02283300 5.200644 5.657689 0.12639751
## 5                     33.33335 0.02145090 2.275327 2.304152 0.02269763
```

```r
# Studentized residual
unusual_points %>%
  filter(abs(stud_res) > 2 | abs(stud_res) > 3) %>%
  head(5)
```

```
##   zip_code estimate_total below_poverty_level below_poverty_level_male
## 1    10018           9678                1492                      637
## 2    10454          38485               18060                     7819
## 3    10469          69058               10244                     4292
## 4    10475          43407                4648                     2016
## 5    11109           4964                 400                      135
##   below_poverty_level_male_under_5_years below_poverty_leve_male_5_years
## 1                                     24                               0
## 2                                   1062                             187
## 3                                    320                              61
## 4                                     91                             132
## 5                                      0                               0
##   below_poverty_leve_female below_poverty_level_female_under_5_years
## 1                       855                                       21
## 2                     10241                                      795
## 3                      5952                                      429
## 4                      2632                                      253
## 5                       265                                        0
##   below_poverty_level_female_5_years total_count sum_capacity
## 1                                  0           1           62
## 2                                261           9          588
## 3                                 93           8          598
## 4                                  0           2          104
## 5                                  0           2           90
##   mean_maximum_capacity mean_workers mean_violation_rate
## 1              62.00000      8.00000            50.00000
## 2              65.33333     10.33333            55.34391
## 3              74.75000     14.75000            46.19046
## 4              52.00000      7.50000            75.00000
## 5              45.00000     10.50000            50.00000
##   mean_health_hazard_violation        h_i stnd_res stud_res      cooks
## 1                     33.33330 0.02248379 2.313757 2.344331 0.02462698
## 2                     36.96649 0.03381068 2.266477 2.294909 0.03595211
## 3                     32.94642 0.02707901 2.196123 2.221544 0.02684716
## 4                     58.33335 0.02283300 5.200644 5.657689 0.12639751
## 5                     33.33335 0.02145090 2.275327 2.304152 0.02269763
```

Introduction

Childcare resources distribute inequality