Assignment 1: Regression

Giorgi Kukishvili

2023-02-26

R Markdown

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'
```

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

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
## arm (Version 1.13-1, built: 2022-8-25)
```

```
## Working directory is /Users/giorgikukishvili
```

```
## Registering fonts with R
```

```
##
## Attaching package: 'boot'
```

```
## The following object is masked from 'package:arm':
##
##     logit
```

```
##  [1] "treatment"  "pri2000s"   "pri2000v"   "t2000"      "t2000r"
##  [6] "pri1994"    "pan1994"    "prd1994"    "pri1994s"   "pri1994v"
## [11] "pan1994s"   "pan1994v"   "prd1994s"   "prd1994v"   "t1994"
## [16] "t1994r"     "votos1994"  "avgpoverty" "pobtot1994" "villages"
```

```
##        0        1
## 63.81483 68.08451
```

```
##        0        1
## 34.48895 38.11145
```

```
##
## Call:
## lm(formula = t2000 ~ treatment, data = progresa)
##
## Coefficients:
## (Intercept)    treatment
##       63.81         4.27
```

```
##
## Call:
## lm(formula = pri2000s ~ treatment, data = progresa)
##
## Coefficients:
## (Intercept)    treatment
##      34.489        3.622
```

```
##
## Call:
## lm(formula = t2000 ~ treatment + avgpoverty + pobtot1994 + votos1994 +
##     pri1994 + pan1994 + prd1994, data = progresa)
##
## Coefficients:
## (Intercept)    treatment   avgpoverty   pobtot1994    votos1994      pri1994
##   64.011735     4.549445     0.310255    -0.001213    -0.026152     0.036055
##     pan1994      prd1994
##    0.026538     0.017575
```

```
##
## Call:
## lm(formula = pri2000s ~ treatment + avgpoverty + pobtot1994 +
##     votos1994 + pri1994 + pan1994 + prd1994, data = progresa)
##
## Coefficients:
## (Intercept)    treatment   avgpoverty   pobtot1994    votos1994      pri1994
##  37.9500862    2.9277395    0.5329801   -0.0004996   -0.0417278    0.0624589
##     pan1994      prd1994
##  -0.0487349   -0.0287363
```

```
##
## Call:
## lm(formula = t2000 ~ treatment + avgpoverty + log(pobtot1994) +
##     t1994 + pri1994s + pan1994s + prd1994s, data = progresa)
##
## Coefficients:
##     (Intercept)        treatment       avgpoverty  log(pobtot1994)
##         19.7984          -0.1530           2.8621          -3.2471
##           t1994         pri1994s         pan1994s         prd1994s
##          0.6605           0.1943           0.6374           0.3065
```

```
##
## Call:
## lm(formula = pri2000s ~ treatment + avgpoverty + log(pobtot1994) +
##     t1994 + pri1994s + pan1994s + prd1994s, data = progresa)
##
## Coefficients:
##     (Intercept)        treatment       avgpoverty  log(pobtot1994)
##        35.85174          0.23547          2.47163         -4.62934
##           t1994         pri1994s         pan1994s         prd1994s
##         0.03257          0.51047         -0.18384         -0.05293
```

```
## [1] 0.006297673
```

```
## [1] 0.006297673
```

```
## [1] 0.06273301
```

```
## [1] 0.2072516
```

```
## [1] 0.6868331
```

```
## [1] 0.5721621
```

#Graded part ###Setting up the data

```
## [1] 417
```

```
## [1] 416
```

Exercise 1:

setting up dataframe for histograms

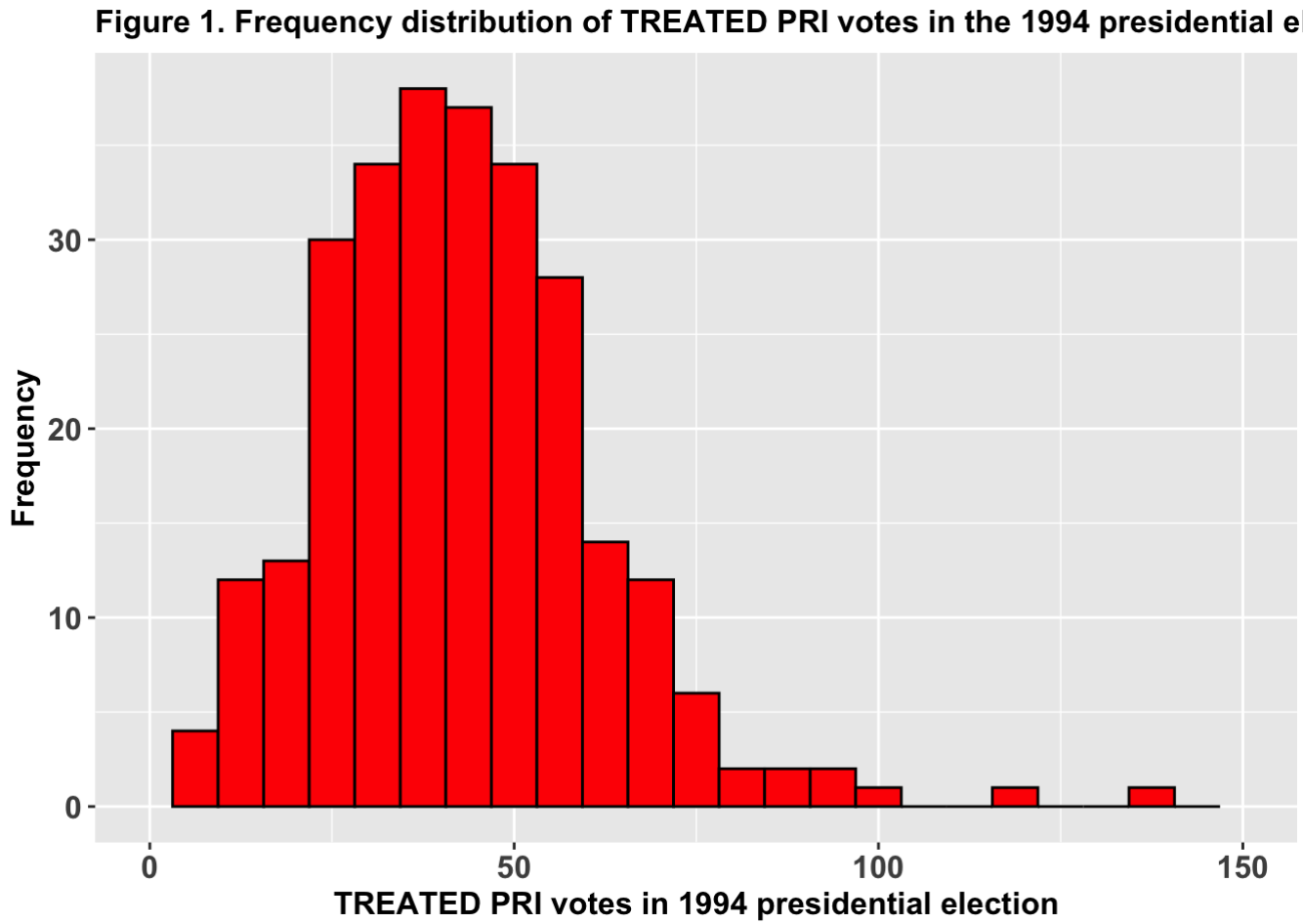
```
## [1] 20
```

```
## [1] 21
```

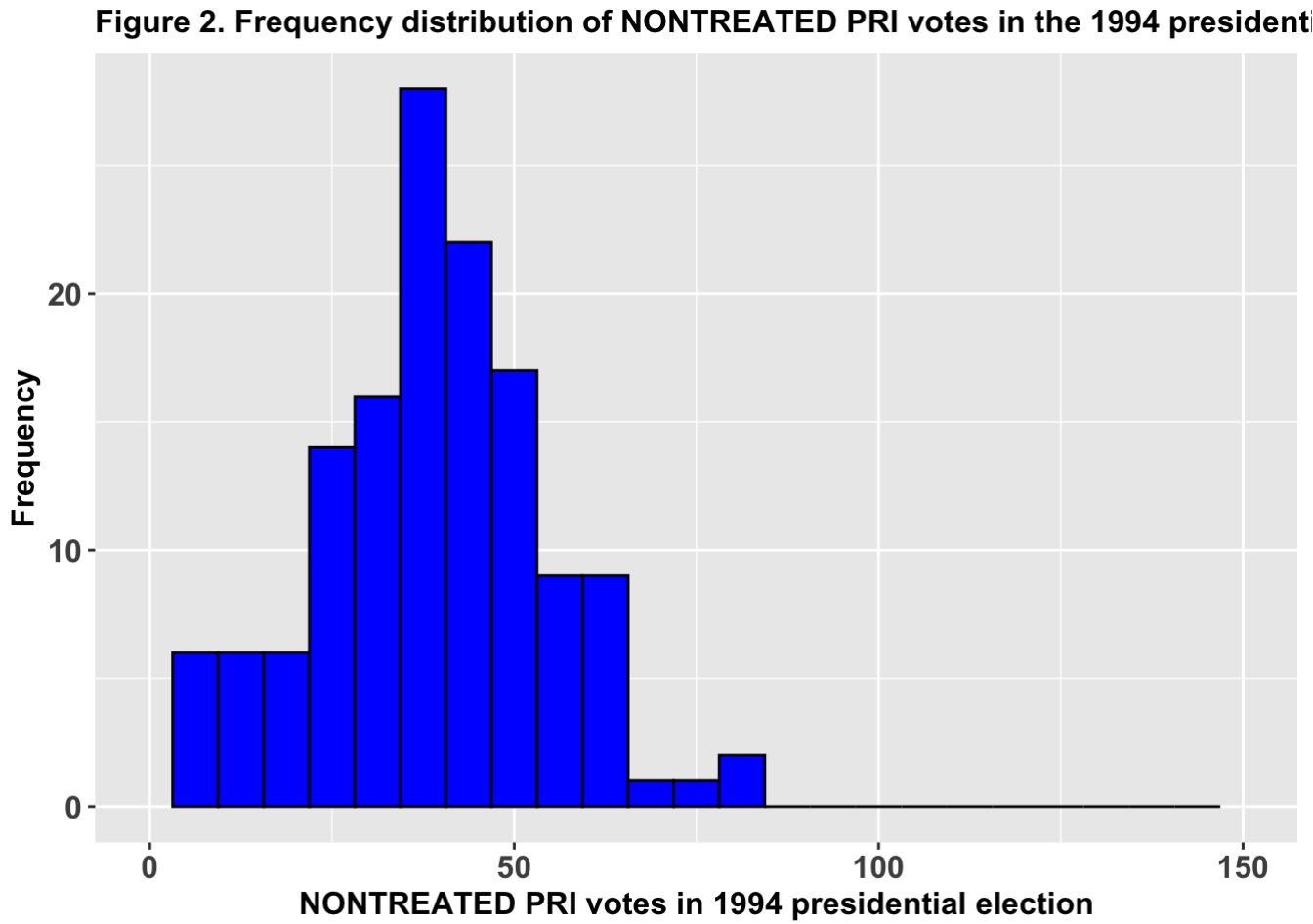
##Examining the treatment and control group distributions for variable pri1994s

```
## Warning: Removed 2 rows containing non-finite values (`stat_bin()`).
```

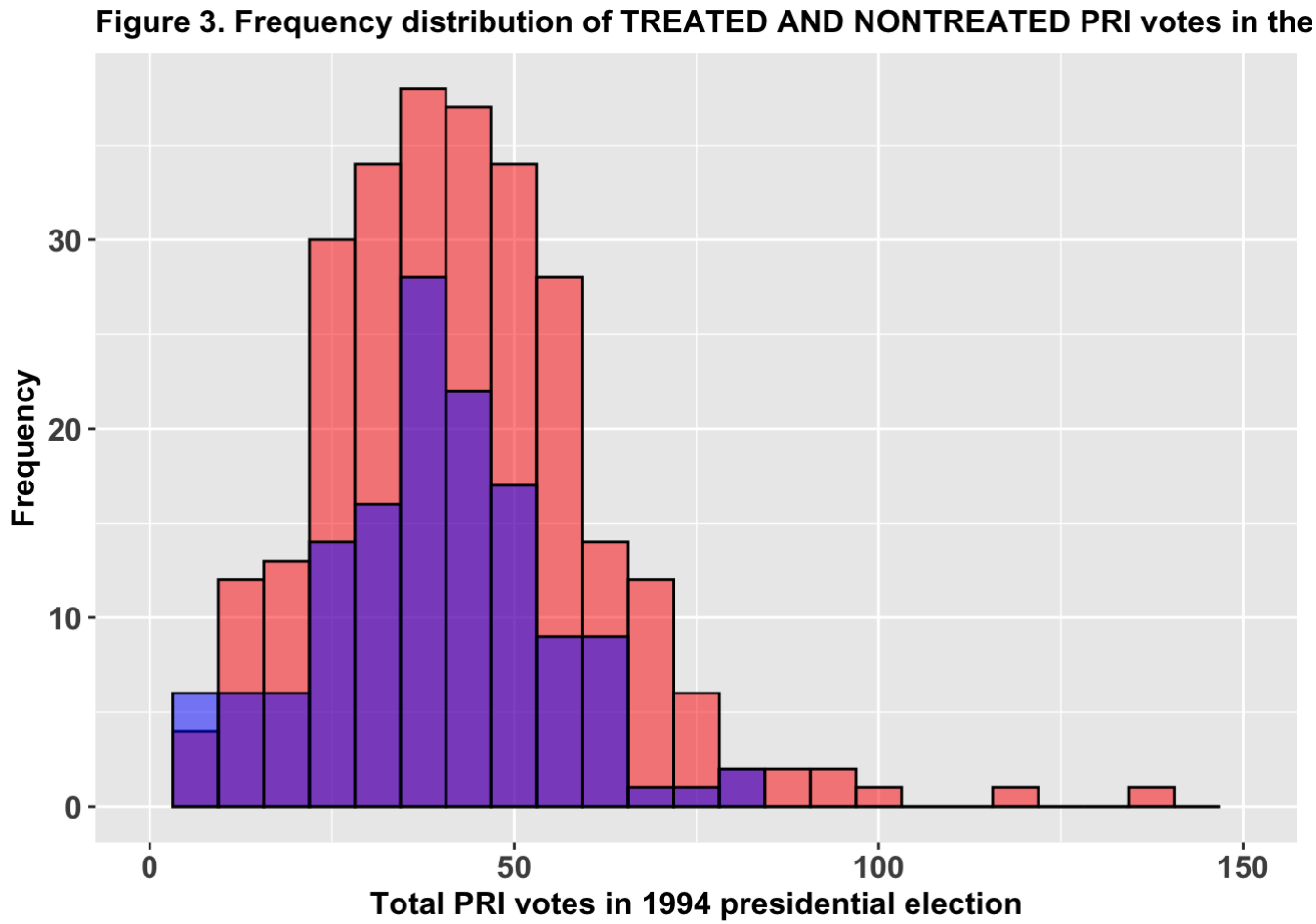
```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```




```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```




```
## Warning: Removed 2 rows containing non-finite values (`stat_bin()`).
## Removed 2 rows containing missing values (`geom_bar()`).
## Removed 2 rows containing missing values (`geom_bar()`).
```

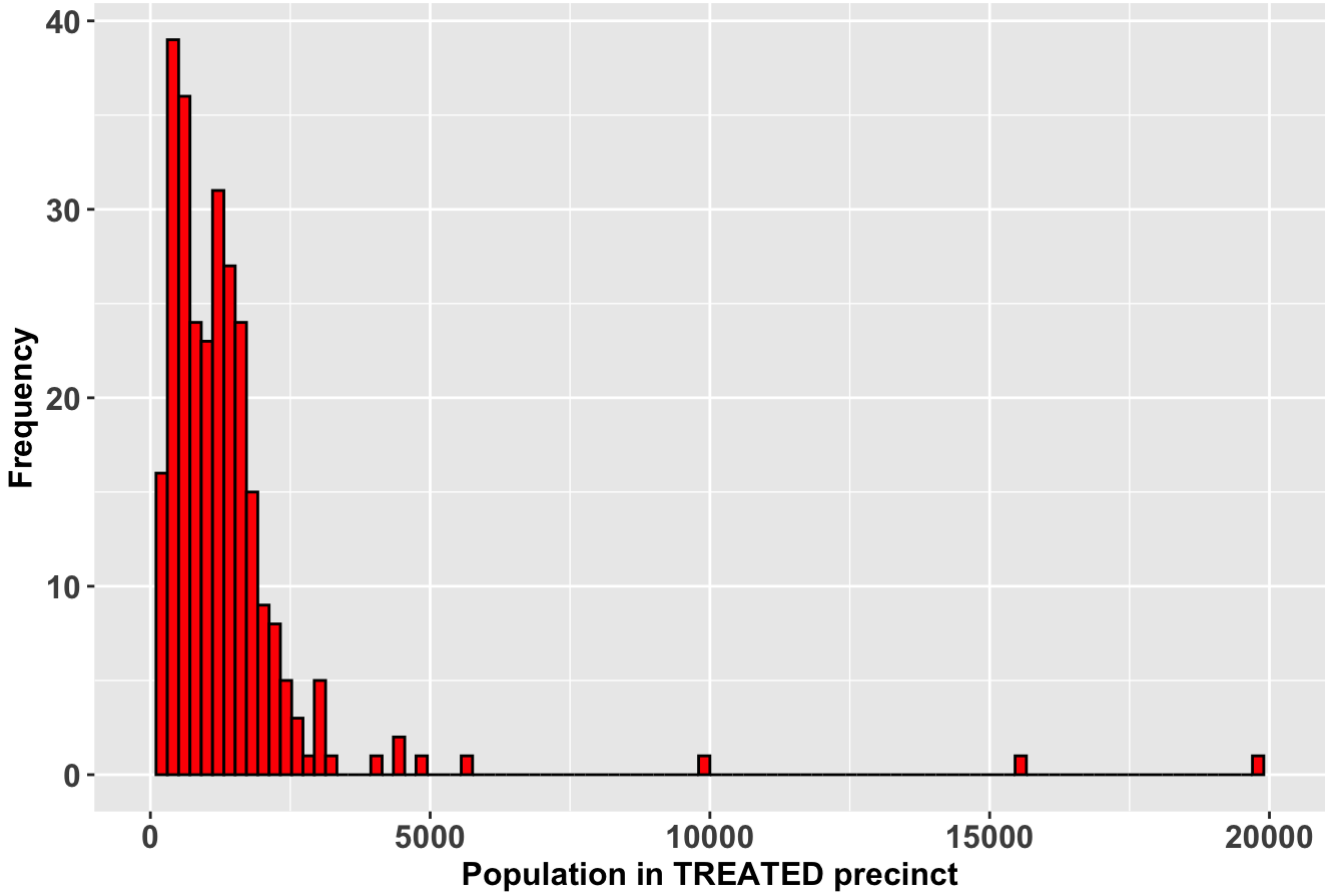


##Examining the treatment and control group distributions for variable pobtot1994

```
## Warning: Removed 3 rows containing non-finite values (`stat_bin()`).
```

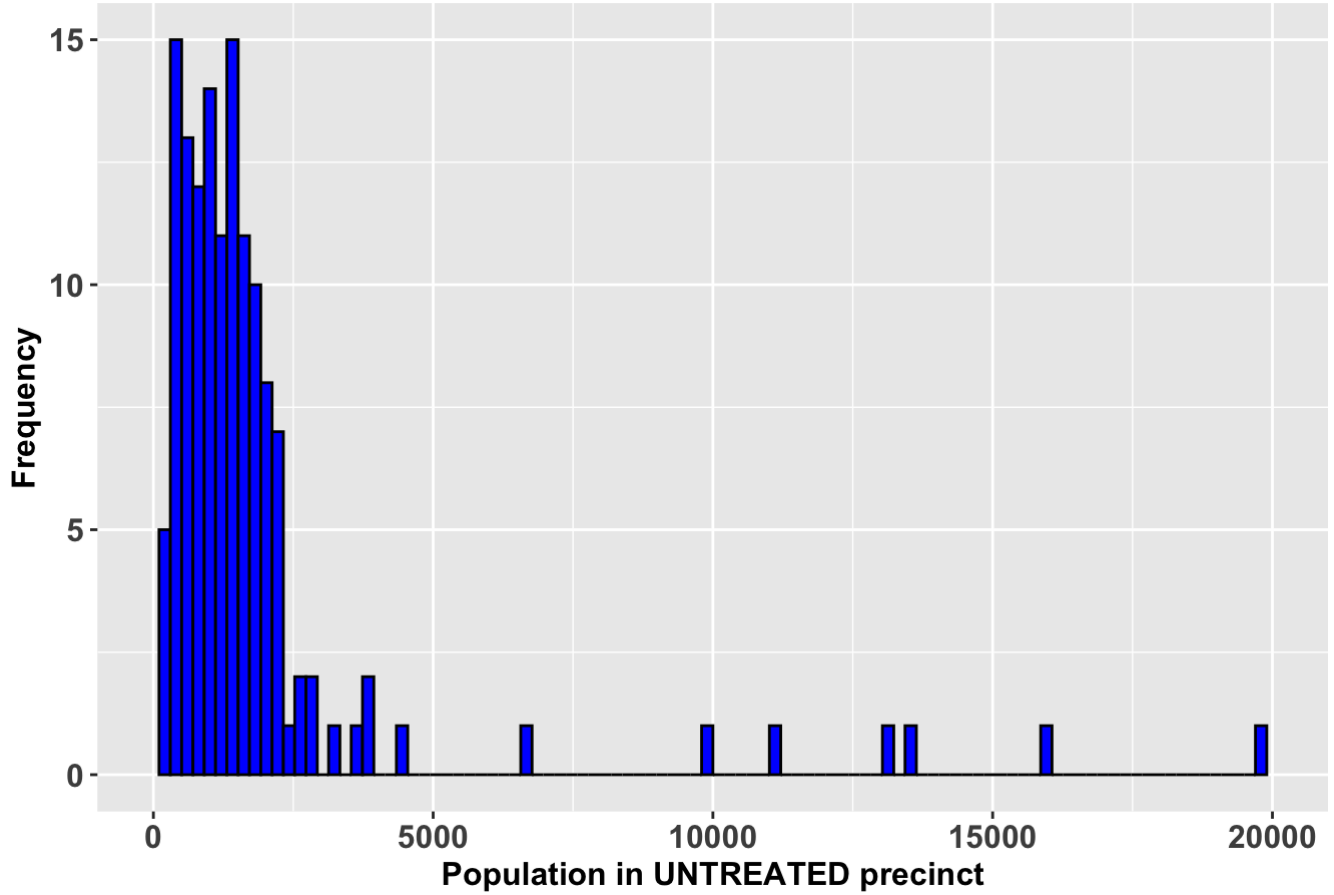
```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```

Figure 4. Frequency distribution of total population in TREATED precinct

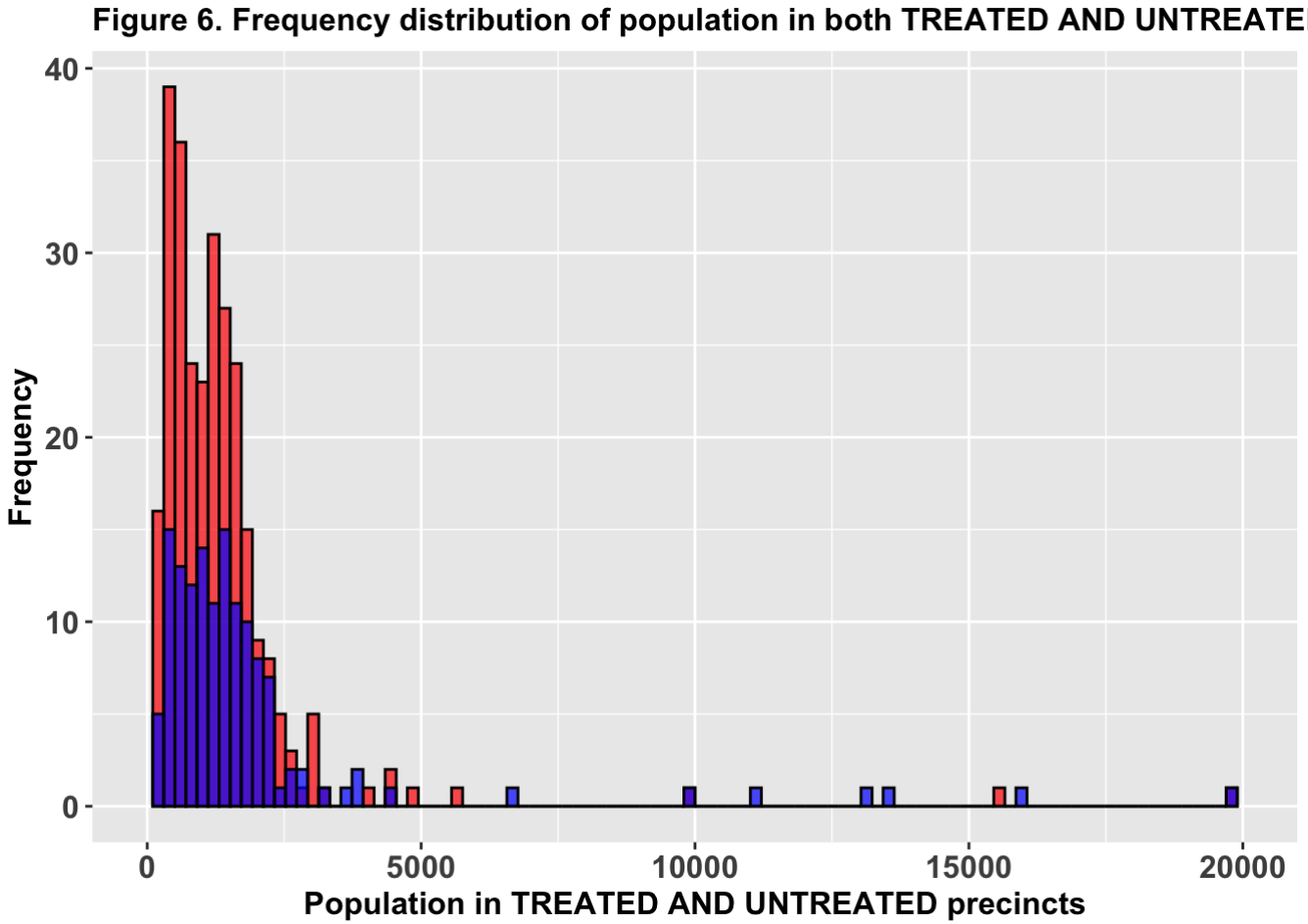



```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```

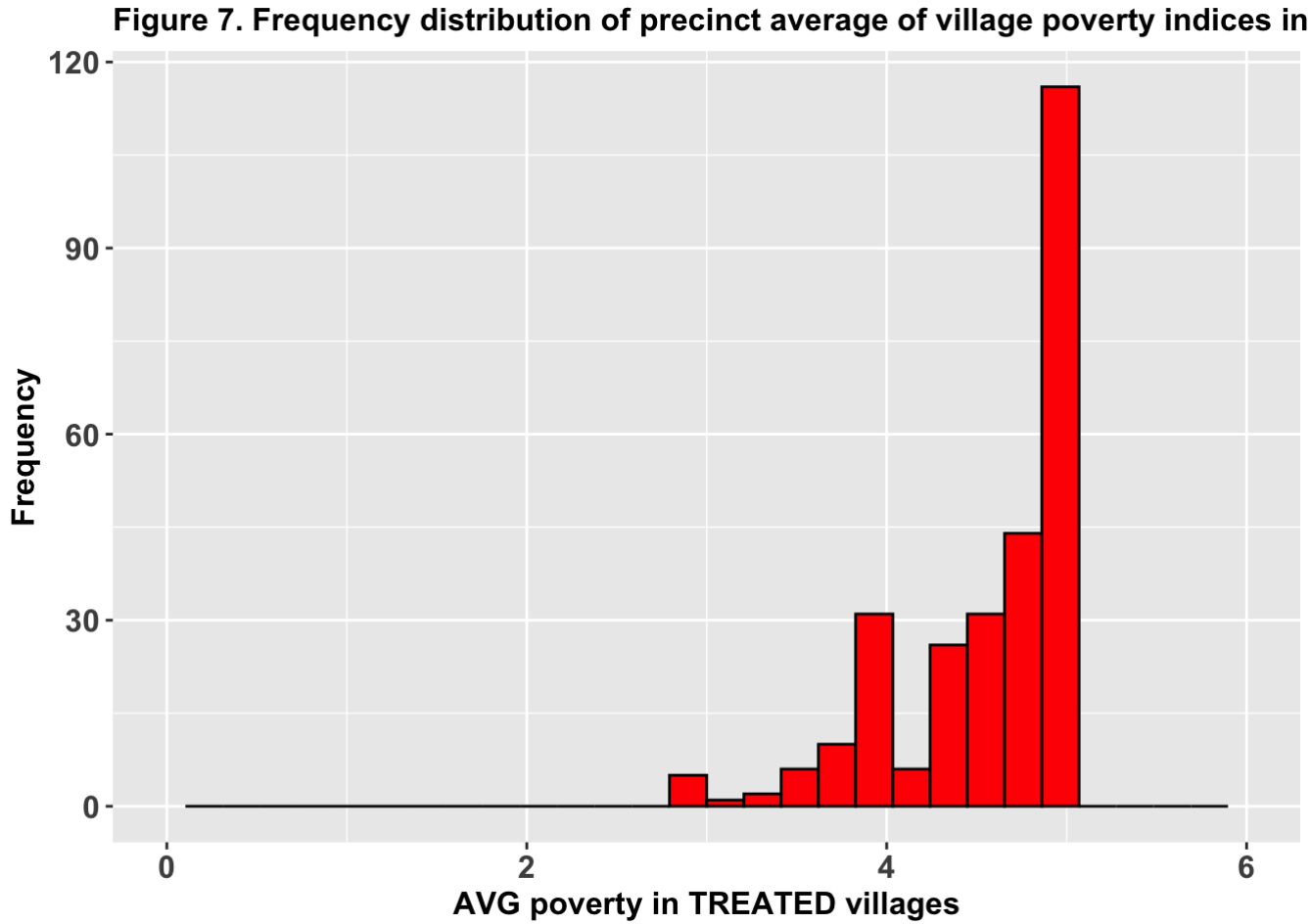
Figure 5. Frequency distribution of total population in UNTREATED precinct



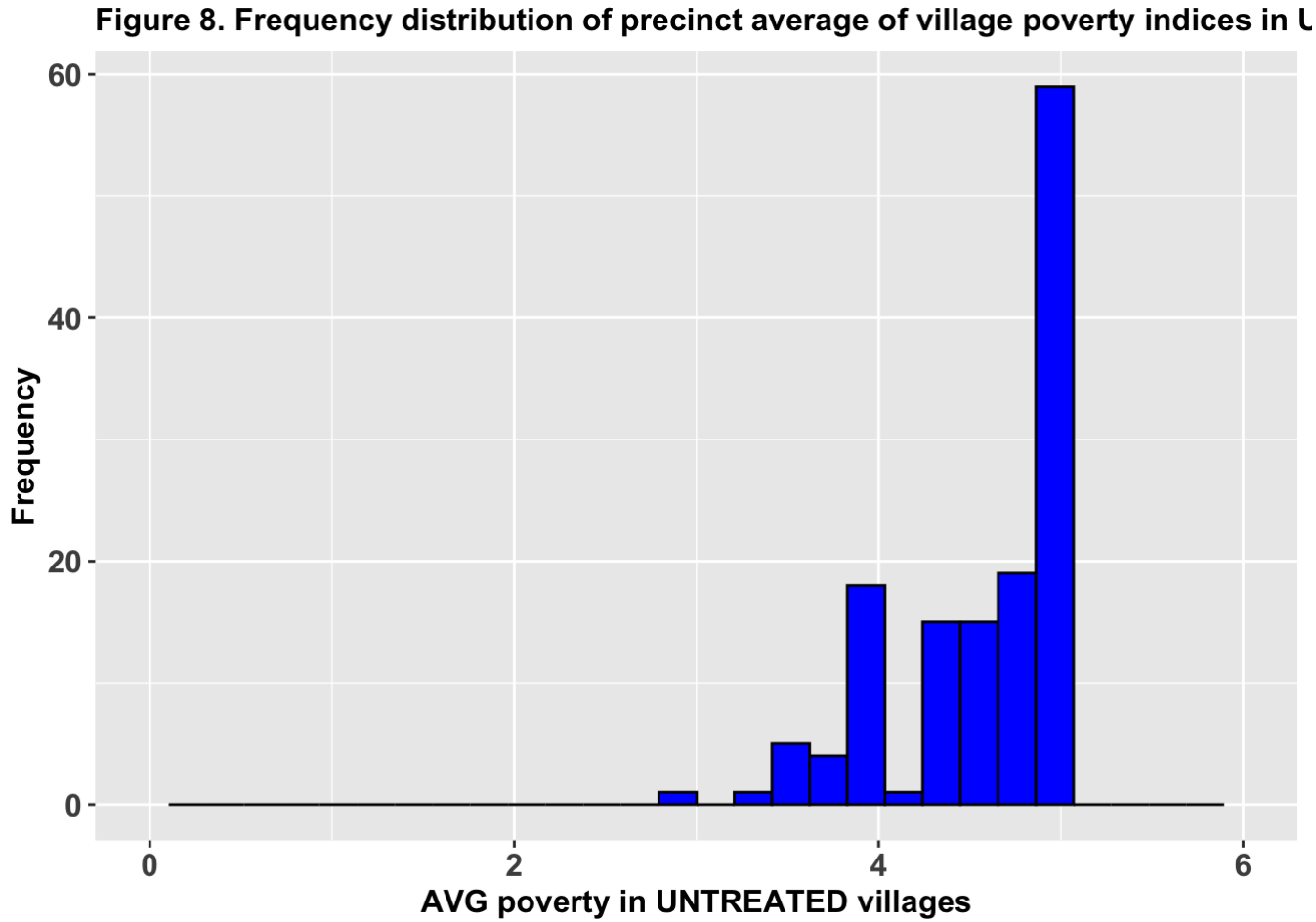

```
## Warning: Removed 3 rows containing non-finite values (`stat_bin()`).
## Removed 2 rows containing missing values (`geom_bar()`).
## Removed 2 rows containing missing values (`geom_bar()`).
```




```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```

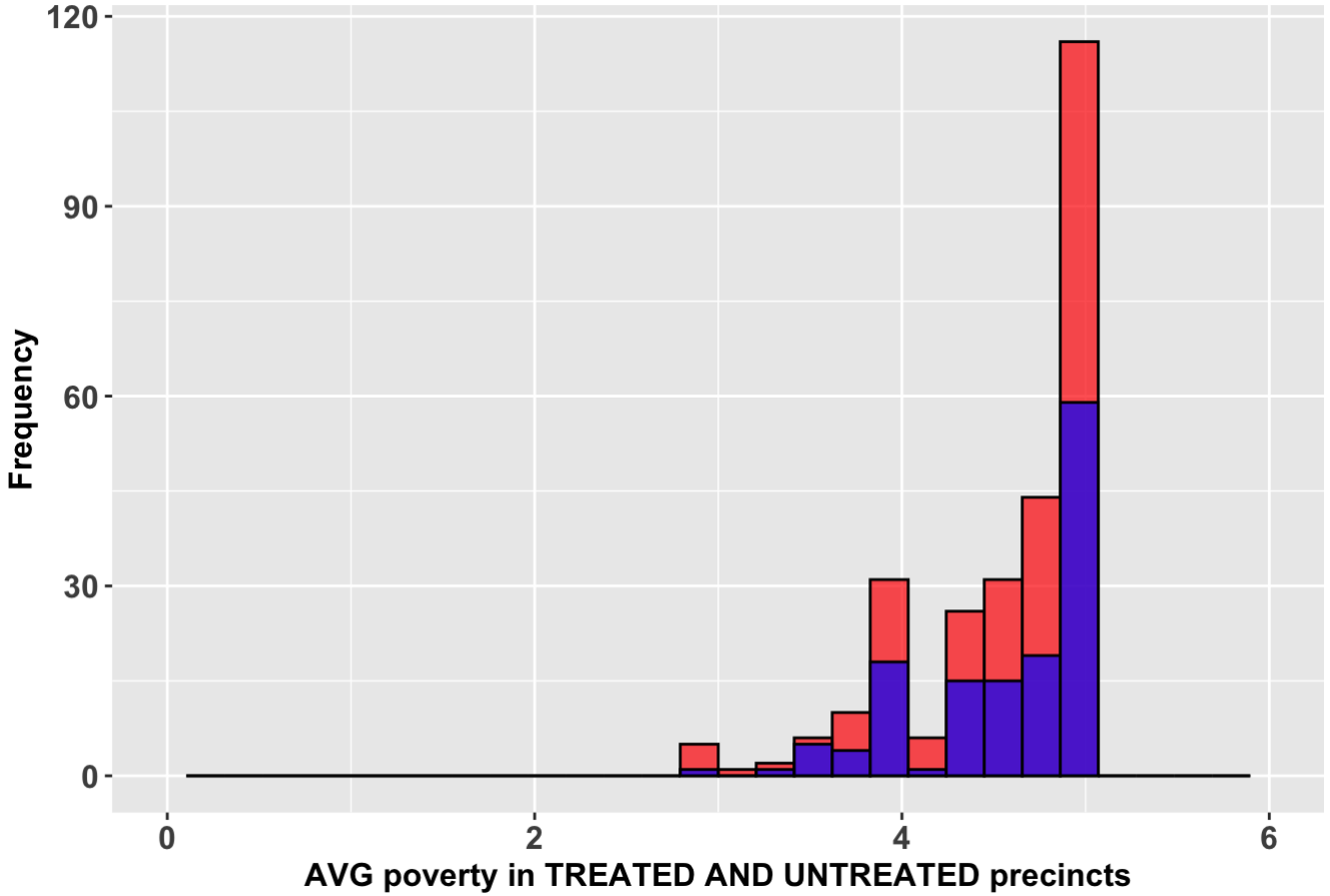



```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```



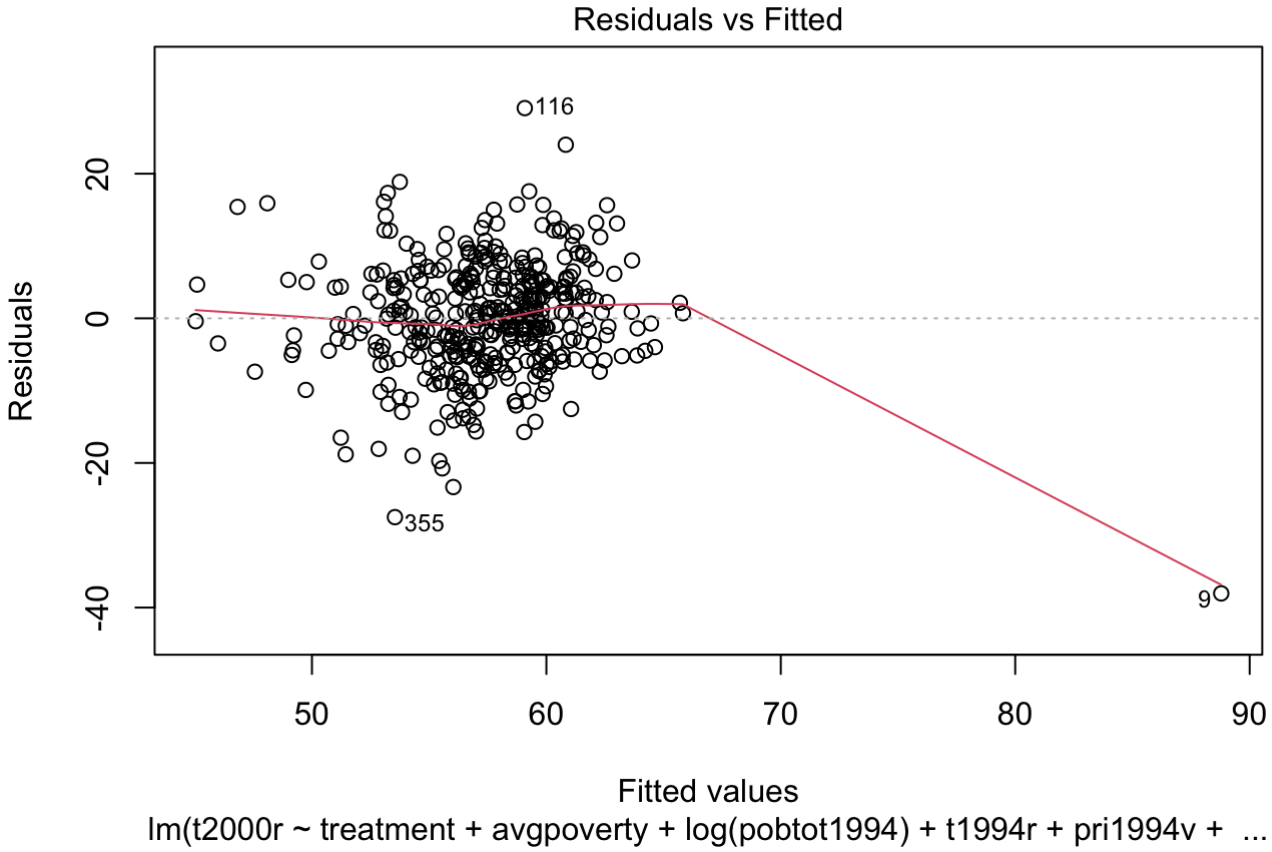

```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
## Removed 2 rows containing missing values (`geom_bar()`).
```

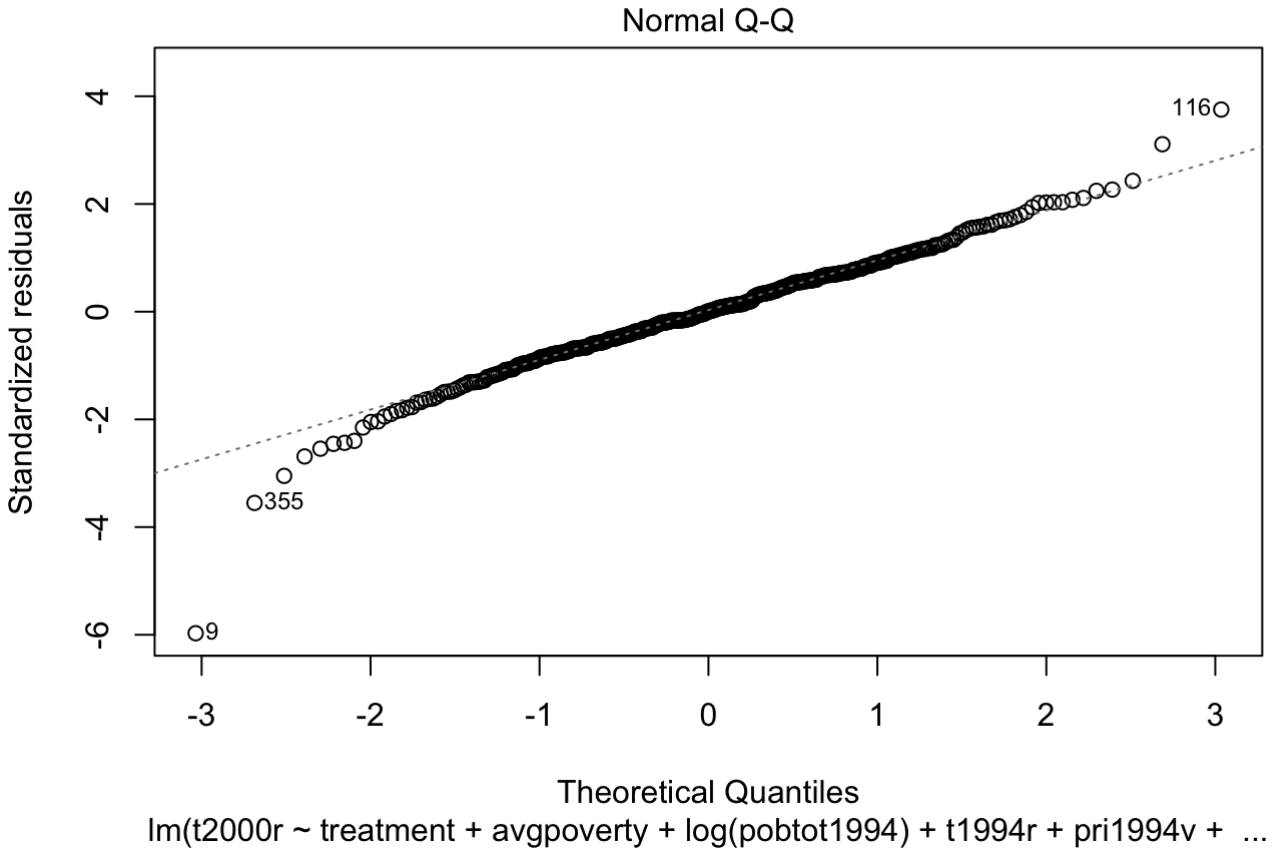
Figure 9. Frequency distribution of precinct average of village poverty indices in

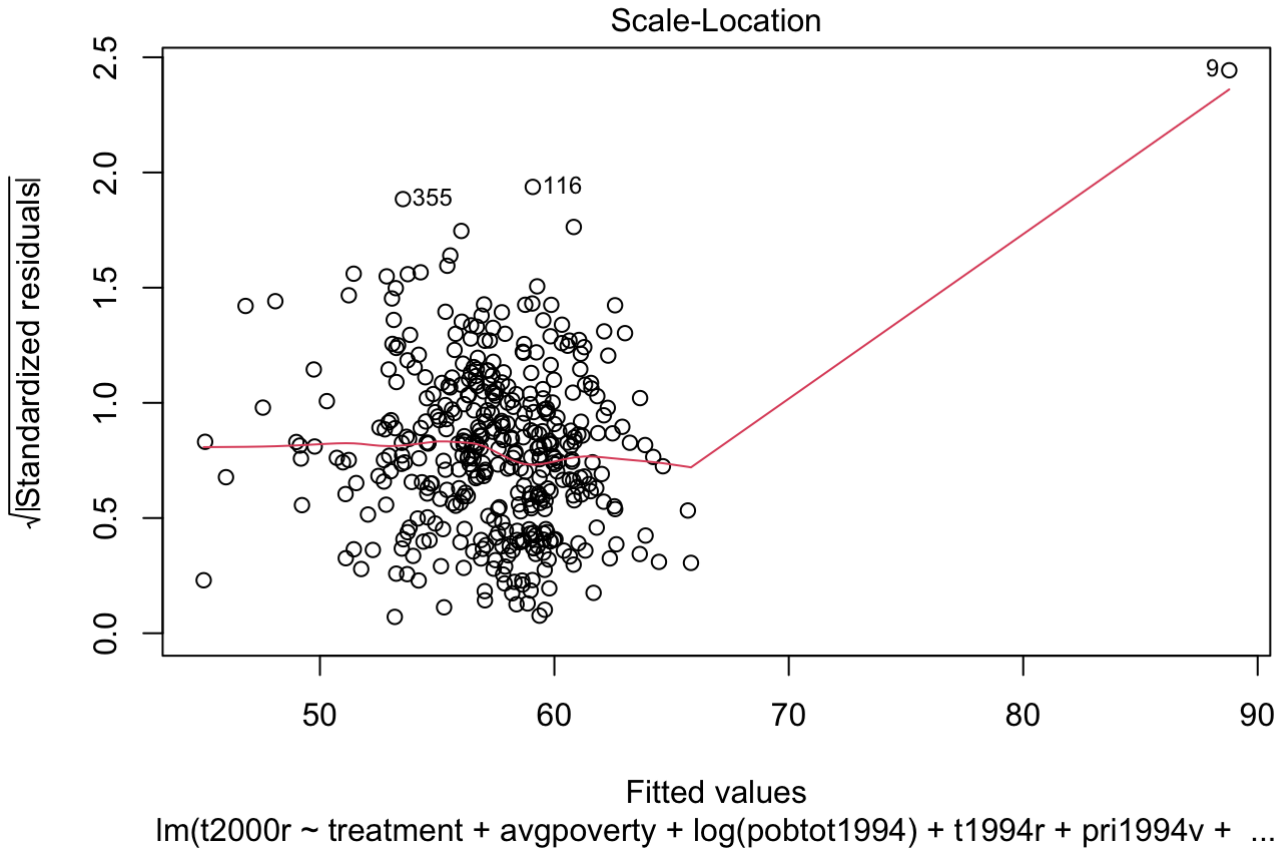


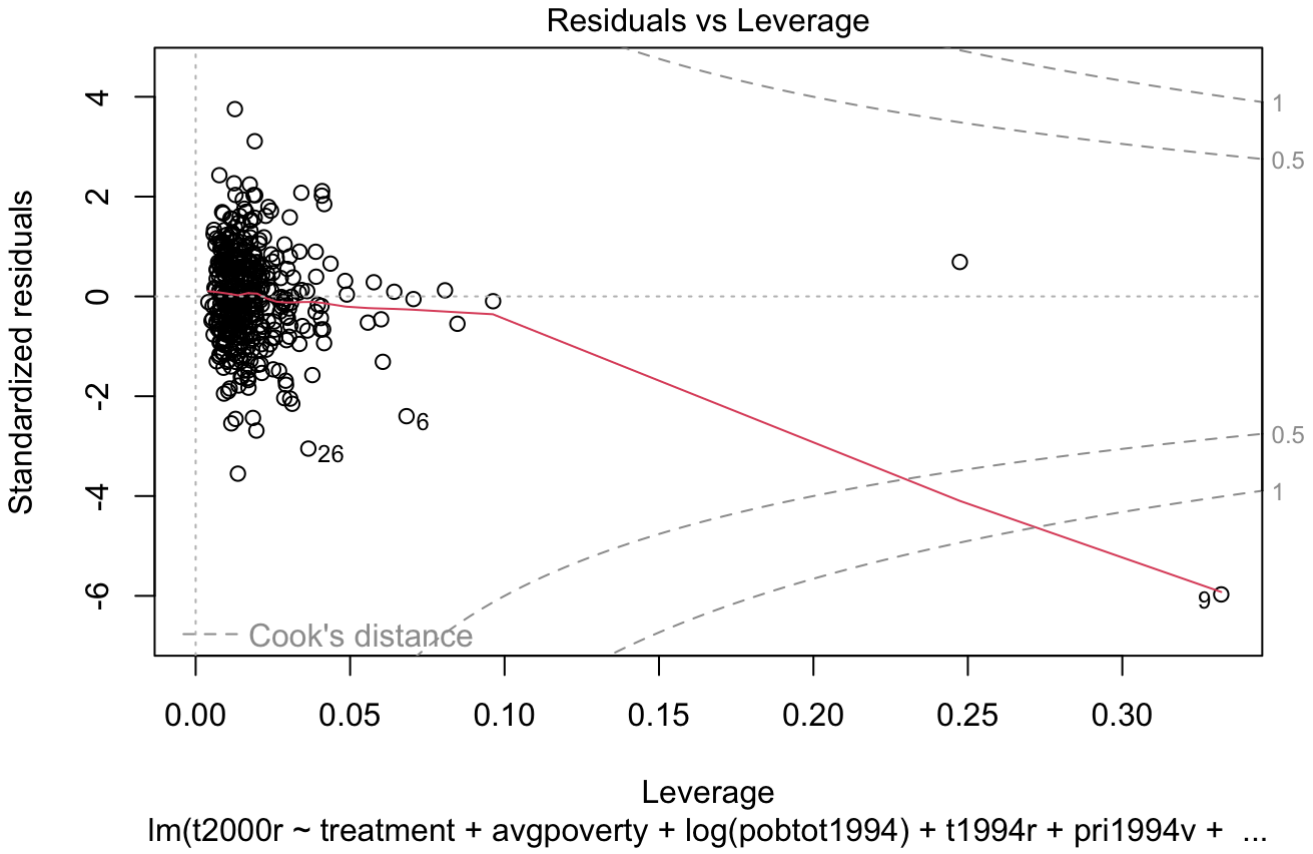
###the results indicate that in the treated precincts 1) there were more PRI woters. 2) Treated precincts were more populated and also 3) poorer. The histograms indicate that there can be a relationship that's worth examining betweenpoverty and birthrates, poverty and support to certain party. Additionally, there can be a nice findings if we look into differences between health levels within treated and untreated precincts.

#Exercise 2 ##1. First regression using t2000r as the outcome variable.



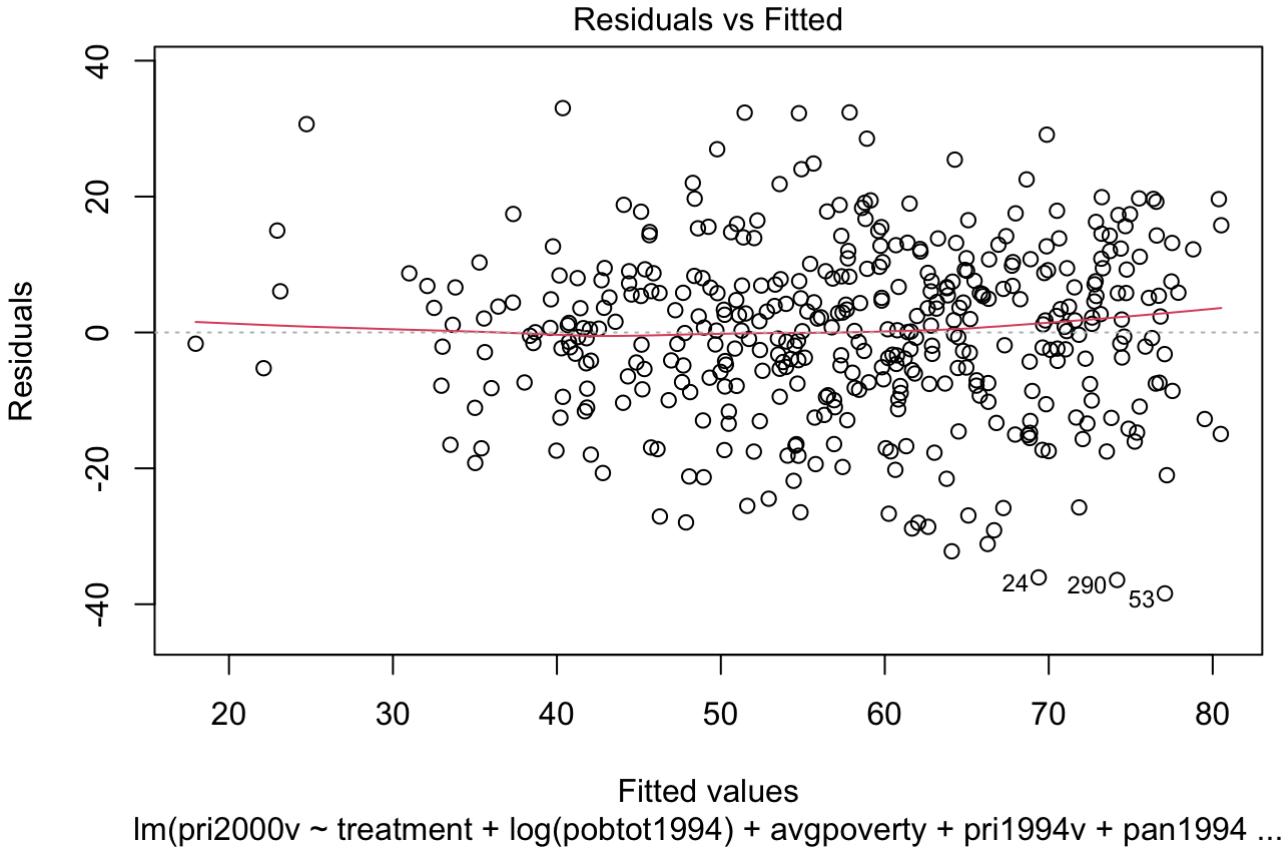

Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(t2000r ~ treatment + avgpoverty + log(pobtot1994) + t1994r + pri1994v + ...
116
355
9
4
2
0
-2
-4
-6
-3
-2
-1
0
1
2
3

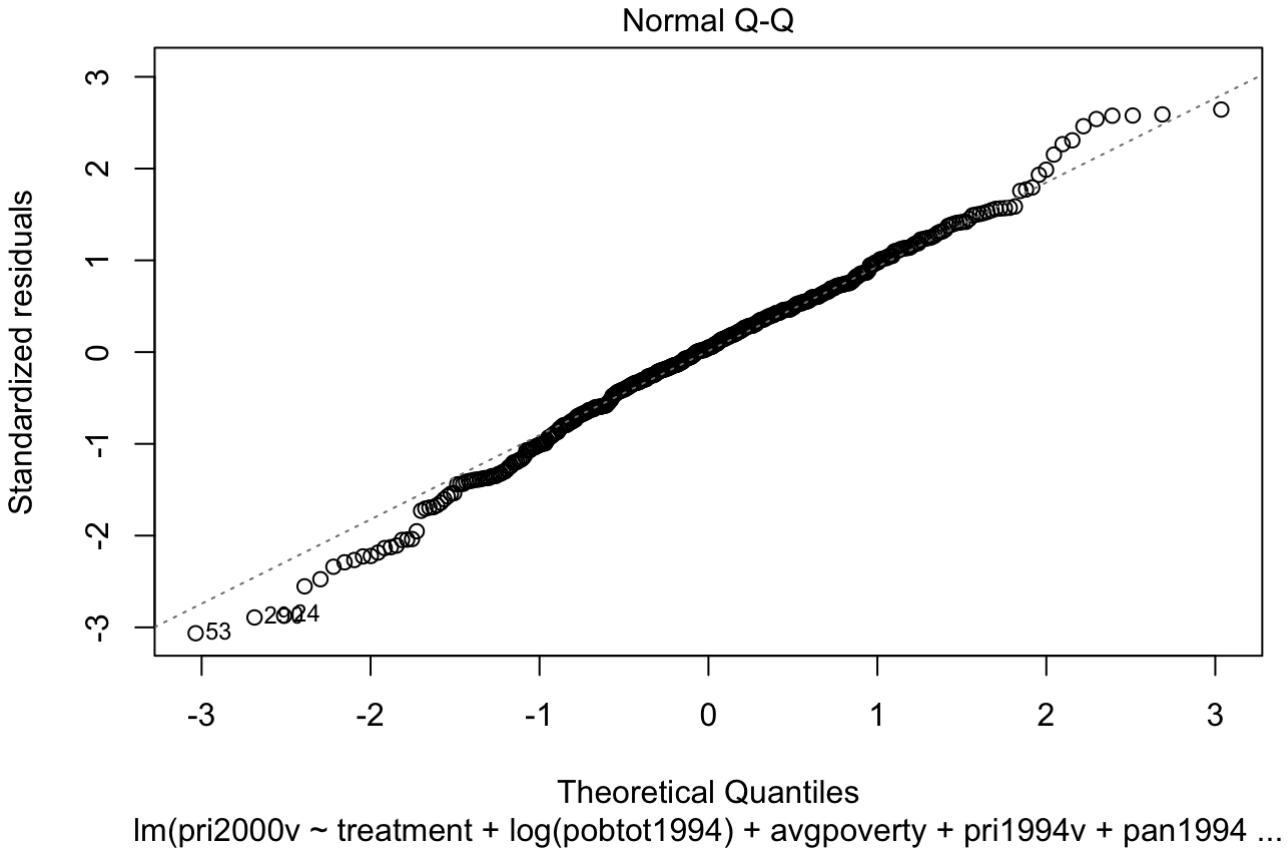

Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
50
60
70
80
90
355
116
9
Fitted values
lm(t2000r ~ treatment + avgpoverty + log(pobtot1994) + t1994r + pri1994v + ...

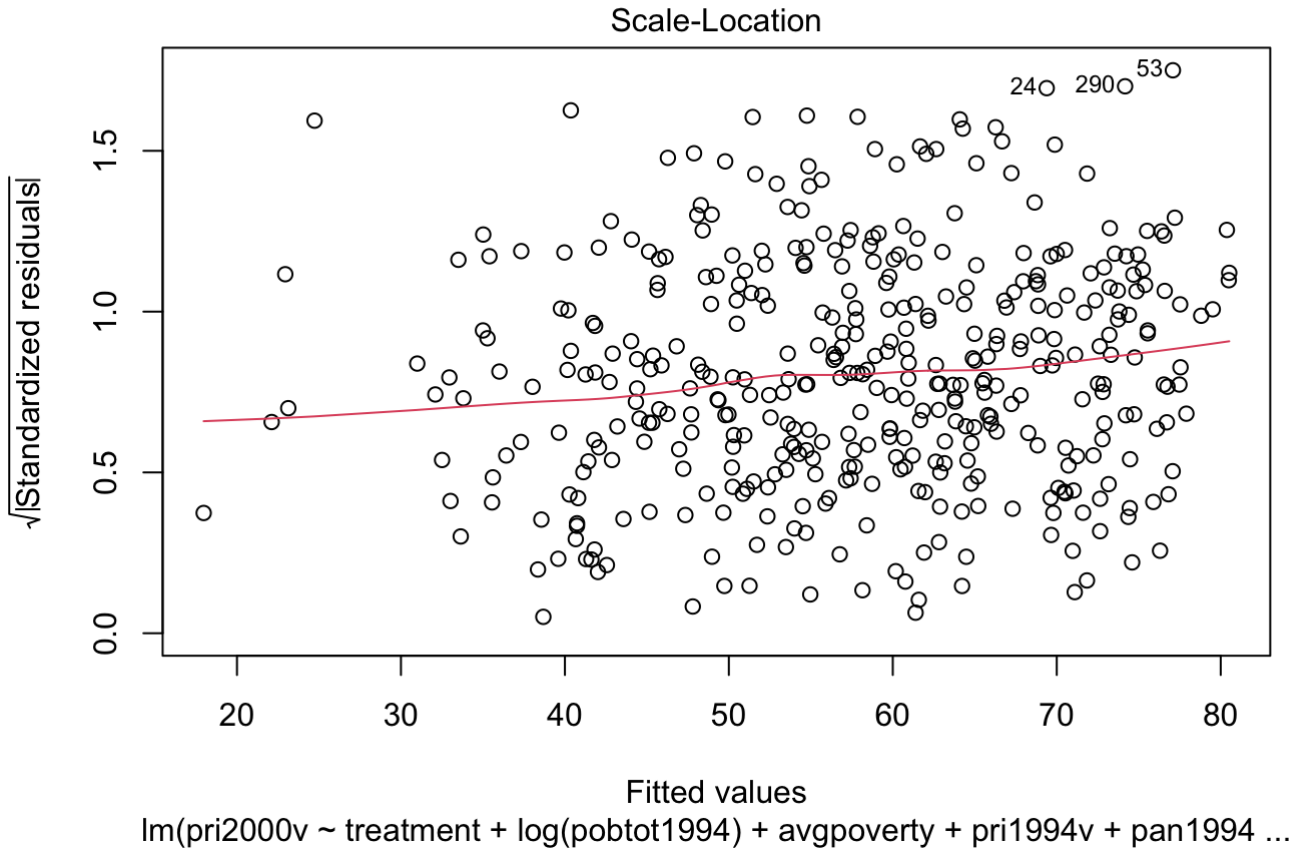

Residuals vs Leverage
Standardized residuals
4
2
0
-2
-4
-6
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(t2000r ~ treatment + avgpoverty + log(pobtot1994) + t1994r + pri1994v + ...
1
0.5
0.5
1
Cook's distance
26
6
9

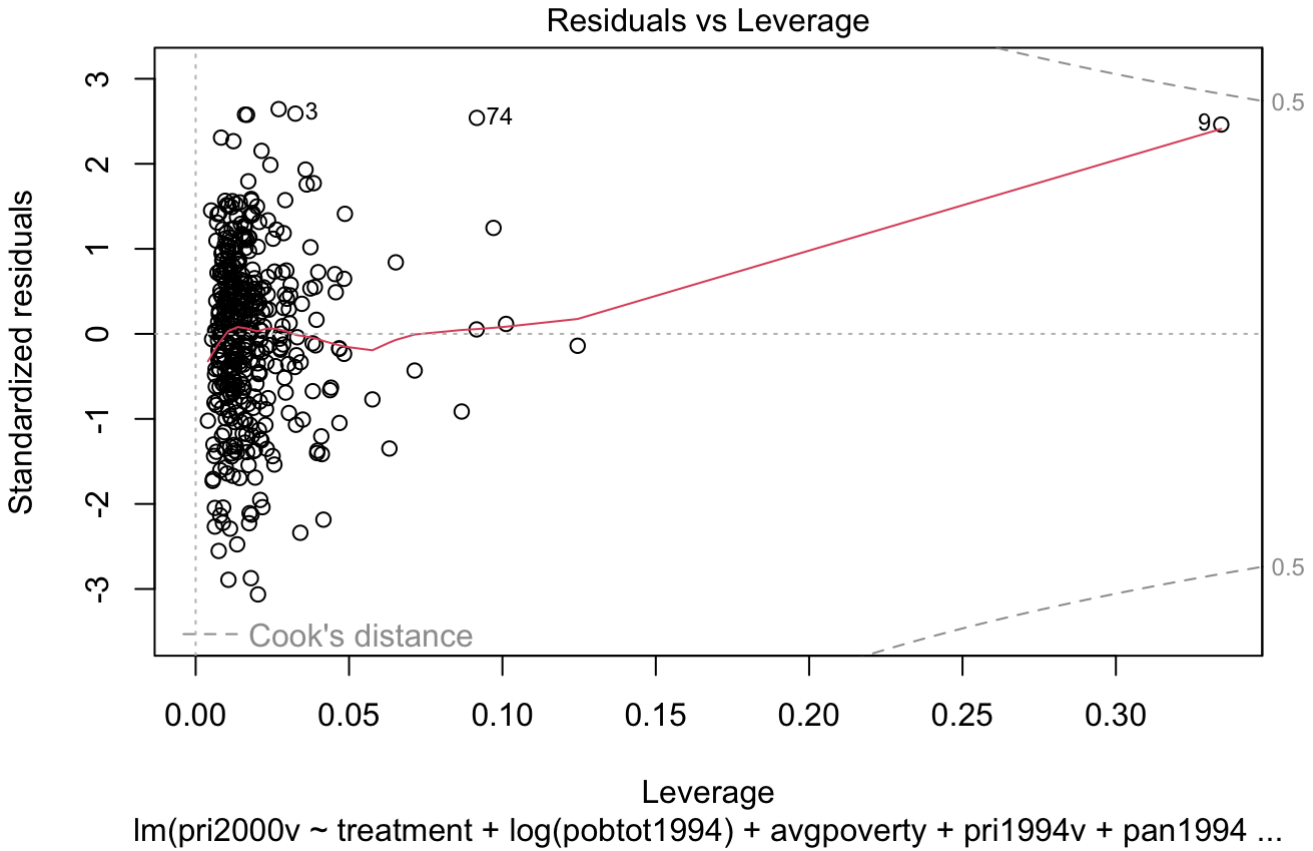
```
##
## Call:
## lm(formula = t2000r ~ treatment + avgpoverty + log(pobtot1994) +
##     t1994r + pri1994v + pan1994v + prd1994v, data = progresa1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.041  -4.555   0.029   5.010  29.069
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    33.47237    8.85553   3.780 0.000180 ***
## treatment      -1.08094    0.81683  -1.323 0.186465
## avgpoverty     -0.27734    0.88768  -0.312 0.754874
## log(pobtot1994) -0.27878    0.47487  -0.587 0.557487
## t1994r          0.22238    0.03102   7.168 3.59e-12 ***
## pri1994v        0.12473    0.05489   2.272 0.023586 *
## pan1994v        0.27356    0.07257   3.770 0.000188 ***
## prd1994v        0.11323    0.05790   1.955 0.051209 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.795 on 408 degrees of freedom
## Multiple R-squared:  0.1809, Adjusted R-squared:  0.1669
## F-statistic: 12.87 on 7 and 408 DF,  p-value: 5.603e-15
```

##2. Second regression using pri2000v as the outcome variable.




Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(pri2000v ~ treatment + log(pobtot1994) + avgpoverty + pri1994v + pan1994 ...
53
2004
-3
-2
-1
0
1
2
3


Scale-Location
24
290
53
√|Standardized residuals|
0.0
0.5
1.0
1.5
20
30
40
50
60
70
80
Fitted values
lm(pri2000v ~ treatment + log(pobtot1994) + avgpoverty + pri1994v + pan1994 ...


Residuals vs Leverage
Standardized residuals
3
2
1
0
-1
-2
-3
3
74
9
0.5
0.5
Cook's distance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Leverage
lm(pri2000v ~ treatment + log(pobtot1994) + avgpoverty + pri1994v + pan1994 ...

```
##
## Call:
## lm(formula = pri2000v ~ treatment + log(pobtot1994) + avgpoverty +
##     pri1994v + pan1994v + prd1994 + t1994r, data = progresa1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.408  -7.648   0.599   7.950  33.020
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     34.69950   11.71229   2.963  0.00323 **
## treatment        0.87742    1.32691   0.661  0.50883
## log(pobtot1994) -1.45615    0.83794  -1.738  0.08301 .
## avgpoverty       2.89820    1.45196   1.996  0.04659 *
## pri1994v         0.44265    0.05440   8.137 4.96e-15 ***
## pan1994v        -0.40133    0.08897  -4.511 8.45e-06 ***
## prd1994         -0.04107    0.01332  -3.083  0.00219 **
## t1994r          -0.07449    0.05085  -1.465  0.14370
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.66 on 408 degrees of freedom
## Multiple R-squared:  0.4873, Adjusted R-squared:  0.4785
## F-statistic: 55.39 on 7 and 408 DF,  p-value: < 2.2e-16
```

```
## Rows: 416
## Columns: 21
## $ treatment  <int> 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 0,…
## $ pri2000s   <dbl> 40.82397, 22.44186, 38.93130, 31.16883, 76.92308, 23.85321,…
## $ pri2000v   <dbl> 73.15436, 73.52381, 87.03072, 71.28713, 78.94737, 65.40881,…
## $ t2000      <dbl> 55.80524, 31.22093, 47.02290, 45.02164, 100.00000, 37.38532…
## $ t2000r     <dbl> 60.56911, 50.28736, 42.52540, 49.75369, 48.40764, 34.79212,…
## $ pri1994    <int> 102, 245, 52, 58, 30, 128, 47, 30, 274, 81, 160, 208, 137, …
## $ pan1994    <int> 3, 18, 8, 6, 0, 0, 8, 1, 0, 4, 4, 9, 0, 0, 1, 1, 6, 0, 1, 3…
## $ prd1994    <int> 23, 14, 4, 7, 38, 20, 9, 99, 30, 14, 352, 202, 26, 2, 9, 2,…
## $ pri1994s   <dbl> 45.759889, 15.115377, 7.988570, 29.162154, 20.170385, 32.81…
## $ pri1994v   <dbl> 77.27273, 65.68365, 50.98039, 73.41772, 43.47826, 51.82186,…
## $ pan1994s   <dbl> 1.3458791, 1.1105175, 1.2290108, 3.0167747, 0.0000000, 0.00…
## $ pan1994v   <dbl> 2.2727273, 4.8257373, 7.8431373, 7.5949367, 0.0000000, 0.00…
## $ prd1994s   <dbl> 10.3184067, 0.8637358, 0.6145054, 3.5195705, 25.5491525, 5.…
## $ prd1994v   <dbl> 17.424242, 3.753351, 3.921569, 8.860759, 55.072464, 8.09716…
## $ t1994      <dbl> 60.56456, 23.81443, 21.66131, 40.22366, 46.39188, 63.33183,…
## $ t1994r     <dbl> 67.00508, 55.83832, 35.54007, 56.83453, 48.59155, 68.61111,…
## $ votos1994  <int> 135, 386, 141, 80, 69, 247, 109, 186, 314, 103, 587, 425, 1…
## $ avgpoverty <dbl> 5.000000, 5.000000, 4.500000, 5.000000, 5.000000, 5.000000,…
## $ pobtot1994 <int> 541, 3289, 1320, 384, 307, 840, 790, 632, 356, 435, 1423, 2…
## $ villages   <int> 4, 2, 2, 1, 1, 6, 1, 2, 2, 2, 9, 8, 8, 1, 1, 2, 1, 2, 3, 2,…
## $ group      <chr> "treatment", "treatment", "treatment", "treatment", "contro…
```

interpretations: first regression shows the effect of the program on the election turnovers. According to this model treatment has a negative effect on t2000r outcome variable. This shows that, holding all other confounders constant the program had small + also statistically insignificant effect on election outcomes.

###contrary to the first regression the second regression shows that the treatment effect had positive effect (0.8) on outcome variable. Meaning that, ceteris paribus change in treatment variable constituted to 0.8 change in PRI2000v variable. But this effect is also statistically insignificant. However, other variables suggest that poverty, total population, and previous party affiliation had significant effects on election turnout.

#Exercise 3 ##Estimating two 95% prediction intervals for a dependent variable for a precinct 11

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

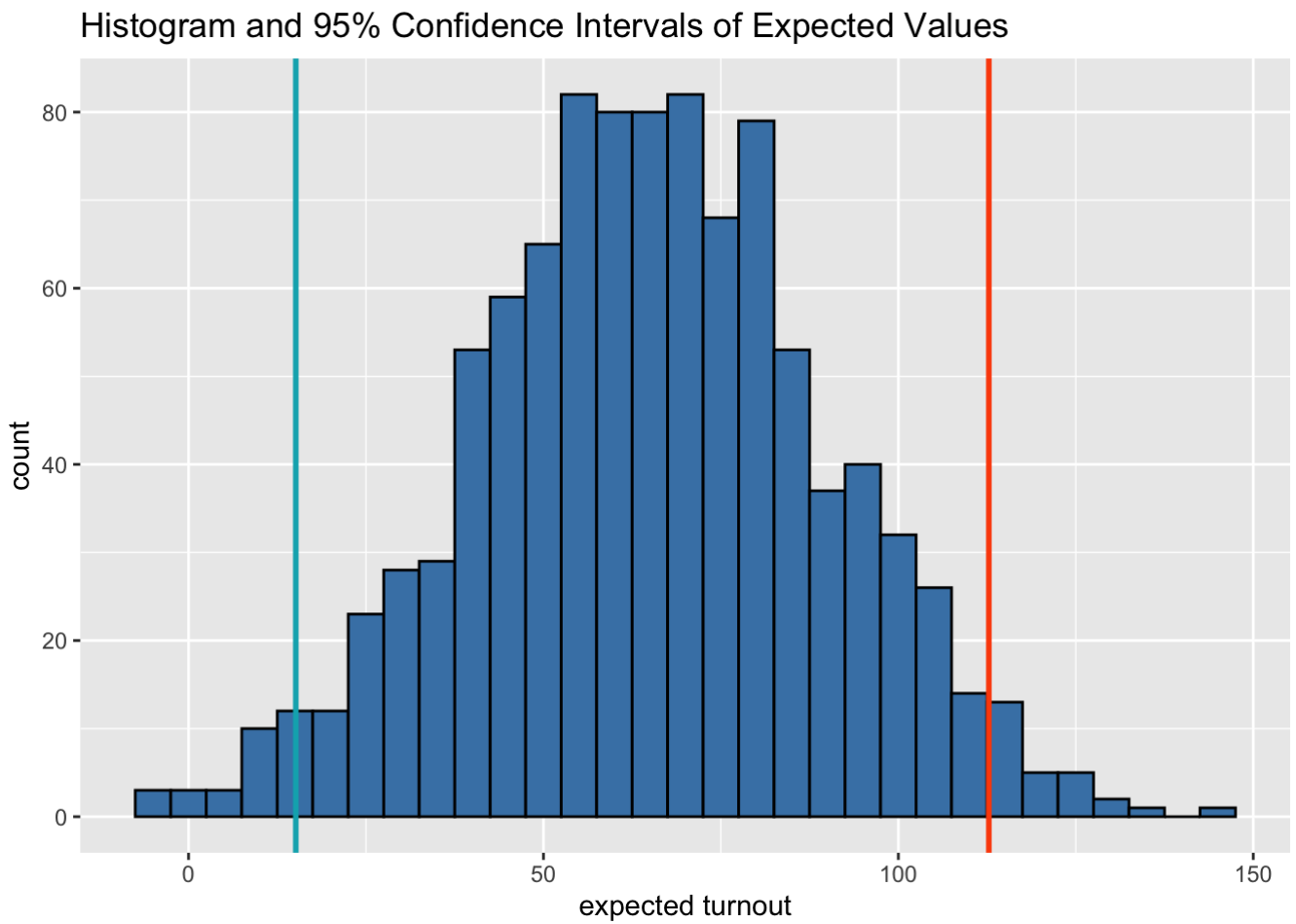
```
## 7.260523
```

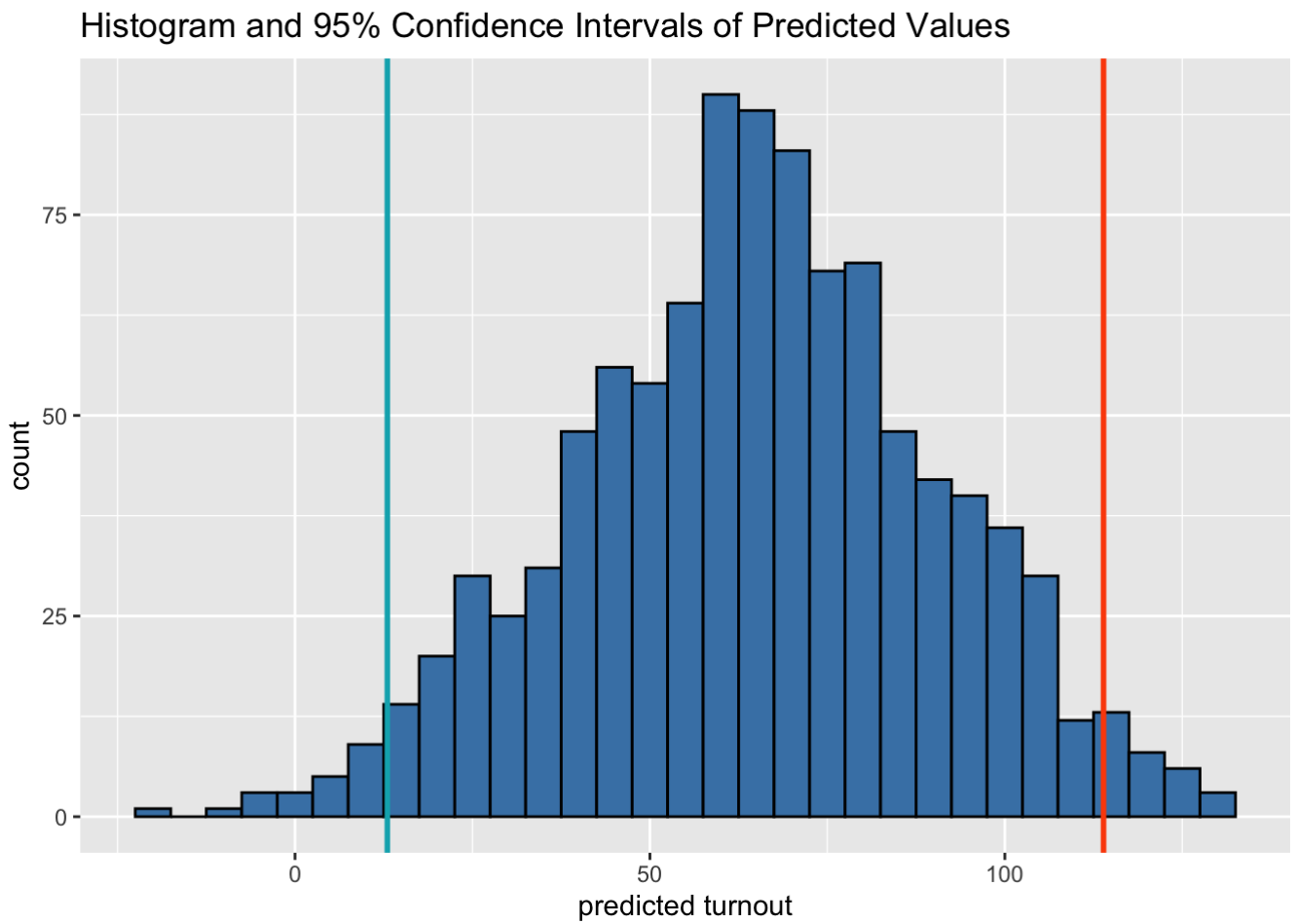
```
## 95% CI of expected values : 15.11148 112.7539
```

```
## 95% CI of predicted values : 13.00744 113.9055
```

###visualization:

```
## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
## ℹ Please use `linewidth` instead.
```


Histogram and 95% Confidence Intervals of Expected Values
count
80
60
40
20
0
0
50
100
150
expected turnout



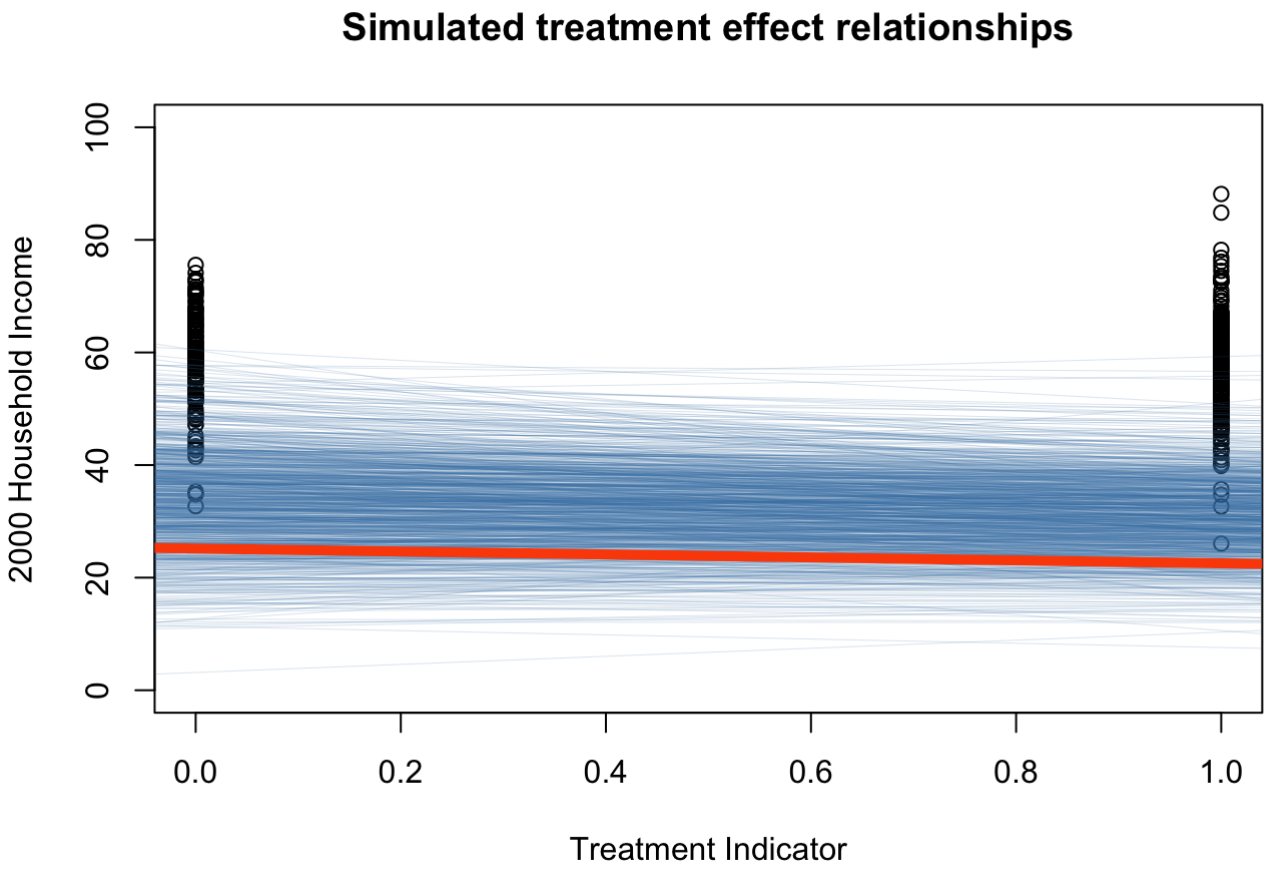
###Simulation shows that the anticipated treatment effect in the control sample of Precinct 11 ranges from 15.11148 to 112.7539, while the predicted treatment effect's 95% confidence interval for the same sample ranges from 13.00744 to 113.9055. The difference in these intervals can be explained by the presence of stochastic term in the predicted simulated values which show greater variance. ###interpretation: The second interval with a 95% confidence level, represents the predicted values of the election turnout, which includes a stochastic term/sigma. The interval spans from 13.00744 to 113.9055, indicating that we can be 95% confident that the true mean of predicted turnout values in the precinct 11 fall within this range. Unlike the first interval, the second interval accounts for the inherent uncertainty in making predictions, which can be influenced by sampling error, model assumptions, and confounders. Thus, the second interval is wider than the first.

#Exercise 4 ##Regression including an interaction term

```
##
## Call:
## lm(formula = t2000r ~ treatment + avgpoverty + log(pobtot1994) +
##     t1994r + pri1994v + pan1994v + prd1994v + interact + I(treatment *
##     interact), data = progresa1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.784  -4.545  -0.003   4.975  29.117
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             34.02324    8.86447   3.838 0.000144 ***
## treatment               -1.40499    0.84984  -1.653 0.099057 .
## avgpoverty              -0.35029    0.88926  -0.394 0.693852
## log(pobtot1994)         -0.29122    0.47581  -0.612 0.540848
## t1994r                   0.22117    0.03103   7.127 4.71e-12 ***
## pri1994v                 0.12767    0.05493   2.324 0.020609 *
## pan1994v                 0.27440    0.07258   3.781 0.000180 ***
## prd1994v                 0.10768    0.06218   1.732 0.084097 .
## interact                -1.96657    2.84329  -0.692 0.489551
## I(treatment * interact)  4.34713    3.10870   1.398 0.162764
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.794 on 406 degrees of freedom
## Multiple R-squared:  0.1851, Adjusted R-squared:  0.167
## F-statistic: 10.25 on 9 and 406 DF,  p-value: 2.831e-14
```

```
##
## Call:
## lm(formula = t2000r ~ treatment + avgpoverty + log(pobtot1994) +
##     t1994r + pri1994v + pan1994v + prd1994v, data = progresa1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.041  -4.555   0.029   5.010  29.069
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     33.47237    8.85553   3.780 0.000180 ***
## treatment       -1.08094    0.81683  -1.323 0.186465
## avgpoverty      -0.27734    0.88768  -0.312 0.754874
## log(pobtot1994) -0.27878    0.47487  -0.587 0.557487
## t1994r           0.22238    0.03102   7.168 3.59e-12 ***
## pri1994v         0.12473    0.05489   2.272 0.023586 *
## pan1994v         0.27356    0.07257   3.770 0.000188 ***
## prd1994v         0.11323    0.05790   1.955 0.051209 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.795 on 408 degrees of freedom
## Multiple R-squared:  0.1809, Adjusted R-squared:  0.1669
## F-statistic: 12.87 on 7 and 408 DF,  p-value: 5.603e-15
```

###The interaction term allows us to see how the treatment effect changes given the value of the prd1994v variable. To achieve this I added the dummy variable that equals 1 when prd1994v was greater than 50. The coefficient of the variable “interact” represents the difference between the cases in which interact=1 and interact=0. ###The algebreac representation of this variable is treatment effect = -1.4 + 4.3*interact* + *error. Meaning that when prd1994v>50 the* interact* variable is responsible for 4.3 times increase in treatment effect holding all other values constant. -1.4 is the treatment coefficient. ###Regarding the statistical significance of the treatment effect given that it equals 0.09 we can conclude that it is slightly insignificant. I believe that the statistical significance of treatment effect is conditional on interaction term, as inclusion of the interaction term lowers the initial significance level of treatment effect from 0.09 to 0.18 (lowers because higher the value less significant) which is even less significant in the areas which had shown high support for PAN party and the areas where average poverty levels were less than the mean.



#Exercise 5 ##part 1

```
##  [1] "treatment"  "pri2000s"   "pri2000v"   "t2000"      "t2000r"
##  [6] "pri1994"    "pan1994"    "prd1994"    "pri1994s"   "pri1994v"
## [11] "pan1994s"   "pan1994v"   "prd1994s"   "prd1994v"   "t1994"
## [16] "t1994r"     "votos1994"  "avgpoverty" "pobtot1994" "villages"
## [21] "group"      "interact"
```

```
##
## Call:  glm(formula = t2000r ~ treatment + log(pobtot1994) + avgpoverty +
##     I(treatment * avgpoverty), data = progresa1)
##
## Coefficients:
##               (Intercept)                  treatment
##                   84.3162                   -10.0886
##           log(pobtot1994)                 avgpoverty
##                   -0.8981                    -4.3319
## I(treatment * avgpoverty)
##                    1.9601
##
## Degrees of Freedom: 415 Total (i.e. Null);  411 Residual
## Null Deviance:       30270
## Residual Deviance: 29220     AIC: 2961
```

```
## [1] 71.78602 71.78416
```

##part 2

```
##
## Call:  glm(formula = t2000r ~ treatment + log(pobtot1994) + avgpoverty +
##     I(avgpoverty^2) + I(treatment * avgpoverty^2), data = progresa1)
##
## Coefficients:
##                 (Intercept)                    treatment
##                     62.0930                     -5.7777
##             log(pobtot1994)                   avgpoverty
##                     -0.8313                      5.8450
##             I(avgpoverty^2)  I(treatment * avgpoverty^2)
##                     -1.1724                      0.2209
##
## Degrees of Freedom: 415 Total (i.e. Null);  410 Residual
## Null Deviance:       30270
## Residual Deviance: 29190     AIC: 2963
```

```
## [1] 71.97892 71.97671
```

###The cross-validation errors for the two models are similar, coming in at about 71.8 and 71.9, respectively, showing that their predictive abilities are similar. The difference maker in these two model is that Avgpoverty^2, and an interaction between treatment and avgpoverty^2 are present in Model 2. This implies that there may not be a linear connection between poverty and outcome and that the treatment's effectiveness may change with the degree of poverty. To predict the outcome t2000r and comprehend the relationship between poverty, treatment, and the outcome variable in this situation, Model 2 might be a better option. However, one would need to examine the covariance between existing variables furtherly for the conclusion to be robust.