Election and Conditional Cash Transfer Program in Mexico

October 15, 2020

In this exercise, we analyze the data from a study that seeks to estimate the electoral impact of 'Progresa', Mexico's *conditional cash transfer program* (CCT program). This program has been a model for similar programs implemented in many countries around the world where the government provides cash to low income families conditionally on their taking some required actions. For the Progresa program the required actions involved attending workshops regarding health behaviors and having children, particularly girls, attend school. The impacts of the program on socioeconomic status and intergenerational transfer of poverty are strong. Here the interest is in other possible side-effects of the program on voting behavior.

This exercise is based on the following articles:

- Ana de la O. (2013). 'Do Conditional Cash Transfers Affect Voting Behavior? Evidence from a Randomized Experiment in Mexico.' *American Journal of Political Science*, 57:1, pp.1-14; and
- Kosuke Imai, Gary King, and Carlos Velasco. (2015). 'Do Nonpartisan Programmatic Policies Have Partisan Electoral Effects? Evidence from Two Large Scale Randomized Experiments.' Working Paper.

The original study relied on a randomized evaluation of the CCT program in which eligible villages were randomly assigned to receive the program either 21 months (Early *Progresa*) or 6 months (Late *Progresa*) before the 2000 Mexican presidential election. The author of the original study hypothesized that the CCT program would mobilize voters, leading to an increase in turnout and and more support for the incumbent party (PRI in this case). The analysis was based on a sample of precincts that contain at most one participating village in the evaluation.

The data we analyze are available as the CSV file `progresa.csv`. The names and descriptions of variables in the data set are:

Name	Description
`treatment`	Whether an electoral precinct contains a village where households received Early *Progresa*
`pri2000s`	PRI votes in the 2000 election as a share of precinct population above 18 (in percentage points)
`t2000`	Turnout in the 2000 election as a share of precinct population above 18 (in percentage points)
`t1994`	Turnout in the 1994 election as a share of precinct population above 18 (in percentage points)
`avgpoverty`	Precinct Avg of Village Poverty Index
`pobtot1994`	Total Population in the precinct
`villages`	Number of villages in the precinct

Each observation in the data represents a precinct, and for each precinct the file contains information about its treatment status, the outcomes of interest, socioeconomic indicators, and other precinct characteristics.

```r
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.3     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------ tidyverse_conflicts() --
```

```
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```r
progresa <- read.csv("data/progresa.csv")
```

Question 1 [5 pts]

Consider the impact of early versus late receipt of the CCT program on voter turnout in the 2000 election.

1a

What is the specific causal question? What are the potential outcomes of each precinct?

##Answer for 1a What is the impact of implementing progresa (CTP) Cash Transfer program relative to not implementing the program on the election turnout rate of randomized precincts within Mexico?

Potential Outcomes:

1. What the election turnout rate would be if a precinct does implement the progress transfer program?
2. What the election turnout rates would be if the precinct does not apply the progressa cash transfer program?

1b

For precincts receiving the CCT program early, what is their average factual outcome and average missing counterfactual outcome?

##answer for 1b

Average Factual OutCome:

The group level factual outcome is the average election turn out rate across all groups (Precincts) implementing the progresa cash transfer program early.

Average Missing CounterFactual:

What the average election turnout rate would have been for the groups(presinct) that implemented the Cash Transfer Program early if they instead implemented the program late but all else remained the same.

1c

How will the average missing counterfactual outcome for the treated precincts be estimated in this study?

The avg missing counterfactual outcome for the treated group will be estimated by measuring the level of the outcome of the groups that did not implement the CCT program.

1d

What do the researchers hypothesize the treatment effect for this outcome will be?

The authors of the study hypothesised that the CCT program would bring out more voters, which in turn would then get the people to show up for the voye and support the PRI party.

Answer 1

Question 2 [7 pts]

2a

Estimate the impact of early versus late receipt of the CCT program on two outcomes: voter turnout in 2000 and support for the incumbent party in 2000. Do so by comparing the average electoral outcomes in the 'treated' (Early *Progresa*) precincts versus the ones observed in 'control' (Late *Progresa*) precincts. Use the turnout and support rates as shares of the voting eligible population (`t2000` and `pri2000s`, respectively). Interpret your results.

##Answer for 2a

We will be comparing the voter turn out rates for the precincts that have implemented the CCT program, the mean pri2000 was 36.11 as compared to the mean t2000 which was 64.33 (Treatment = 1. We can see the growth with the the treated groups. When we compare the mean for pre2000 vs the mean for t2000 with out the treatment (Treatment = 0) group we can see that this group had a growth.

The difference in the treatment group is 28. This group had a larger average change in the support after the cash program was implemented. The difference in the control group is 22.

2b

Consider two pretreatment covariates, poverty level and voter turnout in the 1994 election. Are these pretreatment covariates balanced between the treatment and control groups? Use appropriate summary statistics and figures to explain your answer. Discuss the implications of the distributions of these two baseline covariates for the internal validity of the results you estimated in the first part of this question.

##Answer for 2b

The avg poverty rate for the treatment group and the control group is 4.6, these covarients have approximately the same mena so we can conclude that the internal validity is strong because we can observe that the poverty level is between the treatment and control group.

When comparing the the voter turnout rate between the treatment and the control groups, we can conclude that they are relatively similar to each other. From observing the two pre-treatment covarients we can conclude that the internal validity will hold and the test is randomized, from this the missing counterfactual can be estimated.

Answer 2

```r
progresa %>%
    group_by(treatment) %>%
    summarize(pri2000s = mean(pri2000s),
              t2000   = mean(t2000))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 3
##   treatment pri2000s t2000
##       <int>    <dbl> <dbl>
## 1         0     34.2  56.4
## 2         1     36.1  64.3
```

```r
progresa %>%
    group_by(treatment) %>%
    summarize(avgpoverty = mean(avgpoverty),
              mean1994   = mean(pobtot1994))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 3
##   treatment avgpoverty mean1994
##       <int>      <dbl>    <dbl>
## 1         0       4.59    1919.
## 2         1       4.57    2152.
```

Question 3 [7 pts]

Other pre-treatment variables are associated with voter turnout. Considering only those precincts that received the CCT program later (controls), investigate the linear relationship between voter turnout in the 2000 election (outcome) and the average poverty level in a precinct.

3a

Use a scatterplot, linear correlation, and linear regression to investigate this relationship. Make a scatterplot and add the estiamted linear regression line to this figure. Make a residual plot and add a horizontal zero line to this figure. What do the scatterplot and linear correlation tell us about this bivariate relationship? Is the linearity assumption for linear regression violated or does it appear to hold?

Answer for 3a

We can observe the relationship between the average poverty level vs the voters in the 2000 elections. No association is observed from this comparison. The clumps we see are between 60.5 and 62, as well as 62 and 62.5. From this we can conclude that our original linear assumption does not hold and is violated because of the clusters that we observed.

3b

Interpret the coefficients of the simple linear regression. Interpret the RMSE and R^2 value from the model.

Answer for 3b

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.815 7.744 6.949 1.48e-11 *** progresa$avgpoverty 1.720 1.682 1.023 0.307

The intercept can tell us that when the voter turnout is 0, the avg poverty rate would be 53.815. This is very unlikely, so we cant just use this to predict the voter turnout rate.

Residual standard error: 16.47 on 405 degrees of freedom

The RMSE here is showing us that the voter turn out was on avg 16.47 points away from what was expected based on its linear relationship with the avg poverty rate.

Our model for R^2 is proportional to the variance in t2000 and is observed through the linear relationship with avg poverty rate, which gives information towards how well the regression prediction approximates the real data points.

The validity of our study is observed by the avg poverty rate and the multiple r^2 is 0.003, meaning that .3% of the varience in the t20000 variable can be shown by the variation of the avg poverty rate variable. This percentage is low, we would need a higher value to make a statistical claim.

Answer 3

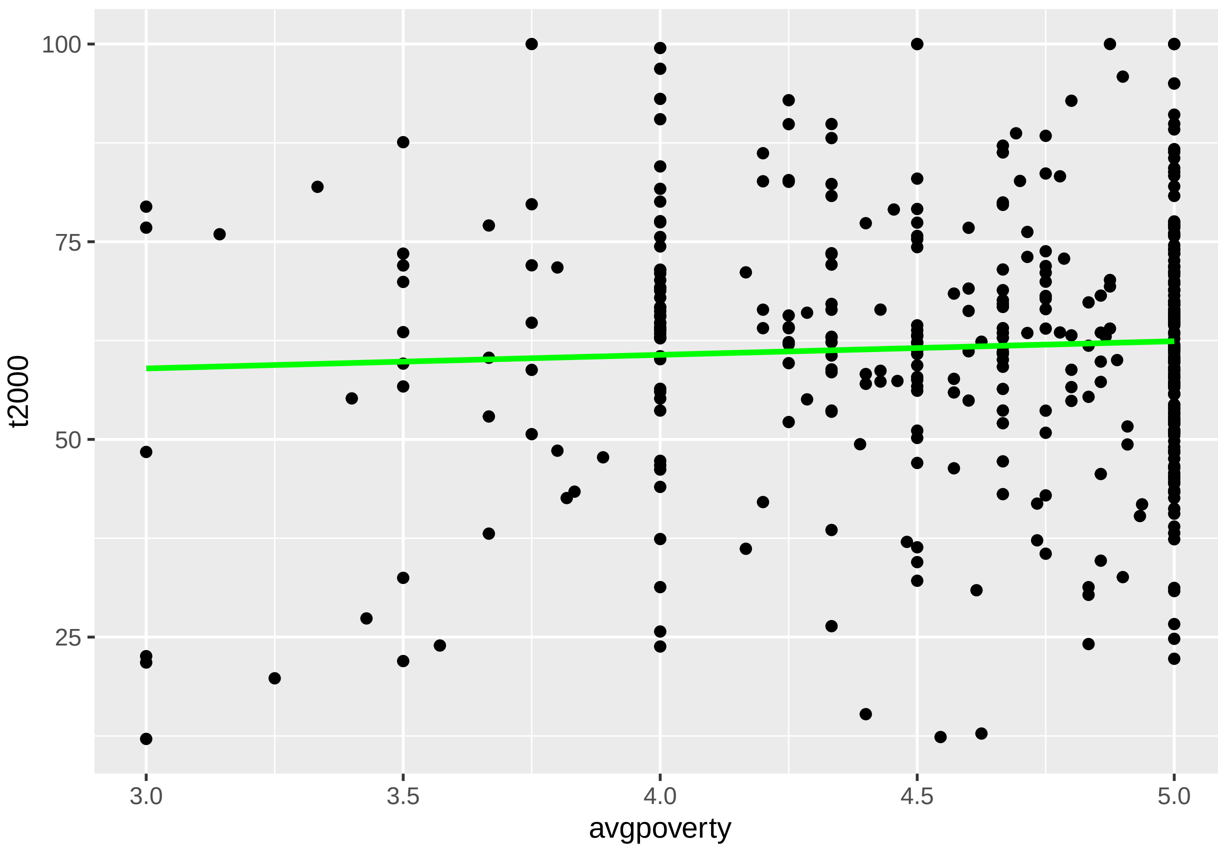
```
progresa %>%
  summary(lm.t2000)
```

```
##    treatment         pri2000s          t1994           avgpoverty
##  Min.   :0.0000   Min.   : 0.741   Min.   :  1.001   Min.   :3.000
##  1st Qu.:0.0000   1st Qu.:25.362   1st Qu.: 50.733   1st Qu.:4.286
##  Median :1.0000   Median :35.227   Median : 62.354   Median :4.750
##  Mean   :0.6658   Mean   :35.470   Mean   : 61.088   Mean   :4.578
##  3rd Qu.:1.0000   3rd Qu.:44.660   3rd Qu.: 72.342   3rd Qu.:5.000
##  Max.   :1.0000   Max.   :87.500   Max.   :100.000   Max.   :5.000
##    pobtot1994        villages          t2000
```

```
##  Min.   :   103   Min.   : 1.000   Min.   : 12.13
##  1st Qu.:   633   1st Qu.: 3.000   1st Qu.: 53.26
##  Median :  1164   Median : 5.000   Median : 62.78
##  Mean   :  2074   Mean   : 5.988   Mean   : 61.69
##  3rd Qu.:  1716   3rd Qu.: 8.000   3rd Qu.: 71.18
##  Max.   :102322   Max.   :14.000   Max.   :100.00
```

```
ggplot(data = progresa, aes(y = t2000, x = avgpoverty )) + geom_point() + geom_smooth(method = lm, se =
```

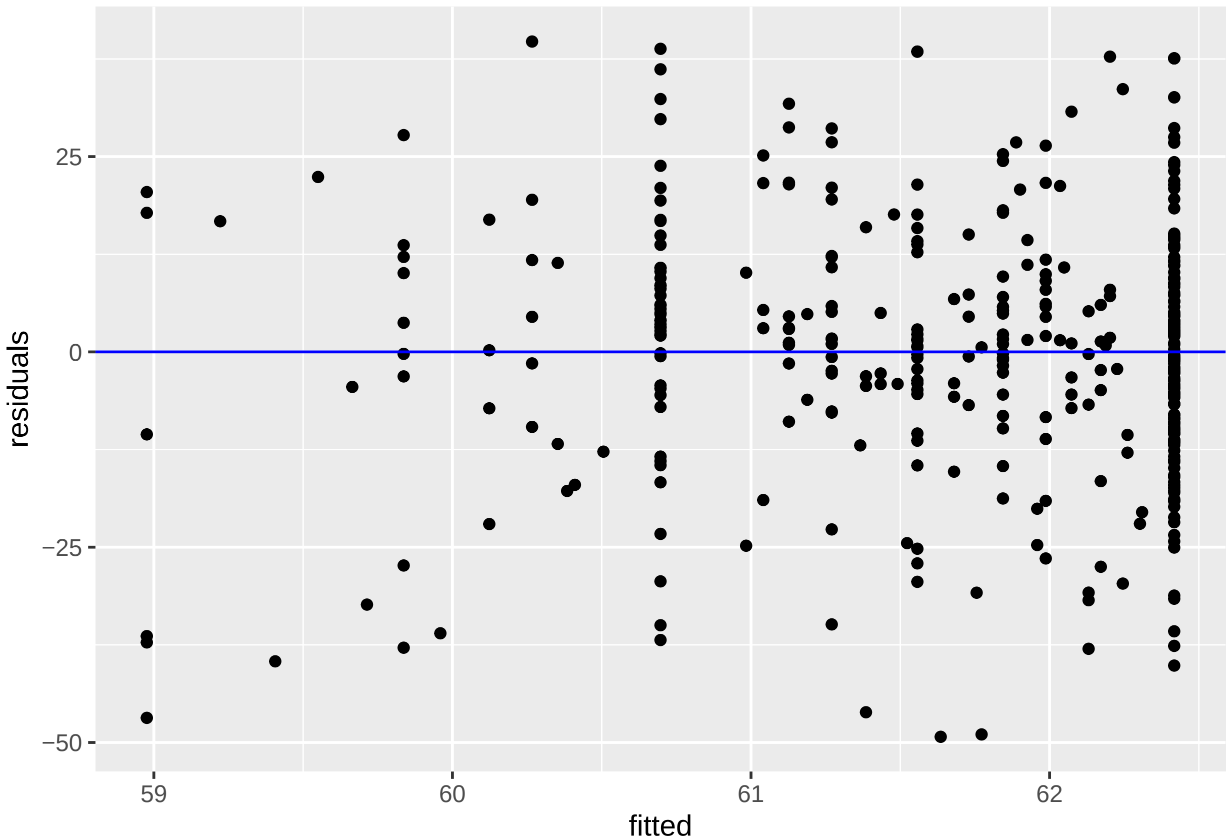
```
## `geom_smooth()` using formula 'y ~ x'
```




```
lm.t2000 <- lm(progresa$t2000 ~ progresa$avgpoverty, data = progresa )

res.dat <- tibble(residuals = lm.t2000$residuals,
                  fitted    = lm.t2000$fitted.values)

ggplot(res.dat, aes(y = residuals, x = fitted)) +
  geom_point() +
  geom_hline(yintercept = 0, col = "blue")
```




```r
summary(lm.t2000)
```

```
##
## Call:
## lm(formula = progresa$t2000 ~ progresa$avgpoverty, data = progresa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.274  -8.967   1.009   9.460  39.733
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)           53.815      7.744   6.949 1.48e-11 ***
## progresa$avgpoverty    1.720      1.682   1.023    0.307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.47 on 405 degrees of freedom
## Multiple R-squared:  0.002576,   Adjusted R-squared:  0.0001134
## F-statistic: 1.046 on 1 and 405 DF,  p-value: 0.307
```

Question 4 [7 pts]

Now lets consider a different pre-treatment variable that may be associated with voter turnout. Considering all precincts in the study, investigate the linear relationship between a precinct's voterturnout in the 2000 election (outcome) and it's voter turnout in the 1994 election.

4a

Use a scatterplot, linear correlation, and linear regression to investigate this relationship. Make a scatterplot and add the estiamted linear regression line to this figure. Make a residual plot and add a horizontal zero line to this figure. What do the scatterplot and linear correlation tell us about this bivariate relationship? Is the linearity assumption for linear regression violated or does it appear to hold?

##Answer for 4a

When we observe the relationship between the voter turnout rate in 1994 vs 2000, we see a positive linear association. In the residual scatterplot we do not see any trends so we can conclude that the linear assumption holds true.

4b

Interpret the coefficients of the simple linear regression. Interpret the RMSE and R^2 value from the model and compare them with the RMSE and R^2 values from the model in Question 3.

##Answer for 4b

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 18.39949 1.56737 11.74 <2e-16 ***t1994 0.70867 0.02449 28.93 <2e-16***

The intercept can tell us that when the voter turnout is 0, the voter turn out rate in t1994 would be 18.39949. This is very unlikely, so we cant just use this to predict the voter turnout rate.

Residual standard error: 9.419 on 405 degrees of freedom

The RMSE here is showing us that the voter turn out was on avg 9.419 points away from what was expected based on its linear relationship with t1994.

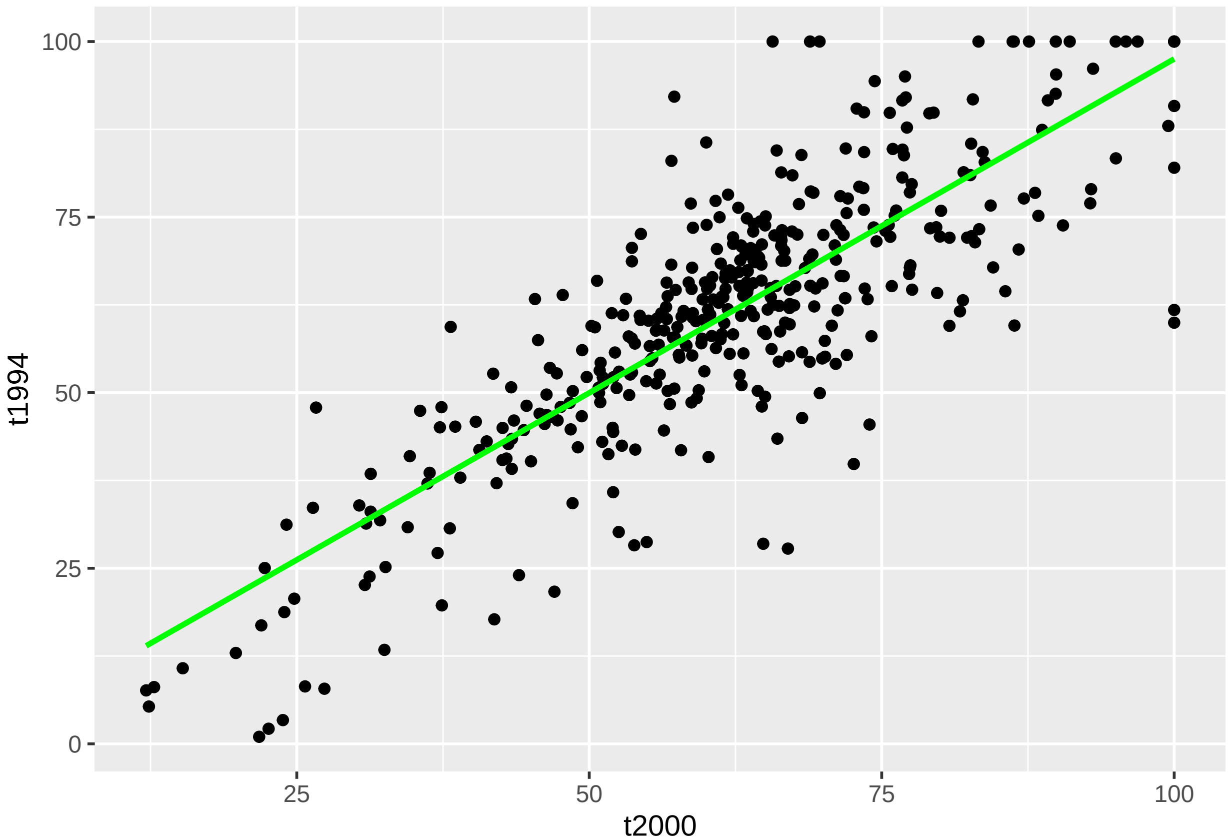
Our model for R^2 is proportional to the variance in t2000 and is observed through the linear relationship with avg poverty rate, which gives information towards how well the regression prediction approximates the real data points.

The validity of our study is observed by 2000 voter turnout rate and the multiple r^2 is 0.674, meaning that 67.4% of the variance in the t20000 variable can be shown by the variation of the t1994 voter turnout rate variable. This percentage is high, we can make a stat claim about the association between the two variables.

Answer 4

```
ggplot(data = progresa, aes(y = t1994, x = t2000 )) + geom_point() + geom_smooth(method = lm, se = FALSE
```

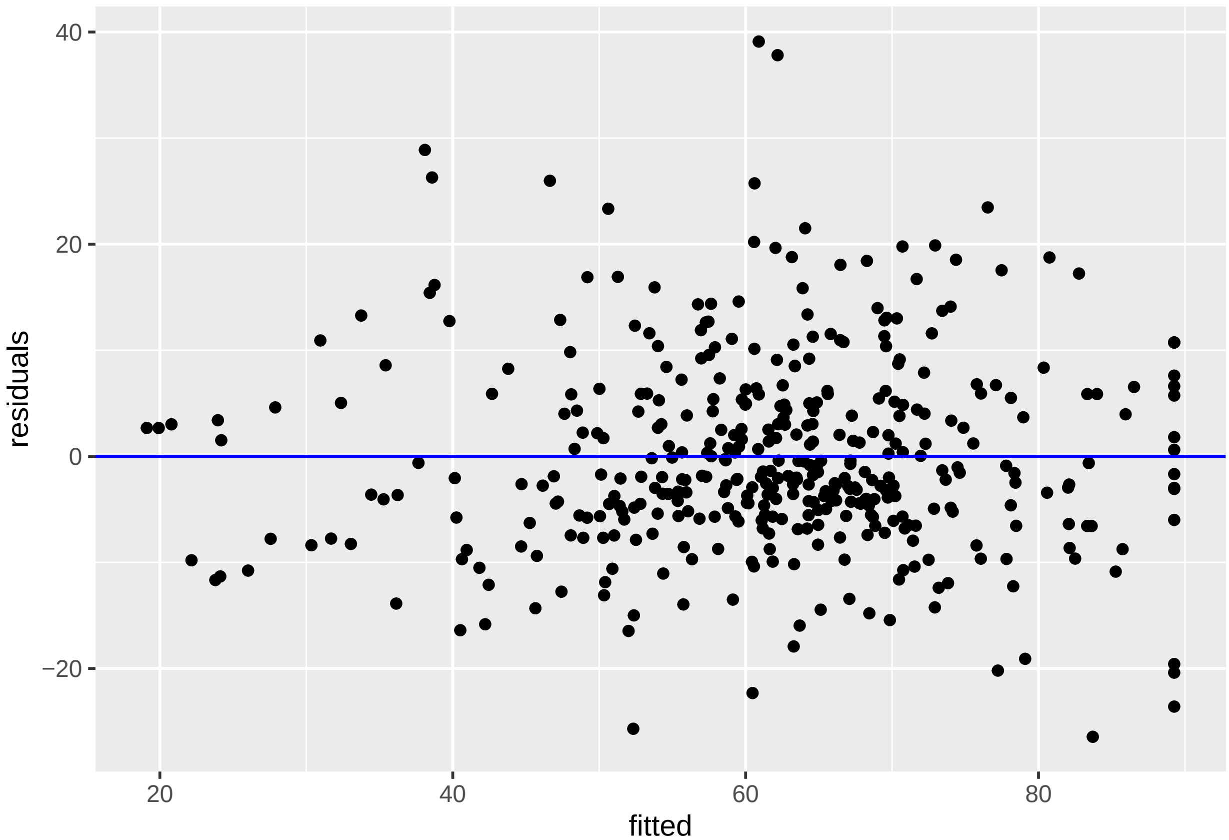
```
## `geom_smooth()` using formula 'y ~ x'
```




```
lm.t1994 <- lm(t2000 ~ t1994, data = progresa )

res.dat <- tibble(residuals = lm.t1994$residuals,
                  fitted    = lm.t1994$fitted.values)

ggplot(res.dat, aes(y = residuals, x = fitted)) +
  geom_point() +
  geom_hline(yintercept = 0, col = "blue")
```




```
summary(lm.t1994)
```

```
##
## Call:
## lm(formula = t2000 ~ t1994, data = progresa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.439  -5.683  -1.848   5.318  39.104
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.39949    1.56737   11.74   <2e-16 ***
## t1994        0.70867    0.02449   28.93   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.419 on 405 degrees of freedom
## Multiple R-squared:  0.674,  Adjusted R-squared:  0.6732
## F-statistic: 837.2 on 1 and 405 DF,  p-value: < 2.2e-16
```

Question 5 [6 pts]

5a

Estimate the impact of early versus late receipt of the CCT program on voter turnout using multiple linear regression. Include two predictors in your model: *treatment*, and turnout in the 1994 election. Create a residual plot and use it to assess the linearity assumption.

##Answer for 5a

When we see the relationship between the two groups we can observe a linear association in both the treatment and control groups. We can see a cluster between 50 and 70 in the residual plot so we can conclude that we can not verify the the linear assumption between the control and treatment.

5b

Write out the multiple regression equation for this model as a single equation and then as a pair of equations, one for each treatment arm. Create a scatterplot of the outcome and the continuous predictor variable. Color the points on this scatterplot by their treatment status. Add the two regression lines to this figure. Or sketch the figure described here by hand, take a picture and include it in your HW document.

##Answer for 5b

Y_i = alpha_hat + beta_hat* X_i _ epsilon_i_hat Y_i = 8.44 + treatment * 15.83 + epsilon_i_hat Y_i= 8.44 + t1994* .80567 + epsilon_i_hat Y_i = 8.44 + treatment:t1994 *(-0.15817)+ epsilon_i_hat Y_i=25.32 + treatment* 15.83+ t1994* .80567 + treatment:t1994 *(-0.15817))+ epsilon_i_hat

5c

Interpret all three of the model coefficients for this multiple regression equation. Interpret the RMSE and the R^2 value for the model. Compare them to the RMSE and R^2 for the model in Question 4. What does this model tell you about whether the timing of the CCT program had the hypothesized effect?

##Answer for 5c

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.44778 2.50423 3.373 0.000814 ***treatment 15.83088 3.09763 5.111 4.97e-07*** t1994 0.80567 0.04007 20.105 < 2e-16 * **treatment:t1994 -0.15817 0.04898 -3.229 0.001343**

The intercept can tell us that when the voter turnout is 0, the voter turn out rate in t1994 would be 18.39949. This is very unlikely, so we cant just use this to predict the voter turnout rate.

Residual standard error: 8.847 on 403 degrees of freedom

The RMSE here is showing us that the voter turn out was on avg 9.419 points away from what was expected based on its linear relationship with t1994.

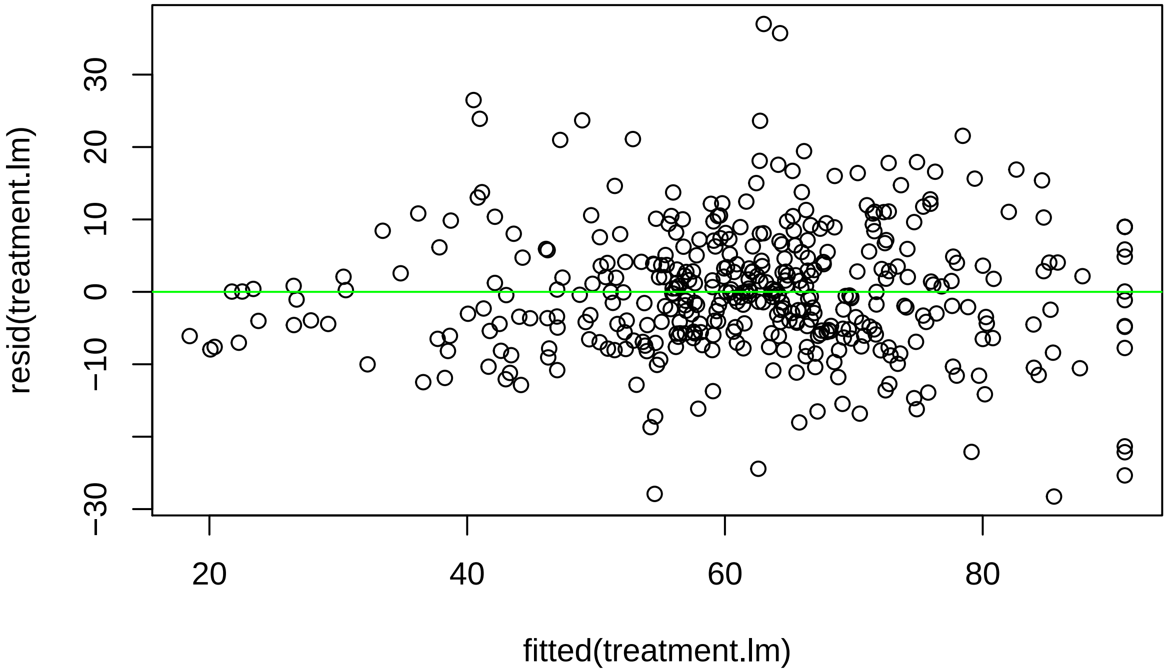
Our model for R^2 is proportional to the variance in t2000 and is observed through the linear relationship with avg poverty rate, which gives information towards how well the regression prediction approximates the real data points.

The validity of our study is observed by treatment and t1994 and treatment:1994voter turnout rate and the multiple r^2 is 0.7138, meaning that 71.38% of the variance in the variables can be shown by the variation of the 3 variables in the voter turnout rate variable. This percentage is high, we can make a stat claim about the association between the three variables. This is similar to our model in question 4.

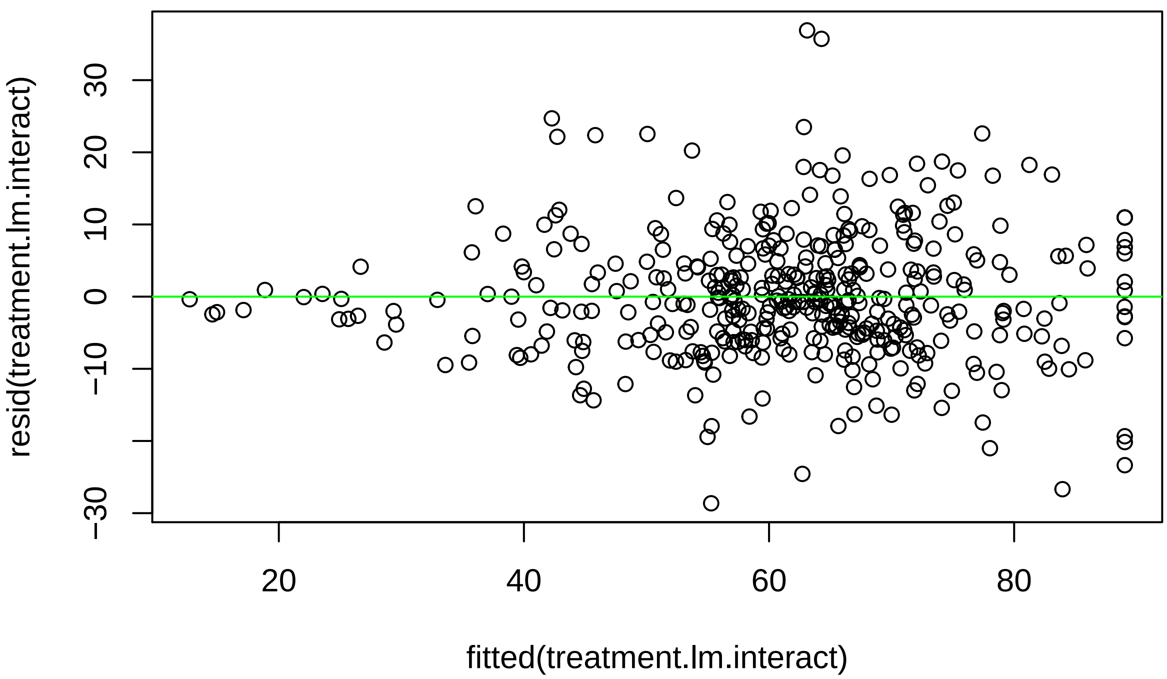
Answer 5

```
treatment.lm <- lm(t2000 ~ treatment + t1994, data = progresa)
treatment.lm.interact <- lm(t2000 ~  treatment + t1994 + treatment*t1994, data = progresa)
```

```r
plot(fitted(treatment.lm), resid(treatment.lm))
abline(h = 0, col = "green")
```




```r
plot(fitted(treatment.lm.interact), resid(treatment.lm.interact))
abline(h = 0, col = "green")
```




```r
summary(treatment.lm)
```

```
##
## Call:
## lm(formula = t2000 ~ treatment + t1994, data = progresa)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.265  -5.577  -0.374   4.792  36.992
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.75296    1.58622   9.301  < 2e-16 ***
## treatment    6.29070    0.94203   6.678 8.04e-11 ***
## t1994        0.69980    0.02331  30.021  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.95 on 404 degrees of freedom
## Multiple R-squared:  0.7064, Adjusted R-squared:  0.7049
## F-statistic: 485.9 on 2 and 404 DF,  p-value: < 2.2e-16
```

```r
summary(treatment.lm.interact)
```

```
##
## Call:
## lm(formula = t2000 ~ treatment + t1994 + treatment * t1994, data = progresa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.632  -5.396  -0.741   4.704  36.893
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.44778    2.50423   3.373 0.000814 ***
## treatment       15.83088    3.09763   5.111 4.97e-07 ***
## t1994            0.80567    0.04007  20.105  < 2e-16 ***
## treatment:t1994 -0.15817    0.04898  -3.229 0.001343 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.847 on 403 degrees of freedom
## Multiple R-squared:  0.7138, Adjusted R-squared:  0.7116
## F-statistic:   335 on 3 and 403 DF,  p-value: < 2.2e-16
```

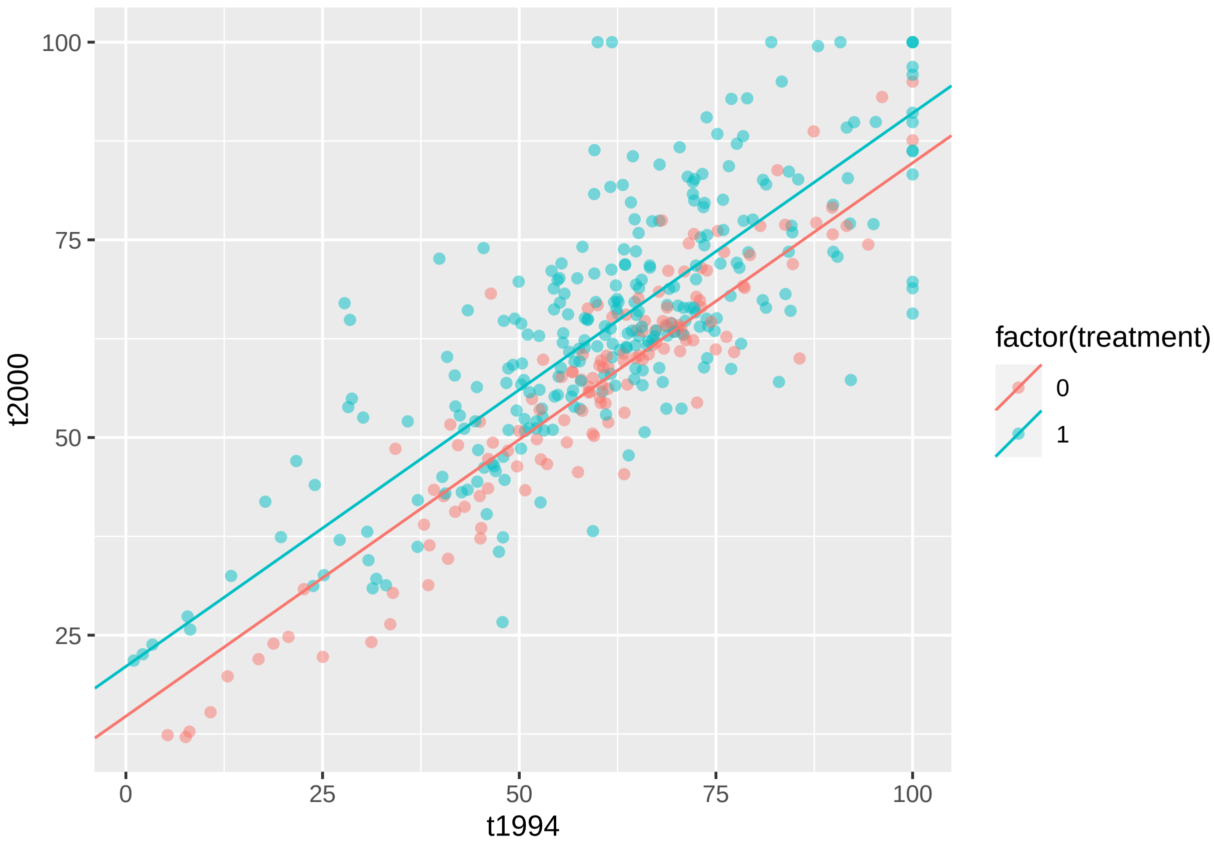
```r
intercepts <- c(coef(treatment.lm)["(Intercept)"],
                coef(treatment.lm)["(Intercept)"] + coef(treatment.lm)["treatment"])

lines.df <- data.frame(intercepts = intercepts,
                       slopes = rep(coef(treatment.lm)["t1994"], 2),
                       treatment.1 = levels(factor(progresa$treatment)))
lines.df
```

```
##   intercepts    slopes treatment.1
## 1   14.75296 0.6997991           0
## 2   21.04366 0.6997991           1
```

```r
progresa %>%
  ggplot(aes(x = t1994, y = t2000, color = factor(treatment))) +
  geom_point(alpha = 0.5) +
  geom_abline(aes(intercept = intercepts,
```

```
                     slope = slopes,
                     color = factor(treatment.1)), data = lines.df)
```



Question 6 [6 pts]

Now, we will explore whether early versus late receipt of the CCT program affects 2000 voter turnout differently for precincts that had different voter turnout in the prior 1994 election.

##Answer for

6a

Add an interaction term to your model from Question 5 between 1994 voter turnout and the treatment variable. Create a residual plot and use it to assess the linearity assumption.

##Answer for 6a

From observing the relationship between the voter turnout in 1994 vs 1994:treatment, a linear association is sean in both treatment and control groups. Around 60 we can observer a cluster looking at the residual plot, from this we can not claim that the association between the 1994 vs 1994:treatment. We can not verify an assumption based on the cluster.

6b

Write out the multiple regression equation for this model as a single equation and then as a pair of equations, one for each treatment arm. Create a scatterplot of the outcome and the continuous predictor variable. Color the points on this scatterplot by their treatment status. Add the two regression lines to this figure. Or sketch the figure described here by hand, take a picture and include it in your HW document.

#Answer for 6b

Y_i = alpha_hat + beta_hat* X_i _ epsilon_i_hat Y_i = 0.6535661+ treatment * (-.009967146) + epsilon_i_hat Y_i= 0.6535661 + control* (-.009967146) + epsilon_i_hat

Y_i=1.3071322 + treatment * (-.009967146) + control* (-.009967146) + epsilon_i_hat

6c

Interpret all four of the model coefficients for this multiple regression equation. Interpret the RMSE and the R^2 value for the model. Compare them to the RMSE and R^2 for the model in Question 5. What does this model tell you about whether the timing of the CCT program had more or less of an effect on precincts with prior low voter turnout rates relative to precincts with prior high voter turnout rates?

##Answer for 6c

Residual standard error: 8.847 on 403 degrees of freedom

This is telling us that the voter turnout rate in 1994 was an approximate avg 8.85% away from what it is expected to be based on its linear relationship with the avg voter turnout rate in 1994:treatment

R^2 value is proportion of the variance in 1994 and is like the linear relationship with the avg voter turnout rate in the group of 1994:treatment, which gives us info for the real data that predicts the approximation.

Multiple R-squared: 0.7138, Adjusted R-squared: 0.7116

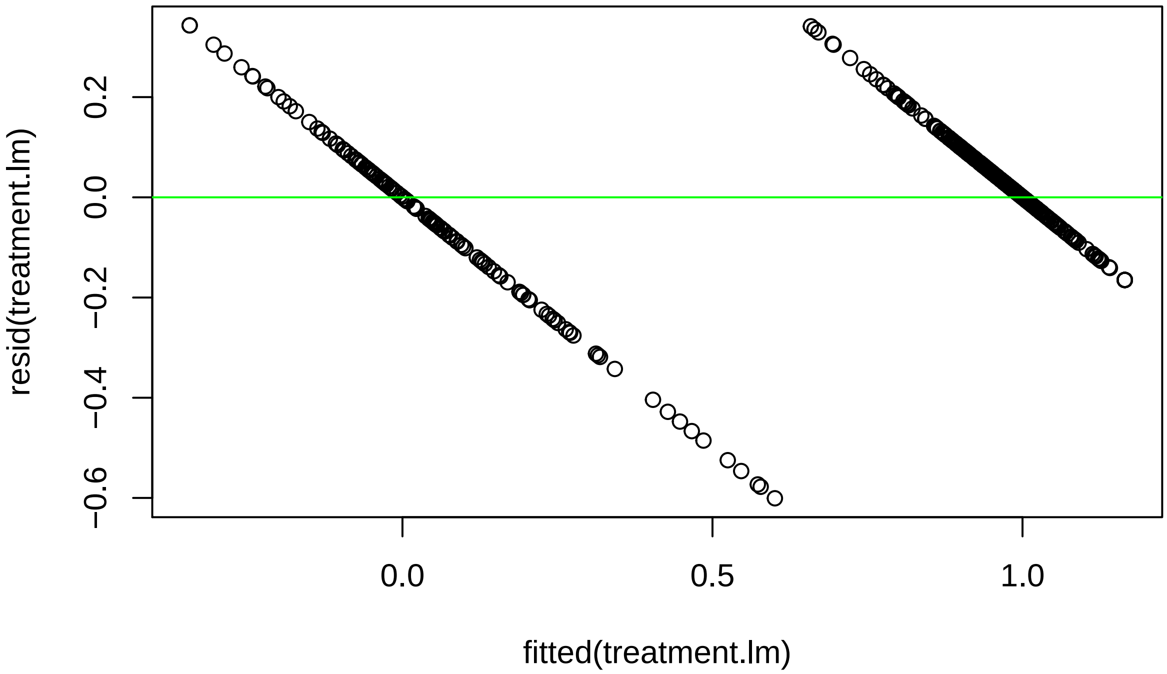
The variable of the voter turn out rate in 1994 and is taken into consideration by the avg voter turnout rate in 1994:treatment. From our multiple r^2 being at 71.4% of the variation in t1994 and which can be explained by the variation in the voter turnout rate of 1994:treatment. This value is high so we can make a cliam from the association from these variables.

By looking at the answer from question 4, the rmse is slightly smaller indicating a better fit of our data set. The R^2 is larger demonstrating larger confidence level in the data.

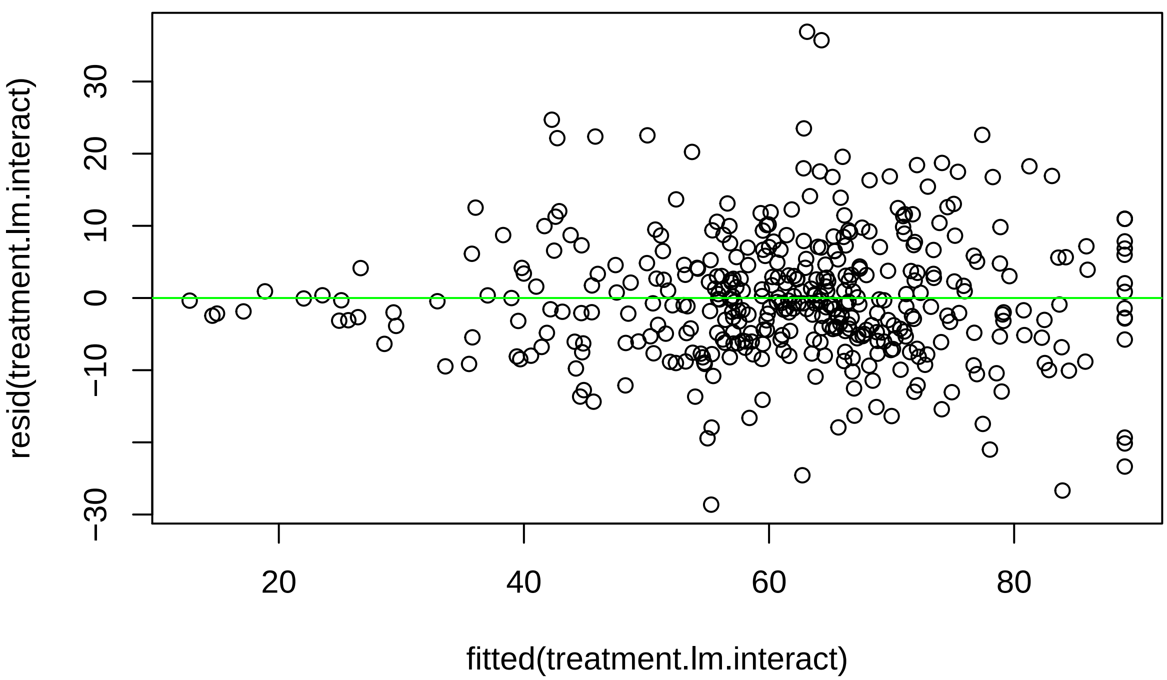
We can conclude that the treatment and the control had no affect on the CCT program, this can also be observed with the fact that we have no overlap in the graph.

Answer 6

```r
treatment.lm <- lm(formula = treatment ~ t1994:treatment + t1994, data = progresa)
treatment.lm.interact <- lm(t2000 ~  treatment + t1994 + treatment*t1994, data = progresa)
plot(fitted(treatment.lm), resid(treatment.lm))
abline(h = 0, col = "green")
```




```r
plot(fitted(treatment.lm.interact), resid(treatment.lm.interact))
abline(h = 0, col = "green")
```



```r
summary(treatment.lm)
```

```
##
## Call:
## lm(formula = treatment ~ t1994:treatment + t1994, data = progresa)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60066 -0.05493  0.01492  0.07538  0.34315
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.6535661  0.0236736   27.61   <2e-16 ***
## t1994           -0.0099671  0.0004103  -24.29   <2e-16 ***
## treatment:t1994  0.0150810  0.0002365   63.77   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1421 on 404 degrees of freedom
## Multiple R-squared:  0.9099, Adjusted R-squared:  0.9095
## F-statistic:  2040 on 2 and 404 DF,  p-value: < 2.2e-16
```

```r
summary(treatment.lm.interact)
```

```
##
## Call:
## lm(formula = t2000 ~ treatment + t1994 + treatment * t1994, data = progresa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.632  -5.396  -0.741   4.704  36.893
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.44778    2.50423   3.373 0.000814 ***
## treatment       15.83088    3.09763   5.111 4.97e-07 ***
## t1994            0.80567    0.04007  20.105  < 2e-16 ***
## treatment:t1994 -0.15817    0.04898  -3.229 0.001343 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.847 on 403 degrees of freedom
## Multiple R-squared:  0.7138, Adjusted R-squared:  0.7116
## F-statistic:   335 on 3 and 403 DF,  p-value: < 2.2e-16
```

```r
intercepts <- c(coef(treatment.lm)["(Intercept)"],
                coef(treatment.lm)["(Intercept)"] + coef(treatment.lm)["treatment"])
lines.df <- data.frame(intercepts = intercepts,
                       slopes = rep(coef(treatment.lm)["t1994"], 2),
                       treatment.1 = levels(factor(progresa$treatment)))
lines.df
```

```
##   intercepts       slopes treatment.1
## 1  0.6535661 -0.009967146           0
## 2         NA -0.009967146           1
```

```r
progresa %>%
  ggplot(aes(x = t1994, y = t2000, color = factor(treatment))) +
  geom_point(alpha = 0.5) +
  geom_abline(aes(intercept = intercepts,
                  slope = slopes,
                  color = factor(treatment.1)), data = lines.df)
```

```
## Warning: Removed 1 rows containing missing values (geom_abline).
```