# Election and Conditional Cash Transfer Program in Mexico

October 12, 2023

In this HW, 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 the 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 the intended outcomes, socioeconomic status and inter-generational transfer of poverty, are strong.

Here the interest is in a possible positive side-effect of the program on improving voter turnout and perhaps impacting which party citizens voted for.

This exercise is based on the following two 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 government did not have the resources to provide the *Progresa* program to all eligible villages when it was first decided upon. For this reason, the government decided it was ethical to randomize provision of the program to begin early or later - this is an example of a lagged randomized study design. The treatment is Early *Progresa* and the alternative is Late *Progresa*.

The author of the original study hypothesized that the CCT program would mobilize voters, leading to an increase in turnout and more support for the incumbent party (PRI in this case).The analysis was based on a sample of precincts that each contain at most one village participating 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* |

| Name | Description |
|---|---|
| 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 for Villages in the Precinct |
| 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.

```
library(tidyverse)
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse
2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.0
## ✓ ggplot2   3.4.3     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── ──
tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all
conflicts to become errors
```

```
progresa <- read.csv("data2/progresa.csv")
```

## Question 1 [7 pts]

Consider the impact of early versus late receipt of the CCT program on *voter turnout in the 2000 election*. No data analysis is needed for this question.

### 1a [3 points]

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

### 1b [2 points]

For precincts receiving the CCT program Early(treat), what is their average missing counterfactual outcome?

### 1c

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

### 1d

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

## Answer 1

### Answer 1a

what is the impact on the voter turnout in the 2000 election for a precinct that experiences the early progresa relative to a precinct that experiences late progresa for a precinct?

potential outcomes #1: what the voter turnout would be for the precinct that experienced early progresa (21 months)of the cct program. potential outcome #2 : what the voter turnout would be for a precinct that experienced late receipt (6 months) of the progresa cct program. ### Answer 1b the average MCF: the average MCF for treatment precincts : what the average voter turnout would have been in the 2000 election for all precincts that implemented early cct program, if instead they had received late cct program but all else remained the same. ### Answer 1c we would use the average outcome (voter turnout) for the precincts that receive (control) late progresa(6 months) precincts to estimate the average MCF for the treated/early(21 months) cct program precincts. ### Answer 1d We would then estimate the treatment effect of the intervention by taking the difference between the average factual outcome(voter turnout) for the early treated precincts and this estimate of their average MCF. the researcher hypothesized that the CCT program would mobilize voters, leading to an increase in turnout and more support for the incumbent party (PRI in this case).

## Question 2 [10 pts]

### 2a [4 points]

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 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.

### 2b [6 points]

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 results you estimated in the first part of this question. Use the term *confounder* in your answer.

## Answer 2

### Answer 2a

the treatment group(early progresa) has a mean for voter turnout share in percent points is 64.3 and the median is 64.41 percent points. 50 % of the share of voter turnouts in treatment precincts lie between 55.56% and 73.49%. The alternative group(late progresa) has a mean of 56.43% share of voter turnout.the middle 50 percent of values for share of voter turnout lies between 48.91 % and 65.72%. The treatment difference between early and late progresa is 7.87% in voter turnout share with the treatment group(early progresa) having the higher turnout. There is also an observable difference in the IQR's 55.56 - 77.49(treat) and 48.91 - 65.72(alternative). there seems to be greater voter turnout outcome for the treatment group.

the alternate group(late progresa) has percentage of pri votes as 34.18 and the treatment group(early progresa) has the percentage of pri votes as 36.11%, so treatment group has a higher percentage in support for the party by 1.93 percent points.

```
progresa_treat <- progresa %>% filter(progresa$treatment== 1)
progresa_alt <- progresa %>% filter(progresa$treatment== 0)
summary(progresa_treat$t2000)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.78   55.56   64.41   64.33   73.49  100.00
```

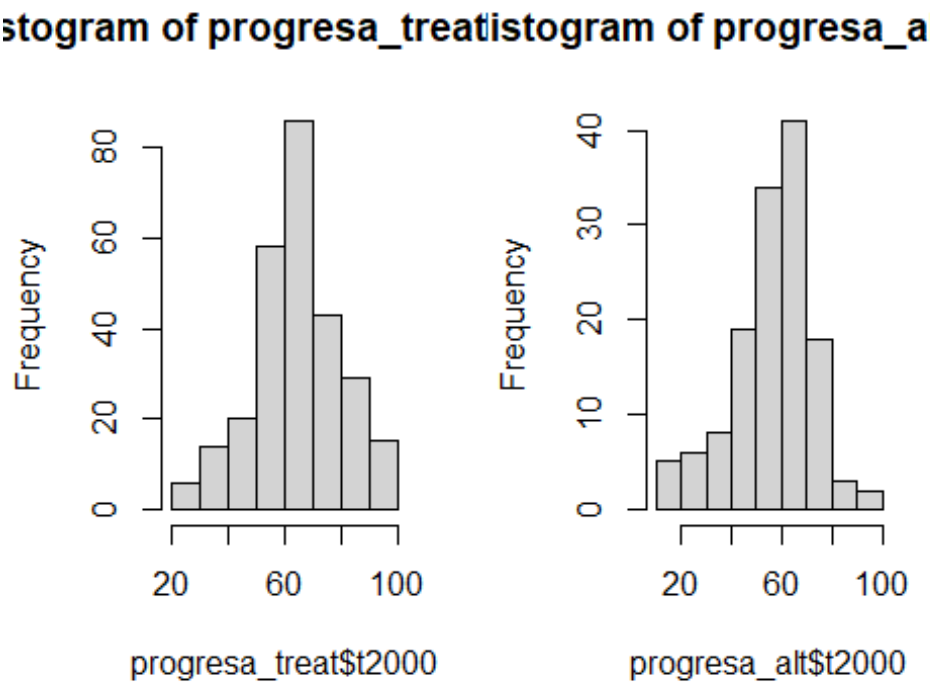
```
summary(progresa_treat$pri2000s)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.741  25.700  35.496  36.112  45.022  87.500
```

```
hist(progresa_treat$pri2000s)
```


```r
hist(progresa_alt$pri2000s)
```

```
par(mfrow=c(1,2))
hist(progresa_treat$t2000)
hist(progresa_alt$t2000)
```




```
par(mfrow=c(3,4))
summary(progresa_alt$t2000)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   12.13   48.91   59.72   56.43   65.72   95.00

summary(progresa_alt$pri2000s)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.942  25.137  35.067  34.189  44.035  76.923
```

### Answer 2b

AVG_POVERTY: the covariate average poverty variable is balanced between the treatment(early progresa) and alternative group( late progresa) with a range of 3-5 for both and the mean 4.57 for early progresa and 4.59 for late progresa and the IQRs for early treatment - 50% of the values for the average poverty rate is between 4.26-5 and 4.3-5 boxplot also suggests that there are two outliers for the treatment variable and two for the alternative variable that are at 3 and 3.15 approximately for both groups. From the histogram, the spread is similar and it is highly skewed to the left. Both are unimodal.

T-1994: mean for the treatment group is 61.85 and alternate is : 59.554 The covariate that represents the share of voters in precincts that voted in 1994 the treatment group(early

progresa) and the alternative group(late progresa) have a similar distribution but are slightly different and both are left skewed. They are both unimodal. the range for the early progresa( treatment group)1-100 while the range for late progressa is 5.3-100. there are a few outliers at the tail end of the distribution.

TREAT(early progresa): 0-15 the outliers are between 0-15 alternative(late progresa): the outliers are between 3-15 the outliers are all at a distance of 1/4 th IQR's from each other and are 1.5 Iqr's and 2 Iqr's away from the 3 rd quartile.

For the most part this will give an unbiased estimate of outcome that is the voting share of the precincts considering our distribution of the covariates between both groups are balanced. As we've defined earlier for our MCF to be unbiased we have to consider that the baseline covariates between both the groups are not systematically different and those that can impact or effect the outcome. The 2 covariates average poverty and voter share percent in 1994 should not be confounders that bias the outcome and for the two covariates to not be confounders they would have to be balanced and no systemic differences that can confound the outcome or are not directly impact the outcome.

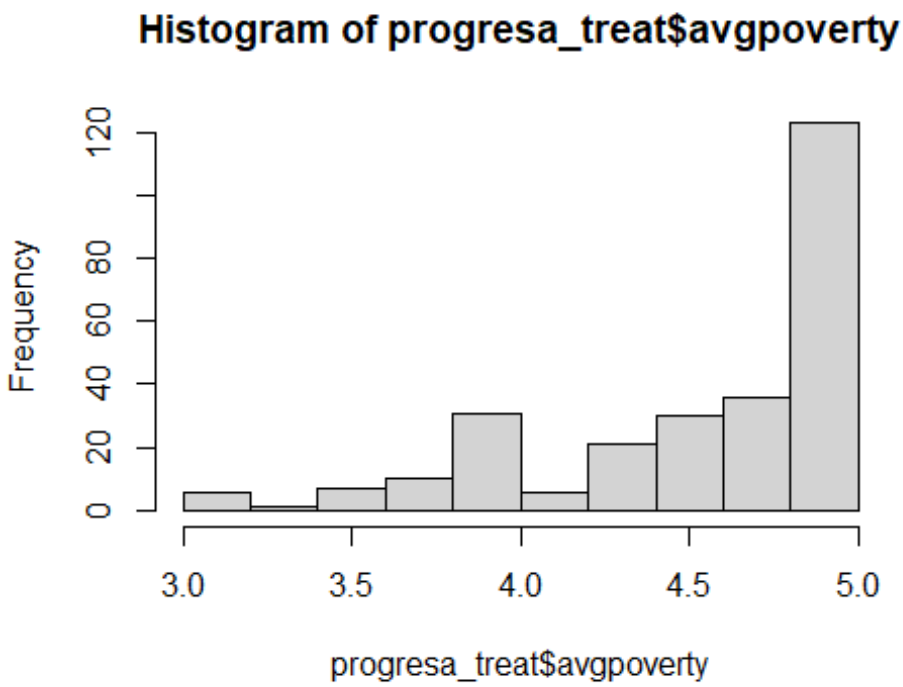
```
summary(progresa_treat$avgpoverty)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.268   4.733   4.571   5.000   5.000
```

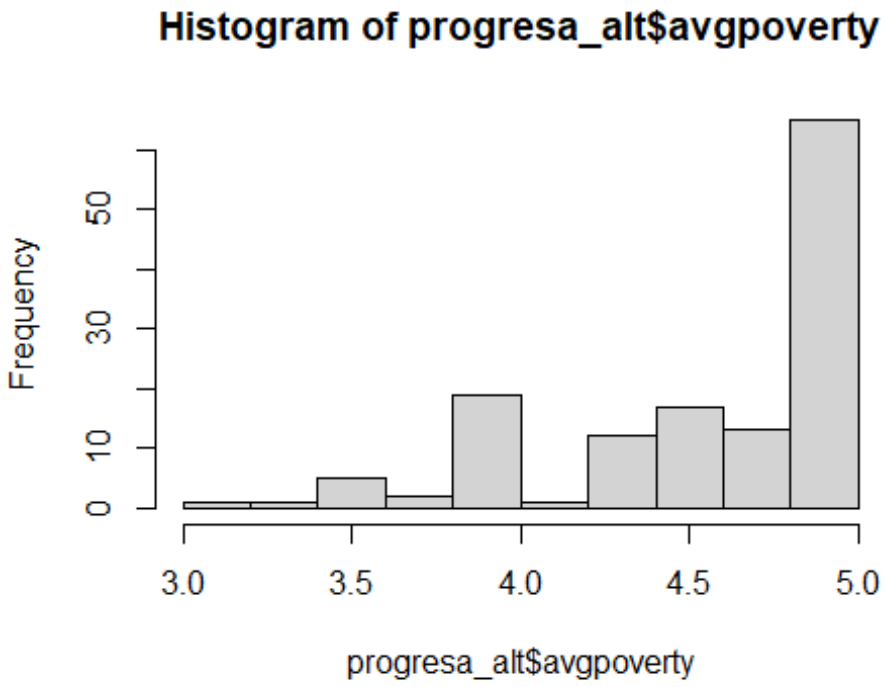
```
summary(progresa_alt$avgpoverty)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.321   4.750   4.592   5.000   5.000
```

```
hist(progresa_treat$avgpoverty)
```




```r
hist(progresa_alt$avgpoverty)
```



```r
par(mfrow=c(1,2))
summary(progresa_treat$t1994)
```

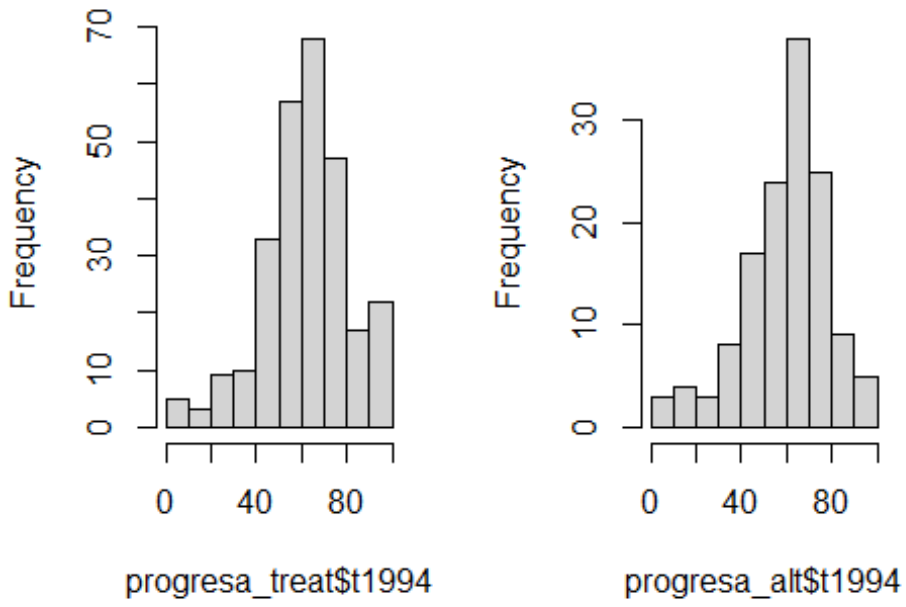
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.001  51.159  62.621  61.858  73.162 100.000
```

```r
summary(progresa_alt$t1994)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.308  49.913  61.309  59.554  70.970 100.000
```

```r
hist(progresa_treat$t1994)
hist(progresa_alt$t1994)
```

**stogram of progresa_treatistogram of progresa_alt$**



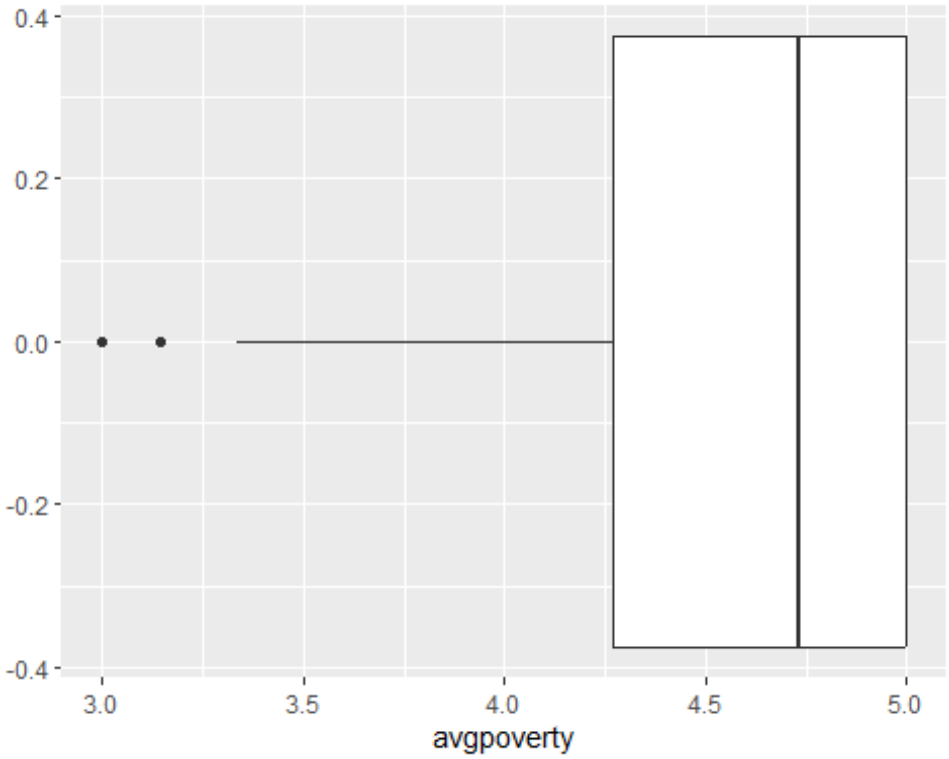

```r
par(mfrow=c(1,2))

tapply(progresa$t1994, progresa$treatment, summary)
```

```
## $`0`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.308  49.913  61.309  59.554  70.970 100.000
##
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.001  51.159  62.621  61.858  73.162 100.000
```

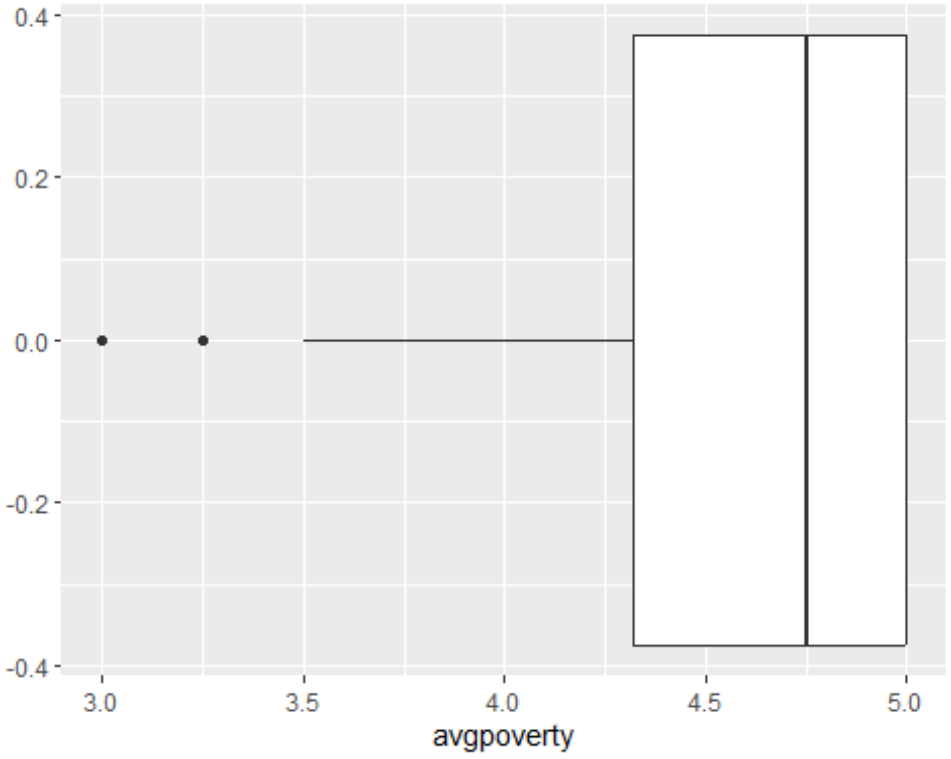
```r
tapply(progresa$avgpoverty, progresa$treatment, summary)
```

```
## $`0`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.321   4.750   4.592   5.000   5.000
##
## $`1`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.268   4.733   4.571   5.000   5.000
```

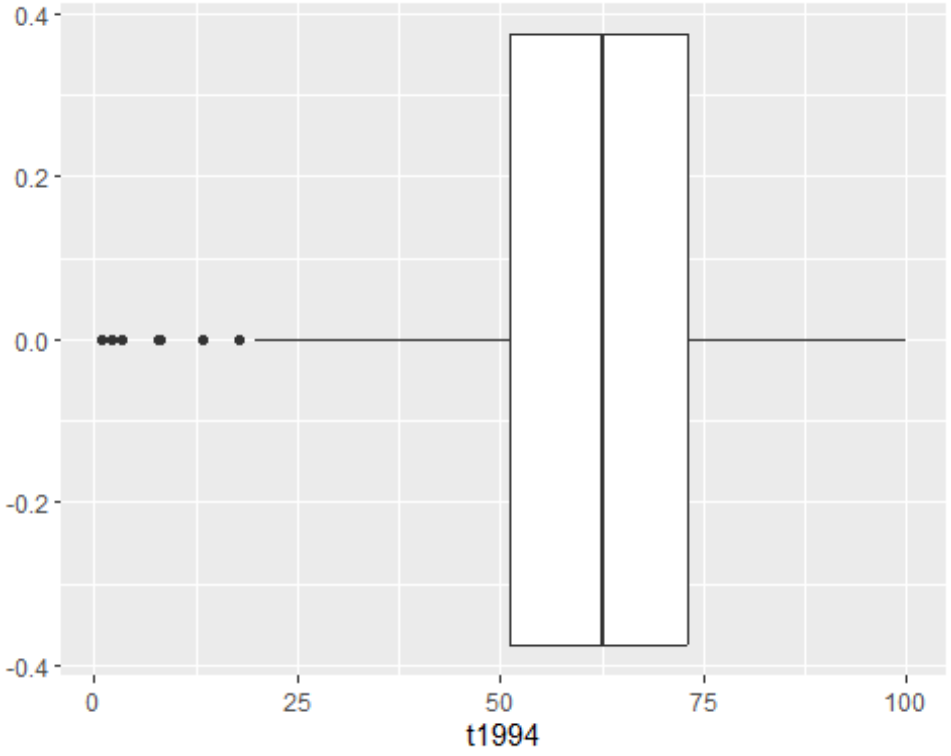
```
progresa_treat %>%
  ggplot(aes(x=avgpoverty))+
  geom_boxplot()
```



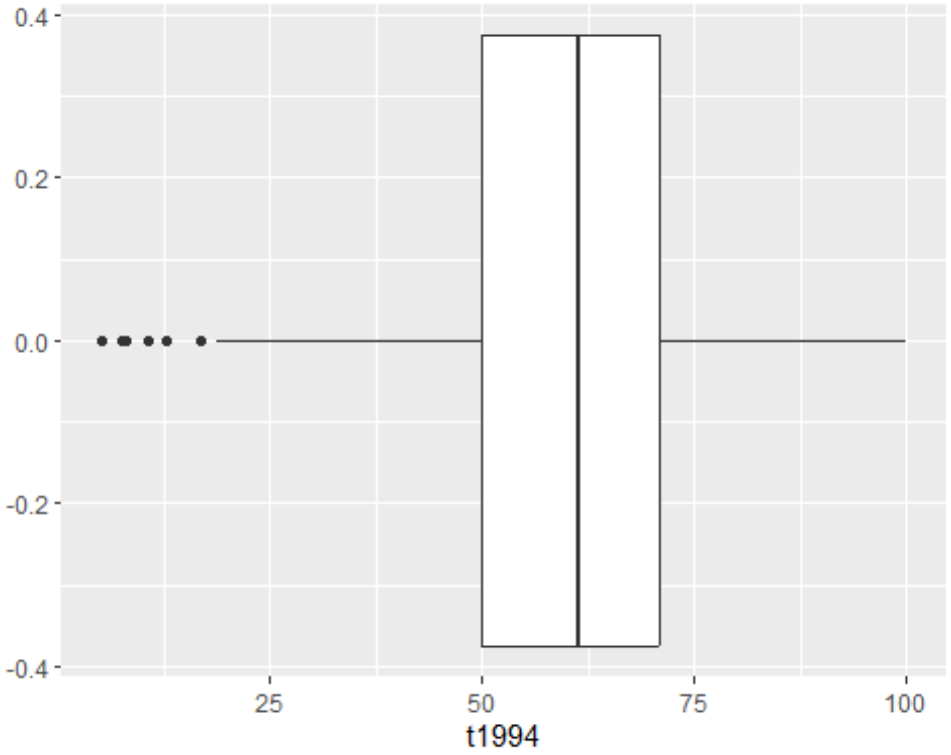

```
progresa_alt %>%
  ggplot(aes(x=avgpoverty))+
  geom_boxplot()
```




```
progresa_treat %>%
  ggplot(aes(x=t1994))+
  geom_boxplot()
```



```
progresa_alt %>%
  ggplot(aes(x=t1994))+
  geom_boxplot()
```




```
summary(progresa_treat$t1994)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.001  51.159  62.621  61.858  73.162 100.000
```

```
summary(progresa_alt$t1994)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.308  49.913  61.309  59.554  70.970 100.000
```

## Question 3 [10 pts]

Considering all precincts in the study, in this Question we will investigate the linear relationship between a precinct's voter turnout in the 2000 election (outcome) and it's voter turnout in the 1994 election (predictor) by answering the questions below.

### 3a [6 points]

Run the simple linear regression model described above. Make a scatterplot and add the estimated linear regression line to this figure. Make a residual plot and add a horizontal zero line to this figure. What do the scatterplot and residual plot tell us about the shape of this bivariate relationship? In particular, is the linearity assumption for linear regression violated or does it appear to hold? What do you see in the figure that tells you that it does or does not hold?

### 3b [4 points]

Regardless of your answer to 3a, interpret the estimated coefficients (y-intercpet and slope) of the simple linear regression. Interpret the RMSE and $R^2$ value from the model.

## Answer 3

### Answer 3a

The scatterplot of between the predictor variable and the outcome variable with the regression line shows a positive corelation between the two variables. A high voting outcome in 1994 seems to have a high voting outcome in the year 2000 as well. There are two outliers with the value of their y outcome( 2000 voting share) at 100 percent for 61 and 62 percent in their x predictor variable(t1994).

The mean of the residuals appears to be zero in each segment as each section has the same amount of data above and below the zero line in 4 segments from the residual plot and the linearity assumption holds. The residual plot between 0-25 seems to have a mean that is close to zero with the points above and below the zero line being balanced. Between 25 and 50, there are more densely located points below the zero line but points above the zero line are a little further away from the mean the mean looks to be zero. Similarly we observe the same between 50 and 75 with the points above and below the zero line being balanced which can indicate a mean of zero Similarly, the last segment has equal distribution above and below and the mean is zero. we can say that the linearity assumption holds as the mean of the residuals appears to be zero.

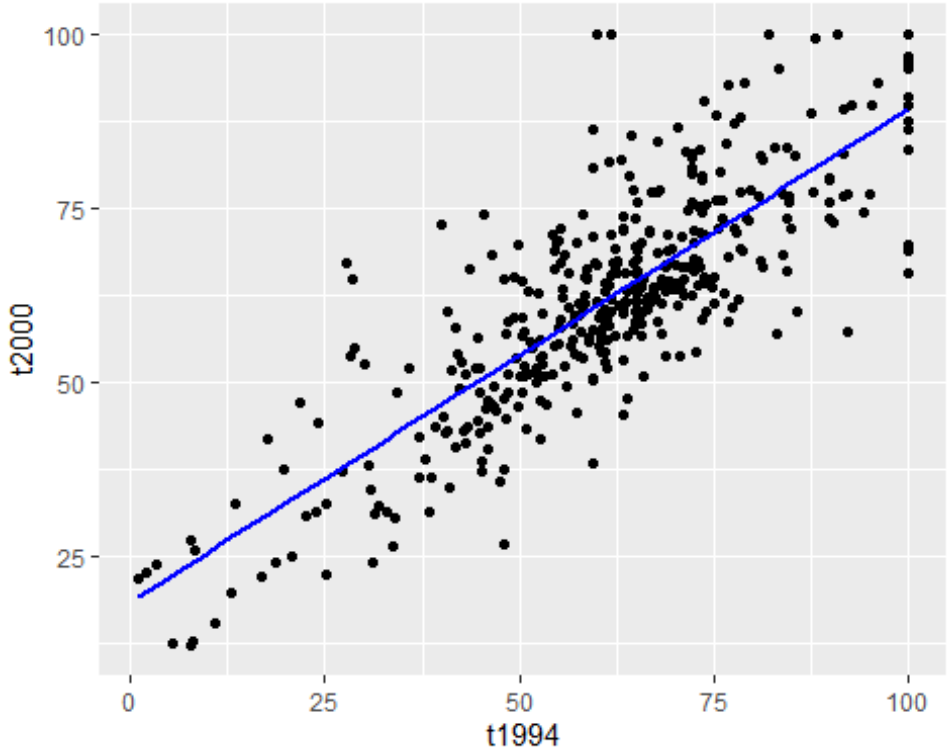
```
progresa_mod <- lm(t2000 ~ t1994, data = progresa)
summary(progresa_mod)
```

```
##
## 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
```

```
progresa %>%
ggplot(aes(y = t2000, x = t1994)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE, color = "blue")
```

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

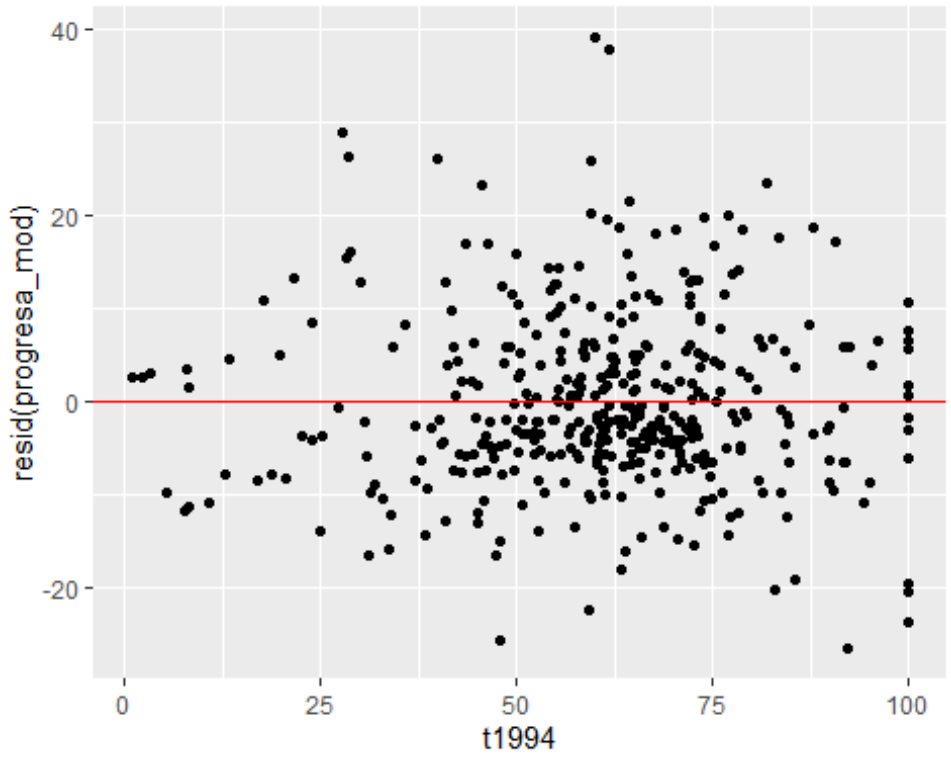



```
cor(progresa$t2000, progresa$t1994)
```

```
## [1] 0.8209515
```

```
progresa %>%
   ggplot(aes(x = t1994, y = resid(progresa_mod))) +
```

```
  geom_point()+
  geom_hline(yintercept=0 ,color= "red")
```




### Answer 3b

(t2000)Yi= 18.40+ 0.70 Xi(t1994) + E^i

intercept: when the value of Xi= 0 which is the 1994 voter turnout percentage, the outcome variable which is voter tunout percentage in 2000 is 18.40 %. It could be meaningful as there could be precincts that have 0% as their (t1994) percentage of voter turnout in the year 1994.

the slope indicates that for a 1 percentage point increase in the 1994 voter turnout(t1994) (1 %), we expect the precincts to have an average 0.70 percent point increase in the percentage of voter turnout in 2000 . This means a 10 percent point increase in the percentage of the precinct that voted in 1994 is associated on average with a 7 percent point increase in the precinct percentage of voter turnout in 2000.

RMSE:9.4% : we should expect this prediction of the voter turnout outcome(t2000) to be off (above or below) by about 9.4 percent points this is the average prediction error. Half the time, the true value for the percent voter turnout in 2000 will be no more than 5.6 percent points below and 5.31 percent points above the prediction we make with this regression model. The worst of our predictions could be as far as 26.4 percent points below the true value and 39.104 (percent points ) above the true value.

The R-squared value is 0.67 which means that there is a 67% of the variability in percentage of voter outcome in 2000 can be explained by the percent of voter outcome in 1994 and the linear relationship between the two variables.(t2000 and t1994)

## Question 4 [22 pts]

### 4a [4 points]

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.

### 4b [8 points]

Write out the multiple regression equation for this model first 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 regression lines described here by hand, take a picture and include it in your HW document.

### 4c [10 points]

Interpret the model coefficients for this multiple regression equation (the three from the single 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 3. What does this model tell you about whether the timing of the CCT program had the hypothesized effect?

## Answer 4

### Answer 4a

The linearity assumption appears to hold. Our evidence for this conclusion is that the mean of the residuals seems to be zero in all regions of the residual plot as you move from left to right.

```
progresa_mod2 = lm( formula = t2000 ~ treatment + t1994, data = progresa)
summary(progresa_mod2)
```

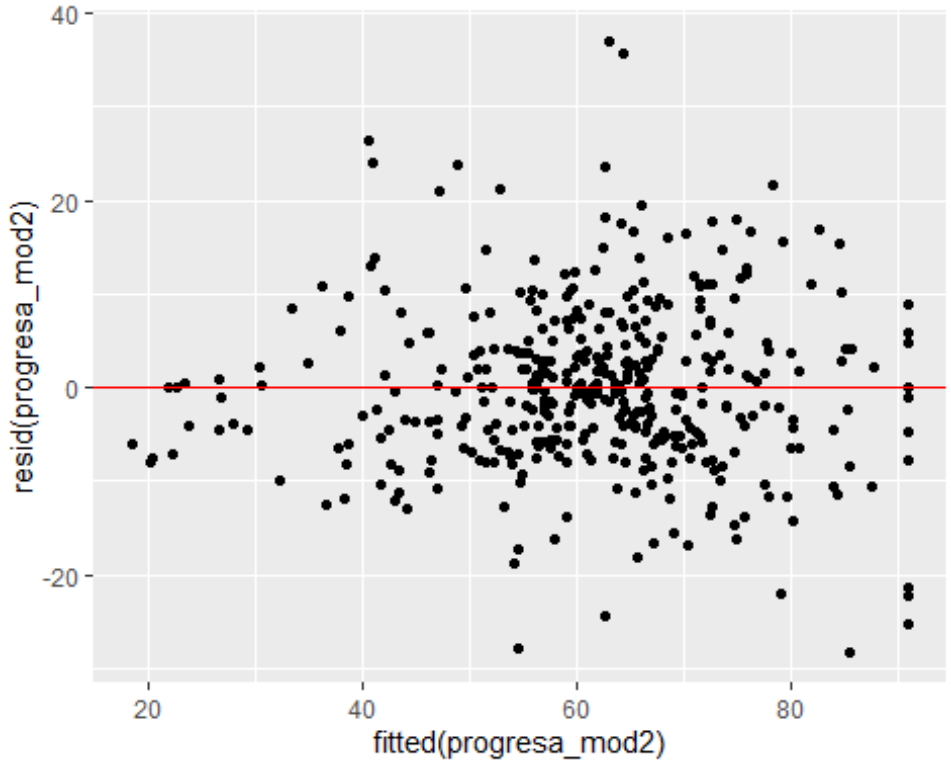
```
##
## Call:
## lm(formula = t2000 ~ treatment + t1994, data = progresa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.265  -5.577  -0.374   4.792  36.992
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.75296    1.58622   9.301  < 2e-16 ***
## 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
```

```
summary(resid(progresa_mod2))
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -28.2652  -5.5769  -0.3742   0.0000   4.7916  36.9916
```

```
progresa %>%
   ggplot(aes(x = fitted(progresa_mod2), y = resid(progresa_mod2))) +
  geom_point()+
  geom_hline(yintercept=0 ,color= "red")
```




### Answer 4b

the general multiple regression equation:

$$Yi(outcomeofvoterturnoutin2000) = 14.75 + 6.29Xi(treatment) + 0.70Xi1(t1994) + \grave{E}i$$

equation for when treatment =0(late progresa 6 months CCT)

$$Yi(outcomeofvoterturnoutin2000) = 14.75 + 0.70Xi1(t1994) + \grave{E}i$$

equation for when treatment =1(early progresa 21 months CCT)

$$Yi(outcomeofvoterturnoutin2000) = 21.04 + 0.70Xi1(t1994) + \grave{E}i$$

```
progresa_mod2 = lm(formula = t2000 ~ treatment + t1994, data = progresa)
summary(progresa_mod2)
```

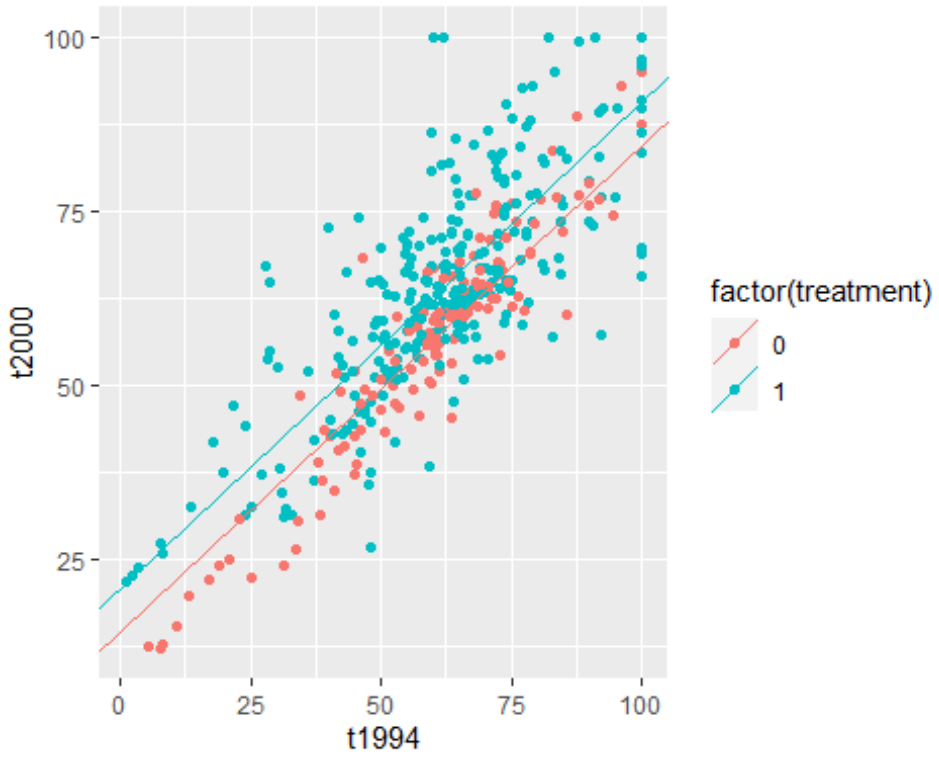
```
##
## Call:
## lm(formula = t2000 ~ treatment + t1994, data = progresa)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.265  -5.577  -0.374   4.792  36.992
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 14.75296    1.58622   9.301  < 2e-16 ***
## 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
```

```
progresa_mod2$coef
```

```
## (Intercept)   treatment       t1994
##  14.7529579   6.2907049   0.6997991
```

```
#since no treatment for precincts with the late cct progresa is coded as 0,
the regression model intercept
# is the intercept for precicnts that recieved late cct progresa alternative
int_notreat = progresa_mod2$coef['(Intercept)']

#since treatment for precincts with the early cct progresa(treatment)is coded
as 1, the intercept for precincts with early cct (treat) is the intercept of
this regression model
#PLUS the coefficient on the dichotomous variable
int_yestreat = progresa_mod2$coef['(Intercept)'] +
progresa_mod2$coef['treatment']

#since there is no interaction effect, both lines have the same slope
slope = progresa_mod2$coef['t1994']  ##coefficient on the continuous variable

progresa %>%
  ggplot(aes(y = t2000, x = t1994, color = factor(treatment))) +
  geom_point() +
  geom_abline(aes(intercept = int_notreat, slope = slope, color = '0')) +
```

```
  #regression line for no treatment(late progresa)
  geom_abline(aes(intercept = int_yestreat, slope = slope, color = '1'))
```




```
  #regression line for countries that had the treatment(early progresa)
```

### Answer 4c

within each treatment arm , The estimated slope means that for one percent point increase in the voting turnout in year 1994, we expect the percentage of voting turnout in 2000 to increase by 0.70 percent points for precincts . Equivalently, we expect that for 10 percent point increase in the voting turnout in year 1994 we expect a 7 percent point increase in the percentage of voting outcome in 2000.

the coefficient of the treatment group variable means that the precincts that experience the treatment (early progresa) program will have an increase of 6.29 percent points when compared to those that experience the alternative irrespective of their voter turnout in 1994.

Late progresa (control): for those precincts with age 0 in the late progresa arm of the study , we expect that their voting outcome in 2000 is 14.75 %.

early progresa(treatment) intercept : for those precincts with age 0 in the early progresa arm of the study, we expect that their voting outcome in 2000 is 21.04%.

RMSE: On average, the percentage of voter turnout in 2000 is approximately 8.95 percent points more or less than the expected value predicted based on the turnout percentage in 1994 and the treatment type(early or late progresa).

multiple R sqrd: the voting outcome percentage in 1994 and the treatment status (early or late progresa) and their linear association with the voting outcome percentage in 2000 can explain 70 percent of the variability in the percentage of voting outcome in 2000.

Adding the dichotomous variable (early or late progresa) to the model reduced the RMSE slightly from 9.412 to 8.95 and the variability in percentage share of voters in 2000 has increased from 67.4% to 70.6% with and is explained by the age , the dichotomous variable(treatment type).this shows that there is a small impact between the early and late progresa.

## Question 5 [19 pts]

Now, we will explore whether Early versus Late receipt of the CCT program affects 2000 voter turnout *differently* for precincts that had lower versus higher voter turnout in the prior 1994 election.

### 5a [2 points]

Add an interaction term to your model from Question 4 between 1994 voter turnout and the treatment variable. Write out the multiple regression equation for this model first as a single equation and then as a pair of equations, one for each treatment arm. Interpret all four of the model coefficients for this multiple regression equation, or interpret the four model coefficients in the two separate simple linear regression equations.

### 5b [9 points]

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 regression lines described here by hand, take a picture and include it in your HW document. Briefly describe the two regression lines. 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.

### 5c [8 points]

What does this multiple linear regression model estimate the average effect of Early versus Late CCT program receipt is on 2000 voter turnout for precincts with a 75% turnout rate in the 1994 election? (Calculate the mean 2000 voter turnout for precincts with 75% voter turnout in 1994 with Early CCT Program and the same value for those with Late CCT Program and take the difference.) What does this multiple linear regression model estimate the effect of Early versus Late CCT program receipt is on 2000 voter turnout for precincts with a 25% turnout rate in the 1994 election? Summarize what this model tells 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 5

### Answer 5a

For late progresa precincts, we expect that for every 1 percent point increase in the voter percent turnout in 1994, there is a 0.80 INCREASE in voter outcome percent in 2000. On a more meaningful scale for a 10 percent point increase in the voter percent turnout there is a 8 percent point increase in the voter outcome percent in 2000. The intercept means that for a precinct under late progresa (treatment=0) and voter percent share in 1994 “0%”, We expect the average voter share outcome in 2000 to be 8.44 % .

For early progresa precincts, we expect that for every one percent increase in the voter percent turnout in 1994, there is an 0.65 percent point increase in the voter outcome percentage in 2000.On a more meaningful scale for a 10 percent point increase in voter turnout percentage in 1994, there is a 6.5 percent point increase in the voter turnout

percentage in 2000. The intercept means that for a precinct with percentage share of voters in 1994 that is 0%," we expect the average share of voter outcome in 2000 to be 24.24%. This number has no practical interpretation.

$$voteroutcomein2000(Yi) = 8.44 + 0.80(t1994Xi) + 15.8(treatmentX1i) - 0.15(treatmentX1i)(t1994Xi) + E\hat{}i$$

for early progresa -> treatment=1

$$voteroutcomein2000(Yi) = 24.27 + 0.65(t1994Xi) + E\hat{}i$$

for late progresa -> treatment =0

$$voteroutcomein2000(Yi) = 8.44 + 0.80(t1994Xi) + E\hat{}i$$

```
progresa_mod3 <- lm(t2000 ~ t1994 + treatment + treatment*t1994, data =
progresa)
summary(progresa_mod3)

##
## Call:
## lm(formula = t2000 ~ t1994 + treatment + 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 ***
## t1994            0.80567    0.04007  20.105  < 2e-16 ***
## treatment       15.83088    3.09763   5.111 4.97e-07 ***
## t1994:treatment -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
```

### Answer 5b

based on the regression lines, there seems to be an overlap between the treatment and the alternative groups, at a approximately 100 percent value of the voter turnout in 1994, the alternative takes over and t 2000 voter turnout begins to get greater for the alternative.Initially the treatment effect seems to be high between the early and late progresa and as the voting percentage increases in 1994, the difference and gap between the two groups starts to decrease as we move along the regression lines.

On average, we can expect the percentage of voter turnout in 2000 to be approximately 8.84 percent points more or less than the expected value based on t1994, treatment and their interaction (this is the RMSE or residual standard error estimate)

precinct voter turnout in 1994, treatment group(late or early progresa ), and their interaction (along with their linear association with precinct voter turnout in 2000) explain 71.3% percent of the variability in the percent share outcome of voters in 2000.

As compared to the 4 th question results, we can observe that introducing interaction between the voter turnout in 1994 and the treatment effect reduces the RMSE slightly from 8.95 % to 8.84 %. however, the R squared variability has increased slightly from 70.6% to 71.3 %.

```
progresa_mod3$coef

##     (Intercept)           t1994       treatment t1994:treatment
##       8.4477812       0.8056721      15.8308777      -0.1581699

##since late progresa is coded as 0...
#the intercept for the late progresa line is just the regression model
intercept
int_latep = progresa_mod3$coef['(Intercept)']
#the slope for the late progresa is just the regression model coefficient on
the continuous variable
slope_latep = progresa_mod3$coef['t1994']
##since early progresa is coded as 1...
#the intercept for the straight engine line is the regression intercept PLUS
the coefficient on the dichotomous variable
int_earlyp = progresa_mod3$coef['(Intercept)'] +
progresa_mod3$coef['treatment']
#the slope for the straight engine line is the coefficient on the continuous
variable PLUS the coefficient on the interaction term
slope_earlyp = progresa_mod3$coef['t1994'] +
progresa_mod3$coef['t1994:treatment']

progresa %>%
ggplot(aes(y = t2000, x = t1994, color = factor(treatment))) +
geom_point() +
geom_abline(aes(intercept = int_latep, slope = slope_latep, color = '0')) +
#regression line for late progresa
geom_abline(aes(intercept = int_earlyp, slope = slope_earlyp, color = '1'))
#regression line for early progresa
```


### Answer 5c

the treatment effect: for precincts that are selected to go through the early progresa (treat=1), on average they have a 72.99 percentage in voter outcome(t2000) for a 75% of voter percentage outcome in 1994.

for precincts that undergo the late progresa with treat=0 , on average have a voter outcome percent 68.44% in 2000 which is associated with a 75 % in their voter outcome in 1994.

therefore, the treatment of early progresa has a Positive impact and average association with voter outcome percentage share with the voting outcome in 1994 as 75% and the difference between early and late progresa is 4.55 percent points. the early progresa has a slightly more postive impact compared to late progresa.

for precincts that undergo the early progresa treatment=1, on average they have a 40.45 percentage in voter outcome in 2000 which is associated with a 25 % in their voter outcome in 1994.

the precincts that undergo late progresa=0, on average have a 28.44% in voter outcome associated with a 25% voter outcome in 1994 for the year 2000.

therefore it is clear that the treatment of early progresa has a postive impact with average association with voter outcome in 1994 as 25 % and the treatment affect between the early and late progresa groups is 12.05%. Therefore, the early progresa precincts have an advantage.

for precincts that had a lower voting share in 1994 the treatment significantly altered the outcome for those in the treatment group, increased the voting percentage by approximately 12.05 percent points. However, for precincts that had a higher percentage(75%) in voting in 1994, precincts in the treatment arm had an improved outcome in 2000 but only by 4.55 percent points.

it is clear that as the 1994 voting outcome percentage didn't significantly change as much for the higher percentage voter turnout in 1994 (75%) it significantly improved the outcome in percentage of votes in 2000 for the precincts that had a lower number of voting in 1994(25%).

```
# For a 75 % 1994 voter turnout in Early CCT program recipient villages
t2000_75_earl <- 24.24 + 0.65 * 75
# For a 75 % 1994 voter turnout in Late CCT program recipient villages
t2000_75_late <- 8.44 + 0.8 * 75

# For a 25 % 1994 voter turnout in Early CCT program recipient villages
t2000_25_earl <- 24.24 + 0.65 * 25
# For a 25 % 1994 voter turnout in Late CCT program recipient villages
t2000_25_late <- 8.44 + 0.80 * 25

t2000_75_earl - t2000_75_late

## [1] 4.55

t2000_25_earl - t2000_25_late

## [1] 12.05
```