PSTAT 120C Presidential Poll Project

Kris Hao

2023-03-05

Question 1: For the presidential poll in 2016, explore the poll in Michigan, Georgia and North Carolina from August 1, 2016 to November 2 in 2016. Use the data to answer the following questions.

```
# a. Who is ahead in each of these three states? What is the percentage
# difference for each state?
index_michigan_2016 = polls_data_2016$state=='Michigan'
michigan_total_michigan_2016 <- polls_data_2016[index_michigan_2016,]
polls_data_2016_enddate = mdy(polls_data_2016$enddate[polls_data_2016$state=="Michigan"])
polls_data_2016_startdate = mdy(polls_data_2016$startdate[polls_data_2016$state=="Michigan"])
michigan_total_enddate_2016 <- michigan_total_michigan_2016[polls_data_2016_enddate <= "2016-11-02",]
michigan_total_2016 <- michigan_total_enddate_2016[polls_data_2016_startdate>="2016-08-01",]

total_michigan_clinton_2016 <- sum(michigan_total_2016$total.clinton, na.rm=T); total_michigan_clinton_
```

```
## [1] 66664
```

```
total_michigan_trump_2016 <- sum(michigan_total_2016$total.trump, na.rm=T); total_michigan_trump_2016
```

```
## [1] 62240.44
```

```
percen_dif_michigan_2016 <- ((total_michigan_clinton_2016 - total_michigan_trump_2016)/
                      (total_michigan_clinton_2016 + total_michigan_trump_2016));percen_dif_michigan_20
```

```
## [1] 0.03431655
```

```
# In Michigan Clinton is ahead with 66,664 total votes whereas Trump received
# 62,240.44 total votes from August 1, 2016 to November 2, 2016. The percentage
# difference for Michigan is 0.03431655, meaning Clinton is ahead of Trump in
# Michigan for 3.43% of the votes.

index_georgia_2016 = polls_data_2016$state=='Georgia'
georgia_total_georgia_2016 <- polls_data_2016[index_georgia_2016,]
polls_data_2016_enddate = mdy(polls_data_2016$enddate[polls_data_2016$state=="Georgia"])
polls_data_2016_startdate = mdy(polls_data_2016$startdate[polls_data_2016$state=="Georgia"])
georgia_total_enddate_2016 <- georgia_total_georgia_2016[polls_data_2016_enddate <= "2016-11-02",]
georgia_total_2016 <- georgia_total_enddate_2016[polls_data_2016_startdate>="2016-08-01",]

total_georgia_clinton_2016 <- sum(georgia_total_2016$total.clinton, na.rm=T); total_georgia_clinton_201
```

```
## [1] 62739.56
```

```
total_georgia_trump_2016 <- sum(georgia_total_2016$total.trump, na.rm=T); total_georgia_trump_2016
```

```
## [1] 71390.73
```

```
percen_dif_georgia_2016 <- ((total_georgia_clinton_2016 - total_georgia_trump_2016)/
                      (total_georgia_clinton_2016 + total_georgia_trump_2016));percen_dif_georgia_2016
```

```
## [1] -0.06449828
```

```
# In Georgia Trump is ahead with 71,390.73 total votes whereas Clinton received
# 62,739.56 total votes. The percentage difference for Georgia is -0.06449828,
# meaning Trump is ahead of Clinton in Georgia by 6.44% of the votes.

index_NC_2016 = polls_data_2016$state=='North Carolina'
NC_total_NC_2016 <- polls_data_2016[index_NC_2016,]
polls_data_2016_enddate = mdy(polls_data_2016$enddate[polls_data_2016$state=="North Carolina"])
polls_data_2016_startdate = mdy(polls_data_2016$startdate[polls_data_2016$state=="North Carolina"])
NC_total_enddate_2016 <- NC_total_NC_2016[polls_data_2016_enddate <= "2016-11-02",]
NC_total_2016 <- NC_total_enddate_2016[polls_data_2016_startdate>="2016-08-01",]

total_NC_clinton_2016 <- sum(NC_total_2016$total.clinton, na.rm=T); total_NC_clinton_2016
```

```
## [1] 103216.2
```

```
total_NC_trump_2016 <- sum(NC_total_2016$total.trump, na.rm=T); total_NC_trump_2016
```

```
## [1] 103578.1
```

```
percen_dif_NC_2016 <- ((total_NC_clinton_2016 - total_NC_trump_2016)/
                      (total_NC_clinton_2016 + total_NC_trump_2016));percen_dif_NC_2016
```

```
## [1] -0.001750234
```

```
# In North Carolina Trump is ahead with 103,578.1 total votes whereas Clinton
# received 103,216.2 total votes. The percentage difference for North Carolina
# is -0.001750234, meaning Trump is ahead of Clinton in North Carolina by .175%
# of the votes.
```

b. Run a paired t test of the counts in polls for each of the state. Who is in favor of winning based on the test? Is the test significant? Is there potential problem?

We let d be the difference between the number of votes for Clinton and Trump per poll with a level of significance of 0.05.

$$H_o : d = 0$$
$$H_a : d > 0$$

```
t.test(michigan_total_2016$total.clinton,
       michigan_total_2016$total.trump, paired=T, alternative='greater')
```

```
##
##  Paired t-test
##
## data:  michigan_total_2016$total.clinton and michigan_total_2016$total.trump
## t = 10.36, df = 170, p-value < 2.2e-16
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  21.73896      Inf
## sample estimates:
## mean difference
##        25.86875
```

```
# Based on the test, since the p value is < 2.2e-16 and much less
# than an acceptable level of significance 0.05, we reject the null hypothesis
# and conclude the true mean differences between Trump and Clinton's total votes
# is greater than 0. Therefore there is significant test evidence that Clinton
# is favored in winning against Trump for Michigan.
```

$$H_o : d = 0$$
$$H_a : d < 0$$

```
t.test(georgia_total_2016$total.clinton,
       georgia_total_2016$total.trump, paired=T, alternative='less')
```

```
##
##  Paired t-test
##
## data:  georgia_total_2016$total.clinton and georgia_total_2016$total.trump
## t = -19.242, df = 167, p-value < 2.2e-16
## alternative hypothesis: true mean difference is less than 0
## 95 percent confidence interval:
##       -Inf -47.06848
## sample estimates:
## mean difference
##       -51.49507
```

```
# Based on the test, since the p value is < 2.2e-16 and much less than an acceptable
# level of significance 0.05, we reject the null hypothesis and conclude
# the true mean differences between Trump and Clinton's total votes is less
# than 0. Therefore there is significant test evidence that Trump
# is favored in winning against Clinton for Georgia.
```

$$H_o : d = 0$$
$$H_a : d < 0$$

```
t.test(NC_total_2016$total.clinton,
       NC_total_2016$total.trump, paired=T, alternative='less')
```

```
##
##  Paired t-test
##
## data:  NC_total_2016$total.clinton and NC_total_2016$total.trump
## t = -0.64049, df = 275, p-value = 0.2612
## alternative hypothesis: true mean difference is less than 0
## 95 percent confidence interval:
##      -Inf 2.067764
## sample estimates:
## mean difference
##       -1.311372
```

```
# Based on the test, since the p value of 0.2612 is  greater than an acceptable
# level of significance 0.05, we fail to reject the null hypothesis that the
# true mean differences between Trump and Clinton's total votes is 0. There is
# significant test evidence that Trump and Clinton are equally favored in
# winning for North Carolina.
```

A potential problem with using the paired t-test is although Trump and Clinton's paired nature is designed from the same subject as pairs of observations, we cannot fully state that the voters' polls are paired where voters are matched on related variables, as they are independent polls. Therefore voter polls might not be considered paired and the t-test might not be suitable.

c. Run a Wilcoxon signed-rank test of the counts in polls for each of the state. Who is in favor of winning based on the test? Is the test significant? Is there potential problem of the test?

We let p be the proportion of votes in favor of Clinton over Trump per poll with a level of significance of 0.05.

$$H_o : p = \frac{1}{2}$$
$$H_a : p > \frac{1}{2}$$

```
wilcox.test(michigan_total_2016$total.clinton, michigan_total_2016$total.trump,
            alternative = "greater")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  michigan_total_2016$total.clinton and michigan_total_2016$total.trump
## W = 16609, p-value = 0.01483
## alternative hypothesis: true location shift is greater than 0
```

```
# Based on the test, since the p value of 0.01483 is less than an acceptable
# level of significance 0.05, we reject the null hypothesis that the true
# location shift is 0. There is significant test evidence to suggest that the
# true location shift is greater than 0 and that Hillary is in favor of winning
# in Michigan.
```

$$H_o : p = \frac{1}{2}$$
$$H_a : p < \frac{1}{2}$$

```
wilcox.test(georgia_total_2016$total.clinton, georgia_total_2016$total.trump,
            alternative = "less")
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  georgia_total_2016$total.clinton and georgia_total_2016$total.trump
## W = 9869, p-value = 9.43e-07
## alternative hypothesis: true location shift is less than 0
```

```
# Based on the test, since the p value is 9.43e-07 and much less than an
# acceptable level of significance 0.05, we reject the null hypothesis that the true
# location shift is 0. There is significant test evidence to suggest that
# the true location shift is less than 0 and that Trump is in favor of winning
# in Georgia.
```

$$H_o : p = \frac{1}{2}$$
$$H_a : p < \frac{1}{2}$$

```
wilcox.test(NC_total_2016$total.clinton, NC_total_2016$total.trump,
            alternative = "less")
```

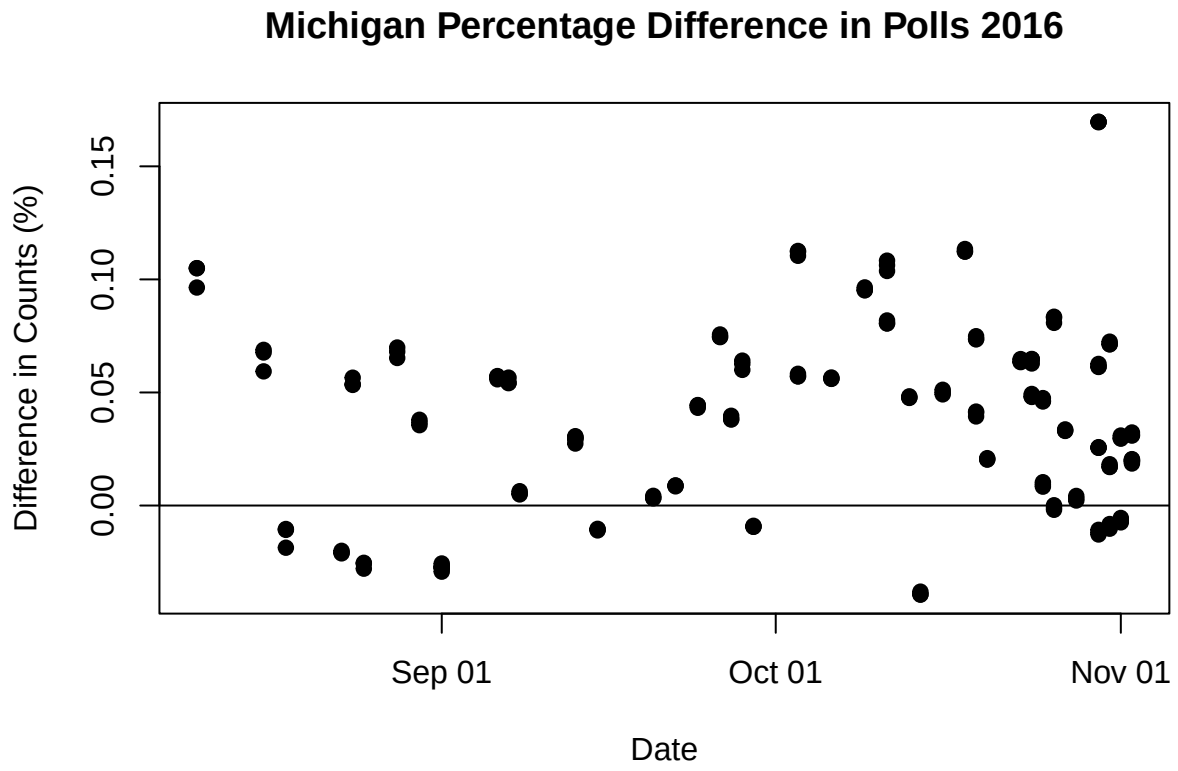
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  NC_total_2016$total.clinton and NC_total_2016$total.trump
## W = 37150, p-value = 0.3084
## alternative hypothesis: true location shift is less than 0
```

```
# Based on the test, since the p value is 0.3084 and more than an acceptable
# level of significance 0.05, we fail to reject the null hypothesis that the true
# location shift is 0. There is significant test evidence to suggest that
# the true location shift is 0 and that Trump and Clinton are equally in favor
# of winning in Georgia.
```

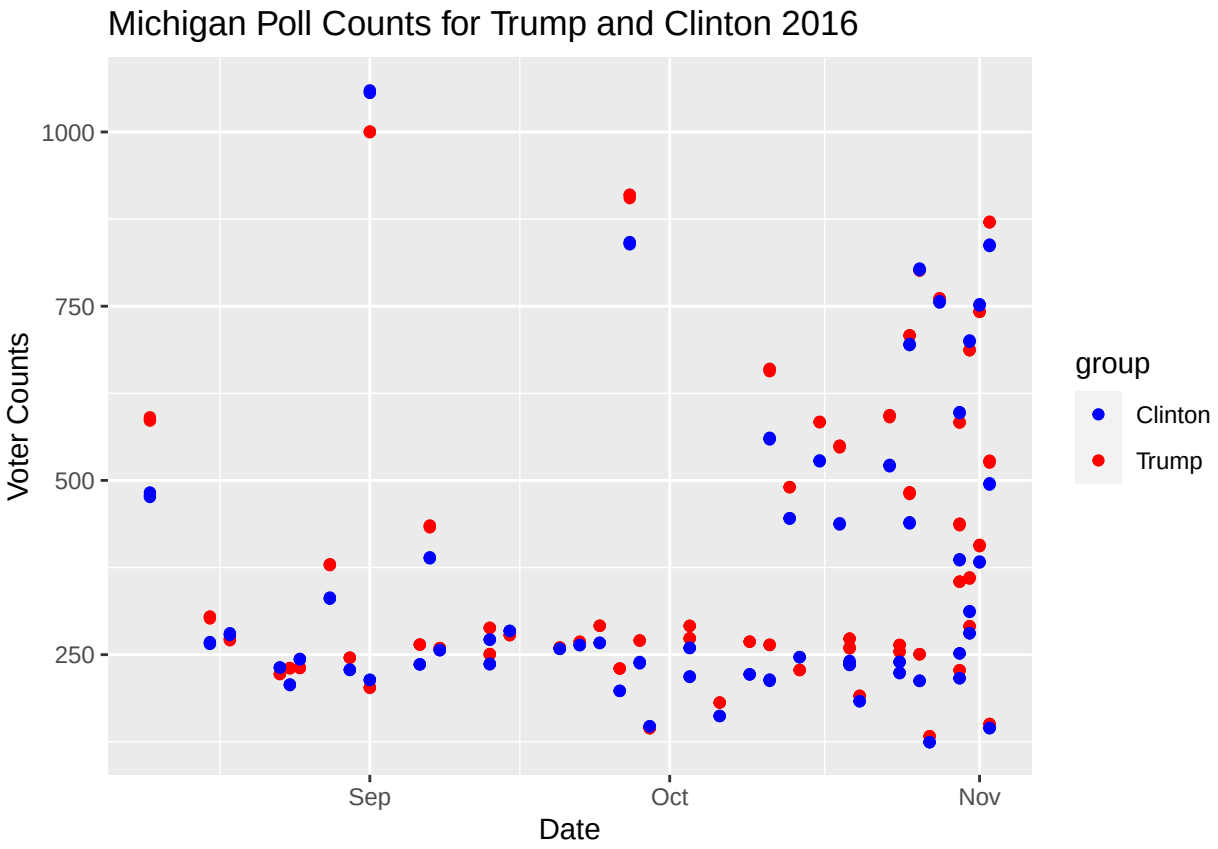
A potential problem with using the Wilcoxon signed-rank test is similar to the t-test in how the voters are unpaired on related variables as they are independent poll decisions. Since the Wilcoxon signed-rank test assesses the location shift between the paired differences, if the voter poll observations are not paired this test might not be as accurate.

```
    # d. Fit a linear model of the percentage difference with respect to date of
    # the polls separately for each of these states. Show a plot of the
    # observations of the polls, fitted values and confidence interval of the
    # fitted line for each of these state. From the linear model and
```

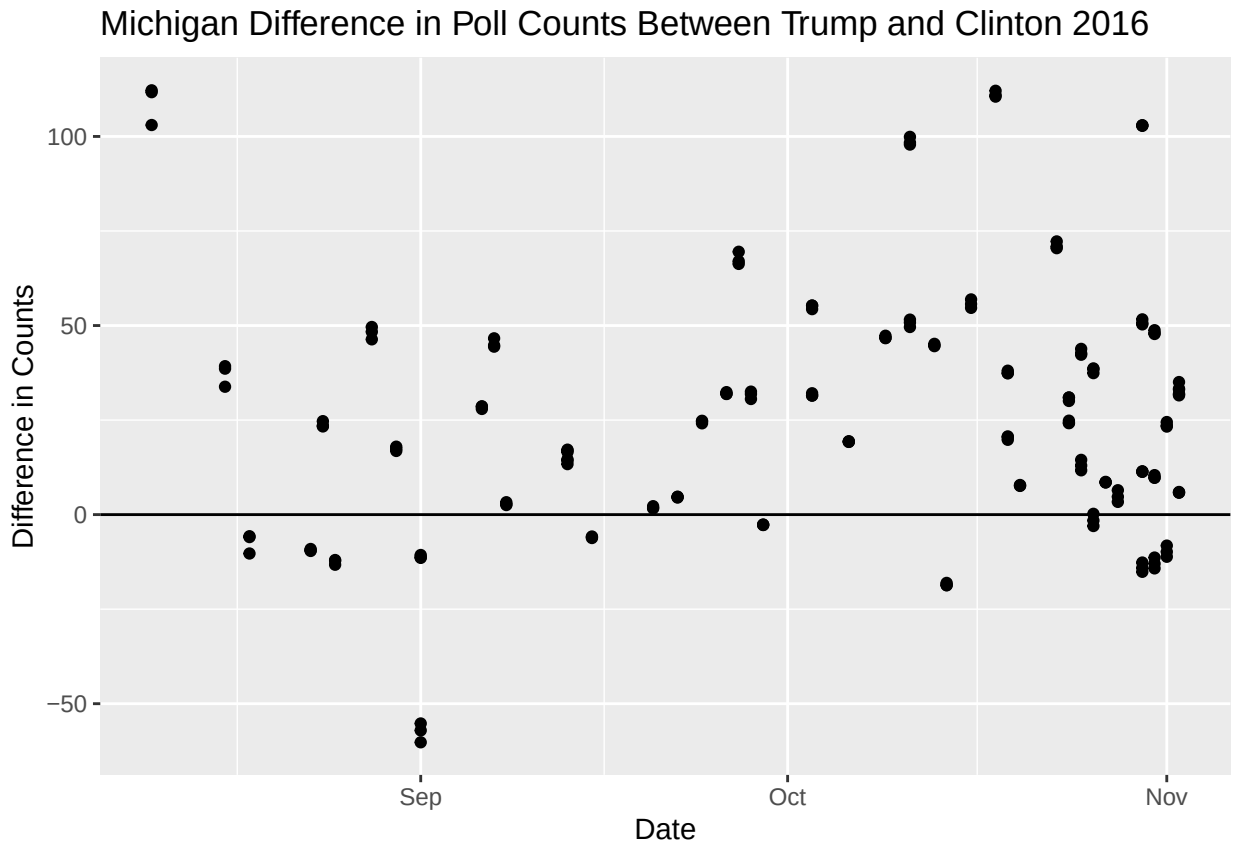
```r
    # observations, which state may have the closest election (in terms of
    # percentage difference)?

# MICHIGAN
    # Percentage difference no ggplot method
date_michigan_2016=mdy(michigan_total_2016$enddate)

percentage_diff_michigan_2016 = (michigan_total_2016$total.clinton - michigan_total_2016$total.trump)/(

plot((na.omit(date_michigan_2016)), (na.omit(percentage_diff_michigan_2016)), col='black',
     pch=19, type='p', xlab='Date', ylab='Difference in Counts (%)',
     main='Michigan Percentage Difference in Polls 2016', );abline(a=0,b=0)
```



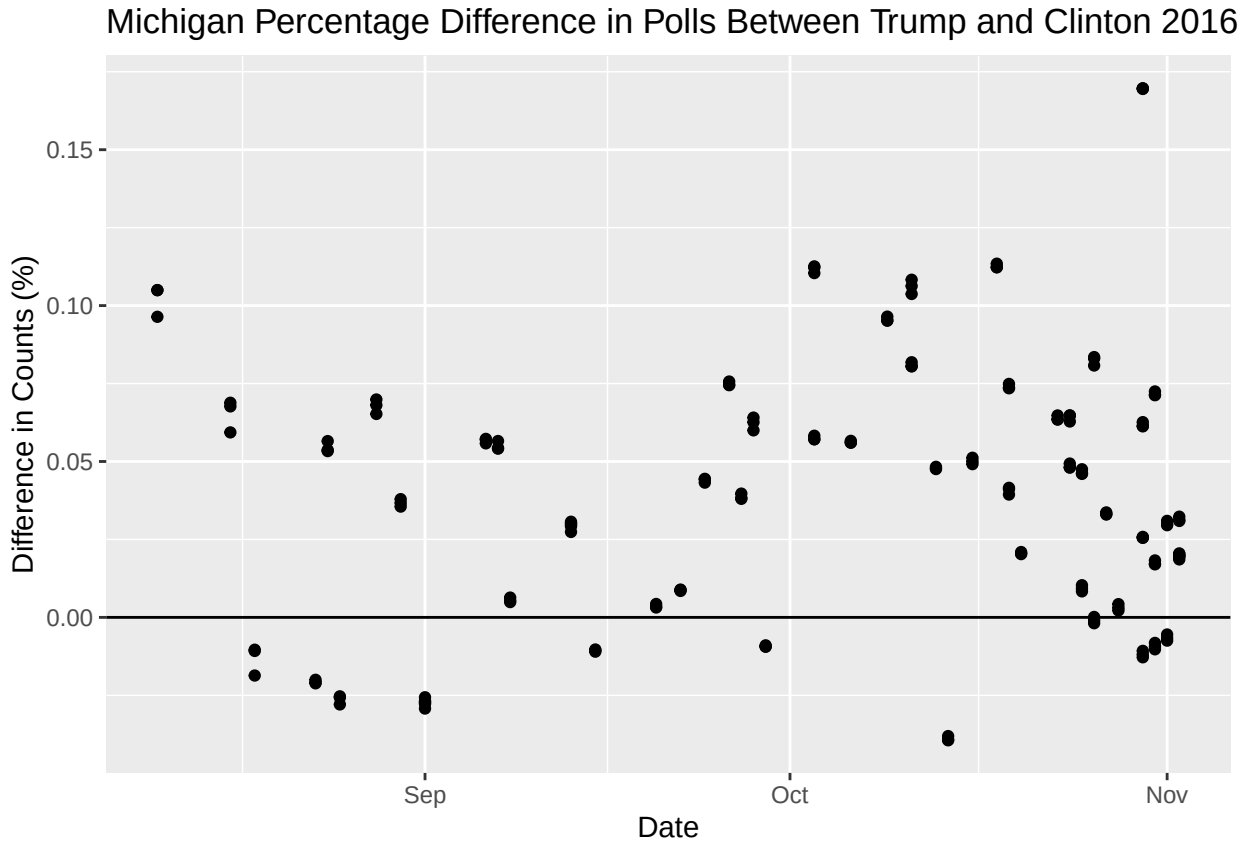

```r
    # Plot observations of polls
counts_michigan_2016 <- data.frame(data_date = c(date_michigan_2016, date_michigan_2016),
  counts = c(michigan_total_2016$total.clinton, michigan_total_2016$total.trump),
  group = c(rep('Trump', length(date_michigan_2016)), rep('Clinton',length(date_michigan_2016))))

ggplot(data=counts_michigan_2016, aes(x=data_date, y=counts, col=group)) +
  geom_point() +
  scale_color_manual(values = c("blue","red")) + labs(x='Date') +
  labs(y='Voter Counts') + labs(title='Michigan Poll Counts for Trump and Clinton 2016')
```




```
counts_michigan_separate_2016 = data.frame(data_date = date_michigan_2016,
  Trump = michigan_total_2016$total.trump,
  Clinton = michigan_total_2016$total.clinton)

ggplot(data = counts_michigan_separate_2016, aes(x=data_date, y=Clinton-Trump)) + geom_point() + xlab('
```




```
    # Percentage difference ggplot method
ggplot(data = counts_michigan_separate_2016, aes(x = data_date, y=(Clinton-Trump)/(Clinton+Trump))) + g
```




```
    # Linear model of the percentage difference with respect to date of the polls
counts_michigan_for_lm_2016 = data.frame(data_date = date_michigan_2016,
      percentage_diff = ((michigan_total_2016$total.clinton - michigan_total_2016$total.trump) / (michig

lm_model_michigan_2016 = lm(percentage_diff ~ (data_date), data = counts_michigan_for_lm_2016); lm_model
```

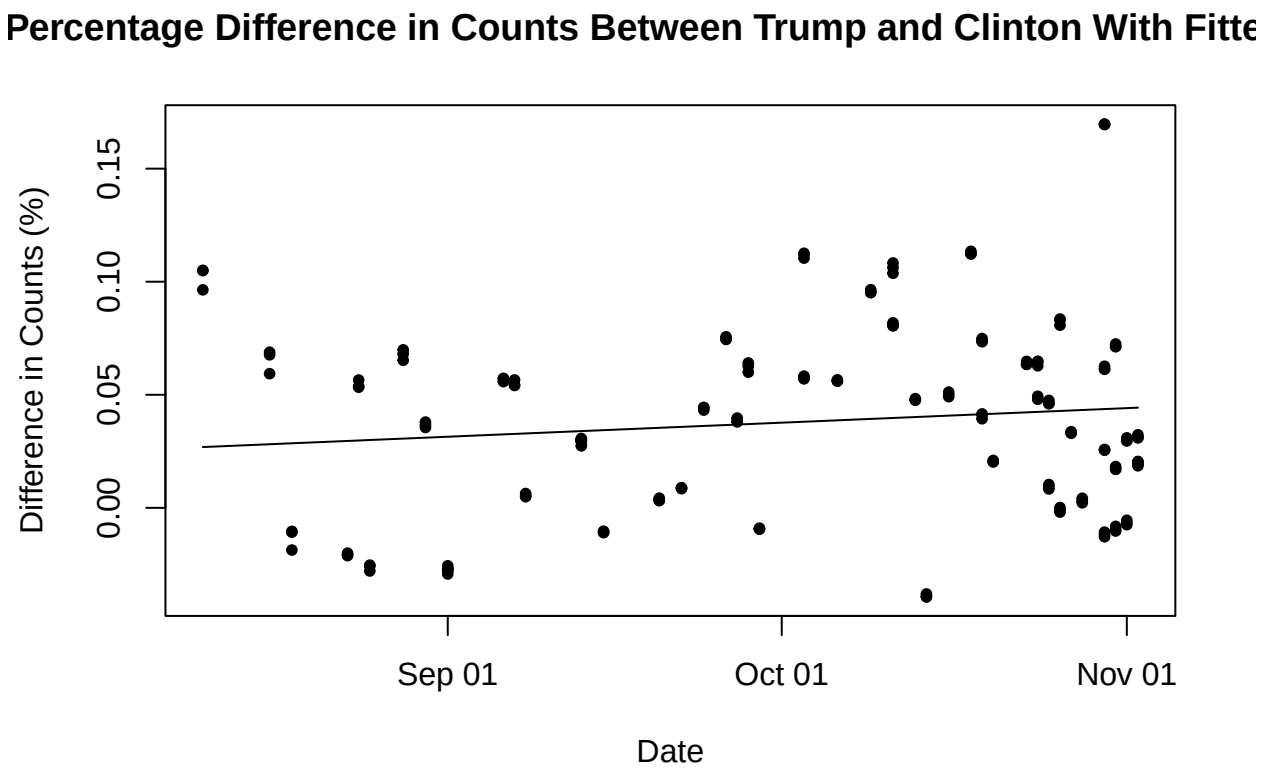
```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_michigan_for_lm_2016)
##
## Coefficients:
## (Intercept)    data_date
##  -3.5019113    0.0002073
```

```
summary(lm_model_michigan_2016)
```

```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_michigan_for_lm_2016)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.079786 -0.032089  0.001431  0.024643  0.126006
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.5019113  2.1601115  -1.621    0.107
## data_date    0.0002073  0.0001265   1.639    0.103
##
## Residual standard error: 0.04154 on 169 degrees of freedom
##   (45 observations deleted due to missingness)
## Multiple R-squared:  0.01565,    Adjusted R-squared:  0.009821
## F-statistic: 2.686 on 1 and 169 DF,  p-value: 0.1031
```

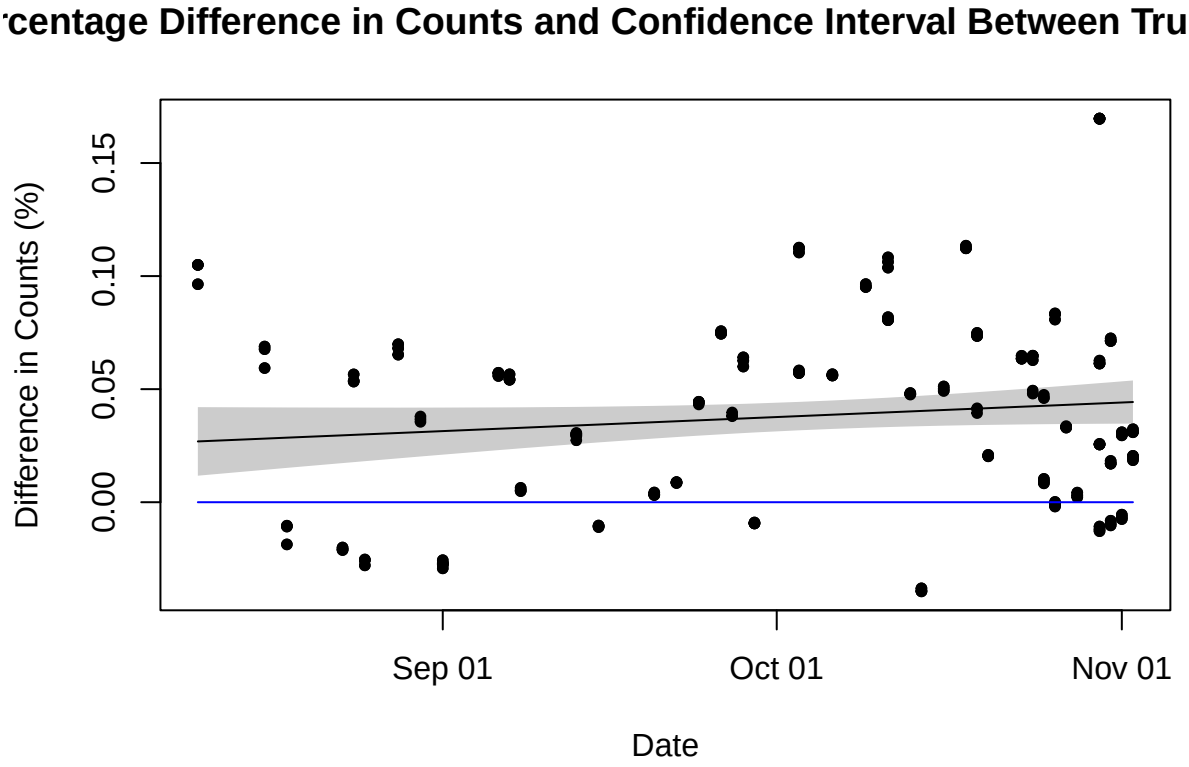
```
    # Plot fitted values of the fitted line
counts_michigan_2016 <- data.frame(data_date = c(date_michigan_2016, date_michigan_2016),
  counts = c(michigan_total_2016$total.clinton, michigan_total_2016$total.trump),
  group = c(rep('Trump', length(date_michigan_2016)), rep('Clinton',length(date_michigan_2016))))

plot(counts_michigan_for_lm_2016$data_date, counts_michigan_for_lm_2016$percentage_diff,
     col='black', pch=20, type='p', xlab='Date', ylab='Difference in Counts (%)',
     main='Michigan Percentage Difference in Counts Between Trump and Clinton With Fitted Values 2016')
col='black', pch=20, type='l', xlab='Date', ylab='Difference in Counts (%)',
main='Michigan')
```




```
    # Plot the confidence interval of the fitted line
fitted_CI_michigan_2016 = predict(lm_model_michigan_2016,
                            newdata = counts_michigan_for_lm_2016,
                            interval = "confidence", level = 0.95)
summary(fitted_CI_michigan_2016)
```

```
##       fit                lwr                upr
##  Min.   :0.02689   Min.   :0.01174   Min.   :0.04184
##  1st Qu.:0.03394   1st Qu.:0.02572   1st Qu.:0.04215
##  Median :0.03974   Median :0.03325   Median :0.04623
##  Mean   :0.03834   Mean   :0.02970   Mean   :0.04698
##  3rd Qu.:0.04285   3rd Qu.:0.03455   3rd Qu.:0.05115
##  Max.   :0.04430   Max.   :0.03477   Max.   :0.05383
##  NA's   :45        NA's   :45        NA's   :45
```

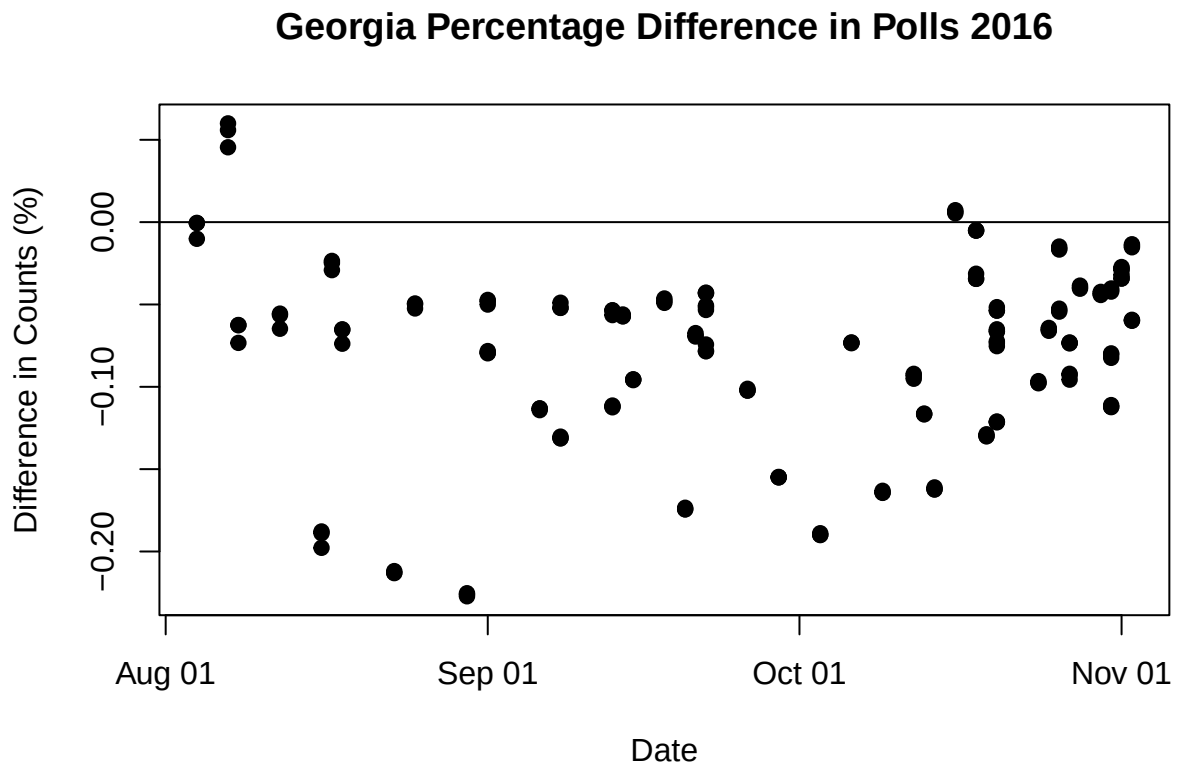
```
plot(counts_michigan_for_lm_2016$data_date, counts_michigan_for_lm_2016$percentage_diff,
     col='black', pch=20, type='p', xlab = 'Date', ylab = 'Difference in Counts (%)',
     main = 'Michigan Percentage Difference in Counts and Confidence Interval Between Trump and Clinton
```



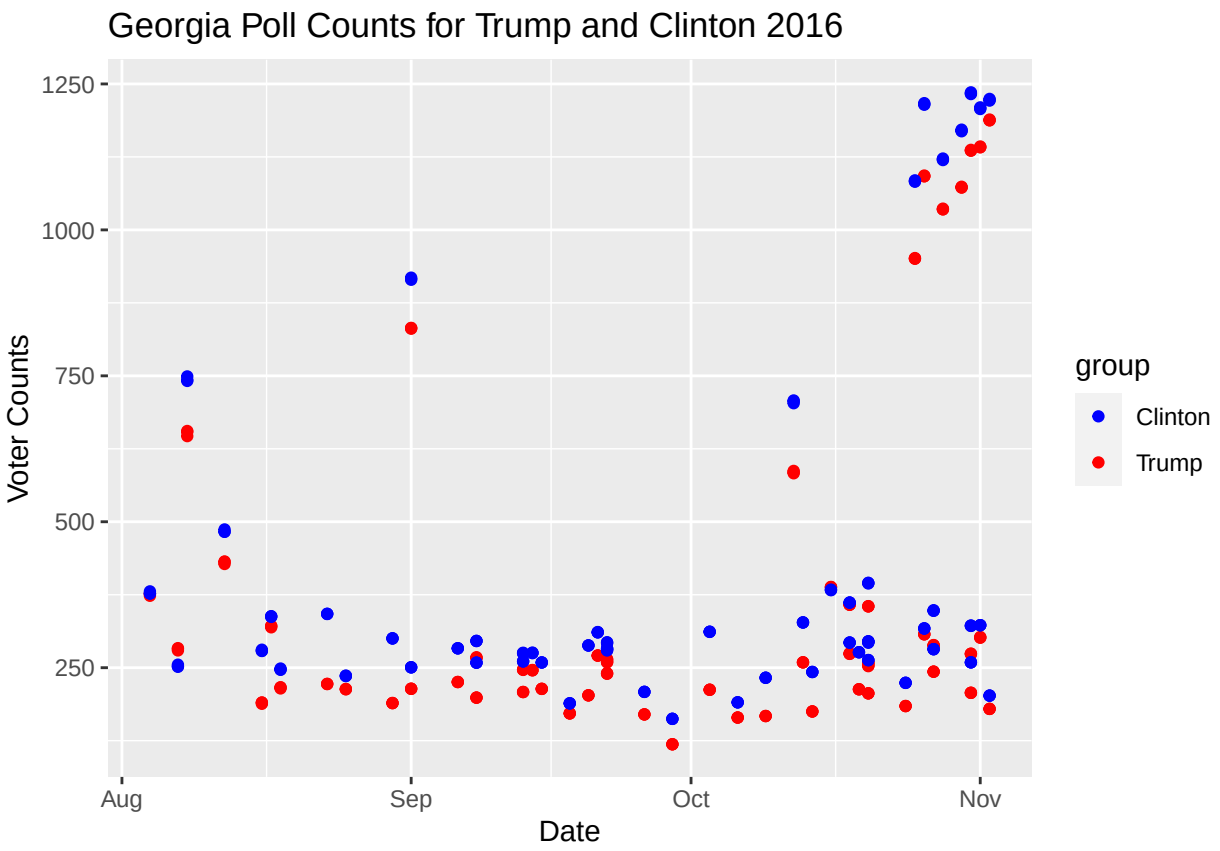
From our plot of fitted values we see evidence of a trend in difference in Michigan counts % and date. We expect that early polls do not have as much impact as recent polls as most polls are concentrated on more recent months. From our linear model of the percentage difference with respect to date of the polls for Michigan we see a p value of 0.1031 which is more than the acceptable level of significance 0.05, meaning we don't have sufficient evidence to reject the null hypothesis and conclude the percentage difference in counts for Michigan is not affected by dates.

From our plot with a confidence interval for the fitted line the confidence interval doesn't contain 0, as the values are positive and above 0 indicating a positive difference in counts % for Trump and Clinton. This means with repeated trails we are expecting a difference in Trump and Clinton's Michigan Difference in Count % with respect to dates of the polls, and this trial is most likely not indicative of being the closest election with the least difference in percentage difference.

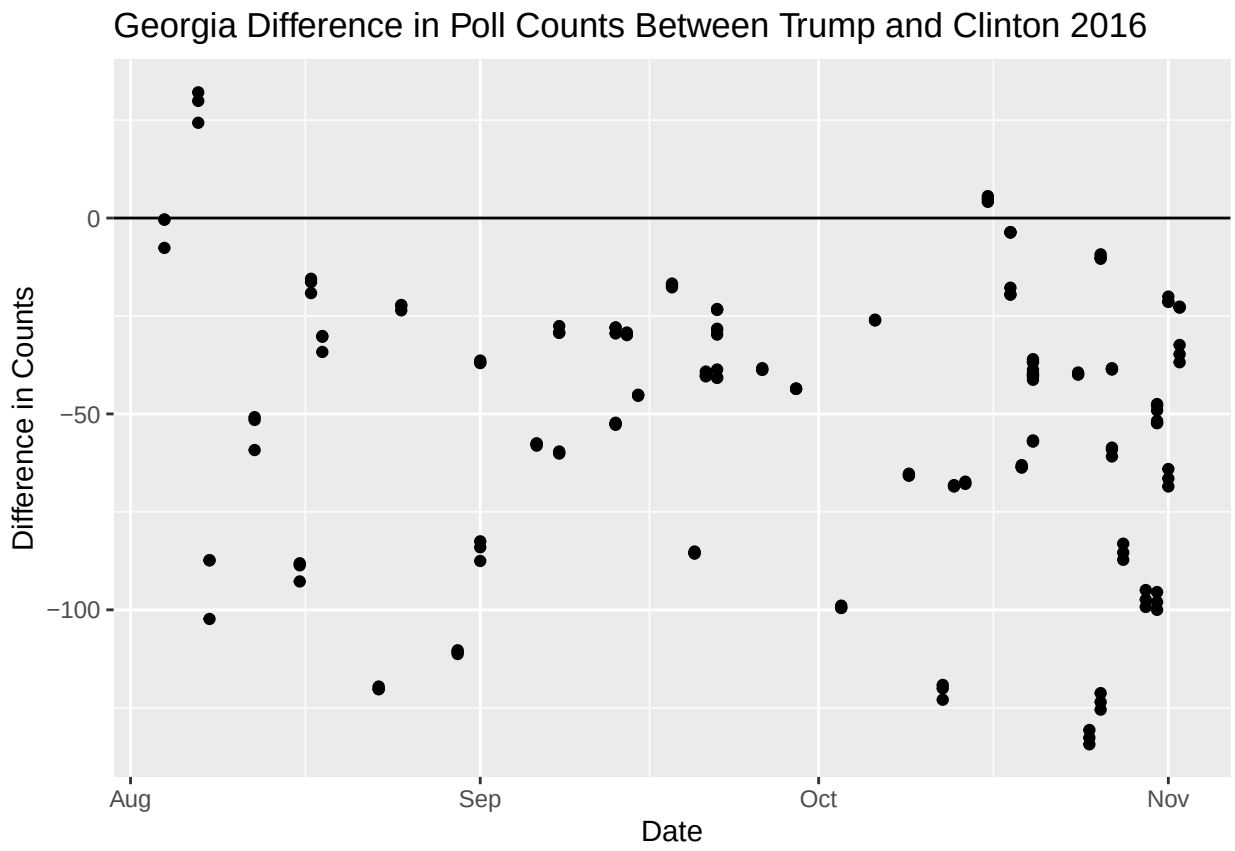
```r
  # d. continued

# GEORGIA
    # Percentage difference no ggplot method
date_georgia_2016=mdy(georgia_total_2016$enddate)

percentage_diff_georgia_2016 = (georgia_total_2016$total.clinton - georgia_total_2016$total.trump)/(geo

plot((na.omit(date_georgia_2016)), (na.omit(percentage_diff_georgia_2016)), col='black',
     pch=19, type='p', xlab='Date', ylab='Difference in Counts (%)',
     main='Georgia Percentage Difference in Polls 2016', );abline(a=0,b=0)
```



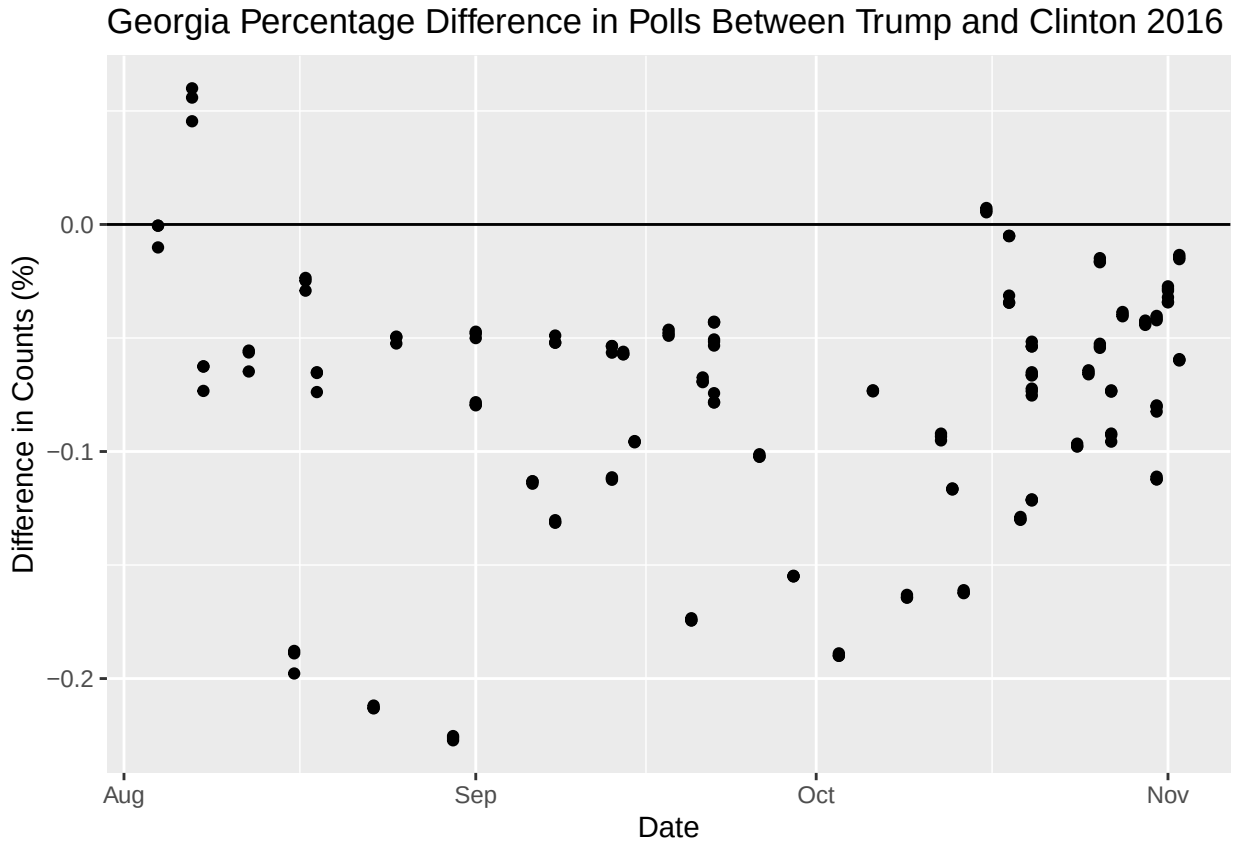

```r
    # Plot observations of polls
counts_georgia_2016 <- data.frame(data_date = c(date_georgia_2016, date_georgia_2016),
  counts = c(georgia_total_2016$total.clinton, georgia_total_2016$total.trump),
  group = c(rep('Trump', length(date_georgia_2016)), rep('Clinton',length(date_georgia_2016))))

ggplot(data=counts_georgia_2016, aes(x=data_date, y=counts, col=group)) +
  geom_point() +
  scale_color_manual(values = c("blue","red")) + labs(x='Date') +
  labs(y='Voter Counts') + labs(title='Georgia Poll Counts for Trump and Clinton 2016')
```




```
counts_georgia_separate_2016 = data.frame(data_date = date_georgia_2016,
  Trump = georgia_total_2016$total.trump,
  Clinton = georgia_total_2016$total.clinton)

ggplot(data = counts_georgia_separate_2016, aes(x=data_date, y=Clinton-Trump)) + geom_point() + xlab('Da
```




```
    # Percentage difference ggplot method
ggplot(data = counts_georgia_separate_2016, aes(x = data_date, y=(Clinton-Trump)/(Clinton+Trump))) + geo
```




```
    # Linear model of the percentage difference with respect to date of the polls
counts_georgia_for_lm_2016 = data.frame(data_date = date_georgia_2016,
      percentage_diff = ((georgia_total_2016$total.clinton - georgia_total_2016$total.trump) /
  (georgia_total_2016$total.clinton + georgia_total_2016$total.trump)))

lm_model_georgia_2016 = lm(percentage_diff ~ (data_date), data = counts_georgia_for_lm_2016); lm_model_
```

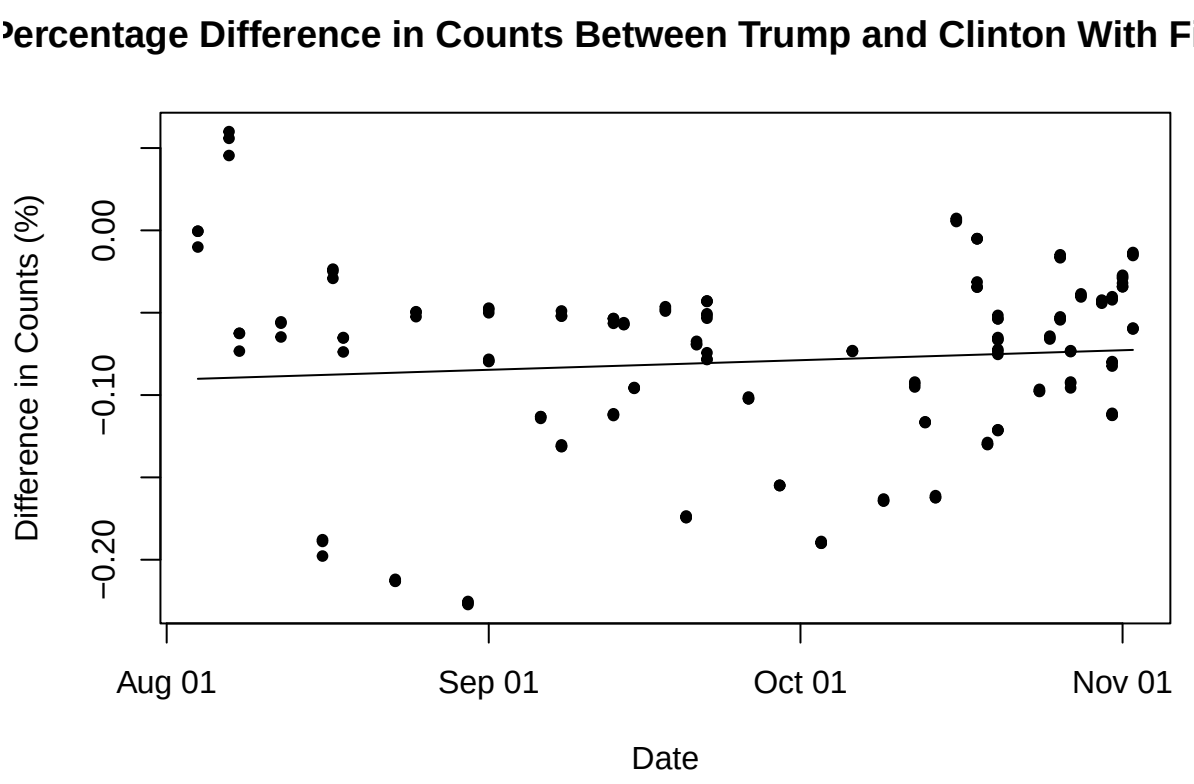
```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_georgia_for_lm_2016)
##
## Coefficients:
## (Intercept)    data_date
##  -3.4065220    0.0001949
```

```
summary(lm_model_georgia_2016)
```

```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_georgia_for_lm_2016)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14213 -0.02947  0.01069  0.03370  0.14951
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.4065220  2.6898032  -1.266    0.207
## data_date    0.0001949  0.0001575   1.237    0.218
##
## Residual standard error: 0.05589 on 166 degrees of freedom
##   (42 observations deleted due to missingness)
## Multiple R-squared:  0.009134,   Adjusted R-squared:  0.003165
## F-statistic:  1.53 on 1 and 166 DF,  p-value: 0.2178
```

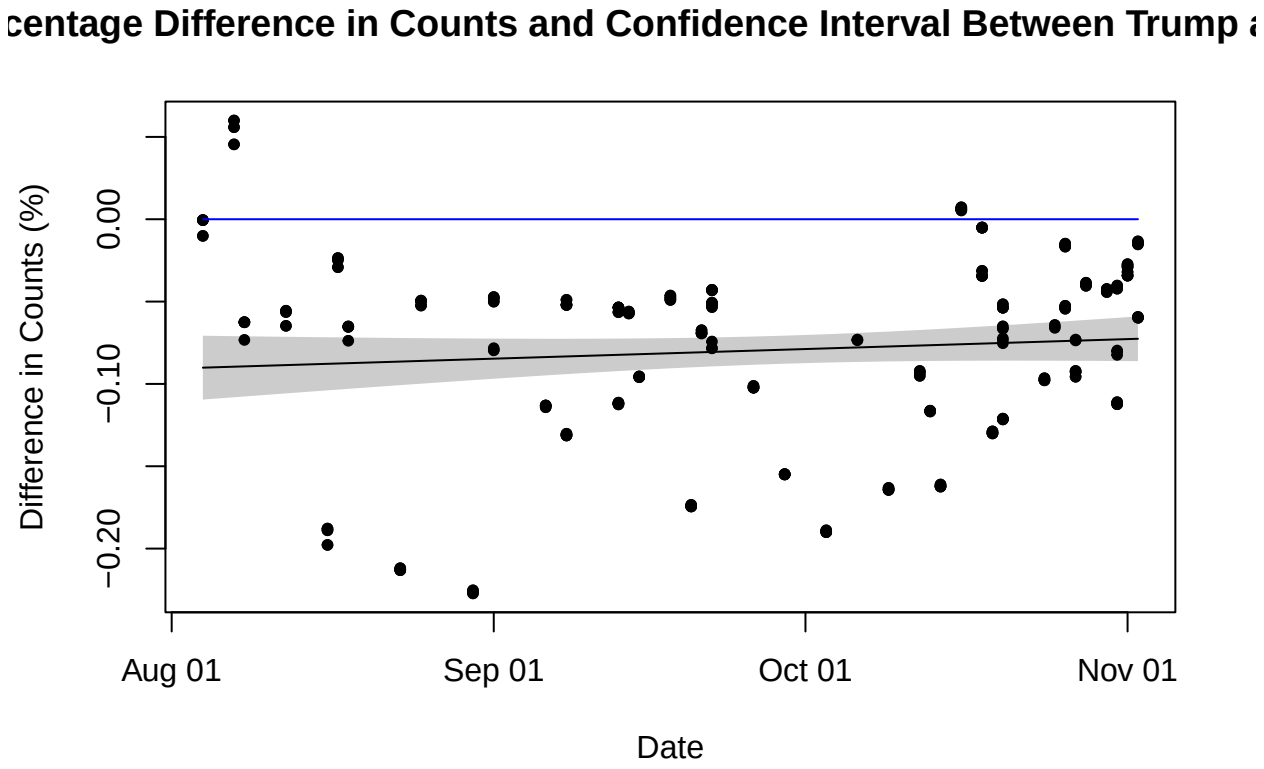
```
    # Plot fitted values of the fitted line
counts_georgia_2016 <- data.frame(data_date = c(date_georgia_2016, date_georgia_2016),
  counts = c(georgia_total_2016$total.clinton, georgia_total_2016$total.trump),
  group = c(rep('Trump', length(date_georgia_2016)), rep('Clinton',length(date_georgia_2016))))

plot(counts_georgia_for_lm_2016$data_date, counts_georgia_for_lm_2016$percentage_diff,
     col='black', pch=20, type='p', xlab='Date', ylab='Difference in Counts (%)',
     main='Georgia Percentage Difference in Counts Between Trump and Clinton With Fitted Values 2016');l
col='black', pch=20, type='l', xlab='Date', ylab='Difference in Counts (%)',
main='Georgia')
```




```
    # Plot the confidence interval of the fitted line
fitted_CI_georgia_2016 = predict(lm_model_georgia_2016,
                                 newdata = counts_georgia_for_lm_2016,
                                 interval = "confidence", level = 0.95)
summary(fitted_CI_georgia_2016)
```

```
##       fit                lwr                upr
##  Min.   :-0.09013   Min.   :-0.10951   Min.   :-0.07258
##  1st Qu.:-0.08331   1st Qu.:-0.09404   1st Qu.:-0.07193
##  Median :-0.07815   Median :-0.08684   Median :-0.06945
##  Mean   :-0.07922   Mean   :-0.09096   Mean   :-0.06748
##  3rd Qu.:-0.07430   3rd Qu.:-0.08597   3rd Qu.:-0.06271
##  Max.   :-0.07259   Max.   :-0.08586   Max.   :-0.05901
##  NA's   :42         NA's   :42         NA's   :42
```

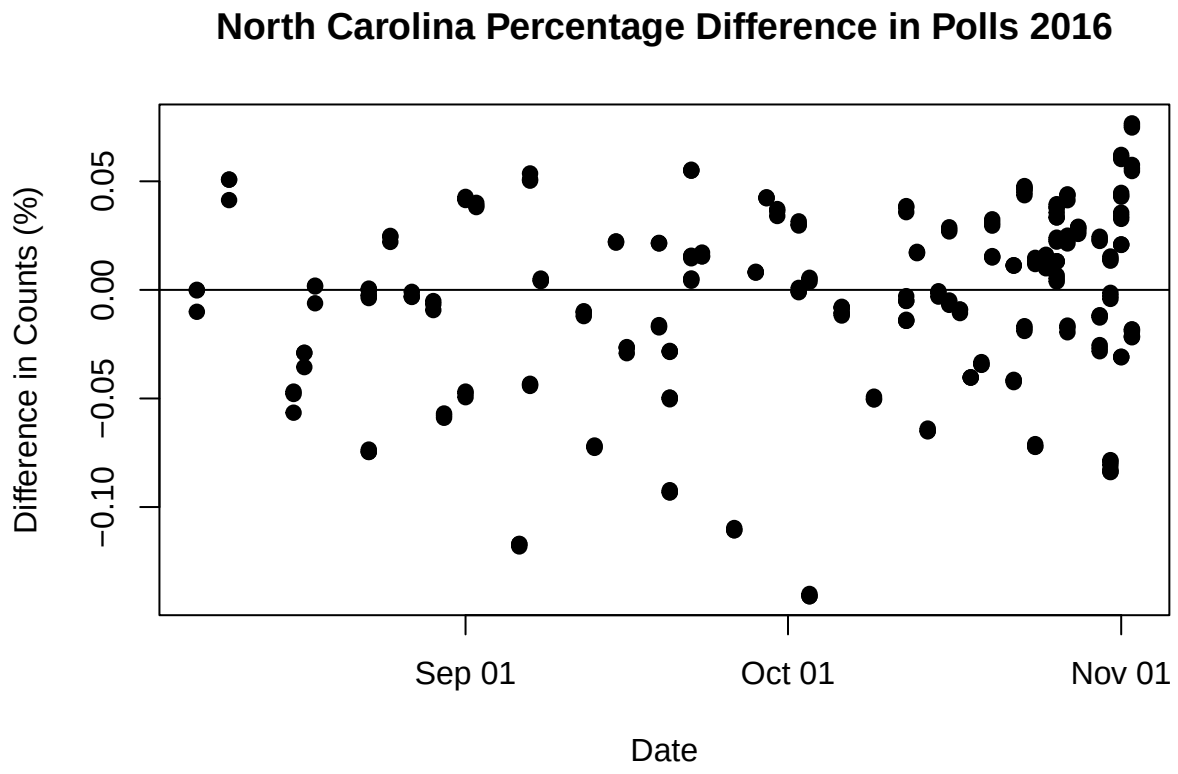
```r
plot(counts_georgia_for_lm_2016$data_date, counts_georgia_for_lm_2016$percentage_diff,
     col='black', pch=20, type='p', xlab = 'Date', ylab = 'Difference in Counts (%)',
     main = 'Georgia Percentage Difference in Counts and Confidence Interval Between Trump and Clinton
```



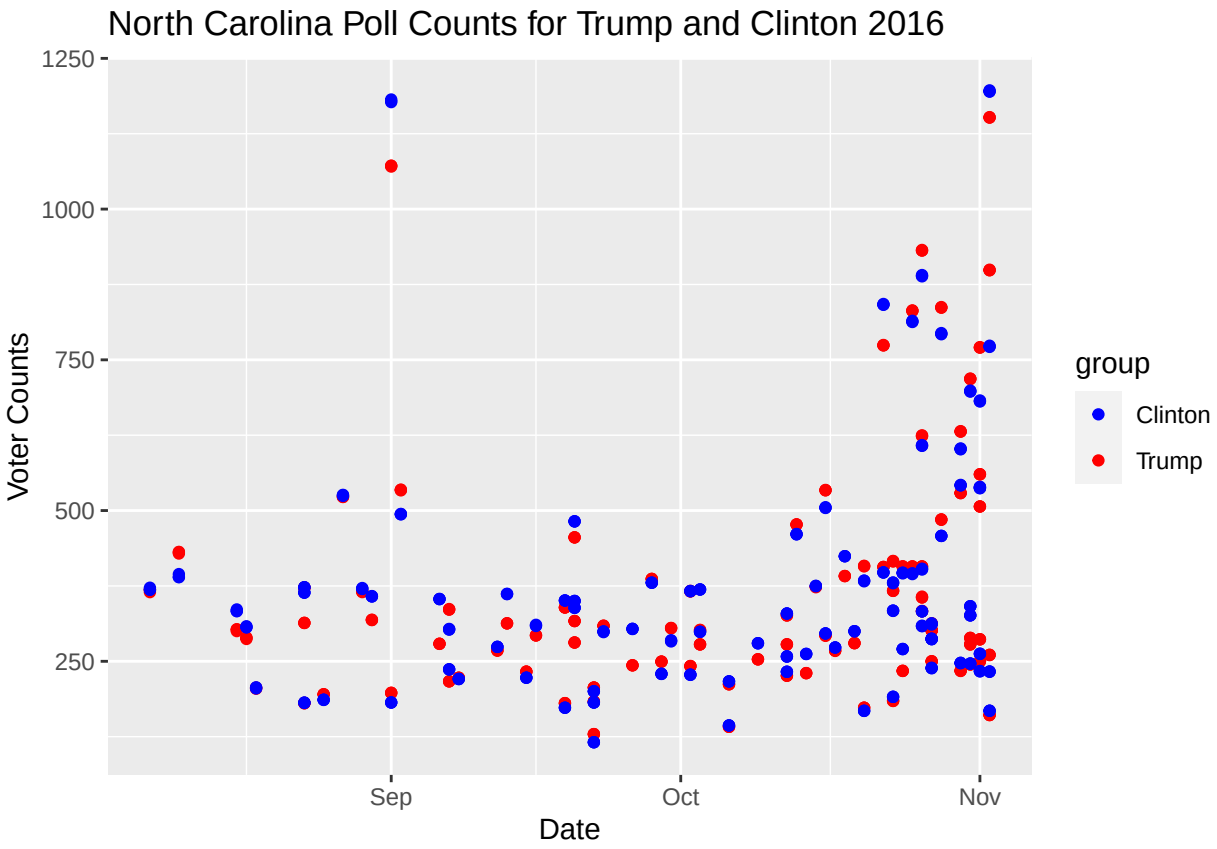
From our plot of fitted values we see evidence of a trend in difference in Georgia counts % and date. From our linear model of the percentage difference with respect to date of the polls for Georgia we see a p value of 0.2178 which is more than the acceptable level of significance 0.05, meaning we don't have sufficient evidence to reject the null hypothesis and conclude the percentage difference in counts for Georgia is not affected by dates.

From our plot with a confidence interval for the fitted line the confidence interval doesn't contain 0, as the values are negative and below 0 indicating a negative difference in counts % for Trump and Clinton. This means with repeated trails we are expecting a difference in Trump and Clinton's Michigan Difference in Count % with respect to dates of the polls, and this trial is most likely not indicative of being the closest election with the least difference in percentage difference.

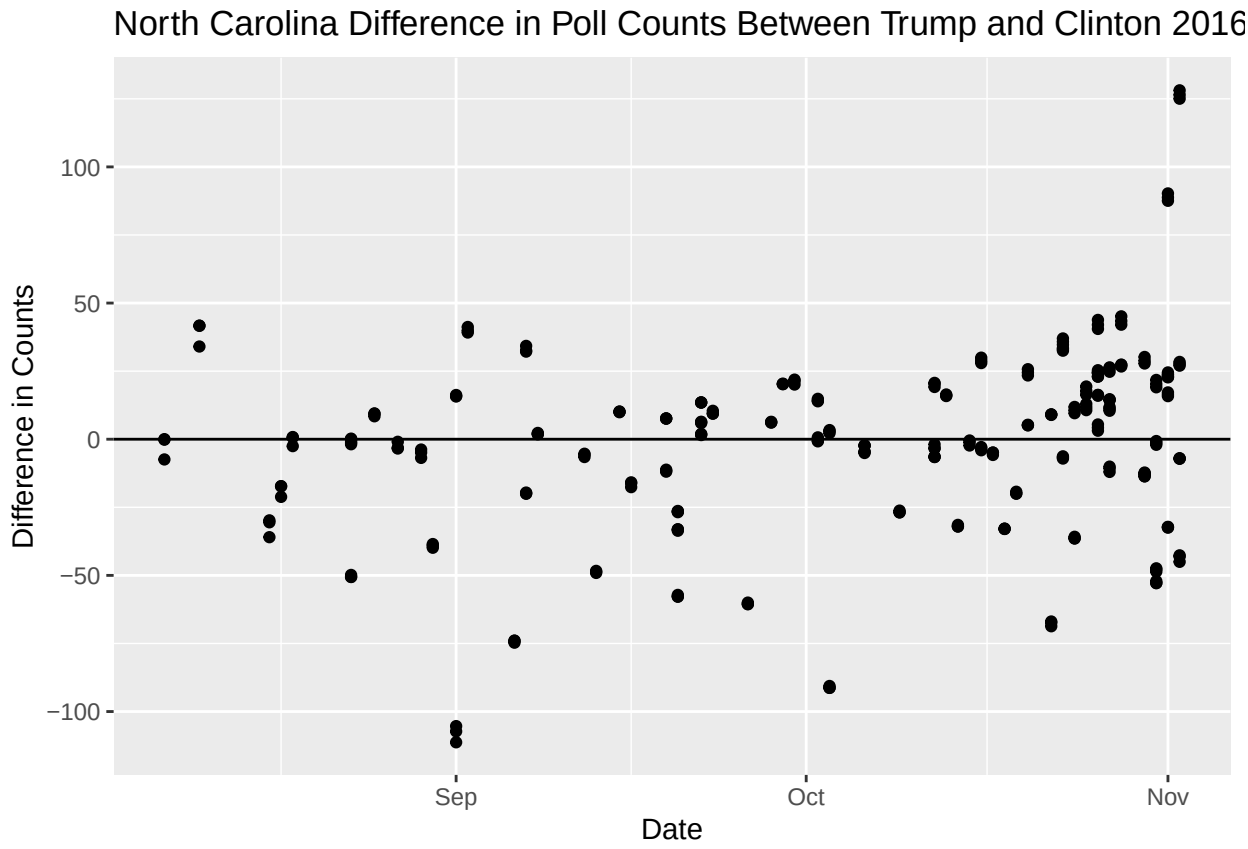
```r
  # d. continued

# NORTH CAROLINA
    # Percentage difference no ggplot method
date_NC_2016=mdy(NC_total_2016$enddate)

percentage_diff_NC_2016 = (NC_total_2016$total.clinton - NC_total_2016$total.trump)/(NC_total_2016$total

plot((na.omit(date_NC_2016)), (na.omit(percentage_diff_NC_2016)), col='black',
     pch=19, type='p', xlab='Date', ylab='Difference in Counts (%)',
     main='North Carolina Percentage Difference in Polls 2016', );abline(a=0,b=0)
```



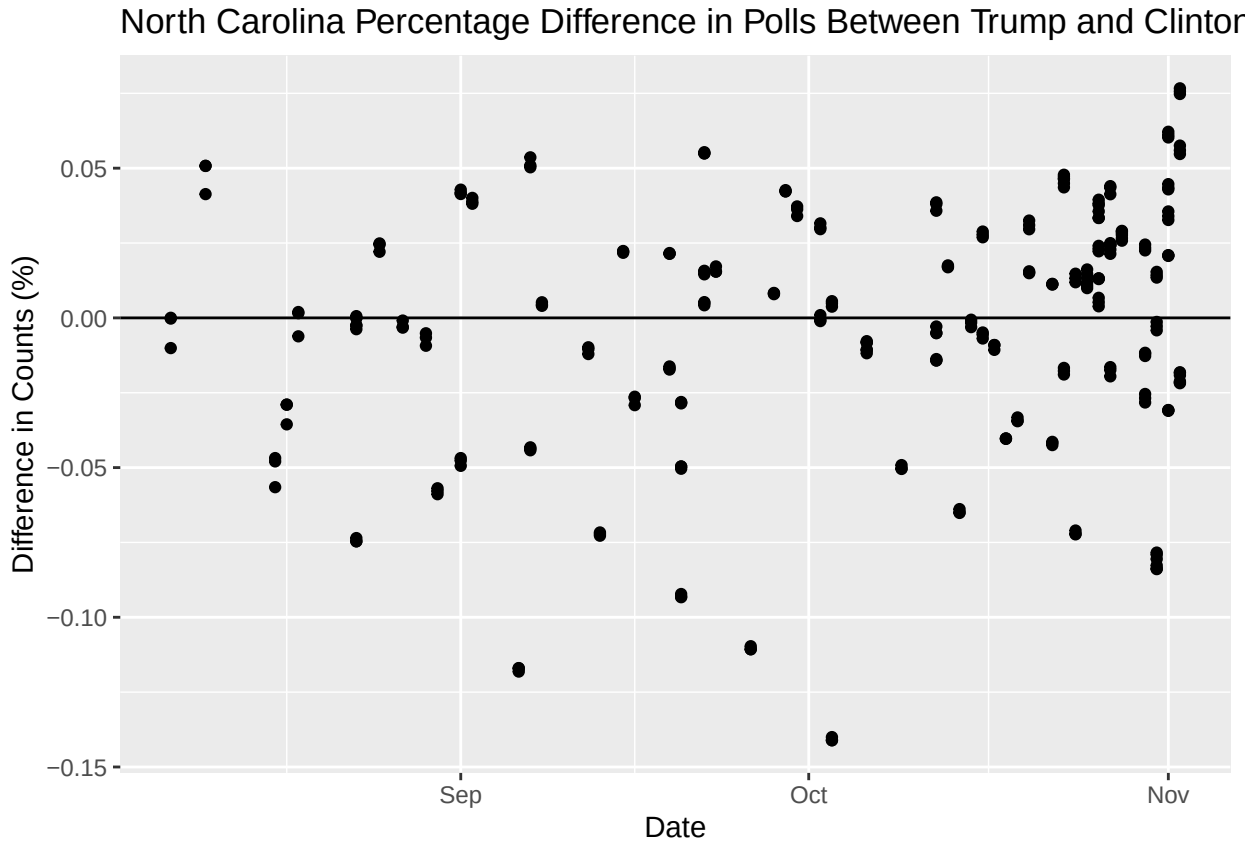

```r
    # Plot observations of polls
counts_NC_2016 <- data.frame(data_date = c(date_NC_2016, date_NC_2016),
  counts = c(NC_total_2016$total.clinton, NC_total_2016$total.trump),
  group = c(rep('Trump', length(date_NC_2016)), rep('Clinton',length(date_NC_2016))))

ggplot(data=counts_NC_2016, aes(x=data_date, y=counts, col=group)) +
  geom_point() +
  scale_color_manual(values = c("blue","red")) + labs(x='Date') +
  labs(y='Voter Counts') + labs(title='North Carolina Poll Counts for Trump and Clinton 2016')
```




```
counts_NC_separate_2016 = data.frame(data_date = date_NC_2016,
  Trump = NC_total_2016$total.trump,
  Clinton = NC_total_2016$total.clinton)

ggplot(data = counts_NC_separate_2016, aes(x=data_date, y=Clinton-Trump)) + geom_point() + xlab('Date')
```




```
    # Percentage difference ggplot method
ggplot(data = counts_NC_separate_2016, aes(x = data_date, y=(Clinton-Trump)/(Clinton+Trump))) + geom_po
```




```
    # Linear model of the percentage difference with respect to date of the polls
counts_NC_for_lm_2016 = data.frame(data_date = date_NC_2016,
      percentage_diff = ((NC_total_2016$total.clinton - NC_total_2016$total.trump) /
  (NC_total_2016$total.clinton + NC_total_2016$total.trump)))

lm_model_NC_2016 = lm(percentage_diff ~ (data_date), data = counts_NC_for_lm_2016); lm_model_NC_2016
```

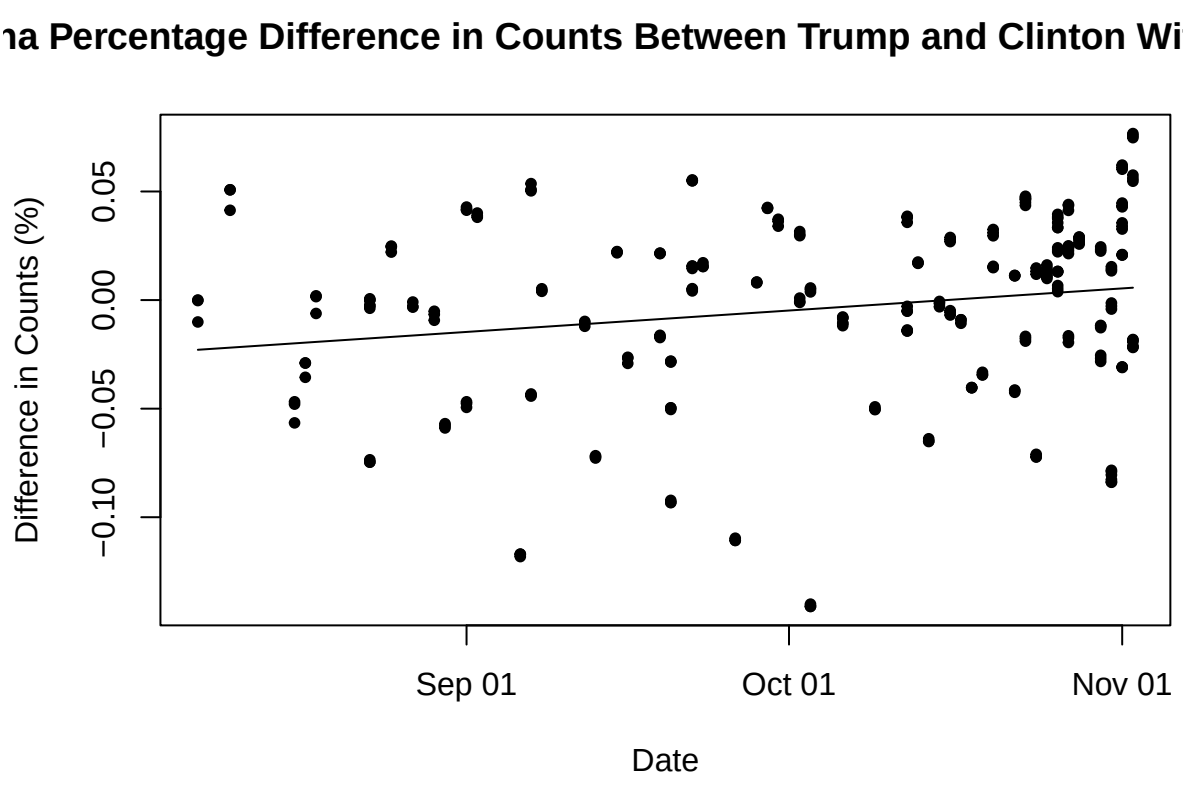
```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_NC_for_lm_2016)
##
## Coefficients:
## (Intercept)    data_date
##  -5.6154509    0.0003286
```

```
summary(lm_model_NC_2016)
```

```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_NC_for_lm_2016)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.136934 -0.023409  0.009753  0.027811  0.072702
##
```

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.6154509  1.7184797  -3.268  0.00122 **
## data_date    0.0003286  0.0001006   3.266  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04155 on 274 degrees of freedom
##   (39 observations deleted due to missingness)
## Multiple R-squared:  0.03746,    Adjusted R-squared:  0.03395
## F-statistic: 10.66 on 1 and 274 DF,  p-value: 0.001231
```

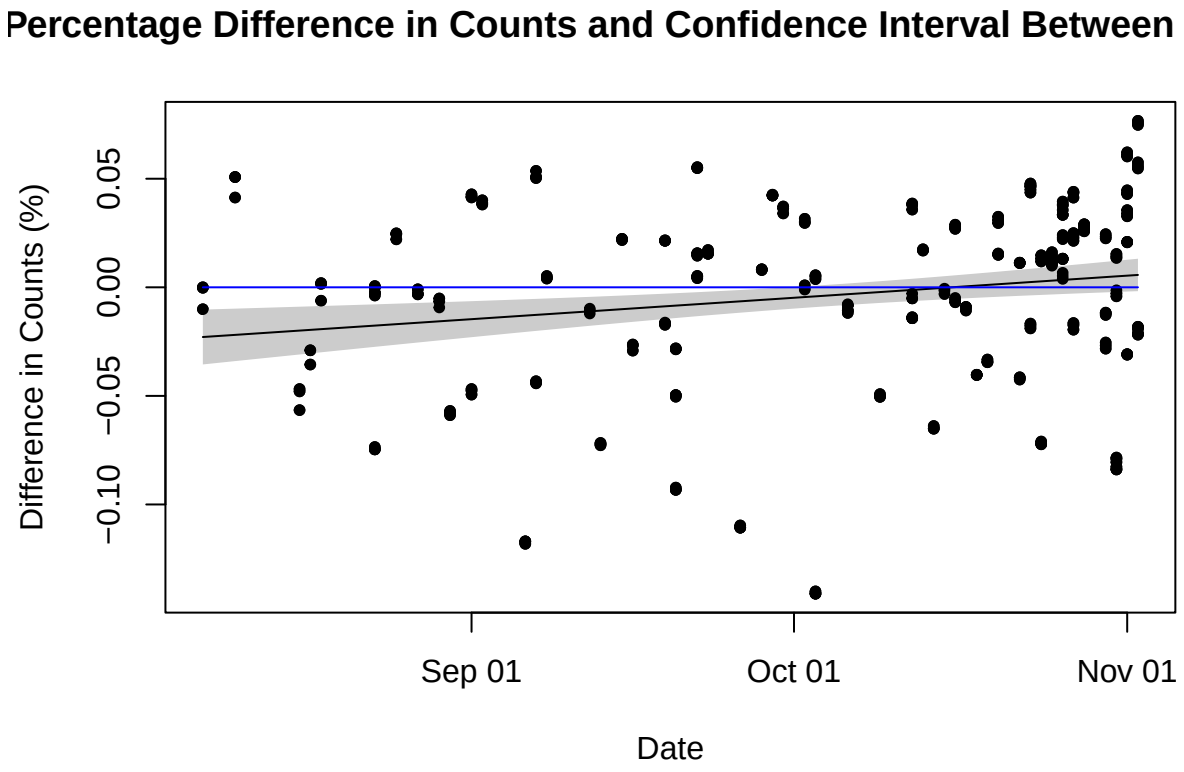
```
    # Plot fitted values of the fitted line
counts_NC_2016 <- data.frame(data_date = c(date_NC_2016, date_NC_2016),
  counts = c(NC_total_2016$total.clinton, NC_total_2016$total.trump),
  group = c(rep('Trump', length(date_NC_2016)), rep('Clinton',length(date_NC_2016))))

plot(counts_NC_for_lm_2016$data_date, counts_NC_for_lm_2016$percentage_diff, col='black', pch=20, type=
col='black', pch=20, type='l', xlab='Date', ylab='Difference in Counts (%)',
main='North Carolina')
```




```
    # Plot the confidence interval of the fitted line
fitted_CI_NC_2016 = predict(lm_model_NC_2016,
                            newdata = counts_NC_for_lm_2016,
                            interval = "confidence", level = 0.95)
summary(fitted_CI_NC_2016)
```

```
##       fit                lwr                upr
##  Min.   :-0.02289   Min.   :-0.03551   Min.   :-0.01026
##  1st Qu.:-0.00900   1st Qu.:-0.01491   1st Qu.:-0.00310
##  Median :-0.00104   Median :-0.00620   Median : 0.00413
##  Mean   :-0.00361   Mean   :-0.01037   Mean   : 0.00316
##  3rd Qu.: 0.00340   3rd Qu.:-0.00309   3rd Qu.: 0.00989
##  Max.   : 0.00570   Max.   :-0.00176   Max.   : 0.01317
##  NA's   :39         NA's   :39         NA's   :39
```

```
plot(counts_NC_for_lm_2016$data_date, counts_NC_for_lm_2016$percentage_diff,
     col='black', pch=20, type='p', xlab = 'Date', ylab = 'Difference in Counts (%)', main = 'North Car
```



From our plot of fitted values we see evidence of a trend in difference in North Carolina counts % and date. From our linear model of the percentage difference with respect to date of the polls for North Carolina we see a p value of 0.001231 which is less than the acceptable level of significance 0.05, meaning we have sufficient evidence to reject the null hypothesis and conclude the percentage difference in counts for North Carolina is affected by dates.

From our plot with a confidence interval for the fitted line the confidence interval contains 0, indicating both a negative and positive difference in counts % for Trump and Clinton. This means with repeated trails we are not expecting a difference in Trump and Clinton's North Carolina Difference in Count % with respect to dates of the polls, and this trial is most likely
indicative of being the closest election with the least difference in percentage difference.

```
sum(na.omit(percentage_diff_michigan_2016)) # 6.556257
```

```
## [1] 6.556257
```

```
sum(na.omit(percentage_diff_georgia_2016)) # -13.30905
```

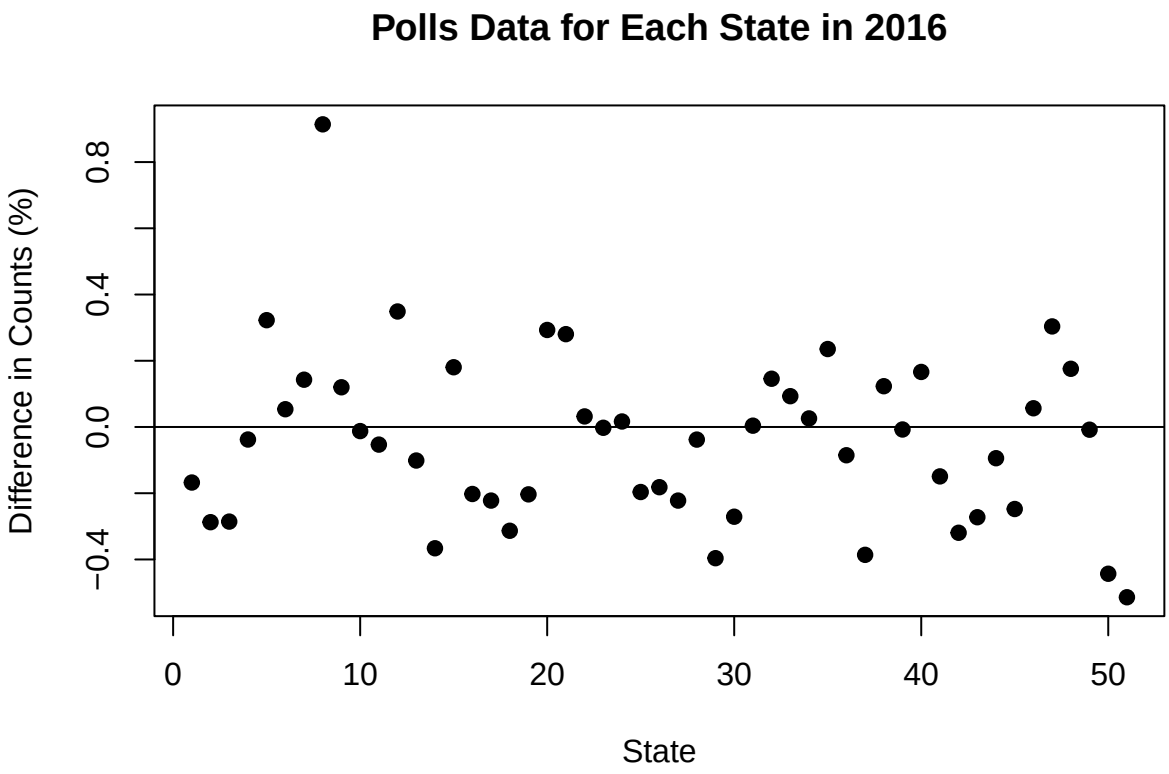
```
## [1] -13.30905
```

```
sum(na.omit(percentage_diff_NC_2016)) #-0.9956127
```

```
## [1] -0.9956127
```

I believe that based on the model and our observations North Carolina would have the closest election in terms of the lowest percentage difference between Trump and Clinton in 2016. The sum of their percentage differences is -.995 which is differs less from 0 than Michigan and Georgia's sum of percentages, meaning North Carolina's sum of percentage differences is closest to 0 and has the closest election, Trump beating Clinton by 0.9956%. Also, since North Carolina's confidence interval model is the only one that contains 0, North Carolina is expected to not have a difference in counts % and date.

```
  # e. From the real results of 2016 election, which state has the smallest
#      margin (in terms of percentage difference)? Discuss at least two reasons
#      that are different than what polls indicate. (You may check Wikipedia for
#      2016 US presidential election to find out the real voting results for each
#      state.)
load("/Users/krishao/Downloads/polls_election/data/pres_results.RData")
polls_data_real_2016=pres_results[pres_results$year=="2016",]

states_to_number <- c(1:51)

polls_data_real_2016_percentage_diff = (polls_data_real_2016$dem - polls_data_real_2016$rep)/(polls_data

polls_data_real_2016_diff_votes = polls_data_real_2016$dem - polls_data_real_2016$rep

plot(states_to_number, polls_data_real_2016_percentage_diff, col='black', pch=19, type='p', xlab='State
```



Calling the percentage differences of the real poll data from 2016 I see that -0.0023534 is the 23rd state with the smallest margin in terms of percentage difference, and our 23rd state is Michigan. This means that Trump beat Clinton by .23% contrary to our polls from part a) predicting Clinton would lead in Michigan. Georgia as our 11th state reported a margin in terms of percentage difference of -0.051313 and North Carolina as our 29th state reported a margin in terms of percentage difference of -0.39618, which followed our poll predictions of Trump being ahead in counts.

Reason the the polls are different is the margins are random variables which are not going to be ranked the same and the polls might be biased. Such as, differing from normal and having sampling biases of only polling certain areas but generalizing to a whole state or only receiving polling data from a certain strongly opinionated group of people.

f. Do polls correctly predict the candidate who wins these states? Discuss the bias of polls in these states. Name a few possible reasons. The polls didn't correctly predict the candidate who wins these states as it predicted Clinton would have more counts but Trump ended up having more counts in Michigan. There are multiple areas of bias. One is how the polling samples aren't independent and identically distributed. The polls aren't independent because votes for Trump and Clinton are dependent on each other as they might change depending on many factors which can change opinions such as debates, current events, and such, so this yields inaccurate prediction results that can vary drastically in different scenarios. Others biases are polling or sampling errors that inevitably occur and nonresponse errors where individuals don't answer and lead to inaccurate data.

Question 2: Redo Question 1 (a)-(f) for the same three states for the presidential polls in from August 1 to November 2 in 2020. (You may check Wikipedia for 2020 US presidential election to find out the real voting re- sults for each state.)

```
polls_data_2020 = read.csv(
"/Users/krishao/Downloads/polls_election/data/president_polls_2020.csv")
```

```r
# a. Who is ahead in each of these three states? What is the percentage
# difference for each state?
index_michigan_2020 = polls_data_2020$state=='Michigan'
michigan_total_michigan_2020 <- polls_data_2020[index_michigan_2020,]
polls_data_2020_enddate = mdy(polls_data_2020$end_date[polls_data_2020$state=="Michigan"])
polls_data_2020_startdate = mdy(polls_data_2020$start_date[polls_data_2020$state=="Michigan"])
michigan_total_enddate_2020 <- michigan_total_michigan_2020[polls_data_2020_enddate <= "2020-11-02",]
michigan_total_2020 <- michigan_total_enddate_2020[polls_data_2020_startdate>="2020-08-01",]
michigan_total_2020=michigan_total_2020[which(michigan_total_2020$pollster_id!=1610 & michigan_total_202

total_michigan_biden_2020 <- sum(michigan_total_2020$ sample_size[michigan_total_2020$answer=="Biden"]*m
```

```
## [1] 42149.71
```

```r
total_michigan_trump_2020 <- sum(michigan_total_2020$ sample_size[michigan_total_2020$answer=="Trump"]*m
```

```
## [1] 36275.64
```

```r
percen_dif_michigan_2020 <- ((total_michigan_biden_2020 - total_michigan_trump_2020)/
                      (total_michigan_biden_2020 + total_michigan_trump_2020));percen_dif_michigan_2020
```

```
## [1] 0.07490019
```

```r
# In Michigan Biden is ahead with 42,149.71 total votes whereas Trump received
# 36,275.64 total votes from August 1, 2020 to November 2, 2020. The percentage
# difference for Michigan is 0.07490019, meaning Biden is ahead of Trump in
# Michigan for 7.49% of the votes.

index_georgia_2020 = polls_data_2020$state=='Georgia'
georgia_total_georgia_2020 <- polls_data_2020[index_georgia_2020,]
polls_data_2020_enddate = mdy(polls_data_2020$end_date[polls_data_2020$state=="Georgia"])
polls_data_2020_startdate = mdy(polls_data_2020$start_date[polls_data_2020$state=="Georgia"])
georgia_total_enddate_2020 <- georgia_total_georgia_2020[polls_data_2020_enddate <= "2020-11-02",]
georgia_total_2020 <- georgia_total_enddate_2020[polls_data_2020_startdate>="2020-08-01",]
georgia_total_2020=georgia_total_2020[which(georgia_total_2020$pollster_id!=1610 & georgia_total_2020$po

total_georgia_biden_2020 <- sum(georgia_total_2020$ sample_size[georgia_total_2020$answer=="Biden"]*geor
```

```
## [1] 22920.43
```

```r
total_georgia_trump_2020 <- sum(georgia_total_2020$ sample_size[georgia_total_2020$answer=="Trump"]*geor
```

```
## [1] 22483.14
```

```r
percen_dif_georgia_2020 <- ((total_georgia_biden_2020 - total_georgia_trump_2020)/
                      (total_georgia_biden_2020 + total_georgia_trump_2020));percen_dif_georgia_2020
```

```
## [1] 0.009631098
```

```
# In Georgia Biden is ahead with 22,920.43 total votes whereas Trump received
# 22,483.14 total votes from August 1, 2020 to November 2, 2020. The percentage
# difference for Michigan is 0.009631098, meaning Biden is ahead of Trump in
# Michigan for .963% of the votes.

index_NC_2020 = polls_data_2020$state=='North Carolina'
NC_total_NC_2020 <- polls_data_2020[index_NC_2020,]
polls_data_2020_enddate = mdy(polls_data_2020$end_date[polls_data_2020$state=="North Carolina"])
polls_data_2020_startdate = mdy(polls_data_2020$start_date[polls_data_2020$state=="North Carolina"])
NC_total_enddate_2020 <- NC_total_NC_2020[polls_data_2020_enddate <= "2020-11-02",]
NC_total_2020 <- NC_total_enddate_2020[polls_data_2020_startdate>="2020-08-01",]
NC_total_2020=NC_total_2020[which(NC_total_2020$pollster_id!=1610 & NC_total_2020$pollster_id!=1193),]

total_NC_biden_2020 <- sum(NC_total_2020$ sample_size[NC_total_2020$answer=="Biden"]*NC_total_2020$pct[
```

```
## [1] 45270.63
```

```
total_NC_trump_2020 <- sum(NC_total_2020$ sample_size[NC_total_2020$answer=="Trump"]*NC_total_2020$pct[
```

```
## [1] 43461.72
```

```
percen_dif_NC_2020 <- ((total_NC_biden_2020 - total_NC_trump_2020)/
                       (total_NC_biden_2020 + total_NC_trump_2020));percen_dif_NC_2020
```

```
## [1] 0.0203862
```

```
# In North Carolina Biden is ahead with 45,270.63 total votes whereas Trump received
# 43,461.72 total votes from August 1, 2020 to November 2, 2020. The percentage
# difference for Michigan is 0.0203862, meaning Biden is ahead of Trump in
# Michigan for 2.04% of the votes.
```

b. Run a paired t test of the counts in polls for each of the state. Who is in favor of winning based on the test? Is the test significant? Is there potential problem?

We let d be the difference between the number of votes for Biden and Trump per poll with a level of significance of 0.05.

$$H_o : d = 0$$
$$H_a : d > 0$$

```
t.test(michigan_total_2020$sample_size[michigan_total_2020$answer=="Biden"]*michigan_total_2020$pct[mic
```

```
##
##  Paired t-test
##
## data:  michigan_total_2020$sample_size[michigan_total_2020$answer == "Biden"] * michigan_total_2020$p
## t = 13.993, df = 98, p-value < 2.2e-16
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  52.29314      Inf
## sample estimates:
## mean difference
##        59.33407
```

```
# Based on the test, since the p value is < 2.2e-16 and much less
# than an acceptable level of significance 0.05, we reject the null hypothesis
# and conclude the true mean differences between Trump and Biden's total votes
# is greater than 0. Therefore there is significant test evidence that Biden
# is favored in winning against Trump for Michigan.
```

$$H_o : d = 0$$
$$H_a : d > 0$$

```
t.test(georgia_total_2020$sample_size[georgia_total_2020$answer=="Biden"]*georgia_total_2020$pct[georgi
```

```
##
##  Paired t-test
##
## data:  georgia_total_2020$sample_size[georgia_total_2020$answer == "Biden"] * georgia_total_2020$pct
## t = 2.257, df = 57, p-value = 0.01393
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  1.954117      Inf
## sample estimates:
## mean difference
##        7.539417
```

```
# Based on the test, since the p value is 0.01393 and less than an acceptable
# level of significance 0.05, we reject the null hypothesis and conclude
# the true mean differences between Trump and Biden's total votes is greater
# than 0. Therefore there is significant test evidence that Biden
# is favored in winning against Trump for Georgia.
```

$$H_o : d = 0$$
$$H_a : d > 0$$

```
t.test(NC_total_2020$sample_size[NC_total_2020$answer=="Biden"]*NC_total_2020$pct[NC_total_2020$answer=
```

```
##
##  Paired t-test
##
## data:  NC_total_2020$sample_size[NC_total_2020$answer == "Biden"] * NC_total_2020$pct[NC_total_2020$a
## t = 7.4991, df = 108, p-value = 9.508e-12
## alternative hypothesis: true mean difference is greater than 0
## 95 percent confidence interval:
##  12.924    Inf
## sample estimates:
## mean difference
##        16.59556
```

```
# Based on the test, since the p value of 9.508e-12 is much less than an acceptable
# level of significance 0.05, we reject the null hypothesis and conclude
# the true mean differences between Trump and Biden's total votes is greater
# than 0. Therefore there is significant test evidence that Biden
# is favored in winning against Trump for Georgia.
```

A potential problem with using the paired t-test is although Trump and Biden's paired nature is designed from the same subject as pairs of observations, we cannot fully state that the voters' polls are paired where voters are matched on related variables, as they are independent polls. Therefore voter polls might not be considered paired and the t-test might not be suitable.

c. Run a Wilcoxon signed-rank test of the counts in polls for each of the state. Who is in favor of winning based on the test? Is the test significant? Is there potential problem of the test?

We let p be the proportion of votes in favor of Biden over Trump per poll with a level of significance of 0.05.

$$H_o : p = \frac{1}{2}$$
$$H_a : p > \frac{1}{2}$$

```
wilcox.test(michigan_total_2020$sample_size[michigan_total_2020$answer=="Biden"]*michigan_total_2020$pc
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  michigan_total_2020$sample_size[michigan_total_2020$answer == "Biden"] * michigan_total_2020$p
## W = 6227, p-value = 0.0005025
## alternative hypothesis: true location shift is greater than 0
```

```
# Based on the test, since the p value of 0.0005025 is less than an acceptable
# level of significance 0.05, we reject the null hypothesis that the true
# location shift is 0. There is significant test evidence to suggest that the
# true location shift is greater than 0 and that Biden is in favor of winning
# against Trump in Michigan.
```

$$H_o : p = \frac{1}{2}$$
$$H_a : p > \frac{1}{2}$$

```
wilcox.test(georgia_total_2020$sample_size[georgia_total_2020$answer=="Biden"]*georgia_total_2020$pct[g
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  georgia_total_2020$sample_size[georgia_total_2020$answer == "Biden"] * georgia_total_2020$pct
## W = 1711, p-value = 0.4375
## alternative hypothesis: true location shift is greater than 0
```

```
# Based on the test, since the p value is 0.4375 and more than an acceptable
# level of significance 0.05, we fail to reject the null hypothesis that the true
# location shift is 0. There is significant test evidence to suggest that
# the true location shift is 0 and insignificant evidence to suggest Biden is in
# favor of winning. Trump and Biden are equally in favor of winning in Georgia.
```

$$H_o : p = \frac{1}{2}$$
$$H_a : p > \frac{1}{2}$$

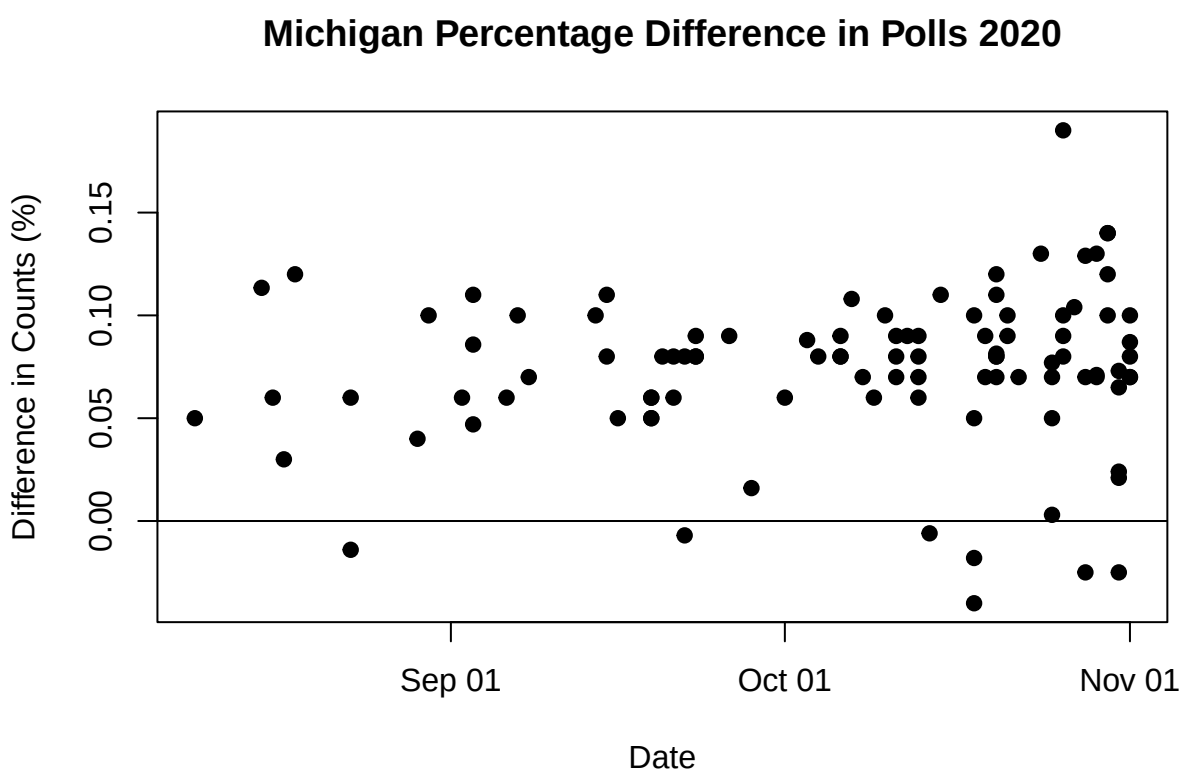
```
wilcox.test(NC_total_2020$sample_size[NC_total_2020$answer=="Biden"]*NC_total_2020$pct[NC_total_2020$ans
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  NC_total_2020$sample_size[NC_total_2020$answer == "Biden"] * NC_total_2020$pct[NC_total_2020$a
## W = 6313, p-value = 0.2122
## alternative hypothesis: true location shift is greater than 0
```

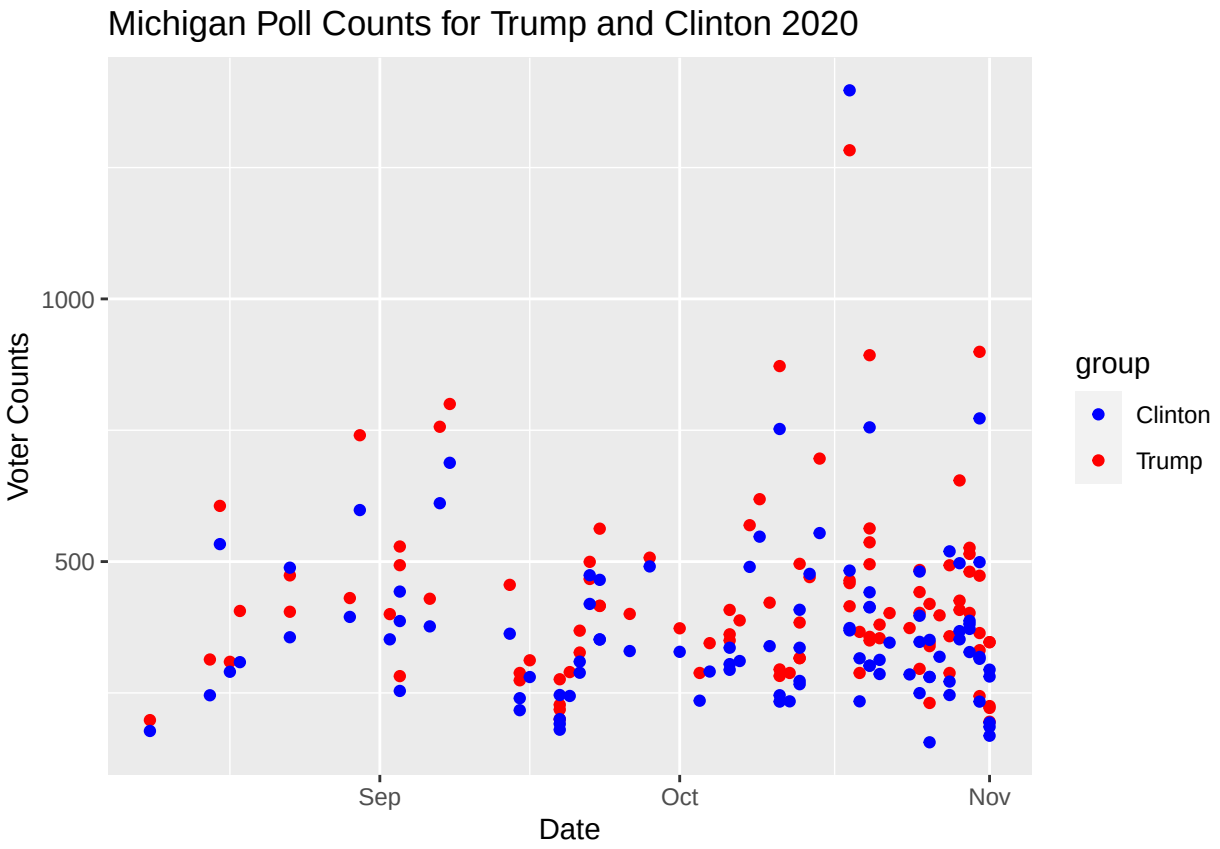
```
# Based on the test, since the p value is 0.2122 and more than an acceptable
# level of significance 0.05, we fail to reject the null hypothesis that the true
# location shift is 0. There is significant test evidence to suggest that
# the true location shift is 0 and insignificant evidence to suggest Biden is in
# favor of winning. Trump and Biden are equally in favor of winning in North Carolina.
```

A potential problem with using the Wilcoxon signed-rank test is similar to the t-test in how the voters are unpaired on related variables as they are independent poll decisions. Since the Wilcoxon signed-rank test assesses the location shift between the paired differences, if the voter poll observations are not paired this test might not be as accurate.

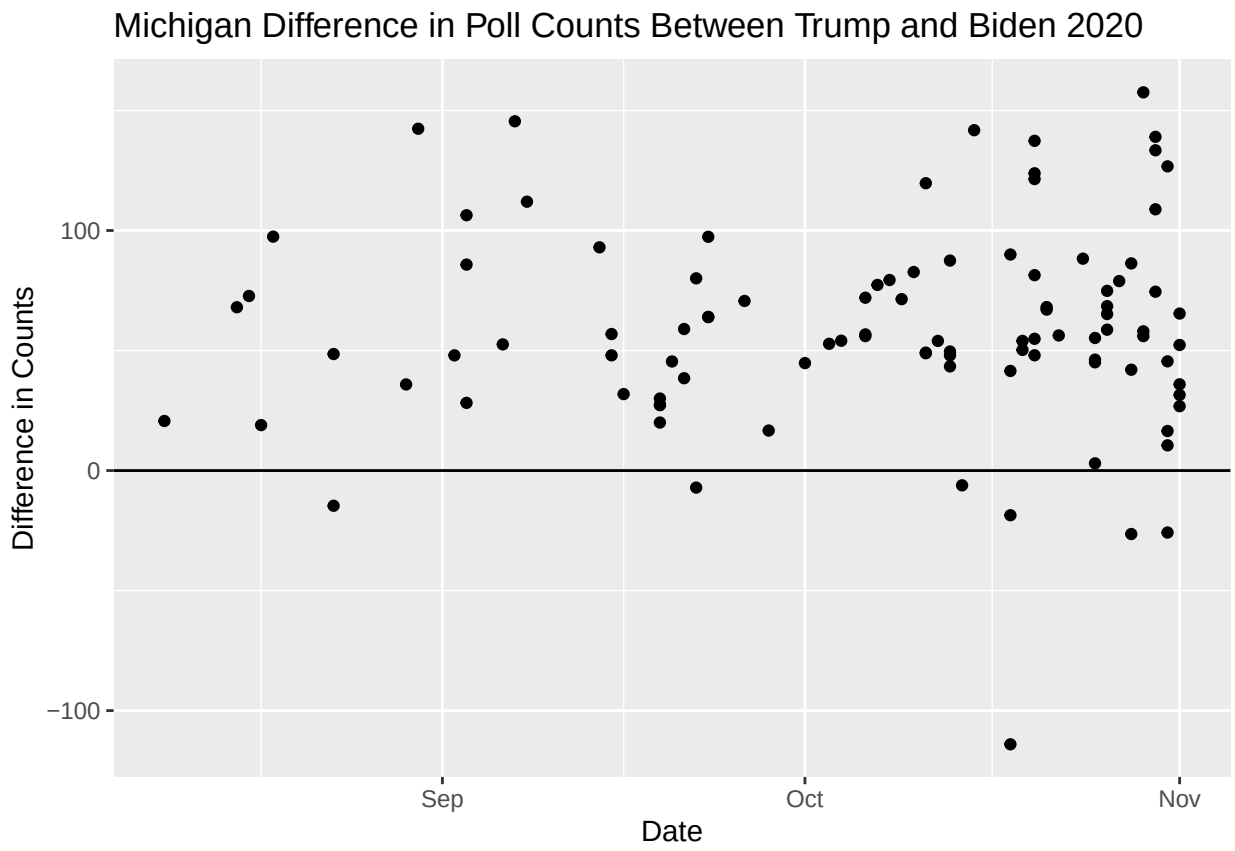
```
    # d. Fit a linear model of the percentage difference with respect to date of
    # the polls separately for each of these states. Show a plot of the
    # observations of the polls, fitted values and confidence interval of the
    # fitted line for each of these state. From the linear model and
    # observations, which state may have the closest election (in terms of
    # percentage difference)?

# MICHIGAN
    # Percentage difference no ggplot method
michigan_total_2020=michigan_total_2020[which(michigan_total_2020$pollster_id!=1610 & michigan_total_202
date_michigan_2020=mdy(michigan_total_2020$end_date[michigan_total_2020$answer=="Biden"])

percentage_diff_michigan_2020 = (michigan_total_2020$pct[michigan_total_2020$answer=="Biden"] - michigan

plot(date_michigan_2020, percentage_diff_michigan_2020, col='black',
     pch=19, type='p', xlab='Date', ylab='Difference in Counts (%)',
     main='Michigan Percentage Difference in Polls 2020', );abline(a=0,b=0)
```



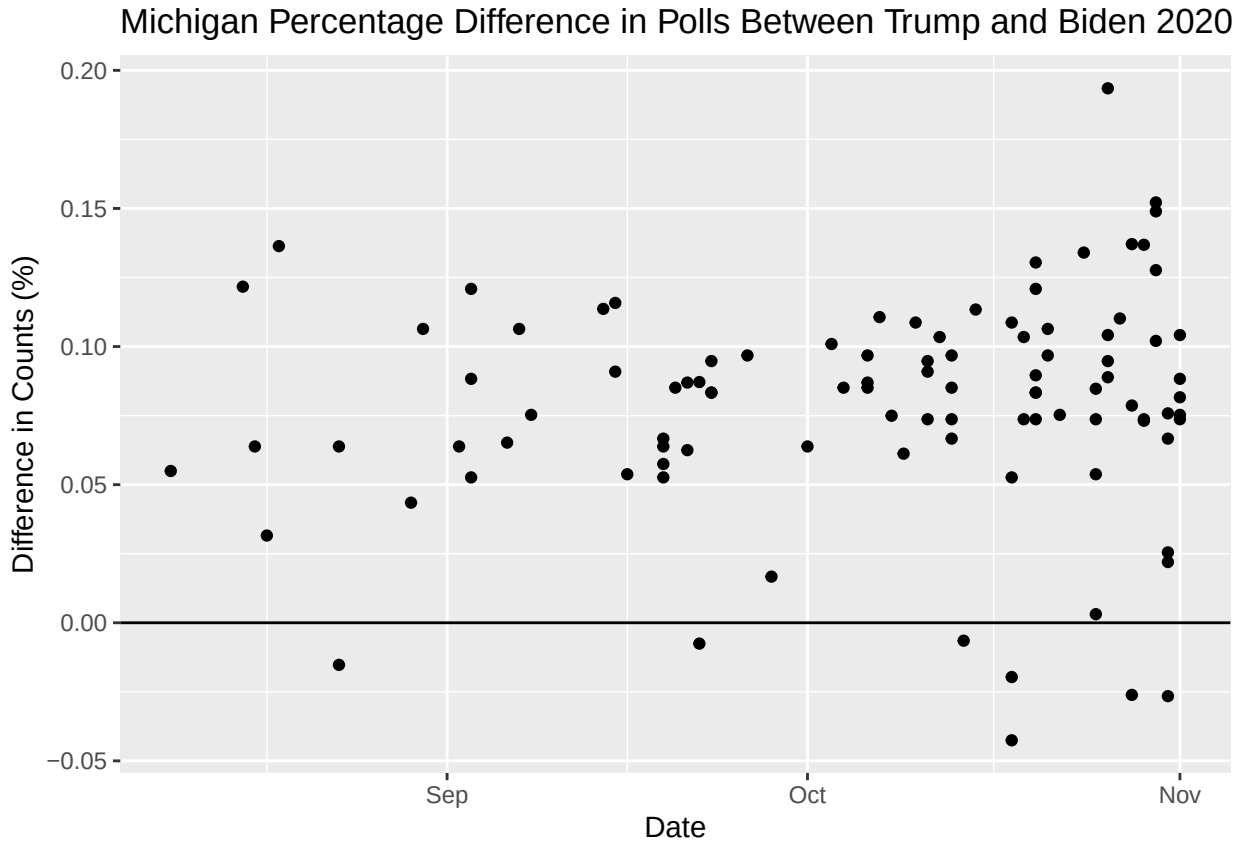

```
    # Plot observations of polls
counts_michigan_2020 <- data.frame(data_date = c(date_michigan_2020, date_michigan_2020),
  counts = c(michigan_total_2020$ sample_size[michigan_total_2020$answer=="Biden"]*michigan_total_2020$p
  group = c(rep('Trump', length(date_michigan_2020)), rep('Clinton',length(date_michigan_2020))))

ggplot(data=counts_michigan_2020, aes(x=data_date, y=counts, col=group)) +
  geom_point() +
  scale_color_manual(values = c("blue","red")) + labs(x='Date') +
  labs(y='Voter Counts') + labs(title='Michigan Poll Counts for Trump and Clinton 2020')
```




```
counts_michigan_separate_2020 = data.frame(data_date = date_michigan_2020,
  Trump = michigan_total_2020$ sample_size[michigan_total_2020$answer=="Trump"]*michigan_total_2020$pct
  Biden = michigan_total_2020$ sample_size[michigan_total_2020$answer=="Biden"]*michigan_total_2020$pct

ggplot(data = counts_michigan_separate_2020, aes(x=data_date, y=Biden-Trump)) + geom_point() + xlab('Da
```




```
    # Percentage difference ggplot method
ggplot(data = counts_michigan_separate_2020, aes(x = data_date, y=(Biden-Trump)/(Biden+Trump))) + geom_
```




```
    # Linear model of the percentage difference with respect to date of the polls
counts_michigan_for_lm_2020 = data.frame(data_date = date_michigan_2020,
      percentage_diff = ((michigan_total_2020$ sample_size[michigan_total_2020$answer=="Biden"]*michigan

counts_michigan_for_lm_2020 = data.frame(data_date = date_michigan_2020,
      percentage_diff = ((michigan_total_2020$pct[michigan_total_2020$answer=="Biden"] - michigan_total_

lm_model_michigan_2020 = lm(percentage_diff ~ (data_date), data = counts_michigan_for_lm_2020); lm_model
```

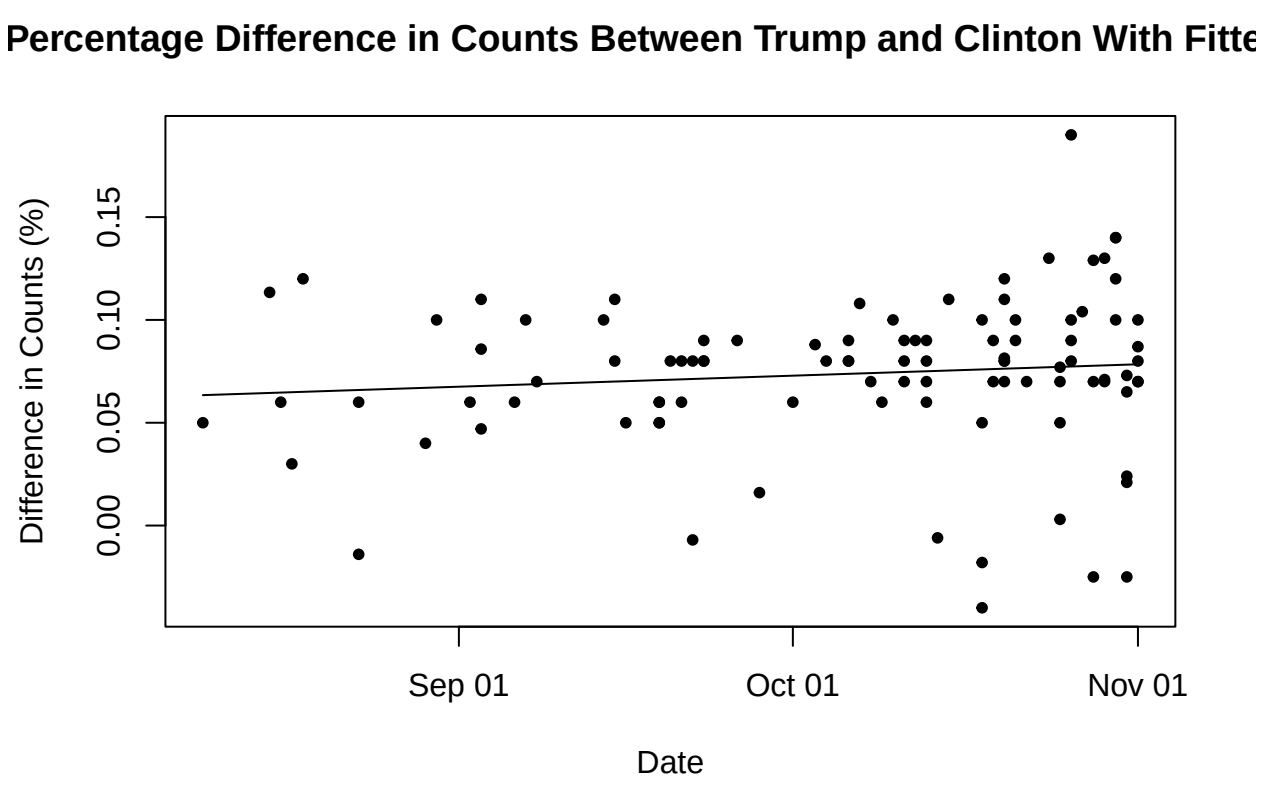
```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_michigan_for_lm_2020)
##
## Coefficients:
## (Intercept)    data_date
##  -3.2391291    0.0001787
```

```
summary(lm_model_michigan_2020)
```

```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_michigan_for_lm_2020)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.115923 -0.010920  0.003719  0.020059  0.112647
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.2391291  3.1175352  -1.039    0.301
## data_date    0.0001787  0.0001681   1.063    0.291
##
## Residual standard error: 0.03792 on 97 degrees of freedom
## Multiple R-squared:  0.01151,    Adjusted R-squared:  0.001318
## F-statistic: 1.129 on 1 and 97 DF,  p-value: 0.2906
```

```
    # Plot fitted values of the fitted line
counts_michigan_2020 <- data.frame(data_date = c(date_michigan_2020, date_michigan_2020),
  counts = c(michigan_total_2020$ sample_size[michigan_total_2020$answer=="Biden"]*michigan_total_2020$p
  group = c(rep('Trump', length(date_michigan_2020)), rep('Biden',length(date_michigan_2020)))))

plot(counts_michigan_for_lm_2020$data_date, counts_michigan_for_lm_2020$percentage_diff,
     col='black', pch=20, type='p', xlab='Date', ylab='Difference in Counts (%)',
     main='Michigan Percentage Difference in Counts Between Trump and Clinton With Fitted Values 2020')
col='black', pch=20, type='l', xlab='Date', ylab='Difference in Counts (%)',
main='Michigan')
```

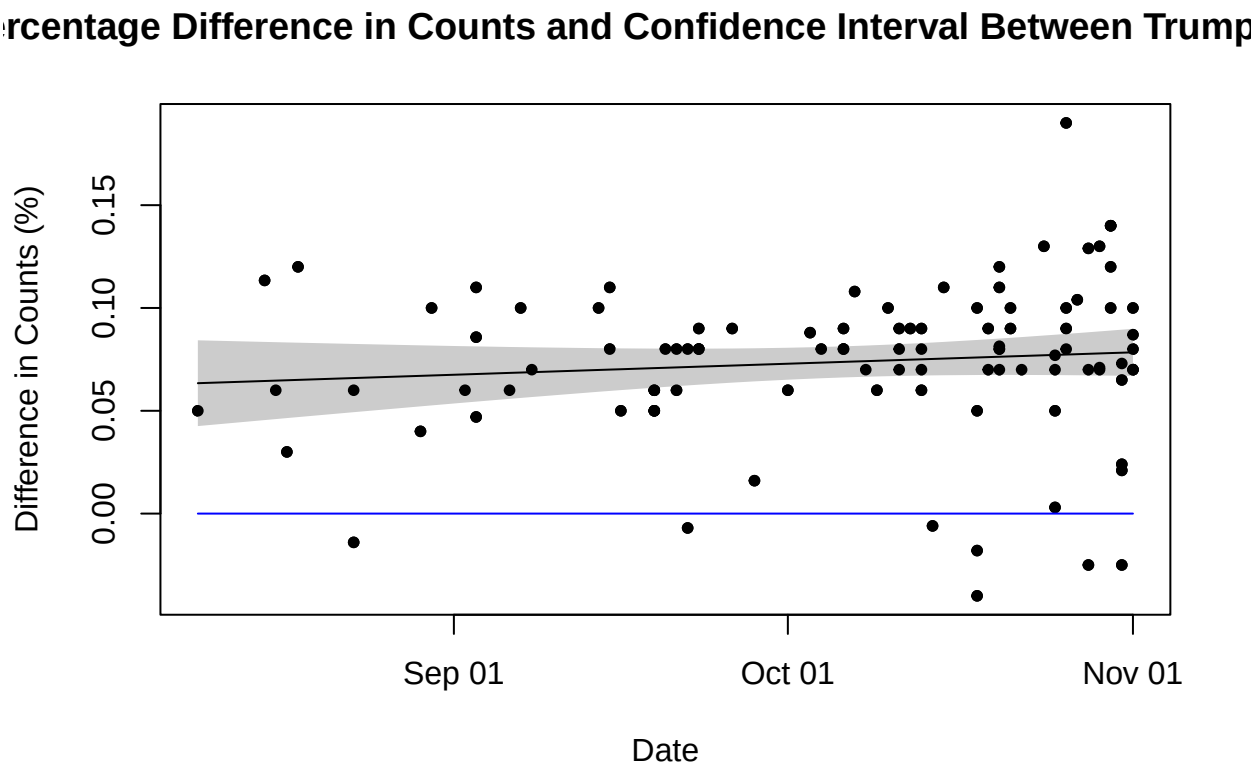



```
    # Plot the confidence interval of the fitted line
fitted_CI_michigan_2020 = predict(lm_model_michigan_2020,
                                  newdata = counts_michigan_for_lm_2020,
```

```
                              interval = "confidence", level = 0.95)
summary(fitted_CI_michigan_2020)
```

```
##       fit                lwr                upr
##  Min.   :0.06342   Min.   :0.04256   Min.   :0.08021
##  1st Qu.:0.07101   1st Qu.:0.06180   1st Qu.:0.08131
##  Median :0.07503   Median :0.06698   Median :0.08324
##  Mean   :0.07382   Mean   :0.06349   Mean   :0.08415
##  3rd Qu.:0.07726   3rd Qu.:0.06728   3rd Qu.:0.08719
##  Max.   :0.07842   Max.   :0.06744   Max.   :0.08987
```

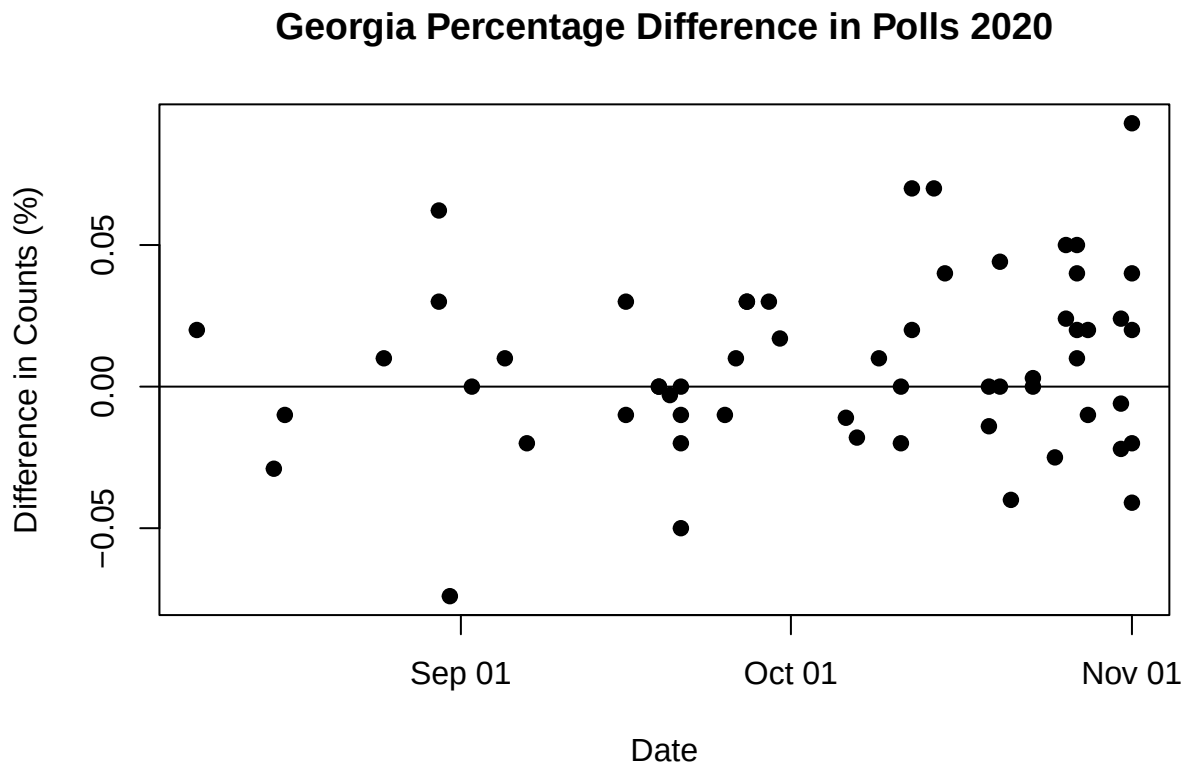
```
plot(counts_michigan_for_lm_2020$data_date, counts_michigan_for_lm_2020$percentage_diff,
     col='black', pch=20, type='p', xlab = 'Date', ylab = 'Difference in Counts (%)',
     main = 'Michigan Percentage Difference in Counts and Confidence Interval Between Trump and Biden 2
```



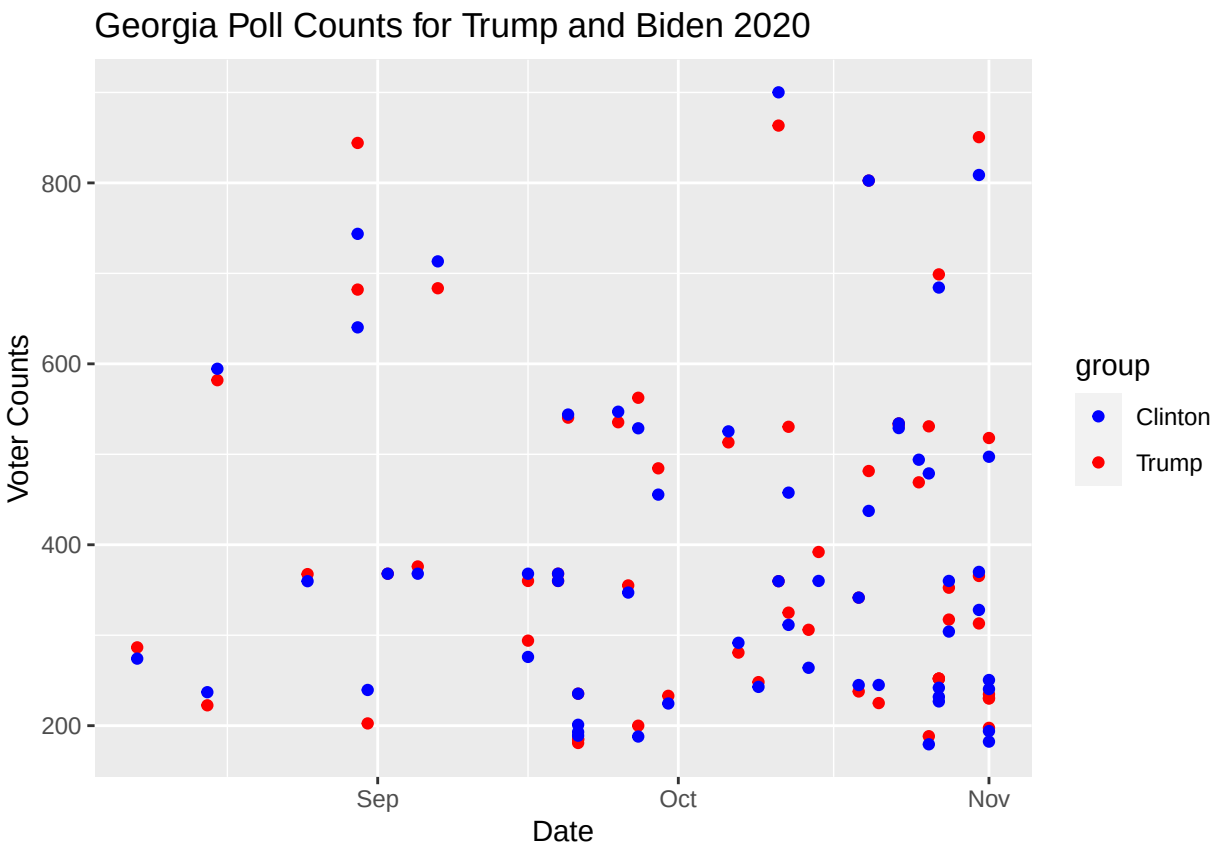
From our plot of fitted values we see evidence of a trend in difference in Michigan counts % and date. We expect that early polls do not have as much impact as recent polls as most polls are concentrated on more recent months. From our linear model of the percentage difference with respect to date of the polls for Michigan we see a p value of 0.2906 which is more than the acceptable level of significance 0.05, meaning we don't have sufficient evidence to reject the null hypothesis and conclude the percentage difference in counts for Michigan is not affected by dates.

From our plot with a confidence interval for the fitted line the confidence interval doesn't contain 0, as the values are positive and above 0 indicating a positive difference in counts % for Trump and Biden. This means with repeated trails we are expecting a difference in Trump and Biden's Michigan Difference in Count % with respect to dates of the polls, and this trial is most likely not indicative of being the closest election with the least difference in percentage difference.

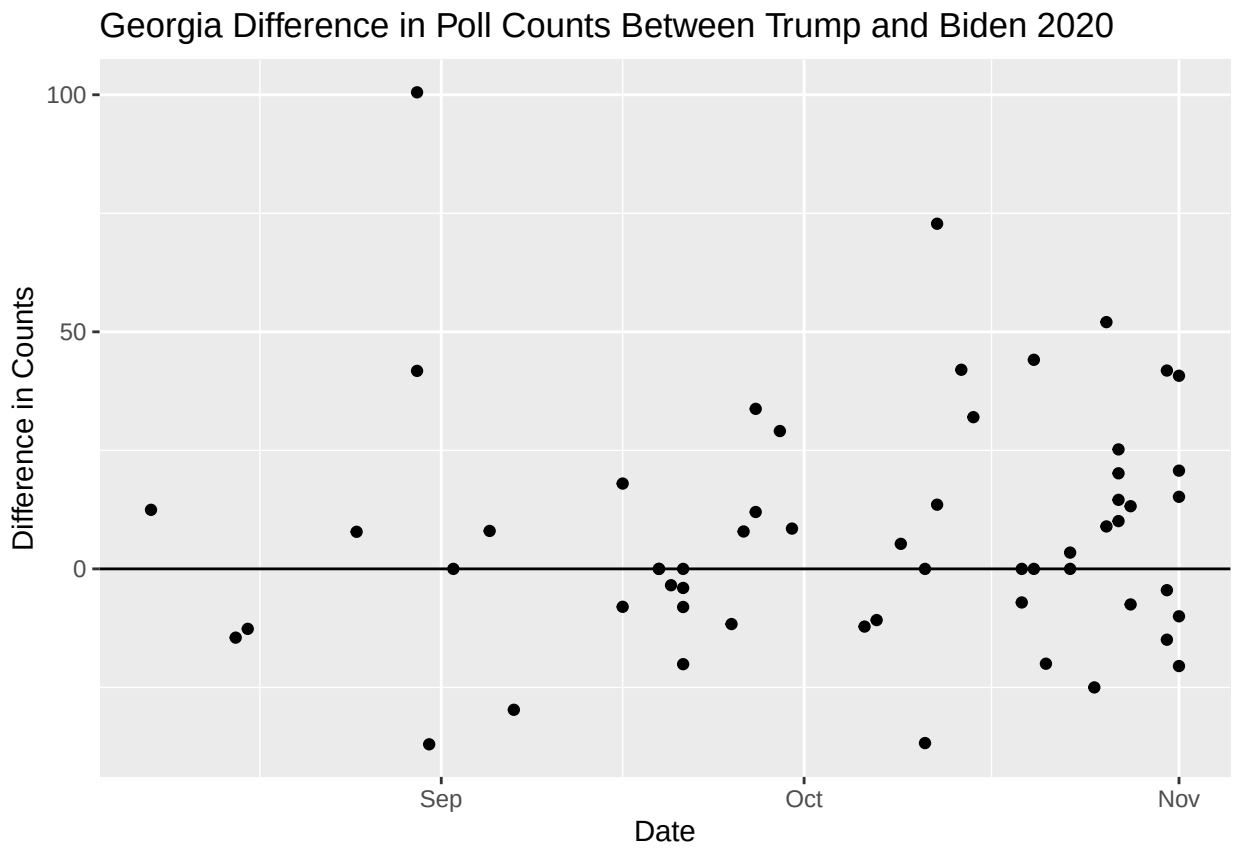
```r
  # d. continued

# GEORGIA
    # Percentage difference no ggplot method
georgia_total_2020=georgia_total_2020[which(georgia_total_2020$pollster_id!=1610 & georgia_total_2020$p
date_georgia_2020=mdy(georgia_total_2020$end_date[georgia_total_2020$answer=="Biden"])

percentage_diff_georgia_2020 = (georgia_total_2020$pct[georgia_total_2020$answer=="Biden"] - georgia_to

plot(date_georgia_2020, percentage_diff_georgia_2020, col='black',
     pch=19, type='p', xlab='Date', ylab='Difference in Counts (%)',
     main='Georgia Percentage Difference in Polls 2020', );abline(a=0,b=0)
```



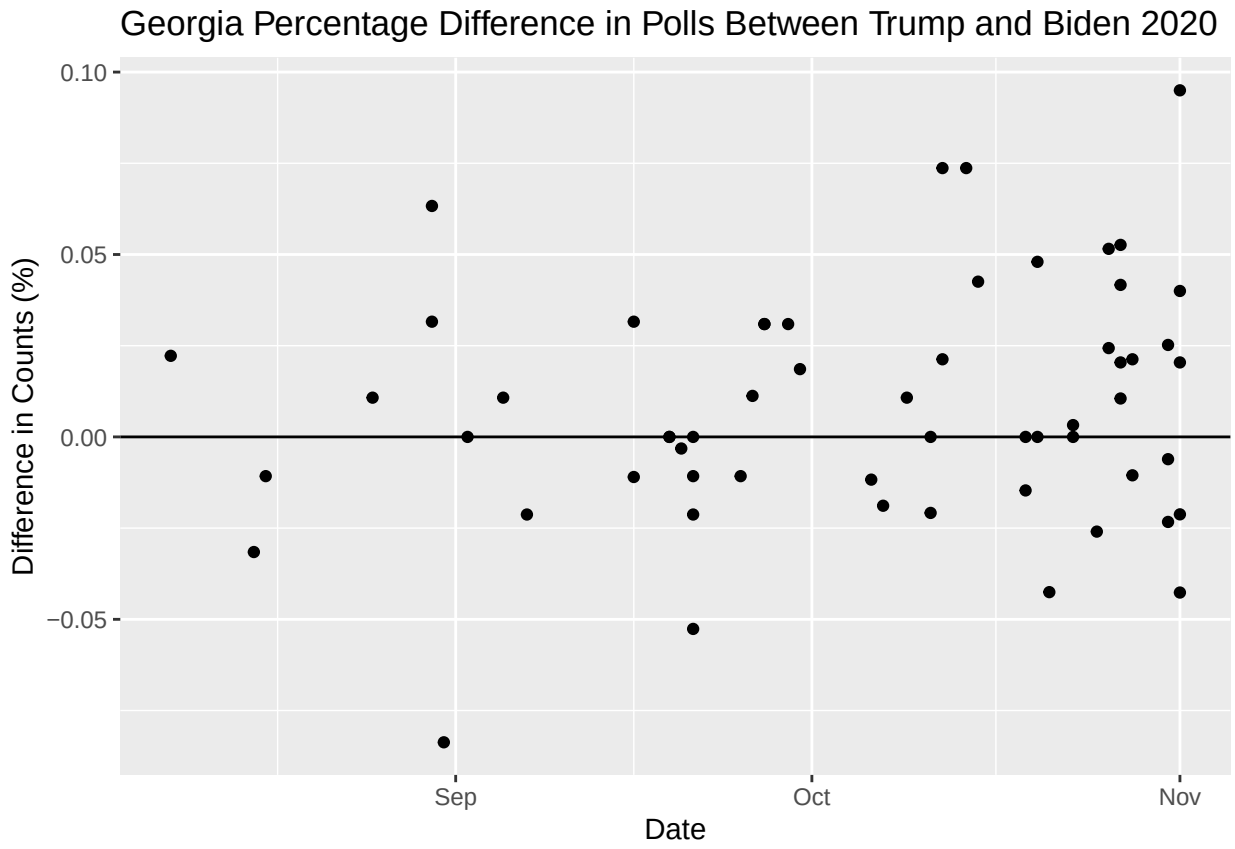

```r
    # Plot observations of polls
counts_georgia_2020 <- data.frame(data_date = c(date_georgia_2020, date_georgia_2020),
  counts = c(georgia_total_2020$ sample_size[georgia_total_2020$answer=="Biden"]*georgia_total_2020$pct
  group = c(rep('Trump', length(date_georgia_2020)), rep('Clinton',length(date_georgia_2020))))

ggplot(data=counts_georgia_2020, aes(x=data_date, y=counts, col=group)) +
  geom_point() +
  scale_color_manual(values = c("blue","red")) + labs(x='Date') +
  labs(y='Voter Counts') + labs(title='Georgia Poll Counts for Trump and Biden 2020')
```




```
counts_georgia_separate_2020 = data.frame(data_date = date_georgia_2020,
  Trump = georgia_total_2020$ sample_size[georgia_total_2020$answer=="Trump"]*georgia_total_2020$pct[ge
  Biden = georgia_total_2020$ sample_size[georgia_total_2020$answer=="Biden"]*georgia_total_2020$pct[ge

ggplot(data = counts_georgia_separate_2020, aes(x=data_date, y=Biden-Trump)) + geom_point() + xlab('Dat
```




```
    # Percentage difference ggplot method
ggplot(data = counts_georgia_separate_2020, aes(x = data_date, y=(Biden-Trump)/(Biden+Trump))) + geom_p
```




```
    # Linear model of the percentage difference with respect to date of the polls
counts_georgia_for_lm_2020 = data.frame(data_date = date_georgia_2020,
      percentage_diff = ((georgia_total_2020$ sample_size[georgia_total_2020$answer=="Biden"]*georgia_t

counts_georgia_for_lm_2020 = data.frame(data_date = date_georgia_2020,
      percentage_diff = ((georgia_total_2020$pct[georgia_total_2020$answer=="Biden"] - georgia_total_20

lm_model_georgia_2020 = lm(percentage_diff ~ (data_date), data = counts_georgia_for_lm_2020); lm_model_
```

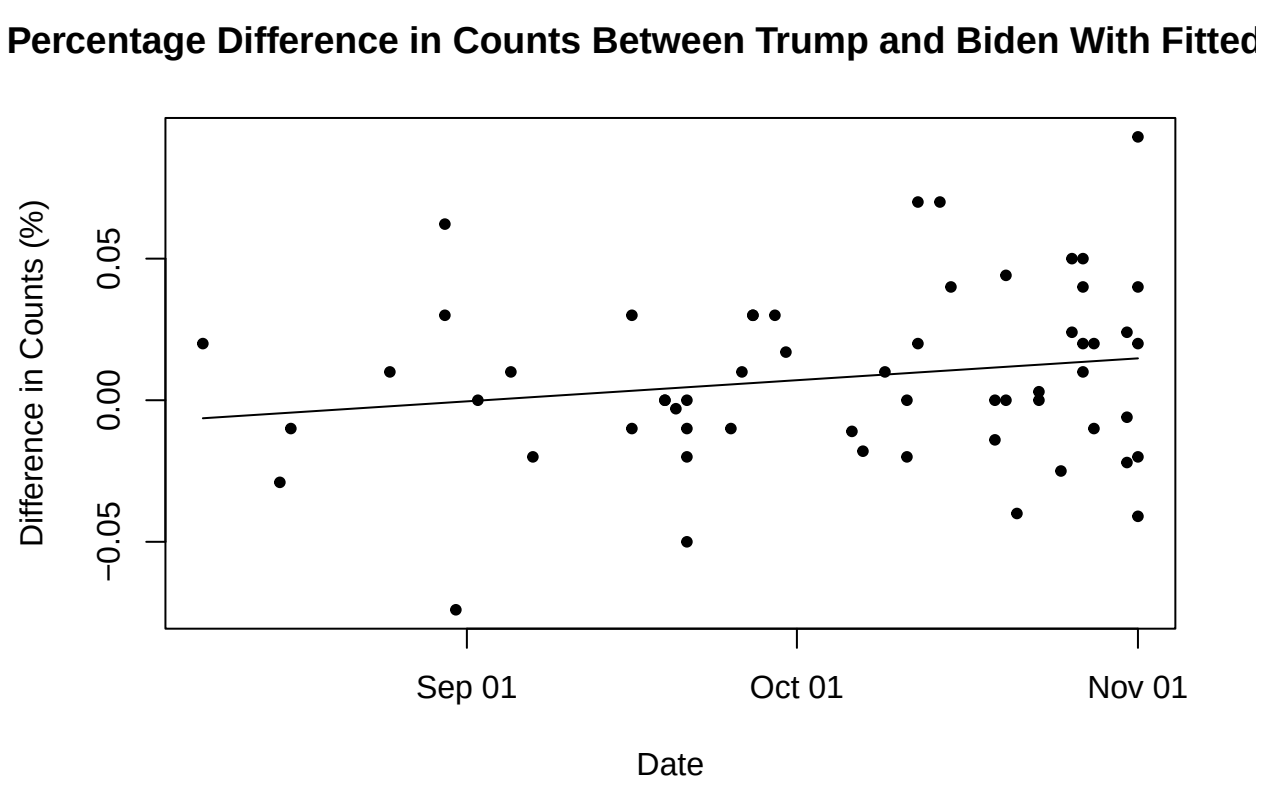
```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_georgia_for_lm_2020)
##
## Coefficients:
## (Intercept)    data_date
##  -4.6064057    0.0002489
```

```
summary(lm_model_georgia_2020)
```

```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_georgia_for_lm_2020)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.073354 -0.020231 -0.003812  0.023801  0.078215
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.6064057  3.2133031  -1.434    0.157
## data_date    0.0002489  0.0001733   1.436    0.157
##
## Residual standard error: 0.03072 on 56 degrees of freedom
## Multiple R-squared:  0.03552,    Adjusted R-squared:  0.01829
## F-statistic: 2.062 on 1 and 56 DF,  p-value: 0.1566
```

```
    # Plot fitted values of the fitted line
counts_georgia_2020 <- data.frame(data_date = c(date_georgia_2020, date_georgia_2020),
  counts = c(georgia_total_2020$ sample_size[georgia_total_2020$answer=="Biden"]*georgia_total_2020$pct
  group = c(rep('Trump', length(date_georgia_2020)), rep('Biden',length(date_georgia_2020)))))

plot(counts_georgia_for_lm_2020$data_date, counts_georgia_for_lm_2020$percentage_diff,
     col='black', pch=20, type='p', xlab='Date', ylab='Difference in Counts (%)',
     main='Georgia Percentage Difference in Counts Between Trump and Biden With Fitted Values 2020');li
col='black', pch=20, type='l', xlab='Date', ylab='Difference in Counts (%)',
main='Georgia')
```

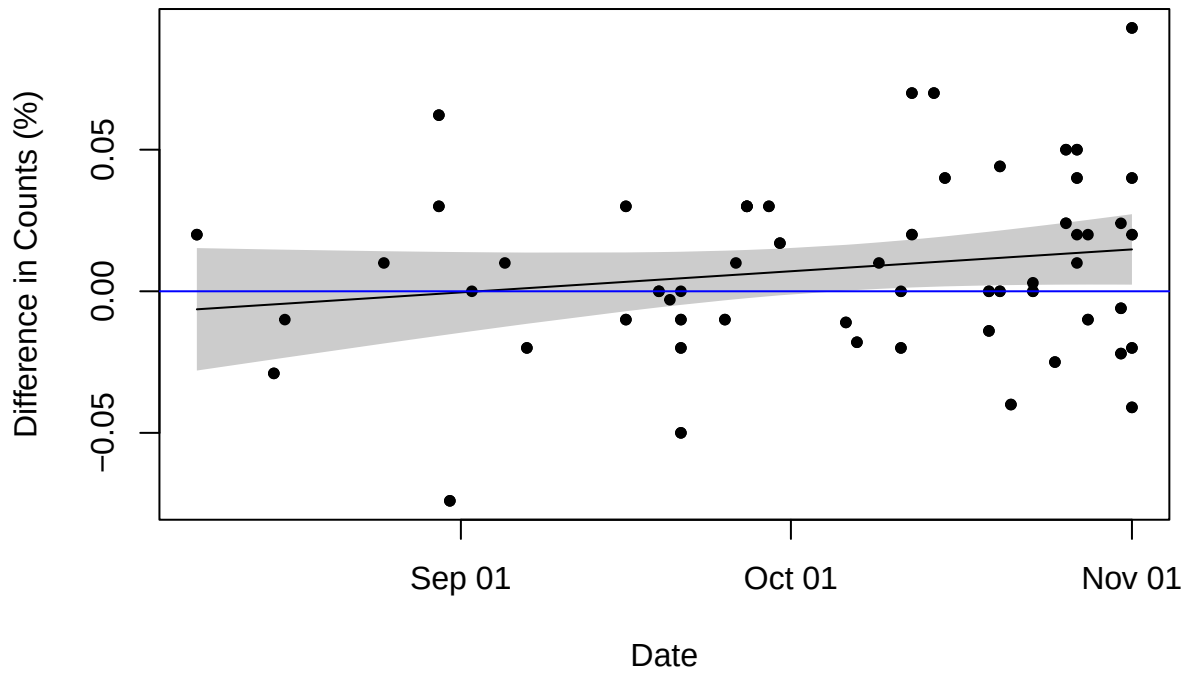



```
    # Plot the confidence interval of the fitted line
fitted_CI_georgia_2020 = predict(lm_model_georgia_2020,
                                 newdata = counts_georgia_for_lm_2020,
```

```
                              interval = "confidence", level = 0.95)
summary(fitted_CI_georgia_2020)
```

```
##       fit                 lwr                 upr
##  Min.   :-0.006371   Min.   :-0.027991   Min.   :0.01367
##  1st Qu.: 0.004394   1st Qu.:-0.005129   1st Qu.:0.01401
##  Median : 0.009559   Median : 0.001193   Median :0.01792
##  Mean   : 0.008005   Mean   :-0.003058   Mean   :0.01907
##  3rd Qu.: 0.013292   3rd Qu.: 0.002346   3rd Qu.:0.02423
##  Max.   : 0.014785   Max.   : 0.002370   Max.   :0.02723
```

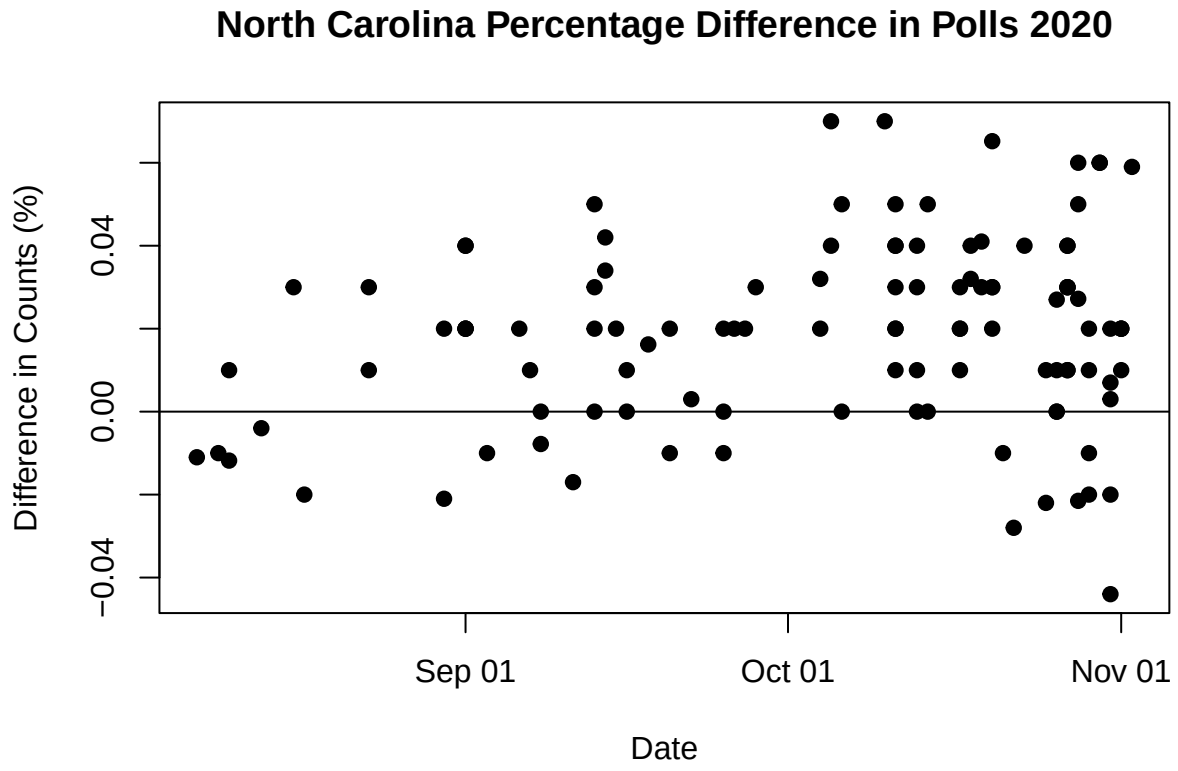
```
plot(counts_georgia_for_lm_2020$data_date, counts_georgia_for_lm_2020$percentage_diff,
     col='black', pch=20, type='p', xlab = 'Date', ylab = 'Difference in Counts (%)',
     main = 'Georgia Percentage Difference in Counts and Confidence Interval Between Trump and Biden 20
```



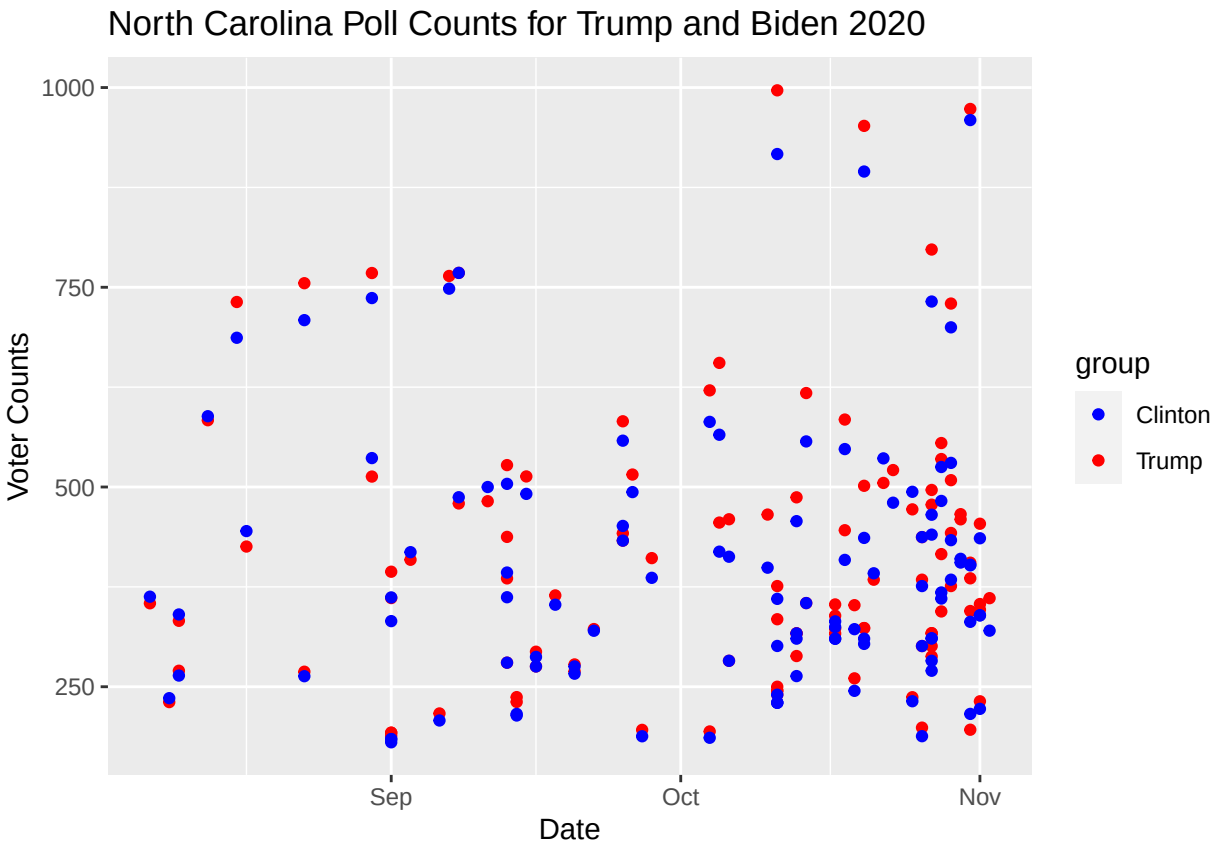
From our plot of fitted values we see evidence of a trend in difference in Georgia counts % and date. We expect that early polls do not have as much impact as recent polls as most polls are concentrated on more recent months. From our linear model of the percentage difference with respect to date of the polls for Georgia we see a p value of 0.1566 which is more than the acceptable level of significance 0.05, meaning we don't have sufficient evidence to reject the null hypothesis and conclude the percentage difference in counts for Georgia is not affected by dates.

From our plot with a confidence interval for the fitted line the confidence interval contains 0, as the values are negative and positive above and below 0, indicating there is no difference in counts % for Trump and Biden. This means with repeated trails we are not expecting a difference in Trump and Biden's Georgia Difference in Count % with respect to dates of the polls, and this trial can be indicative of being the closest election with the least difference in percentage difference.

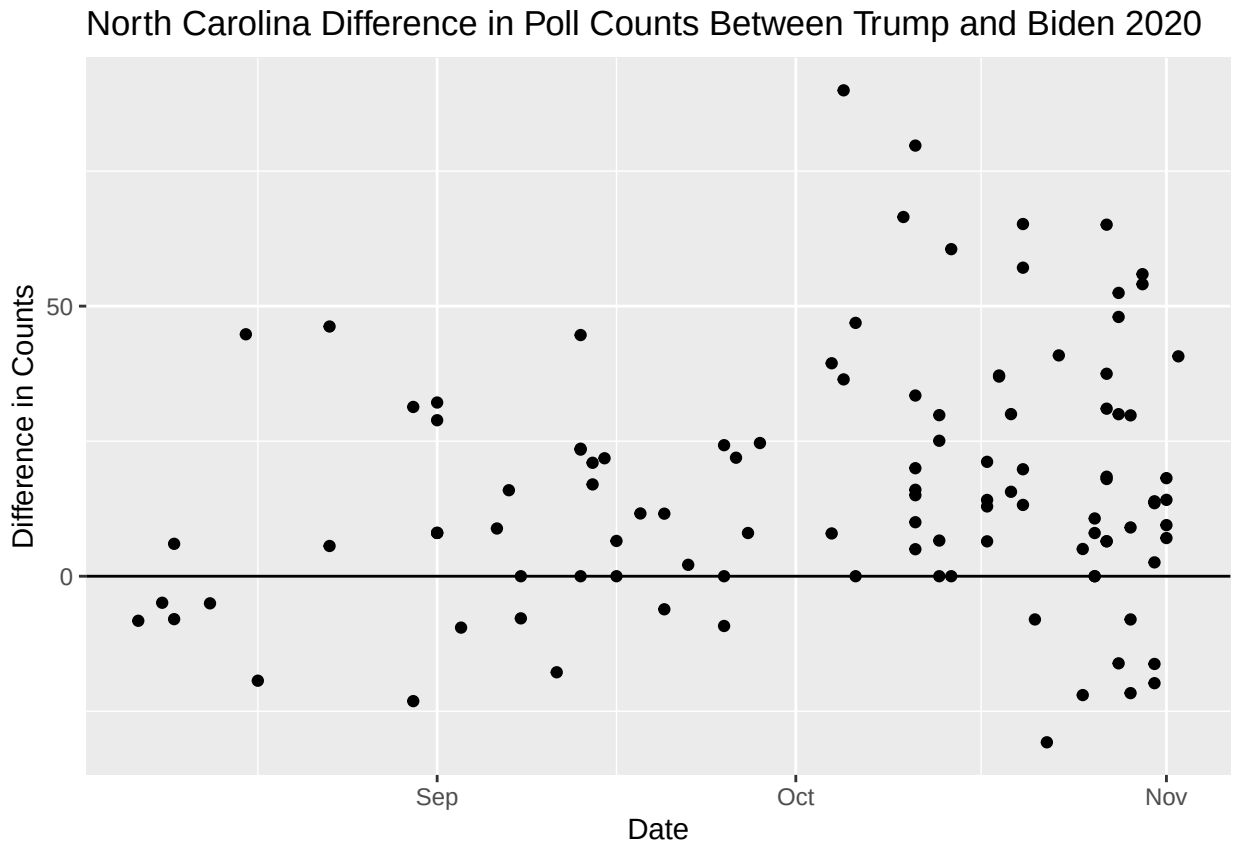
```r
  # d. continued

# NORTH CAROLINA
    # Percentage difference no ggplot method
NC_total_2020=NC_total_2020[which(NC_total_2020$pollster_id!=1610 & NC_total_2020$pollster_id!=1193),]
date_NC_2020=mdy(NC_total_2020$end_date[NC_total_2020$answer=="Biden"])

percentage_diff_NC_2020 = (NC_total_2020$pct[NC_total_2020$answer=="Biden"] - NC_total_2020$pct[NC_total

plot(date_NC_2020, percentage_diff_NC_2020, col='black',
     pch=19, type='p', xlab='Date', ylab='Difference in Counts (%)',
     main='North Carolina Percentage Difference in Polls 2020', );abline(a=0,b=0)
```



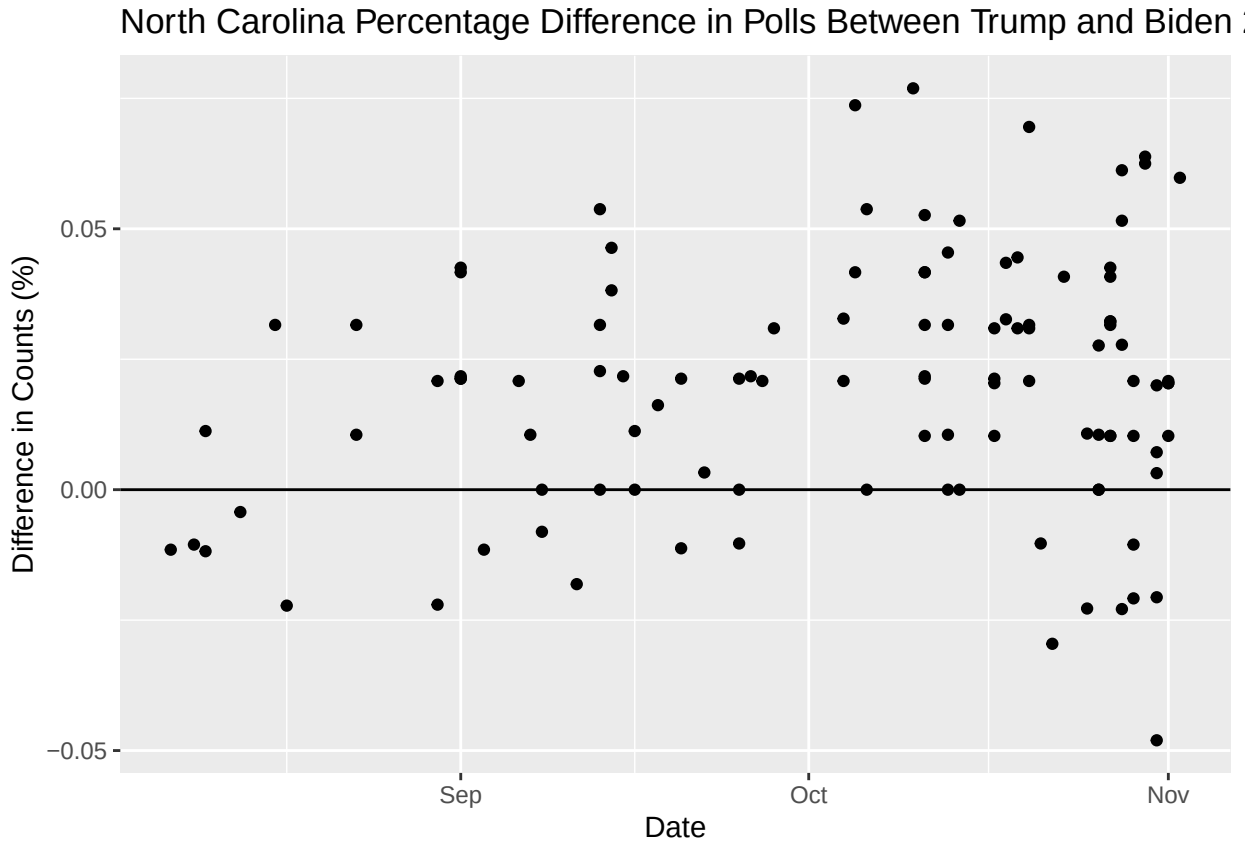

```r
    # Plot observations of polls
counts_NC_2020 <- data.frame(data_date = c(date_NC_2020, date_NC_2020),
  counts = c(NC_total_2020$ sample_size[NC_total_2020$answer=="Biden"]*NC_total_2020$pct[NC_total_2020$a
  group = c(rep('Trump', length(date_NC_2020)), rep('Clinton',length(date_NC_2020))))

ggplot(data=counts_NC_2020, aes(x=data_date, y=counts, col=group)) +
  geom_point() +
  scale_color_manual(values = c("blue","red")) + labs(x='Date') +
  labs(y='Voter Counts') + labs(title='North Carolina Poll Counts for Trump and Biden 2020')
```




```
counts_NC_separate_2020 = data.frame(data_date = date_NC_2020,
  Trump = NC_total_2020$ sample_size[NC_total_2020$answer=="Trump"]*NC_total_2020$pct[NC_total_2020$ansu
  Biden = NC_total_2020$ sample_size[NC_total_2020$answer=="Biden"]*NC_total_2020$pct[NC_total_2020$ansu

ggplot(data = counts_NC_separate_2020, aes(x=data_date, y=Biden-Trump)) + geom_point() + xlab('Date') +
```




```
    # Percentage difference ggplot method
ggplot(data = counts_NC_separate_2020, aes(x = data_date, y=(Biden-Trump)/(Biden+Trump))) + geom_point()
```




```
    # Linear model of the percentage difference with respect to date of the polls
counts_NC_for_lm_2020 = data.frame(data_date = date_NC_2020,
      percentage_diff = ((NC_total_2020$ sample_size[NC_total_2020$answer=="Biden"]*NC_total_2020$pct[NC

counts_NC_for_lm_2020 = data.frame(data_date = date_NC_2020,
      percentage_diff = ((NC_total_2020$pct[NC_total_2020$answer=="Biden"] - NC_total_2020$ pct[NC_total

lm_model_NC_2020 = lm(percentage_diff ~ (data_date), data = counts_NC_for_lm_2020); lm_model_NC_2020
```

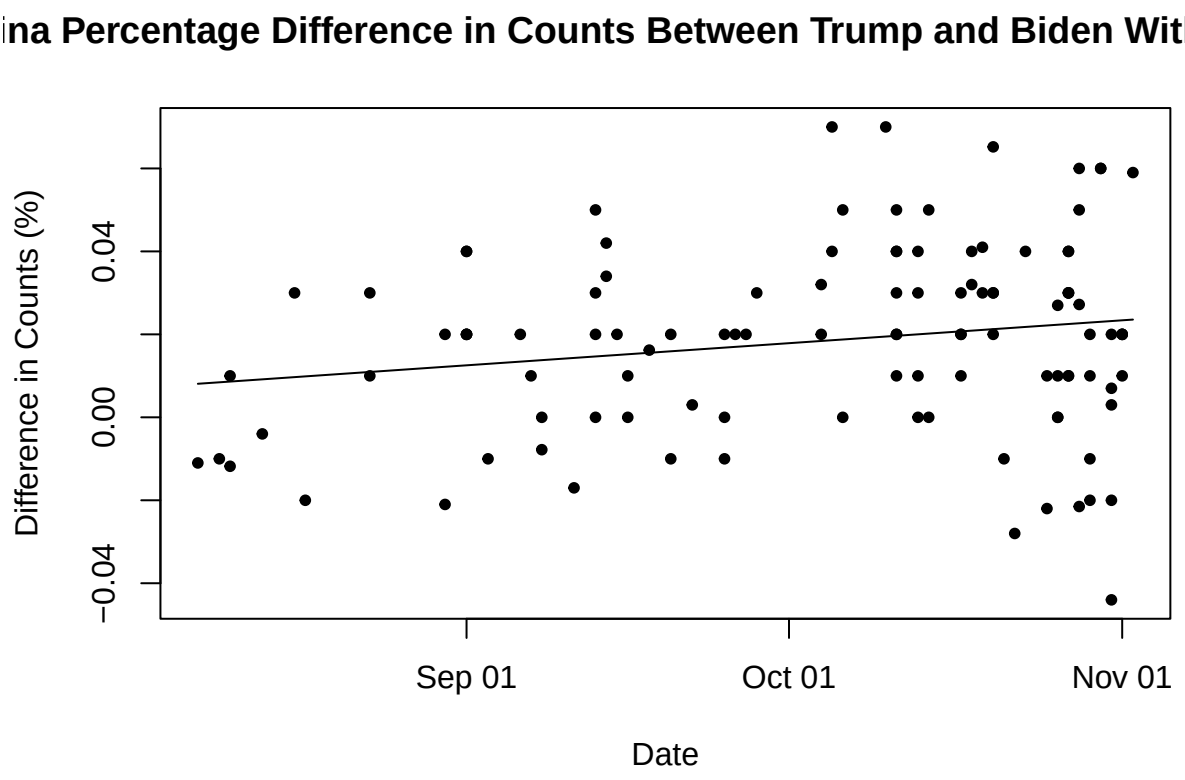
```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_NC_for_lm_2020)
##
## Coefficients:
## (Intercept)    data_date
##  -3.2790761    0.0001779
```

```
summary(lm_model_NC_2020)
```

```
##
## Call:
## lm(formula = percentage_diff ~ (data_date), data = counts_NC_for_lm_2020)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
```

```
## -0.067208 -0.014671  0.001377  0.015329  0.051416
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.279e+00  1.638e+00  -2.002   0.0478 *
## data_date    1.779e-04  8.835e-05   2.013   0.0466 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02256 on 107 degrees of freedom
## Multiple R-squared:  0.0365, Adjusted R-squared:  0.02749
## F-statistic: 4.053 on 1 and 107 DF,  p-value: 0.0466
```

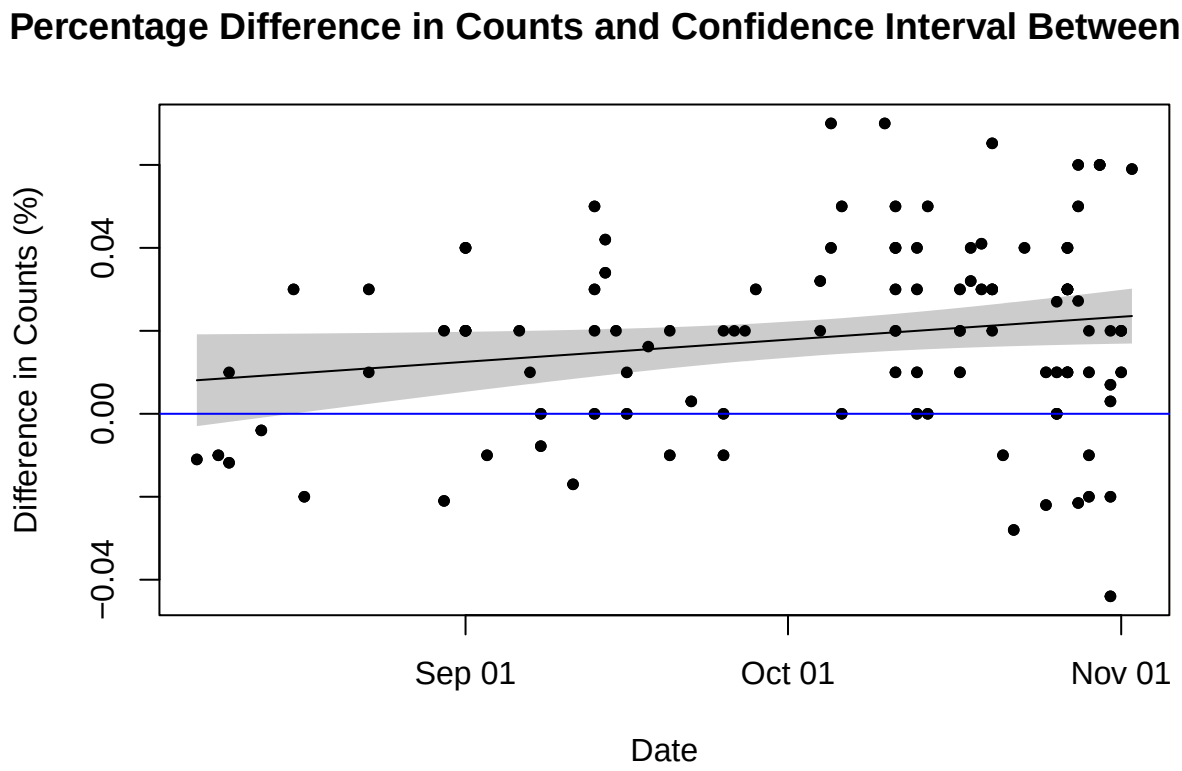
```
    # Plot fitted values of the fitted line
counts_NC_2020 <- data.frame(data_date = c(date_NC_2020, date_NC_2020),
  counts = c(NC_total_2020$ sample_size[NC_total_2020$answer=="Biden"]*NC_total_2020$pct[NC_total_2020$a
  group = c(rep('Trump', length(date_NC_2020)), rep('Biden',length(date_NC_2020)))))

plot(counts_NC_for_lm_2020$data_date, counts_NC_for_lm_2020$percentage_diff,
     col='black', pch=20, type='p', xlab='Date', ylab='Difference in Counts (%)',
     main='North Carolina Percentage Difference in Counts Between Trump and Biden With Fitted Values 202
col='black', pch=20, type='l', xlab='Date', ylab='Difference in Counts (%)',
main='North Carolina')
```



```
    # Plot the confidence interval of the fitted line
fitted_CI_NC_2020 = predict(lm_model_NC_2020,
                            newdata = counts_NC_for_lm_2020,
                            interval = "confidence", level = 0.95)
summary(fitted_CI_NC_2020)
```

```
##       fit                lwr                upr
##  Min.   :0.008089   Min.   :-0.00297   Min.   :0.01915
##  1st Qu.:0.014848   1st Qu.: 0.00929   1st Qu.:0.02041
##  Median :0.019651   Median : 0.01521   Median :0.02410
##  Mean   :0.018445   Mean   : 0.01257   Mean   :0.02432
##  3rd Qu.:0.022319   3rd Qu.: 0.01658   3rd Qu.:0.02805
##  Max.   :0.023564   Max.   : 0.01695   Max.   :0.03018
```

```
plot(counts_NC_for_lm_2020$data_date, counts_NC_for_lm_2020$percentage_diff,
     col='black', pch=20, type='p', xlab = 'Date', ylab = 'Difference in Counts (%)',
     main = 'North Carolina Percentage Difference in Counts and Confidence Interval Between Trump and B
```



From our plot of fitted values we see evidence of a trend in difference in North Carolina counts % and date. We expect that early polls do not have as much impact as recent polls as most polls are concentrated on more recent months. From our linear model of the percentage difference with respect to date of the polls for North Carolina we see a p value of 0.0466 which is less than the acceptable level of significance 0.05, meaning we have sufficient evidence to reject the null hypothesis and conclude the percentage difference in counts for North Carolina is affected by dates.

From our plot with a confidence interval for the fitted line and looking at the confidence interval values only 6 of the 109 confidence intervals in early August contain 0, as the values are negative and positive above

and below 0. However, since we deduced early polls don't have as much of an impact as recent polls we conclude overall the confidence interval for the fitted line doesn't contain 0, indicating there is a difference in counts % for Trump and Biden. This means with repeated trails we expecting a difference in Trump and Biden's North Carolina Difference in Count % with respect to dates of the polls, and this trial is unlikely to be indicative of being the closest election with the least difference in percentage difference.

```
sum(percentage_diff_michigan_2020) # 7.3086
```

```
## [1] 7.3086
```

```
sum(percentage_diff_georgia_2020) # 0.4643
```

```
## [1] 0.4643
```

```
sum(percentage_diff_NC_2020) # 2.0105
```

```
## [1] 2.0105
```

I believe that based on the model and our observations Georgia would have the closest election in terms of the lowest percentage difference between Trump and Biden in 2016. The sum of their percentage differences is 0.4643 which differs less from 0 than Michigan and North Carolina's sum of percentages, meaning Georgia's sum of percentage differences is closest to 0 and has the closest election, Biden beating Trump by 0.4643%. Also, since Georgia's confidence interval model contains 0 and is pretty evenly distributed above and below 0, Georgia is expected to not have a difference in counts % and have close count % in a close election. We deduced Michigan and North Carolina's confidence interval models don't contain 0, so they are expected to have a difference in counts %.

e. From the real results of 2020 election, which state has the smallest margin (in terms of percentage difference)? Discuss at least two reasons that are different than what polls indicate. (You may check Wikipedia for 2020 US presidential election to find out the real voting results for each state.) Looking at the percentage differences of the real poll data from 2020 on Wikipedia I see that for Michigan the percentage difference was 0.0278 where Biden is ahead of Trump by 2.78%, for Georgia the percentage difference was .0023 where Biden is ahead of Trump by .23%, and for North Carolina the percentage difference was -0.0134 where Trump is ahead of Biden by 1.34%.

Therefore, the real poll data is consistent with our calculation of Georgia in part a) as the state with the lowest percentage difference. However, our poll calculations also predicted Biden would win in all three states whereas in reality Trump was ahead of Biden in North Carolina by 1.34% instead of Biden being ahead of Trump by .963% of votes.

Reasons why the polls are different is the margins are random variables which are not going to be ranked the same and the polls might be biased. Such as, differing from normal and having sampling biases of only polling certain areas but generalizing to a whole state or only receiving polling data from a certain strongly opinionated group of people.

f. Do polls correctly predict the candidate who wins these states? Discuss the bias of polls in these states. Name a few possible reasons. The polls didn't correctly predict the candidate who wins these states since it predicted Biden would have more counts but Trump ended up having more counts in North Carolina. There are multiple areas of bias. One is how the polling samples aren't independent and identically distributed. The polls aren't independent because votes for Trump and Biden are dependent on each other as they might change depending on many factors which can change opinions such as debates, current events, and such, so this yields inaccurate prediction results that can vary drastically in different scenarios. Others biases are

polling or sampling errors that inevitably occur and nonresponse errors where individuals don't answer and lead to inaccurate data.

Question 3: Explore the poll data from September 1, 2016 to November 2, 2016 and September 1, 2020 to November 2, 2020 to answer the following questions.

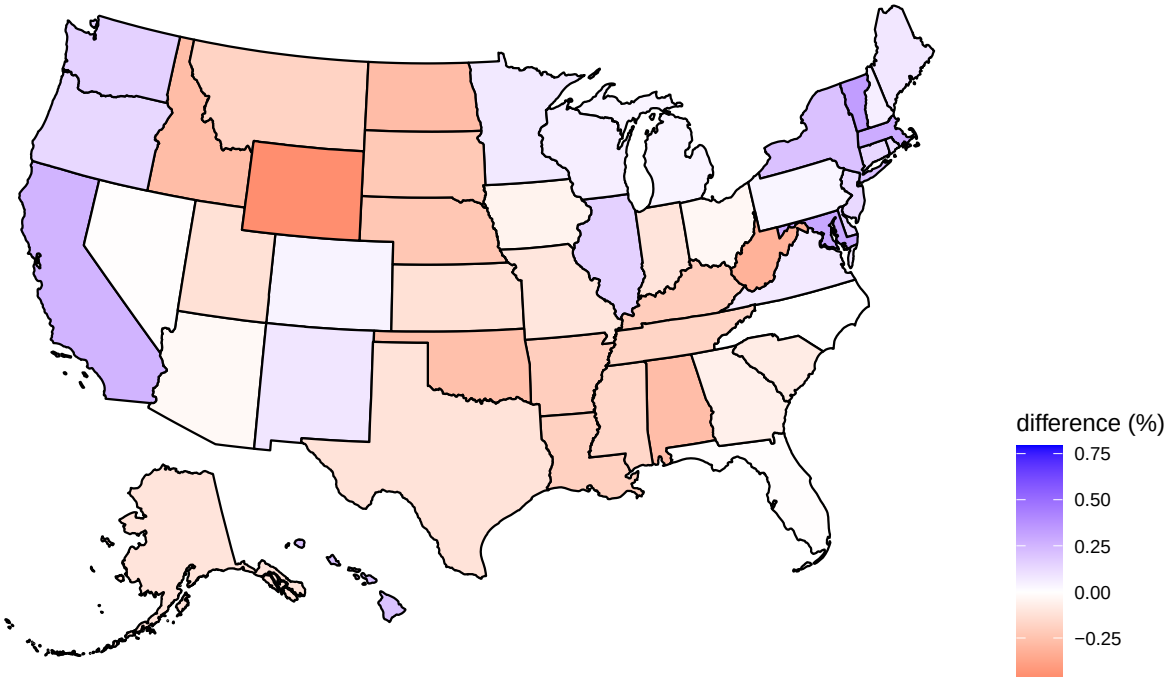
```
  # a. Graph the percentage difference of polls in each state of US for 2016 and
  #    2020. Compare the difference.
polls_data_2016_new <- polls_data_2016[mdy(polls_data_2016$startdate) >= "2016-09-01" & mdy(polls_data_2

poll_state_sum_clinton_2016=aggregate(polls_data_2016_new$total.clinton, by=list(State=polls_data_2016_n
poll_state_sum_trump_2016=aggregate(polls_data_2016_new$total.trump, by=list(State=polls_data_2016_new$s

poll_state_diff_percentage=poll_state_sum_clinton_2016
poll_state_diff_percentage[,2]=(poll_state_sum_clinton_2016[,2]-poll_state_sum_trump_2016[,2])/(poll_sta
delete_index=which((poll_state_diff_percentage[,1])=='U.S.')
poll_state_diff_percentage=poll_state_diff_percentage[-delete_index,]
poll_state_diff_percentage[,1]
```

```
##  [1] "Alabama"              "Alaska"               "Arizona"
##  [4] "Arkansas"             "California"           "Colorado"
##  [7] "Connecticut"          "Delaware"             "District of Columbia"
## [10] "Florida"              "Georgia"              "Hawaii"
## [13] "Idaho"                "Illinois"             "Indiana"
## [16] "Iowa"                 "Kansas"               "Kentucky"
## [19] "Louisiana"            "Maine"                "Maine CD-1"
## [22] "Maine CD-2"           "Maryland"             "Massachusetts"
## [25] "Michigan"             "Minnesota"            "Mississippi"
## [28] "Missouri"             "Montana"              "Nebraska"
## [31] "Nebraska CD-1"        "Nebraska CD-2"        "Nebraska CD-3"
## [34] "Nevada"               "New Hampshire"        "New Jersey"
## [37] "New Mexico"           "New York"             "North Carolina"
## [40] "North Dakota"         "Ohio"                 "Oklahoma"
## [43] "Oregon"               "Pennsylvania"         "Rhode Island"
## [46] "South Carolina"       "South Dakota"         "Tennessee"
## [49] "Texas"                "Utah"                 "Vermont"
## [52] "Virginia"             "Washington"           "West Virginia"
## [55] "Wisconsin"            "Wyoming"
```

```
poll_state_diff_percentage_new <- poll_state_diff_percentage[-c(21,22,31,32,33),]

state_poll_2016 <- data.frame(
  state =poll_state_diff_percentage_new[,1],
  diff_percentage=poll_state_diff_percentage_new[,2])

polls_data_2020_new <- polls_data_2020[mdy(polls_data_2020$start_date) >= "2020-09-01" & mdy(polls_data_

polls_data_2020_new=polls_data_2020_new[which(polls_data_2020$answer=='Biden'|polls_data_2020$answer=='T

index_biden_2020=which(polls_data_2020_new$answer=='Biden')
index_trump_2020=which(polls_data_2020_new$answer=='Trump' )

counts_biden_2020=polls_data_2020$pct[index_biden_2020]*polls_data_2020$sample_size[index_biden_2020]
counts_trump_2020=polls_data_2020$pct[index_trump_2020]*polls_data_2020$sample_size[index_trump_2020]
```

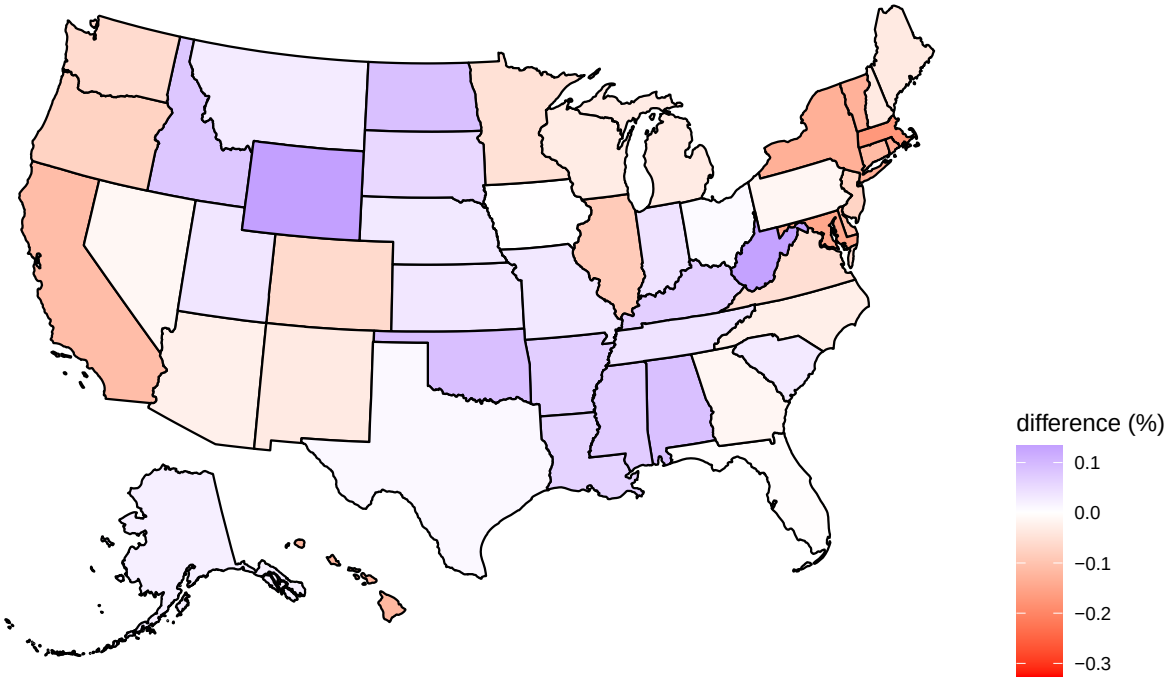
```
polls_data_2020$total.biden=rep(0,dim(polls_data_2020)[1])
polls_data_2020$total.trump=rep(0,dim(polls_data_2020)[1])

polls_data_2020$total.biden[index_biden_2020]=counts_biden_2020
polls_data_2020$total.trump[index_trump_2020]=counts_trump_2020

poll_state_sum_biden_2020=aggregate(polls_data_2020$total.biden, by=list(State=polls_data_2020$state),FU
poll_state_sum_trump_2020=aggregate(polls_data_2020$total.trump, by=list(State=polls_data_2020$state),FU

poll_state_sum_biden_2020=poll_state_sum_biden_2020[-1,]
poll_state_sum_trump_2020=poll_state_sum_trump_2020[-1,]


state_poll_2020 <- data.frame(
  state =poll_state_sum_biden_2020[,1],
  diff_percentage=(poll_state_sum_biden_2020[,2]-poll_state_sum_trump_2020[,2])/(poll_state_sum_biden_2


limit_val=c(min(state_poll_2016$diff_percentage,state_poll_2020$diff_percentage), max(state_poll_2016$di

#install.packages("usmap")
library(usmap)

usmap_2016 <- plot_usmap(data = state_poll_2016, values = "diff_percentage", color = "black") +
  scale_fill_gradient2(name = "difference (%)",   low= "red",
                       mid = "white",
                       high = "blue",
                       midpoint = 0,limits=limit_val)+
  theme(legend.position = "right")+
ggtitle("2016"); usmap_2016
```

2016




```
plot_usmap(data = state_poll_2020, values = "diff_percentage", color = "black") +
  scale_fill_gradient2(name = "difference (%)",   low= "red",
                       mid = "white",
                       high = "blue",
                       midpoint = 0,limits=limit_val)+
  theme(legend.position = "right")+
ggtitle("2020")
```

2020




```
  # b. Name 5 battleground states (states with closest percentage difference
  #    between two candidates) in 2020 based on the plots for (a).
#install.packages("dplyr")
library(dplyr)
state_poll_2020_sorted <- arrange(state_poll_2020, state_poll_2020$diff_percentage)

# Looking at our us map of the percentage difference of polls in each state of the
# US for 2020 we see that the lightest colors indicate the lowest % being Florida,
# Maine, Iowa, Ohio, and Texas. Checking the state_poll_2020 sorted from lowest
# to largest difference in percentage values I see that Ohio has the closest
# percentage difference of 0.0007197018, and the two states above and below
# this value was Maine with Trump winning by .222%, Florida with Trump winning by
# .305%, Ohio with Biden winning by .700%, and Texas with Biden winning by .981%.
```

```
  # c. Compare the difference of the polls in 2016 and in 2020 for states in US.
state_poll_2016$state
```

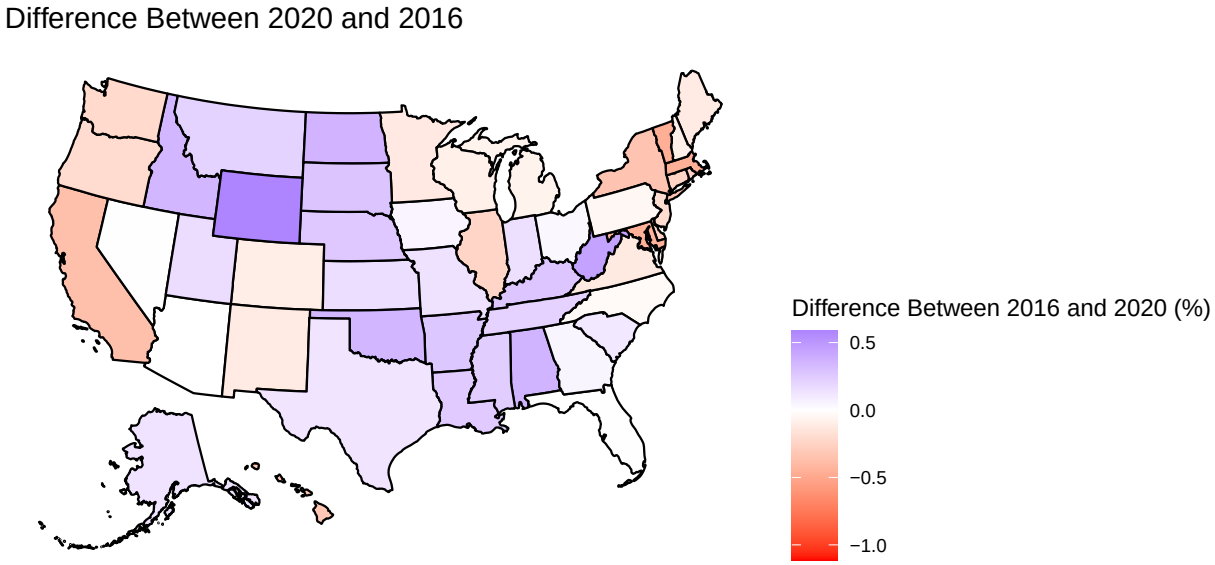
```
##  [1] "Alabama"              "Alaska"               "Arizona"
##  [4] "Arkansas"             "California"           "Colorado"
##  [7] "Connecticut"          "Delaware"             "District of Columbia"
## [10] "Florida"              "Georgia"              "Hawaii"
## [13] "Idaho"                "Illinois"             "Indiana"
## [16] "Iowa"                 "Kansas"               "Kentucky"
## [19] "Louisiana"            "Maine"                "Maryland"
## [22] "Massachusetts"        "Michigan"             "Minnesota"
```

```
## [25] "Mississippi"          "Missouri"            "Montana"
## [28] "Nebraska"             "Nevada"              "New Hampshire"
## [31] "New Jersey"           "New Mexico"          "New York"
## [34] "North Carolina"       "North Dakota"        "Ohio"
## [37] "Oklahoma"             "Oregon"              "Pennsylvania"
## [40] "Rhode Island"         "South Carolina"      "South Dakota"
## [43] "Tennessee"            "Texas"               "Utah"
## [46] "Vermont"              "Virginia"            "Washington"
## [49] "West Virginia"        "Wisconsin"           "Wyoming"
```

```
state_poll_2020$state
```

```
##  [1] "Alabama"              "Alaska"              "Arizona"
##  [4] "Arkansas"             "California"          "Colorado"
##  [7] "Connecticut"          "Delaware"            "District of Columbia"
## [10] "Florida"              "Georgia"             "Hawaii"
## [13] "Idaho"                "Illinois"            "Indiana"
## [16] "Iowa"                 "Kansas"              "Kentucky"
## [19] "Louisiana"            "Maine"               "Maine CD-1"
## [22] "Maine CD-2"           "Maryland"            "Massachusetts"
## [25] "Michigan"             "Minnesota"           "Mississippi"
## [28] "Missouri"             "Montana"             "Nebraska"
## [31] "Nebraska CD-1"        "Nebraska CD-2"       "Nevada"
## [34] "New Hampshire"        "New Jersey"          "New Mexico"
## [37] "New York"             "North Carolina"      "North Dakota"
## [40] "Ohio"                 "Oklahoma"            "Oregon"
## [43] "Pennsylvania"         "Rhode Island"        "South Carolina"
## [46] "South Dakota"         "Tennessee"           "Texas"
## [49] "Utah"                 "Vermont"             "Virginia"
## [52] "Washington"           "West Virginia"       "Wisconsin"
## [55] "Wyoming"
```

```
state_poll_2020=state_poll_2020[-c(21,22,31,32),]

state_poll_2020_2016_diff <- data.frame(
  state =state_poll_2020$state,
  diff=state_poll_2020$diff_percentage-state_poll_2016$diff_percentage)

plot_usmap(data = state_poll_2020_2016_diff, values = "diff", color = "black") +
  scale_fill_gradient2(name = "Difference Between 2016 and 2020 (%)",   low= "red",
                       mid = "white",
                       high = "blue",
                       midpoint = 0)+
  theme(legend.position = "right")+
  ggtitle("Difference Between 2020 and 2016")
```

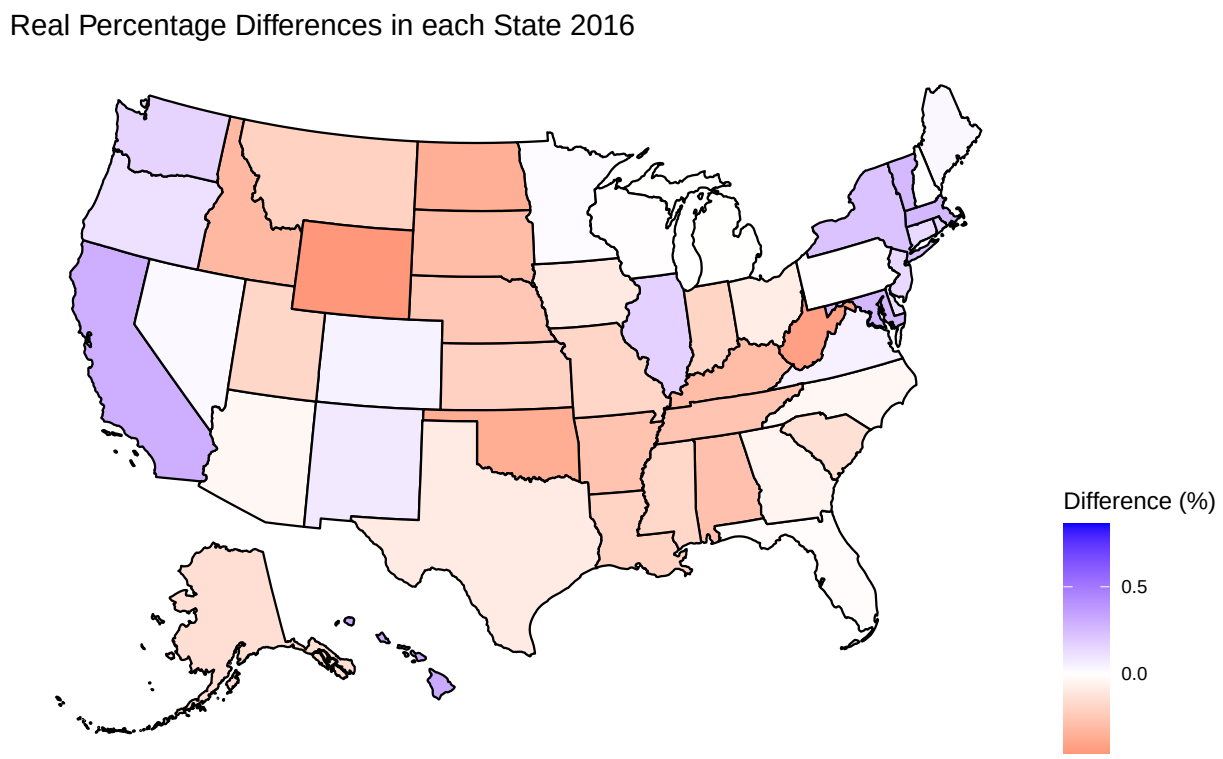



```
state_poll_2020_2016_diff_sorted <- arrange(state_poll_2020_2016_diff, state_poll_2020_2016_diff$diff)
```

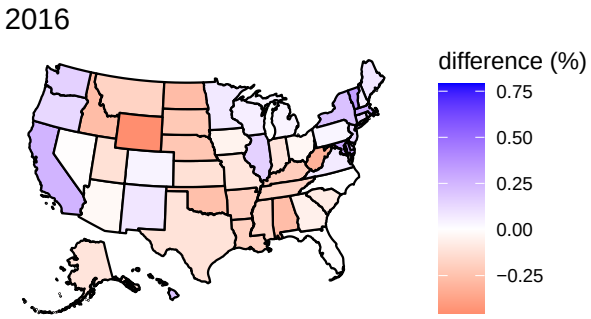
We see that the states with the lowest percentage differences or lightest colors in the US Map between 2020 and 2016 are Arizona, Iowa, and Florida. Checking the state_poll_2020_2016_diff sorted from lowest to largest difference in 2016 and 2020 percentage difference values, we see that Arizona has the closest percentage difference between 2020 and 2016 of 0.0009719987, whereas in Oklahoma the 2020 percentage difference was 4.235% greater compared to 2016 and in Florida the 2020 percentage difference was .43% greater compared to 2016. In the map we see mostly purple values indicated most of the percentage differences between 2016 and 2020 were positive meaning the percentage difference in 2020 was greater than the percentage difference in 2016, and Biden was mostly in the lead in 2020.

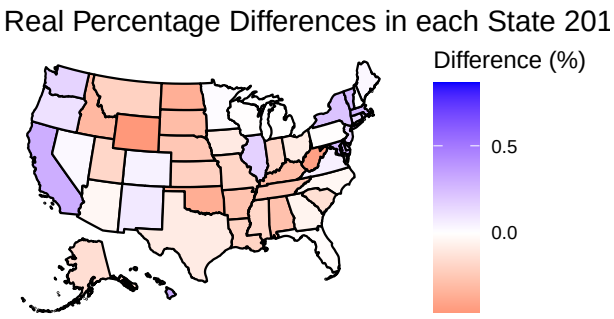
```
  # d. Do polls underestimate the percentage of the real votes (in terms of
  # percentage) received from one candidate in 2016? How about 2020? Discuss
  # some reasons that may explain the bias in polls.
#install.packages("gridExtra")
library(gridExtra)
data_frame_real_2016 <- data.frame(
  state = polls_data_real_2016$state,
  diff = (polls_data_real_2016$dem-polls_data_real_2016$rep)
)

usmap_real_2016 <- plot_usmap(data = data_frame_real_2016, values = "diff", color = "black") +
  scale_fill_gradient2(name = "Difference (%)",   low= "red",
                       mid = "white",
                       high = "blue",
```

```
                      midpoint = 0)+
  theme(legend.position = "right")+
  ggtitle("Real Percentage Differences in each State 2016"); usmap_real_2016
```




```
grid.arrange(usmap_2016, usmap_real_2016, ncol=2)
```






```
state_poll_2016$diff_percentage - data_frame_real_2016$diff
```

```
##  [1] -0.122491050  0.166649264  0.243300853 -0.165172404 -0.047443803
##  [6] -0.009185203 -0.001781699 -0.719226785  0.674990078  0.004626412
## [11] -0.009901390 -0.109612989 -0.177126672  0.471818341 -0.281566352
## [16]  0.141041980  0.081303763  0.092167181  0.005620097 -0.188949969
## [21]  0.034170402  0.250947917  0.039811903  0.059810384  0.021092767
## [26]  0.074710923  0.023564705 -0.194347001  0.349962177  0.306424521
## [31]  0.123692957 -0.057463858  0.125474901 -0.022838999 -0.496120269
## [36]  0.048398091  0.103165758  0.021126216  0.041409298 -0.037372175
## [41]  0.062701560  0.074617628  0.086017557 -0.026687985  0.054714780
## [46]  0.278089825 -0.188945823 -0.001646209 -0.309002595  0.477386747
## [51]  0.001630536
```

Looking at the us maps of real percentage differences in each state for both the real results and polls, we see that more states especially in the North-East part of the U.S. are closer to white than blue in the real percentage difference map compared to the poll percentage differences in 2016. Additionally, the sum of the poll's percentage differences in 2016 is much less with -188.892 compared to the sum of the real percentage differences in 2016. Therefore, we can say that the the polls overestimated the percentage of the real votes in 2016.