Nikhil-Gopal-Exam

Nikhil Gopal

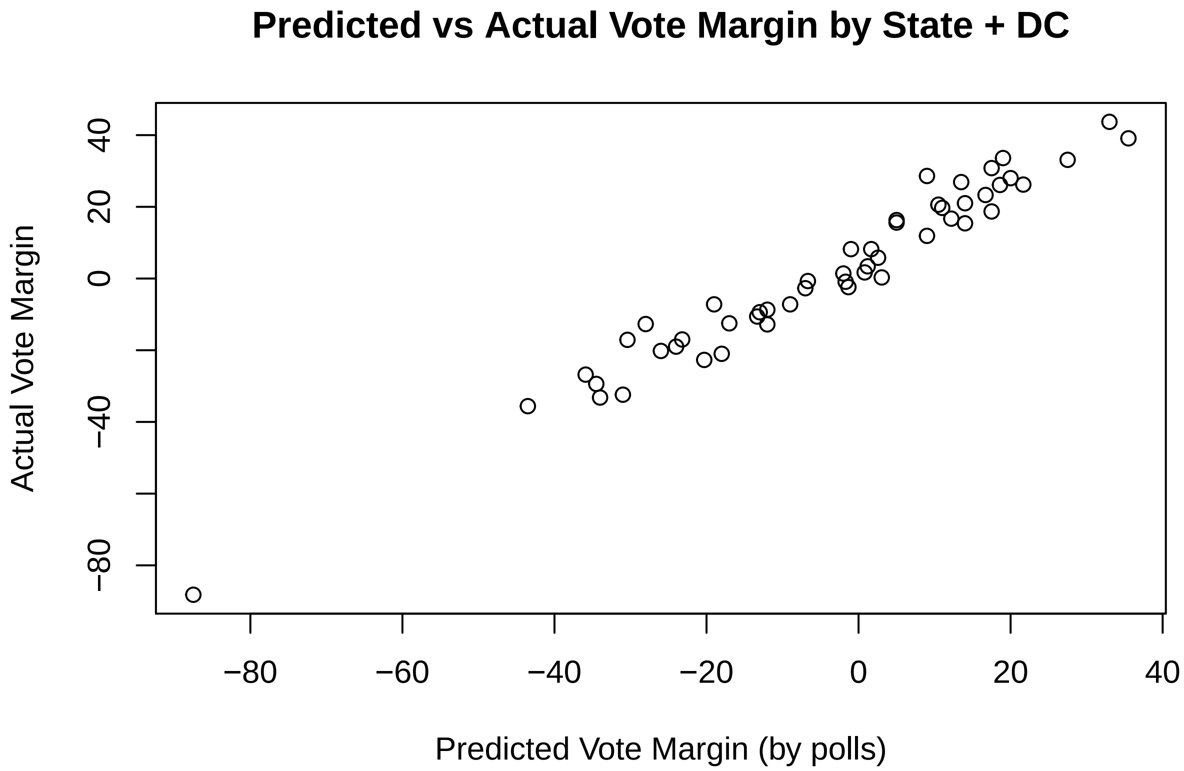
11/6/2020

Question 1.a.i

```r
polls <- read.csv("polls_2020.csv")
pres <- read.csv("vote_2020.csv")

polls$daysleft <- as.Date("2020-11-3")-as.Date(polls$middate)

pres$margin <- pres$trump - pres$biden

#Question 1.a.i Enter command to create variable for the poll margin here
polls$margin <- polls$trump - polls$biden

st.names <- unique(sort(polls$state))
poll.pred <- rep(NA, 51)

for(i in 1:51) {
  state.data <- subset(polls, subset = (state == st.names[i]))
  latest <- state.data$daysleft == min(state.data$daysleft)
  poll.pred[i] <- mean(state.data$margin[latest])
}
```

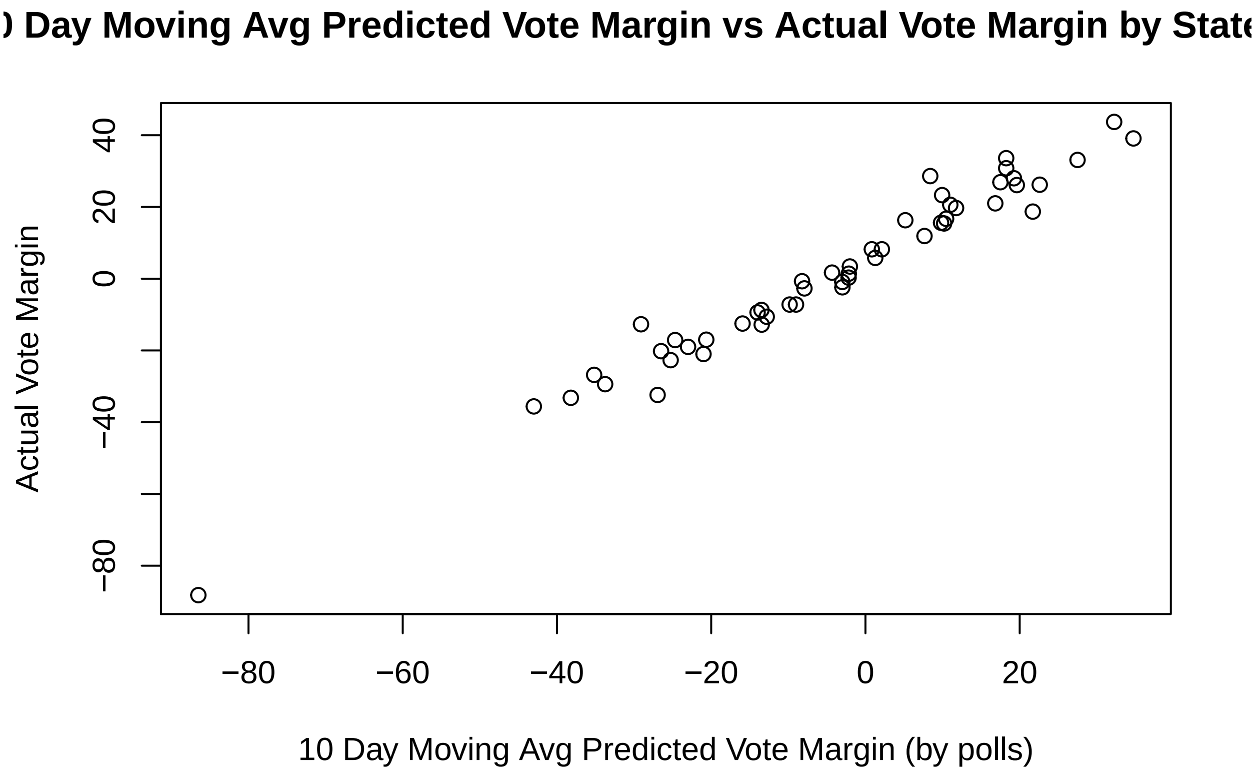
Question 1.a.ii

```r
#Question a ii
plot(x = poll.pred, y=pres$margin, main = "Predicted vs Actual Vote Margin by State + DC", xlab = "Predi
```



Question 1.a.iii

```r
#Question a iii

#empty vector to hold moving averages for states
poll.pred.ten.day.avg <- rep(NA, 51)

for(i in 1:51) {
  #subset for  a state at the given state's index
  state.data <- subset(polls, subset = (state == st.names[i]))
  #object containing the latest poll for each state
  latest <- state.data$daysleft == min(state.data$daysleft)

  poll.pred[i] <- mean(state.data$margin[latest])

  #subset again to only the 10 days before the election for a given state
  avg.data <- subset(state.data, subset = ((daysleft <= 10) & (daysleft >= 1) ))

  #reassign poll prediction variable to the moving average
  poll.pred.ten.day.avg[i] <- mean(avg.data$margin)

}

plot(x = poll.pred.ten.day.avg, y=pres$margin, main = "10 Day Moving Avg Predicted Vote Margin vs Actual
```



The scatter plot does not look that different (only slight differences between the 2)

Question 1.b.i

```
#Question b  i

#empty vector to hold each state's + DC's prediction error
prediction_error_vector <- rep(NA, 51)

state_names <- unique(pres$state)

#iterate thru each state and calculate prediction error, save to vector
for(i in 1:51) {
  state.data <- subset(polls, subset = (state == st.names[i]))
  latest <- state.data$daysleft == min(state.data$daysleft)
  poll.pred[i] <- mean(state.data$margin[latest])

  #create a variable for the actual margin, at the state currently being indexed
  actual_margin <- pres$margin[pres$state==st.names[i]]

  #assign error to vector at the current index being iterated
  prediction_error_vector[i] <- actual_margin - poll.pred.ten.day.avg[i]

}

prediction_error_vector
```

```
##  [1] 20.2000000  6.4750000  8.7500000  0.5888889 -5.4500000  0.6500000
##  [7]  7.5800000 -1.7000000  4.0000000  5.4207500  2.4850000  4.3500000
## [13]  7.3750000 12.5500000  3.4375000  6.2428571  5.2000000  9.4000000
## [19] -3.0000000  5.0000000  8.3800000  4.7857143  5.2133333  2.6307692
## [25]  5.8000000  4.1666667 11.1333333  3.5591429 15.3500000  7.9500000
## [31]  1.8000000  0.0000000  2.2000000  2.1000000 16.4000000  6.0705882
## [37]  5.6000000  3.6400000  6.0333333  6.3000000  4.2470000  3.6000000
## [43] 13.3600000  4.5285714  9.6000000  4.5875000  7.4000000  2.5600000
## [49]  7.5200000  4.3500000 11.4500000
```

Question 1.b.ii

```r
#Question b ii

#Average prediction error

avg_prediction_error <- mean(prediction_error_vector)
#5.72296

#Root mean-squared error

rmse <- sqrt(mean(prediction_error_vector^2))
#7.361366
```

Question 1.b.iii

The polls were not that good at predicting the outcome of the state elections. The average error was about 5.6%, meaning that the polls on average were off by about 5.6% in each state. The root mean standard error was 7.361366, which will tell us if there were lots of big positive values offset by negative ones. This does not appear to be the case.

Possible sources of sampling bias may include Trump voters lying in their responses and saying they would vote for Biden or other candidates. Another source of variability could be Trump voters declining to respond to polls (non response).

Question 2.a.i

```r
#Question 2 a i
blackturnout <- read.csv("blackturnout.csv")

dim(blackturnout)
```

```
## [1] 1237    7
```

```r
#1237 observations, 7 variables
```

There are 1237 observations and 7 variables.

Question 2.a.ii

```r
#Question 2aii

num_districts_with_black_candidates <- 0
```

```r
#iterate thru column and add 1 to total # black candidates variable if candidate is black
for(candidate in blackturnout$candidate){
  if(candidate == 1){
    num_districts_with_black_candidates <- num_districts_with_black_candidates + 1
  }
}

num_districts_with_black_candidates
```

```
## [1] 148
```

```r
#there are 148 black candidates
```

There are 148 districts with black candidates.

Question 2.a.iii

```r
#question 2a iii
length(unique(blackturnout$state))
```

```
## [1] 42
```

```r
#42 states included in the dataset
```

There are 42 states in the dataset.

Question 2.b.i

```r
#Question 2b i

#descriptive stats for entire dataset
summary(blackturnout$turnout)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.07749 0.27149 0.35933 0.40123 0.51671 0.97707
```

```r
#descriptive stats for black candidates
summary(blackturnout$turnout[blackturnout$candidate==1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##  0.1992  0.3261  0.4034  0.4555  0.5859  0.8532
```

```r
#descriptive stats for non-black candidates
summary(blackturnout$turnout[blackturnout$candidate==0])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## 0.07749 0.26569 0.35117 0.39386 0.50966 0.97707
```

```r
#difference in average turnout (of all races, not just black voters) between elections with black and n
mean(blackturnout$turnout[blackturnout$candidate==1]) - mean(blackturnout$turnout[blackturnout$candidat
```

```
## [1] 0.06164014
```

Elections with black candidates have a higher turnout than elections with non-black candidates. Mean turnout was 0.4555 for black candidates vs 0.3938 for non black candidates and 0.4 is the dataset average. Median was also slightly higher in elections with black candidates.

Question 2.b.ii

```r
#Question 2b ii

#variance
var(blackturnout$turnout)
```

```
## [1] 0.0296287
```

```r
#variance black candidates
var(blackturnout$turnout[blackturnout$candidate==1])
```

```
## [1] 0.02780168
```

```r
#variance non black candidates
var(blackturnout$turnout[blackturnout$candidate==0])
```

```
## [1] 0.02944777
```

```r
#standard deviation
sd(blackturnout$turnout)
```

```
## [1] 0.1721299
```

```r
#standard deviation of turnout for black candidates
sd(blackturnout$turnout[blackturnout$candidate==1])
```

```
## [1] 0.1667384
```

```r
#standard deviation of turnout for non black candidates
sd(blackturnout$turnout[blackturnout$candidate==0])
```

```
## [1] 0.1716035
```

```r
#minimum turnout in the dataset
min(blackturnout$turnout)
```

```
## [1] 0.07749077
```

```
#minimum turnout in an election with a black candidate
min(blackturnout$turnout[blackturnout$candidate==1])
```

```
## [1] 0.1992127
```

```
#minimum turnout in an election with a non black candidate
min(blackturnout$turnout[blackturnout$candidate==0])
```

```
## [1] 0.07749077
```

```
#maximum turnout in the dataset
max(blackturnout$turnout)
```

```
## [1] 0.977068
```

```
#maximum turnout in an election with a black candidate
max(blackturnout$turnout[blackturnout$candidate==1])
```

```
## [1] 0.8531716
```

```
#maximum turnout in an election with a non black candidate
max(blackturnout$turnout[blackturnout$candidate==0])
```

```
## [1] 0.977068
```

```
#Turnout IQR in the entire dataset
IQR(blackturnout$turnout)
```

```
## [1] 0.2452179
```

```
#Turnout IQR for districts with black candidates
IQR(blackturnout$turnout[blackturnout$candidate==1])
```

```
## [1] 0.2598133
```

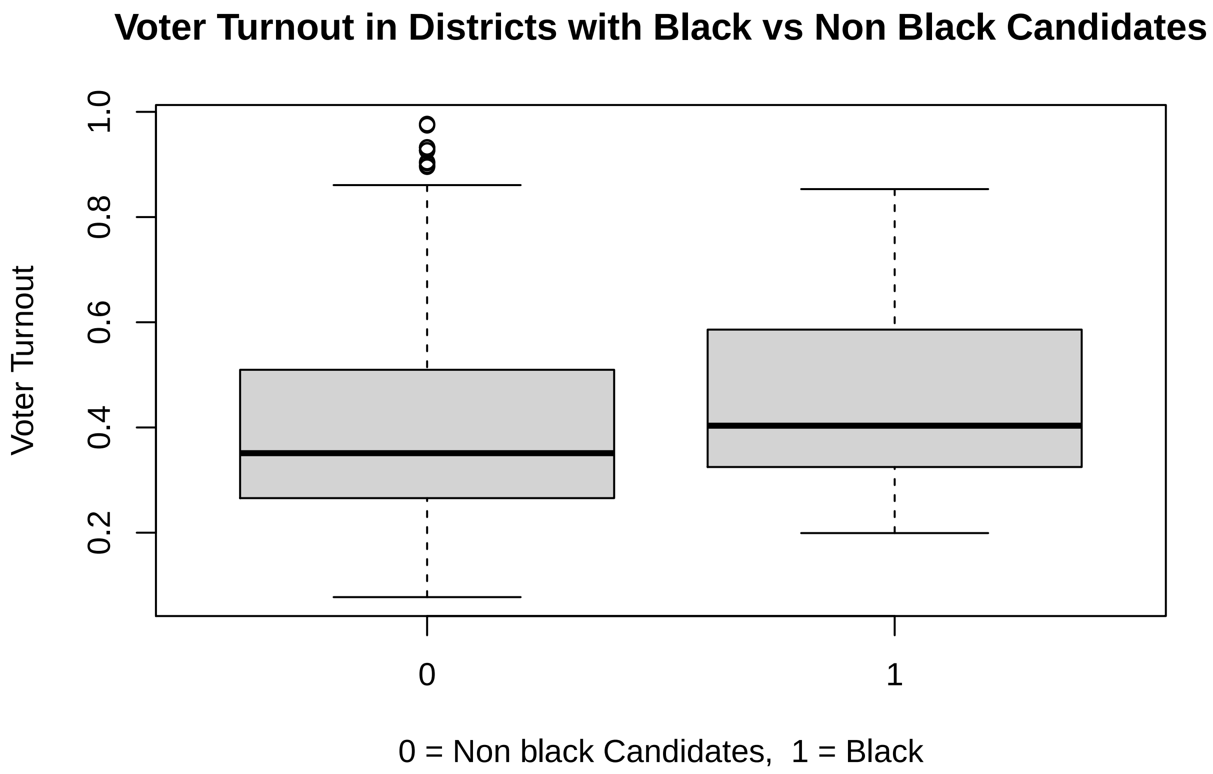
```
#Turnout IQR for districts with non black candidates
IQR(blackturnout$turnout[blackturnout$candidate==0])
```

```
## [1] 0.2439644
```

Turnout variance, IQR and standard deviation were generally the same for districts with black vs non black candidates. Minimum turnout in districts with black candidates was 0.197 compared to 0.077 for districts without black candidates. Maximum turnout was 0.853 in districts with black candidates compared to 0.977 for districts without black candidates.

Question 2.b.iii

```
#Question biii

boxplot(turnout~candidate, data = blackturnout, xlab = "0 = Non black Candidates,  1 = Black", ylab = "V
```



Question 2.c.i

```
#Question c i

turnout_linear_model <- lm(turnout~candidate, data = blackturnout)
```

The outcome variable is the voter turnout and the explanatory variable is candidate race (black vs non black candidates).

Question 2.c.ii

```
#Question c ii

summary(turnout_linear_model)
```

```
##
## Call:
## lm(formula = turnout ~ candidate, data = blackturnout)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.3164 -0.1282 -0.0436  0.1191  0.5832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.393857   0.005183  75.993  < 2e-16 ***
## candidate   0.061640   0.014984   4.114 4.15e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.171 on 1235 degrees of freedom
## Multiple R-squared:  0.01352,    Adjusted R-squared:  0.01272
## F-statistic: 16.92 on 1 and 1235 DF,  p-value: 4.149e-05
```

The estimate of alpha is about 0.39386. This intercept value is what the model estimates turnout to be when the explanatory variable is equal to zero. That is also, what the model estimates the turnout to be when the candidate is not blackturnout

Question 2.c.iii

```r
#Question c iii

summary(turnout_linear_model)
```

```
##
## Call:
## lm(formula = turnout ~ candidate, data = blackturnout)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3164 -0.1282 -0.0436  0.1191  0.5832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.393857   0.005183  75.993  < 2e-16 ***
## candidate   0.061640   0.014984   4.114 4.15e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.171 on 1235 degrees of freedom
## Multiple R-squared:  0.01352,    Adjusted R-squared:  0.01272
## F-statistic: 16.92 on 1 and 1235 DF,  p-value: 4.149e-05
```

The beta coefficient is about 0.06164. This tells us that the model predicts having a black candidate increases turnout by about 0.06164. This value for slope is in accordance with our mean values observed in question 2bi, mean turnout values were higher in elections with black candidates vs those with non black candidates. Thus, the beta coefficient should cause an increase in turnout when the candidate is black.

Question 2.c.iv

We should NOT interpret the estimate of beta as the causal effect of black candidates on the turnout of black voters. First, we do not know the races of the individual voters in the dataset, only the races of the candidates. Thus, black candidates could increase turnout among non-black voters. Additionally, correlation does not equal causation, meaning that there could be a correlation between black candidates and higher turnout, but this correlation does not necessarily mean that black candidates cause higher turnout. This higher turnout could simply be a product of random chance, and our Beta coefficient does not prove or disprove that.

Question 2.c.v

```
#Question c v

summary(turnout_linear_model)
```

```
##
## Call:
## lm(formula = turnout ~ candidate, data = blackturnout)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3164 -0.1282 -0.0436  0.1191  0.5832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.393857   0.005183  75.993  < 2e-16 ***
## candidate   0.061640   0.014984   4.114 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.171 on 1235 degrees of freedom
## Multiple R-squared:  0.01352,    Adjusted R-squared:  0.01272
## F-statistic: 16.92 on 1 and 1235 DF,  p-value: 4.149e-05
```

The R squared of this model was 0.01352, meaning that only about 1.3% of the variability in the data can be explained by the model. This model does not fit the data well at all.

Question d.i

```
#Question d i

turnout_linear_model_with_CVAP <- lm(turnout~candidate+CVAP, data = blackturnout)

summary(turnout_linear_model)
```

```
##
## Call:
## lm(formula = turnout ~ candidate, data = blackturnout)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3164 -0.1282 -0.0436  0.1191  0.5832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.393857   0.005183  75.993  < 2e-16 ***
## candidate   0.061640   0.014984   4.114 4.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.171 on 1235 degrees of freedom
## Multiple R-squared:  0.01352,    Adjusted R-squared:  0.01272
## F-statistic: 16.92 on 1 and 1235 DF,  p-value: 4.149e-05
```

The estimate of beta2 is 0.207392. This means that turnout is predicted to increase by (0.207392*0.10) = 0.0207392 when the black voting age population in a district increase by 0.1.

Question d.ii

```
#Question d ii

summary(turnout_linear_model)
```

```
##
## Call:
## lm(formula = turnout ~ candidate, data = blackturnout)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.3164 -0.1282 -0.0436  0.1191  0.5832
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.393857   0.005183  75.993  < 2e-16 ***
## candidate   0.061640   0.014984   4.114 4.15e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.171 on 1235 degrees of freedom
## Multiple R-squared:  0.01352,    Adjusted R-squared:  0.01272
## F-statistic: 16.92 on 1 and 1235 DF,  p-value: 4.149e-05
```

```
summary(turnout_linear_model_with_CVAP)
```

```
##
## Call:
## lm(formula = turnout ~ candidate + CVAP, data = blackturnout)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30534 -0.12775 -0.04529  0.11750  0.59576
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.375275   0.006677  56.203  < 2e-16 ***
## candidate   -0.007364   0.021703  -0.339    0.734
## CVAP         0.207392   0.047497   4.366 1.37e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1698 on 1234 degrees of freedom
## Multiple R-squared:  0.02853,    Adjusted R-squared:  0.02695
## F-statistic: 18.12 on 2 and 1234 DF,  p-value: 1.756e-08
```

The beta1 coefficient estimate is -0.007364, compared to 0.06164 in the model from question 2.c.iii. In the model that includes CVAP, more weight is placed on the proportion of black voters in a district compared to weather the candidate is black. The estimates differ because one model tries to model the relationship between candidate race and turnout, and the other tries to model the relationship between candidate race and proportion of black voters in the district on voter turnout. Since proportion of black voters in a district appears to have a big affect on turnout, the proportional effect of candidate race decreases, which is why the beta coefficients are different in the two models.