A3_109006240

Jansen

2023-12-21

Package Loading

```
require(lubridate)
```

```
## Loading required package: lubridate
```

```
##
## Attaching package: 'lubridate'
```

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

```
require(gapminder)
```

```
## Loading required package: gapminder
```

```
require(readr)
```

```
## Loading required package: readr
```

```
require(knitr)
```

```
## Loading required package: knitr
```

```
require(dplyr)
```

```
## Loading required package: dplyr
```

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

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

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

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats 1.0.0     v stringr 1.5.0
## v ggplot2 3.4.4     v tibble  3.2.1
## v purrr   1.0.2     v tidyr   1.3.0
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
require(tidyr)
require(ggplot2)
require(broom)
```

```
## Loading required package: broom
```

```
require(Metrics)
```

```
## Loading required package: Metrics
```

```
## Warning: package 'Metrics' was built under R version 4.3.2
```

Question 1

first load the .csv

```
turnout <- read_csv('blackturnout.csv')
```

```
## Rows: 1237 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): state
## dbl (5): year, district, black_turnout, black_share, black_candidate
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
turnout
```

```
## # A tibble: 1,237 x 6
##     year state district black_turnout black_share black_candidate
##    <dbl> <chr>    <dbl>         <dbl>       <dbl>           <dbl>
```

```
##  1  2008 AK           0         0.710      0.0350              0
##  2  2010 AK           0         0.448      0.0323              0
##  3  2010 AK           1         0.448      0.0323              0
##  4  2008 AK           1         0.710      0.0350              0
##  5  2006 AK           1         0.439      0.0318              0
##  6  2010 AL           0         0.397      0.256               0
##  7  2008 AL           0         0.626      0.253               1
##  8  2010 AL           1         0.374      0.261               0
##  9  2006 AL           1         0.266      0.261               0
## 10  2008 AL           1         0.620      0.261               0
## # i 1,227 more rows
```

as we can see below there are 42 unique state and the data include year 2006 2008 and 2010

```
length(unique(turnout$state))
```

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

```
unique(turnout$year)
```

```
## [1] 2008 2010 2006
```

Question 2

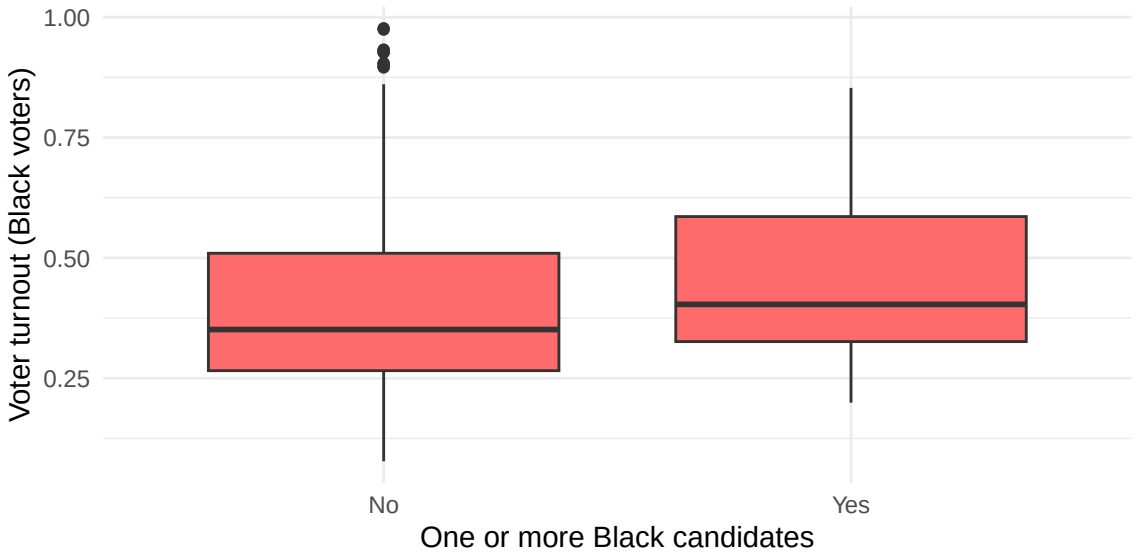
first use mutate and if else function to create a column called is_black_candidate using black_candidate then create a ggplot and create the boxplot using the given fill label and theme

and we can see from the generated boxplot that the rate of yes is higher than no

```
turnout <- turnout %>%
  mutate(is_black_candidate = ifelse(black_candidate == 1, "Yes", "No"))

turnout_box <- ggplot(turnout, aes(x = is_black_candidate, y = black_turnout)) +
  geom_boxplot(fill = "indianred1") +
  labs(x = "One or more Black candidates", y = "Voter turnout (Black voters)") +
  theme_minimal()

turnout_box
```



Question 3

for every unit changed in the black_candidate means that there are 0.05 changes in the black_turnout value. and since the value is positive it means that the turnout will be higher when a co ethnic is running.

the r squared shows how well the model fitted using the data we have. and since the r squared is really low (around 1%) it means that a really small portion of the variability can be explained by the model it self

```
lm_1 <- lm(black_turnout ~ black_candidate, data = turnout)

lm_1 %>%
  broom::tidy() %>%
  select(term, estimate) %>%
  knitr::kable(digits = 2)
```

term	estimate
(Intercept)	0.39
black_candidate	0.06

```
r_squared <- summary(lm_1)$r.squared

r_squared
```

```
## [1] 0.01351812
```

Question 4

1. What does this graph imply about the relationship between Black voting-age population and Black turnout?

   Black turnout -> How many people that can vote and they acttualy vote

Black share -> How many of them can vote

from the graph below we can see a relation between the turn out and the shares. we can see that the higher the share the higher the turnout which means the higher number of people that can vote, there are higher percentage of them that use their vote, i could say this because i treat the data above 0.875 as an outliers because most of the turnout are from 0.125 to 0.0875

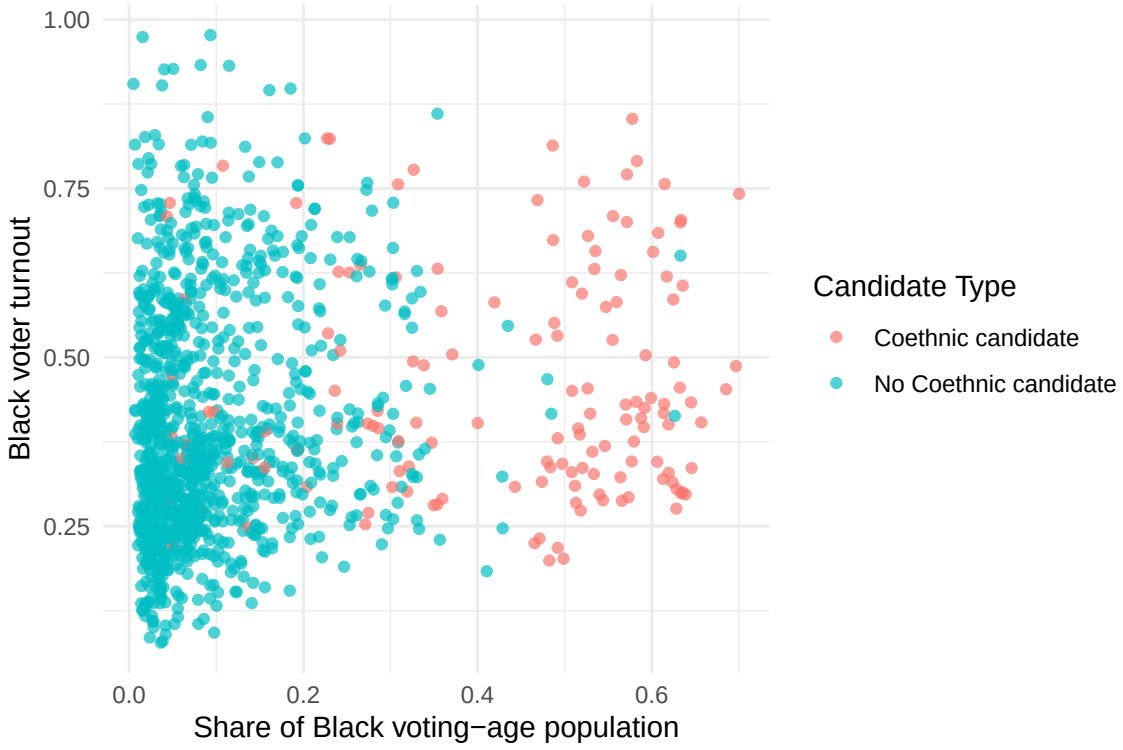
2. What does it inform us about the relationship between Black voting-age population and the presence of a Black candidate?

   existing black candidate actually makes them use their vote. we can see that when there are no co ethnic there are a bunch of turnout that is lower than 0.1875 but as the share growth and there are co ethnic candidate the higher is the turnout.

```
turnout <- turnout %>%
  mutate(candidate_type = ifelse(black_candidate == 1, "Coethnic candidate", "No Coethnic candidate"))

# Create the scatter plot
turnout_scatter <- ggplot(turnout, aes(x = black_share, y = black_turnout, color = candidate_type)) +
  geom_point(alpha = 0.7) +
  labs(x = "Share of Black voting-age population", y = "Black voter turnout", color = "Candidate Type") +
  theme_minimal()

turnout_scatter
```



Question 5

for every unit changed in the black_share means that there are 0.2 changes in the black_turnout value. and since the value is positive it means that the turnout will be higher when the share is higher

the r squared shows how well the model fitted using the data we have. and since the r squared is really low (around 2.8%) it means that a really small portion of the variability can be explained by the model it self

since the RMSE is really low it means that the prediction is close to the actual value and have a high precision which means that the model is very good on predicting

```
lm_2 <- lm(black_turnout ~ black_share, data = turnout)

lm_2 %>%
  broom::tidy() %>%
  select(term, estimate) %>%
  knitr::kable(digits = 2)
```

term	estimate
(Intercept)	0.38
black_share	0.20

```
r_squared_lm2 <- summary(lm_2)$r.squared

r_squared_lm2
```

```
## [1] 0.028437
```

```
rmse_lm1 <- rmse(lm_1$fitted.values, turnout$black_turnout)
rmse_lm2 <- rmse(lm_2$fitted.values, turnout$black_turnout)

rmse_lm1
```

```
## [1] 0.1708934
```

```
rmse_lm2
```

```
## [1] 0.1695962
```

Question 6

for this part now we use black_candidate and black_share as the variable to predict the turnout as we can see that the relationship of black_share remains the same but the relation of the black_candidate changes to negative which means that if we use candidate and shares the higher the share means and the lower the candidate means the higher the turnout

and we can see that the r squared is higher than the 2 previous model which means the model have a better understanding of the given variable

the adjusted r squared generally provides a more conservative estimate of the goodness-of-fit of the model. and we can see that the value is lower which means that when we adjust the r squared the value drop. which means the added predictor doesn't really improve the prediction

```r
lm_3 <- lm(black_turnout ~ black_candidate + black_share, data = turnout)

lm_3 %>%
  broom::tidy() %>%
  select(term, estimate) %>%
  knitr::kable(digits = 2)
```

term	estimate
(Intercept)	0.38
black_candidate	-0.01
black_share	0.21

```r
r_squared_lm3 <- summary(lm_3)$r.squared
adj_r_squared_lm3 <- summary(lm_3)$adj.r.squared

r_squared_lm3
```

```
## [1] 0.02852765
```

```r
adj_r_squared_lm3
```

```
## [1] 0.02695314
```

Question 7

the intercept represent the estimated value when the 2 variable are 0 which in real live scenario is not realistic. its as the base value of the y and these value are also important to the output of the model because without the intercept it means that the base value of y will be 0 and from the data we see if we remove this intercept means that the turnout to be really low and far from the actual turnout and can lead to a misleading prediction.

Question 8

in the question 3 we get the black_candidate as a positive value and in question 6 we get the black_candidate in negative value which means this variable have a different relationship to the turnout as i mentioned before in q3 and q6. while in q3 the higher the black_candidate give higher turnout and in the q6 the higher the black_candidate give a lower turnout. but since the black_candidate is a 0 or 1 value means that in q3 if there are black_candidate the turnout if higher while in q6 if there are black_candidate the turnout is lower

since the question says that i need to focused more on the model fit i would choose the model lm_3 in question 6 because the model has a better fit and we can see that from the higher r squared value. and the higher the r squared means that it have a better understanding

Question 9

if we use a categorical data it will interpret the coefficient as difference of the of the turnout from the omitted data which in this case we can see that it use AK as the base line of the prediction. and when we do the regression with and without the intercept we can see that if we use the intercept the coeff of the other states

are negative and its because it use the AK state as the baseline and if we dont use the intercept the coeff is all positive we kinda can see the relation when we use the intercept or not, let me take one of the state as example :

with intercept :

intercept-> 0.55118 stateAL ->-0.11883

without intercept :

stateAL -> 0.4323

we can see that with intercept if the state is AL the value are 0.55118-0.11883=0.43235 which is the same as stateAL when we dont use intercept so the meaning and the nature of the coefficient are still the same with or without intercept.

```
lm_states <- lm(black_turnout ~ state, data = turnout)

lm_states
```

```
##
## Call:
## lm(formula = black_turnout ~ state, data = turnout)
##
## Coefficients:
## (Intercept)      stateAL      stateAR      stateAZ      stateCA      stateCO
##     0.55118     -0.11883     -0.18848     -0.19542     -0.15493     -0.04704
##     stateCT      stateDE      stateFL      stateGA      stateIA      stateIL
##    -0.12102     -0.01851     -0.12429     -0.13935     -0.03011     -0.18291
##     stateIN      stateKS      stateKY      stateLA      stateMA      stateMD
##    -0.20289     -0.16516     -0.11449     -0.09044     -0.19336     -0.06229
##     stateME      stateMI      stateMN      stateMO      stateMS      stateNC
##     0.35365     -0.03181     -0.08637     -0.16215     -0.14942     -0.10195
##     stateNE      stateNH      stateNJ      stateNM      stateNV      stateNY
##    -0.16040      0.04948     -0.15214     -0.12861     -0.16100     -0.19762
##     stateOH      stateOK      stateOR      statePA      stateRI      stateSC
##    -0.10503     -0.03266      0.13679     -0.21613     -0.12049     -0.11335
##     stateTN      stateTX      stateUT      stateWA      stateWI      stateWV
##    -0.14519     -0.26037     -0.16401     -0.20475     -0.17116     -0.17120
```

```
lm_states_no_intercept <- lm(black_turnout ~ 0 + state, data = turnout)

lm_states_no_intercept
```

```
##
## Call:
## lm(formula = black_turnout ~ 0 + state, data = turnout)
##
## Coefficients:
## stateAK  stateAL  stateAR  stateAZ  stateCA  stateCO  stateCT  stateDE
##  0.5512   0.4323   0.3627   0.3558   0.3962   0.5041   0.4302   0.5327
## stateFL  stateGA  stateIA  stateIL  stateIN  stateKS  stateKY  stateLA
##  0.4269   0.4118   0.5211   0.3683   0.3483   0.3860   0.4367   0.4607
## stateMA  stateMD  stateME  stateMI  stateMN  stateMO  stateMS  stateNC
##  0.3578   0.4889   0.9048   0.5194   0.4648   0.3890   0.4018   0.4492
```

```
## stateNE  stateNH  stateNJ  stateNM  stateNV  stateNY  stateOH  stateOK
##  0.3908   0.6007   0.3990   0.4226   0.3902   0.3536   0.4461   0.5185
## stateOR  statePA  stateRI  stateSC  stateTN  stateTX  stateUT  stateWA
##  0.6880   0.3350   0.4307   0.4378   0.4060   0.2908   0.3872   0.3464
## stateWI  stateWV
##  0.3800   0.3800
```