PLS Model

Janique

11/07/2020

Libraries

```
library(dplyr)
library(purrr)
library(tidyr)
library(jsonlite)
library(plsdepot)
library(ggplot2)
library(foreign)
library(stringr)
library(readr)
```

Data Cleaning

First Dataset

Contains voting, race, gender, age data.

From JSON to tabular format

```
nested <- fromJSON('states.json')

unnest_data <- Vectorize(function(data_list) {
  data_list %>%
    unlist() %>%
    tibble(key = names(.), value = .) %>%
    spread(key, value, convert = TRUE)
})

flat <- nested %>%
  map_dfr(~tibble(data = .), .id = 'state') %>%
  mutate(county = names(data)) %>%
  select(state, county, data) %>%
  mutate(data_df = unnest_data(data)) %>%
  select(-data) %>%
  unnest(data_df)
```

Amending data errors

```
# Changing incorrect numbers of votes
flat$elections.2008.total[1346] <- 16323
flat$elections.2008.dem[727] <- 12368
flat$elections.2008.total[727] <- 12368 + 17019 + 545

# Removing counties with the number of votes missing
flat <- flat[-which(is.na(flat$employed)),]
flat <- flat[-which(flat$state == "Alaska"),]

# Changing column names to more workable ones
new_names <- c("0to4", "10to14", "15to19", "20to24", "25to29", "30to34",
               "35to39", "40to44", "45to49", "5to9", "50to54", "55to59",
               "60to64", "65to69", "70to74", "75to79", "80to84", "85+",
               "AsianF", "AsianM", "BlackF", "BlackM", "HispanicF", "HispanicM",
               "WhiteF",  "WhiteM", "UnemploymentRate")
colnames(flat)[c(4:21, 45:52, 55)] <- new_names

# Introducing ratio, proportion variables for comparative analysis
flat <- flat %>%
  mutate(# Ratio of Republican votes
         gop08 = elections.2008.gop / elections.2008.total,
         gop12 = elections.2012.gop / elections.2012.total,
         gop16 = elections.2016.gop / elections.2016.total,

         # Income logarithm
         IncomeLog = log(avg_income),

         # Adding variables for the change in proportion
         diff0816 = gop16 - gop08,
         diff1216 = gop16 - gop12
         )
```

Second Dataset

Contains poverty, industry data.

```
# Second dataset upload
poverty_data <- read.csv("poverty_data.csv")
poverty_data$County <- tolower(poverty_data$County)

# Fixing encoding issues
poverty_data$County[which(poverty_data$State == "New Mexico")][8] <- "doña ana county"
```

Third Dataset

Contains education, population density data.

```r
# Third dataset upload
edu_data <- read.csv("edu_data.csv", na.strings = c("","NA"))
edu_data$area_name <- tolower(edu_data$area_name)
edu_data <- edu_data[!is.na(edu_data$state_abbreviation), ]

# Adding a column with state names instead of abbreviations
names(state.name) <- state.abb
edu_data$state <- state.name[edu_data$state_abbreviation]

# Data merge
data <- left_join(flat, poverty_data, by = c("county" = "County", "state" = "State"))  %>%
  left_join(edu_data, by = c("county" = "area_name", "state"))

# Removing irrelevant variables
data <- data[c(1:2, 4:21, 45:52, 55, 62:64, 80, 82:86, 105, 120, 122:123)] %>%
  na.omit()
colnames(data)[41] <- "High School"
```

Fourth Dataset

Contains religion data.

```r
relig_data <- read_csv("relig_data.csv")

# Choosing columns with religion rates per 1,000 inhabitants
relig_rate <- relig_data[, grepl("RATE", names(relig_data))]

# Replacing NA with 0
relig_rate[is.na(relig_rate)] <- 0

# Keeping columns with an average higher than 1% of the US population
relig_rate <- relig_rate[, which(colMeans(relig_rate) > 10)]
```

Reasoning behind mergining religions

Evangelical + Sounthern Baptists + Assemblies of God -> all are evangelical protestant.

- Evangelical church - this is not a denominational church, more of a general term, 25% of population identifies with this. Very conservative.
- Southern Baptist Convention - largest protestant denomination in the US, it is a declining church, politically conservative, white, membership mostly in the South (duh).
- Assemblies of God - a union of pentecostal churches which tend to have a very lower class, uneducated following. They believe in speaking in tongues and divine healing.

Mainline + Missouri Synod + Evangelical Lutheran + Methodist -> all are mainline protestant.

- Mainline protestant - a more general term, similarly to evangelical. Mainline protestants tend to be more educated and have a higher social class. Slightly Republican, but includes many Democrats supporters.

- Lutheran Church Missouri Synod - members mostly in the Midwest, second largest Lutheran denomination. Predominantly white church.
- Evangelical Lutheran Church - largest Lutheran denomination. Predominantly white, educated, Midwest.
- United Methodist Church - largest mainline protestant denomination in the US, second largest Protestant church. Mostly Midwest and the South, many members in Texas. Around 54% Republican, 35% Democrat.

```r
# Merging similar religions
relig_rate <- relig_rate %>% mutate(
  Mainline = MPRTRATE + LCMSRATE + ELCARATE + UMCRATE,
  Evangelical = EVANRATE + SBCRATE + AGRATE,
  Catholic = CATHRATE + CTHRATE
) %>% select(Mainline, Evangelical, Catholic, BPRTRATE) %>%
  cbind(select(relig_data, STNAME, CNTYNAME))
colnames(relig_rate)[4:6] <- c("Black Prot", "state", "county")

# Joining with the rest of the datasets
relig_rate$county <- tolower(relig_rate$county)
data <- left_join(data, relig_rate, by = c("county", "state"))
```

Fifth Dataset

Contains the Gini index

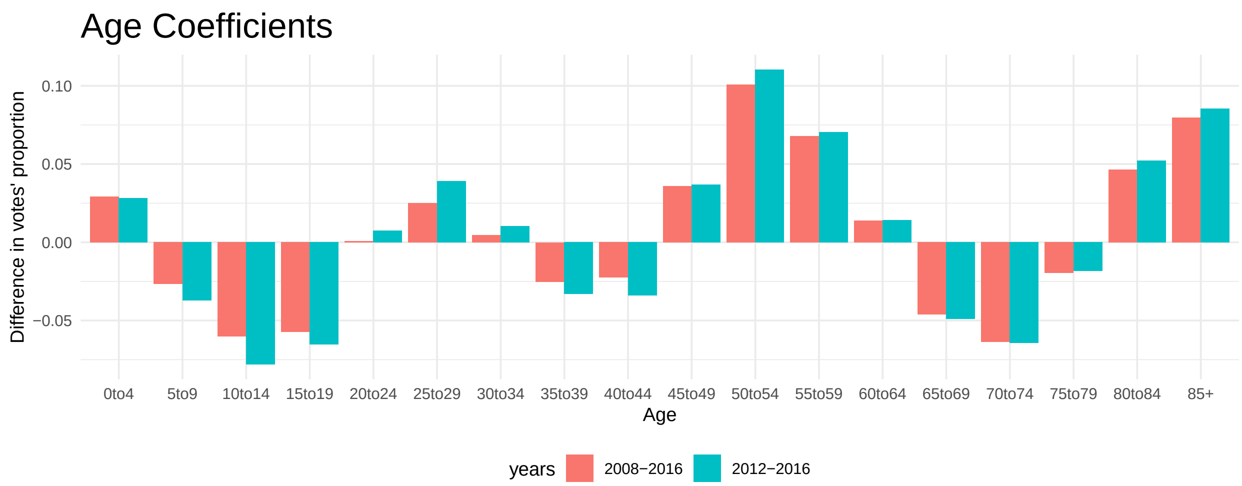
```r
gini_data <- read.csv("gini_data.csv")
index <- nrow(gini_data)

# Splitting state's and county's name into two columns
list_cnt_st <- str_split_fixed(gini_data$county, ", ", 2)
gini_data$county <- list_cnt_st[1:index]
gini_data$state <- list_cnt_st[(index+1):length(list_cnt_st)]
gini_data$county <- tolower(gini_data$county)

# Joining with the rest of datasets
data <- left_join(data, gini_data, by = c("county", "state"))

# Missing values extravaganza
data$no_gini <- ifelse(is.na(data$gini), 0, mean(data$gini, na.rm = TRUE))
data$gini[is.na(data$gini)] <- mean(data$gini, na.rm = TRUE)
```

PLS Model

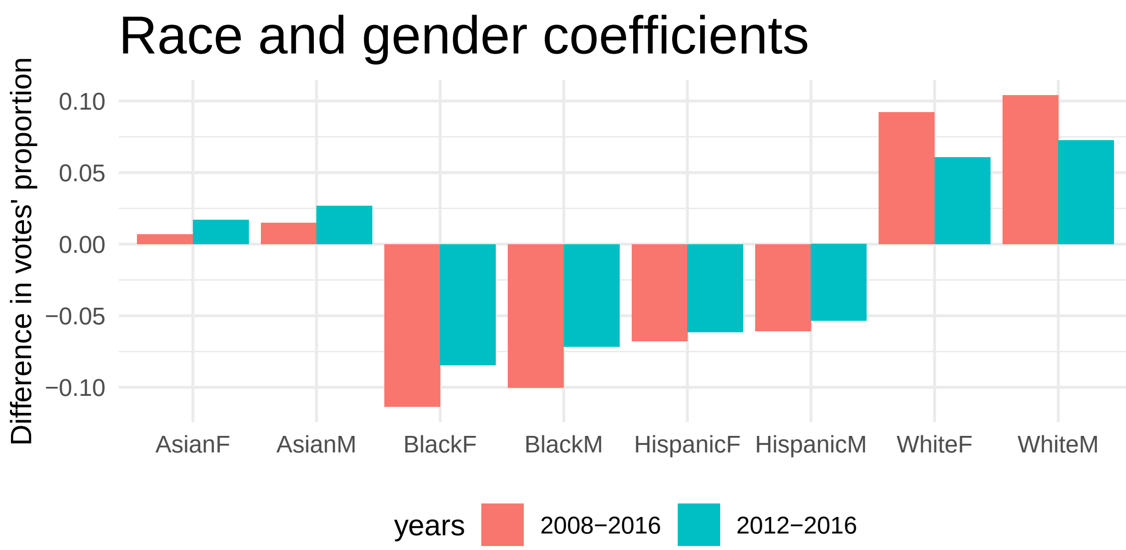
```r
pls_model <- plsreg2(responses = data[31:32], predictors = data[-c(1:2, 31:32)],
                     comps = NULL, crosval = TRUE)
```

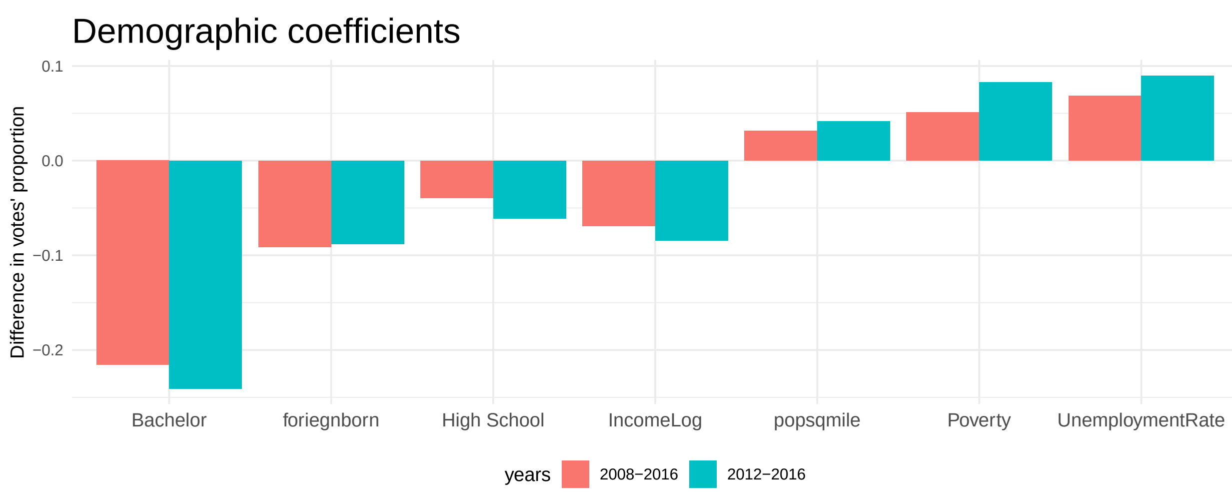
```r
theme_set(theme_minimal())
age_coef <- pls_model$std.coefs[1:18, 1:2]
age_df <- data.frame("diff" = age_coef[1:36],
```

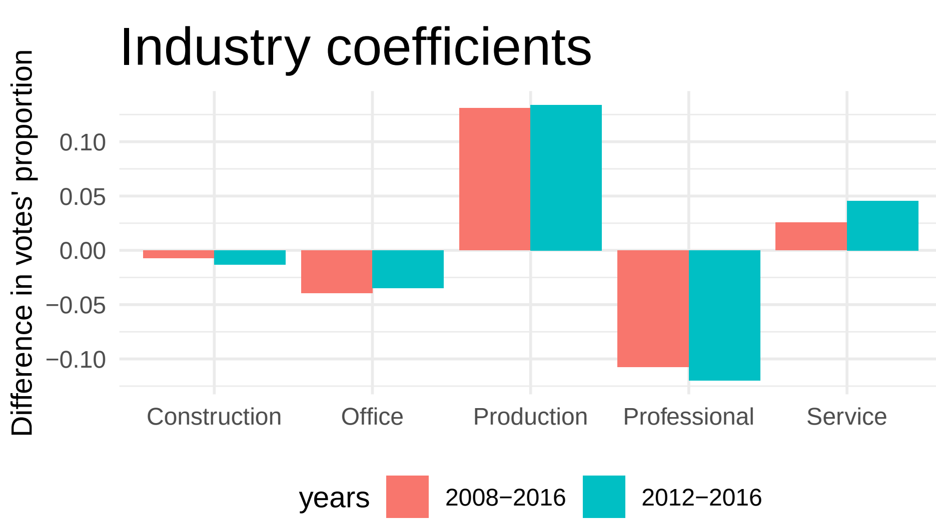
```
                     "years" = c(rep("2008-2016", 18), rep("2012-2016", 18)),
                     "age" = rep(row.names(age_coef), 2),
                     "order" = rep(c(1, 3:10, 2, 11:18), 2))

plt_age <- ggplot(data = age_df, aes(x = reorder(age, order), y = diff,
                                      fill = years)) +
  geom_bar(stat = "identity", position = position_dodge())
```

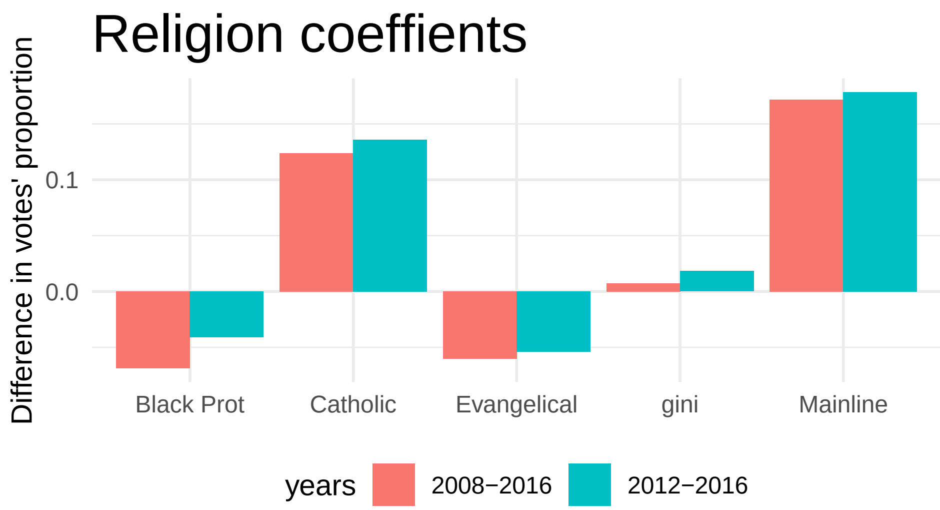



```
make_graph <- function(vec){
  coefs <- pls_model$std.coefs[vec, 1:2]
  df <- data.frame("diff" = coefs[1:length(coefs)],
                   "years" = c(rep("2008-2016", length(coefs)/2),
                               rep("2012-2016", length(coefs)/2)),
                     "value" = rep(row.names(coefs), 2))
  plt <- ggplot(data = df, aes(x = value, y = diff, fill = years)) +
    geom_bar(stat = "identity", position = position_dodge()) +
    theme(legend.position = "bottom", axis.title.x = element_blank(),
          plot.title = element_text(size = 20)) +
    ylab("Difference in votes' proportion")
  return(plt)
}
```


Race and gender coefficients
Difference in votes' proportion
0.10
0.05
0.00
−0.05
−0.10
AsianF
AsianM
BlackF
BlackM
HispanicF
HispanicM
WhiteF
WhiteM
years
2008−2016
2012−2016


Demographic coefficients
Difference in votes' proportion
0.1
0.0
−0.1
−0.2
Bachelor
foriegnborn
High School
IncomeLog
popsqmile
Poverty
UnemploymentRate
years
2008−2016
2012−2016

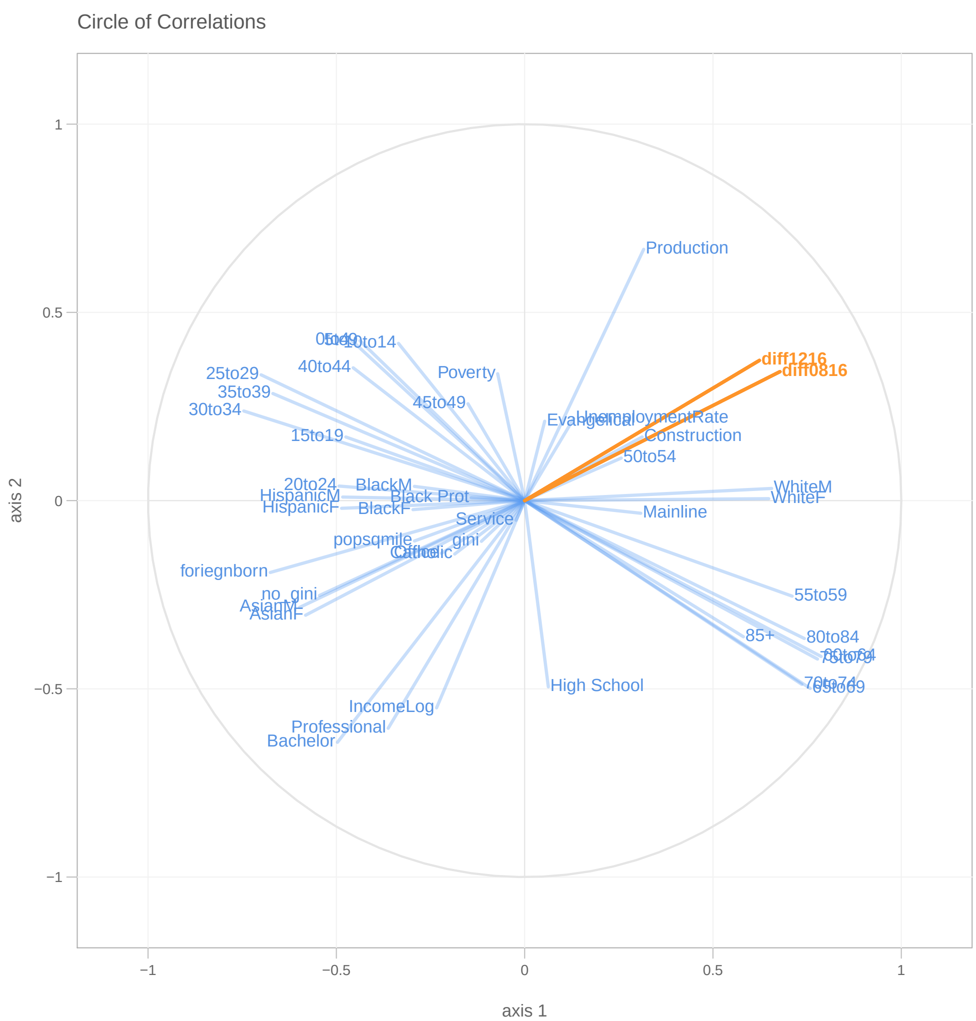

Industry coefficients
Difference in votes' proportion
0.10
0.05
0.00
−0.05
−0.10
Construction
Office
Production
Professional
Service
years
2008−2016
2012−2016



The religion coefficients are quite unexpected. Mainline should be much less conservative than evangelical. Maybe it is included instead in other variables. Maybe it depends on how I joined the religions, we could possibly just treat all of them separately. Maybe mainline protestants liked Trump.

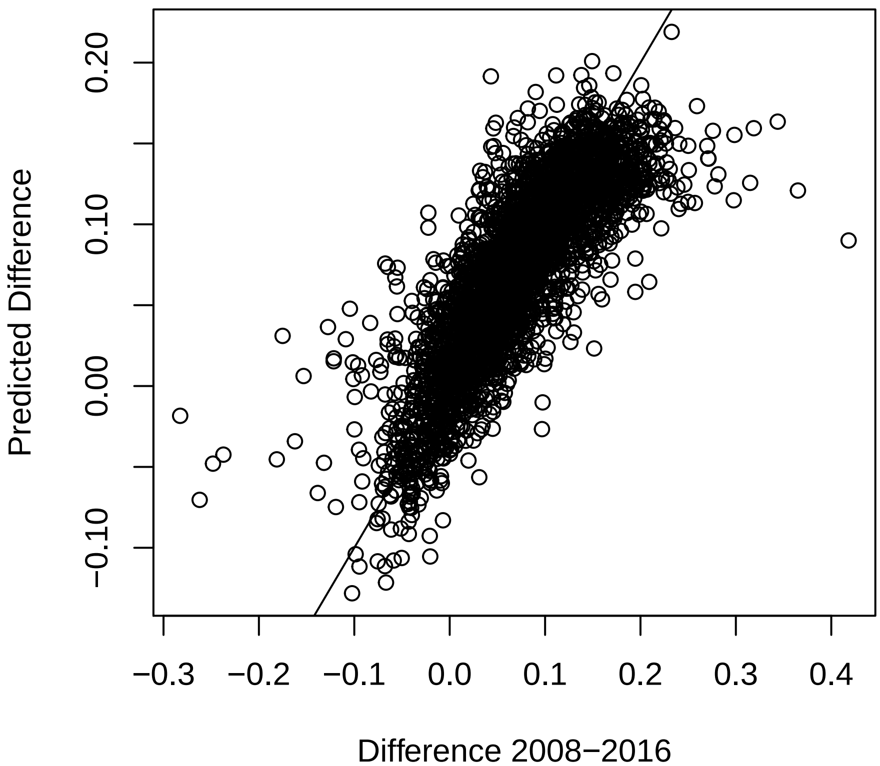
Circle of Correlations

```
plot(pls_model)
```



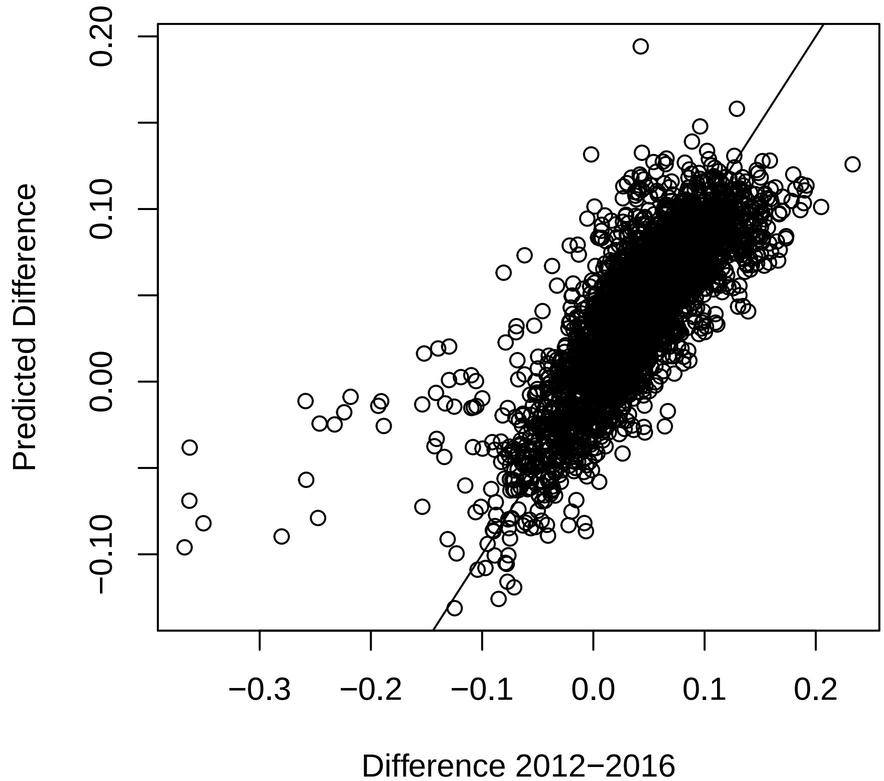
Predictions and Real Values

```r
plot(data$diff0816, pls_model$y.pred[1:(length(pls_model$y.pred)/2), 1],
     xlab = "Difference 2008-2016", ylab = "Predicted Difference") + abline(0, 1)
```




```
## integer(0)
```

```
plot(data$diff1216, pls_model$y.pred[1:(length(pls_model$y.pred)/2), 2],
     xlab = "Difference 2012-2016", ylab = "Predicted Difference") + abline(0, 1)
```




```
## integer(0)
```

Interesting outliers 2008-2016:

- Madison county, Idaho - furthest on the left
- Utah county, Utah - 2nd furthest on the left
- Davis county, Utah - 3rd furthest left
- Cache county, Utah - 4th furthest left
- Laporte county, Indiana - furthest on the right

Interesting outliers 2012-2016:

- Utah county, Utah - furthest on the left
- Madison county, Idaho - 2nd furthest on the left
- Cache county, Utah - 3rd furthest on the left
- Davis couty, Utah - 4rd furthest left

Explained Variance

```r
pls_model$expvar
```

```
##           R2X   R2Xcum         R2Y   R2Ycum
## t1 0.23650434 0.2365043 0.423978864 0.4239789
## t2 0.10304796 0.3395523 0.128018008 0.5519969
## t3 0.09065982 0.4302121 0.054716676 0.6067135
## t4 0.07383549 0.5040476 0.029993312 0.6367069
## t5 0.06339321 0.5674408 0.005817827 0.6425247
```