FinalProjectAlzheimers

Jake Rickord

2/24/2021

Section 2:

A-1.How to import my data

Pretty simple here, all 3 sources of data are found in csv files so I will import all three as separate dataframes and check to ensure their data is pulled in properly.

```
aha<-read.csv("C:/Users/Jake/Desktop/Bellevue Items/Assignments/Stats for DS/Final Project/Alzheimers He
mva<-read.csv("C:/Users/Jake/Desktop/Bellevue Items/Assignments/Stats for DS/Final Project/oasis_longitu
sdcd<-read.csv("C:/Users/Jake/Desktop/Bellevue Items/Assignments/Stats for DS/Final Project/San Diego Co
```

A-2. How to clean my dataset

More complex issue here. I'd like to only keep the rows that we will be investigating in our search here. Let's revisit our list of questions we're aiming to answer here, as well as which dataset has applicable information to make these conclusions for each question.

1. Is dementia & Alzheimer's truly on the rise based on given datasets? DATA FOR THIS TYPE OF ANALYSIS WOULD REQUIRE A TIME COMPONENT. THEREFORE AHA AND SDCD WOULD BE APPLICABLE.
2. Does diet correlate to higher/lower prevalance of subjective cognitive decline or memory loss? THIS WOULD REQUIRE DIET INFORMATION, WHICH IS FOUND ONLY IN AHA.
3. Similarly does exercise correlate to cognitive decline/memory loss? THIS WOULD REQUIRE EXERCISE INFORMATION, WHICH IS FOUND ONLY IN AHA.
4. Does lack of sleep correlate to cognitive decline / memory loss? ONCE AGAIN, ONLY FOUND IN AHA.
5. The individuals I've known to have Alzheimer's and/or dementia have suffered from falls that rapidly increased the speed of mental deterioration. Test this given dataset to see if this rings true. DATA ON FALLS IS FOUND IN AHA.
6. Is a certain sex more pre-disposed to having dementia? ALL THREE DATASETS HAVE MF INDICATORS
7. Does dementia affect certain races more than others? AHA AND SDCD DATASETS HAVE RACE DATA
8. Does a specific state have a higher percentage of individuals with dementia? COMPARING STATES ONLY EXISTS IN AHA
9. Is estimated intracranial volume (eTIV) a solid indicator for predicting dementia? eTIV DATA IS ONLY FOUND IN MVA.

From this data, we can identify which columns are necessary for our analysis for each dataset and filter out the "noise" so to speak.

```
aha<-aha[, c("YearEnd", "LocationAbbr", "Question", "Data_Value_Type", "Data_Value", "StratificationCate
head(aha)
```

```
##   YearEnd LocationAbbr
## 1    2018          NRE
## 2    2016           TX
## 3    2017           AK
## 4    2016           TX
## 5    2017           MD
## 6    2016           MI
##                                                                     Question
## 1          Mean number of days with activity limitations in the past month
## 2 Percentage of older adults who are experiencing frequent mental distress
## 3 Percentage of older adults who are experiencing frequent mental distress
## 4 Percentage of older adults who are experiencing frequent mental distress
## 5 Percentage of older adults who are experiencing frequent mental distress
## 6 Percentage of older adults who are experiencing frequent mental distress
##   Data_Value_Type Data_Value StratificationCategory1   Stratification1
## 1            Mean        5.7               Age Group 65 years or older
## 2      Percentage        7.4               Age Group           Overall
## 3      Percentage        6.4               Age Group           Overall
## 4      Percentage        8.5               Age Group       50-64 years
## 5      Percentage       14.6               Age Group       50-64 years
## 6      Percentage       10.8               Age Group       50-64 years
##   StratificationCategoryID2 StratificationID2
## 1                   OVERALL           OVERALL
## 2                    GENDER              MALE
## 3                    GENDER              MALE
## 4                    GENDER              MALE
## 5                    GENDER            FEMALE
## 6                    GENDER              MALE
```

```
mva<-mva[,c("Subject.ID", "Group", "M.F", "eTIV")]
head(mva)
```

```
##   Subject.ID       Group M.F eTIV
## 1  OAS2_0001 Nondemented   M 1987
## 2  OAS2_0001 Nondemented   M 2004
## 3  OAS2_0002    Demented   M 1678
## 4  OAS2_0002    Demented   M 1738
## 5  OAS2_0002    Demented   M 1698
## 6  OAS2_0004 Nondemented   F 1215
```

```
sdcd<-sdcd[,c("Year","Total_Male", "Total_Female", "White_Total", "Black_Total", "Hispanic_Total", "API_
head(sdcd)
```

```
##   Year Total_Male Total_Female White_Total Black_Total Hispanic_Total API_Total
## 1 2017        534          684         917          42            164        76
## 2 2017         55           69          71          16             23        11
```

```
## 3 2017         24           25           34          NA              6         NA
## 4 2017         16           26           31          NA             NA         NA
## 5 2017         18           24           12          10             14          5
## 6 2017         94          148          195          12             28         NA
##   AIAN_Total Other_notAIAN_Total
## 1         NA                   9
## 2         NA                  NA
## 3         NA                  NA
## 4         NA                  NA
## 5         NA                  NA
## 6         NA                  NA
```

Next I'd like to address missing values. First, for each dataframe we need to identify these columns containing missing values.

```
colnames(aha)[colSums(is.na(aha))>0]
```

```
## [1] "Data_Value"
```

```
colnames(mva)[colSums(is.na(mva))>0]
```

```
## character(0)
```

```
colnames(sdcd)[colSums(is.na(sdcd))>0]
```

```
## [1] "Total_Male"          "Total_Female"        "White_Total"
## [4] "Black_Total"         "Hispanic_Total"      "API_Total"
## [7] "AIAN_Total"          "Other_notAIAN_Total"
```

From this breakdown, we can see that some questions had an N/A % response in the survey for groups for AHA's dataframe. Being that there were 0% values and 0 mean average values entered elsewhere in the survey results, this likely means the question was either left blank or could not be determined. In these cases, I don't believe adding in 0's makes sense, as this would provide outliers in sections with data that may be illogical. As such, I'd lean towards removing these entries altogether.

The mva dataframe has no N/A values presented, so all good there.

For the SDCD dataframe, we find that there were not 0 values added in for any of these sections, and the addition of the other column values add up to the total overall number of respondents, so in this case, it likely makes sense to enter in 0 values for those listed as N/A.

```
aha<-aha[complete.cases(aha), ]
head(aha)
```

```
##   YearEnd LocationAbbr
## 1    2018          NRE
## 2    2016           TX
## 3    2017           AK
## 4    2016           TX
## 5    2017           MD
## 6    2016           MI
##                                                                     Question
```

```
## 1          Mean number of days with activity limitations in the past month
## 2 Percentage of older adults who are experiencing frequent mental distress
## 3 Percentage of older adults who are experiencing frequent mental distress
## 4 Percentage of older adults who are experiencing frequent mental distress
## 5 Percentage of older adults who are experiencing frequent mental distress
## 6 Percentage of older adults who are experiencing frequent mental distress
##   Data_Value_Type Data_Value StratificationCategory1   Stratification1
## 1            Mean        5.7               Age Group 65 years or older
## 2      Percentage        7.4               Age Group           Overall
## 3      Percentage        6.4               Age Group           Overall
## 4      Percentage        8.5               Age Group       50-64 years
## 5      Percentage       14.6               Age Group       50-64 years
## 6      Percentage       10.8               Age Group       50-64 years
##   StratificationCategoryID2 StratificationID2
## 1                   OVERALL           OVERALL
## 2                    GENDER              MALE
## 3                    GENDER              MALE
## 4                    GENDER              MALE
## 5                    GENDER            FEMALE
## 6                    GENDER              MALE
```

```r
sdcd[is.na(sdcd)] <-0

sdcd$Total<-0
head(sdcd)
```

```
##   Year Total_Male Total_Female White_Total Black_Total Hispanic_Total API_Total
## 1 2017        534          684         917          42            164        76
## 2 2017         55           69          71          16             23        11
## 3 2017         24           25          34           0              6         0
## 4 2017         16           26          31           0              0         0
## 5 2017         18           24          12          10             14         5
## 6 2017         94          148         195          12             28         0
##   AIAN_Total Other_notAIAN_Total Total
## 1          0                   9     0
## 2          0                   0     0
## 3          0                   0     0
## 4          0                   0     0
## 5          0                   0     0
## 6          0                   0     0
```

```r
for(i in 1:nrow(sdcd)){
  sdcd[i, "Total"]<-sdcd[i, "Total_Male"]+sdcd[i, "Total_Female"]
}
head(sdcd)
```

```
##   Year Total_Male Total_Female White_Total Black_Total Hispanic_Total API_Total
## 1 2017        534          684         917          42            164        76
## 2 2017         55           69          71          16             23        11
## 3 2017         24           25          34           0              6         0
## 4 2017         16           26          31           0              0         0
## 5 2017         18           24          12          10             14         5
## 6 2017         94          148         195          12             28         0
```

```
##   AIAN_Total Other_notAIAN_Total Total
## 1          0                   9  1218
## 2          0                   0   124
## 3          0                   0    49
## 4          0                   0    42
## 5          0                   0    42
## 6          0                   0   242
```

Now we can test again to ensure that the dataframes no longer contain NA values that can affect modeling.

```
colnames(aha)[colSums(is.na(aha))>0]
```

```
## character(0)
```

```
colnames(sdcd)[colSums(is.na(sdcd))>0]
```

```
## character(0)
```

```
head(sdcd)
```

```
##   Year Total_Male Total_Female White_Total Black_Total Hispanic_Total API_Total
## 1 2017        534          684         917          42            164        76
## 2 2017         55           69          71          16             23        11
## 3 2017         24           25          34           0              6         0
## 4 2017         16           26          31           0              0         0
## 5 2017         18           24          12          10             14         5
## 6 2017         94          148         195          12             28         0
##   AIAN_Total Other_notAIAN_Total Total
## 1          0                   9  1218
## 2          0                   0   124
## 3          0                   0    49
## 4          0                   0    42
## 5          0                   0    42
## 6          0                   0   242
```

All set for that piece. Last thing I would look at before jumping into the analysis is examing the dataframes to see if they're in a format that is conducive to modeling. I'd say that the mva and sdcd dataframes are now in easily interpretable and malleable forms. However, I'd say that the aha dataframe is still slightly ugly. To start there are an incredible number of questions involved, not all of which pertain to our investigations here. So I'd reduce those down to the applicable ones for our questions above.

```
aha<-subset(aha, Question=="Percentage of older adults who reported subjective cognitive decline or memo
unique(aha[c("Question")])
```

```
##
## 27
## 85
## 99
## 192                                                                             Percentage of older ad
## 2647                                                                                         Percentag
## 50318            Percentage of older adults who reported subjective cognitive decline or memory loss t
## 64274 Percentage of older adults who reported subjective cognitive decline or memory loss that interf
## 85227              Percentage of older adults who reported that as a result of subjective cognitive d
```

```r
aha<-unique(aha)

searchIndex1<-function(year, loc, class1, class2){
  row<-which(aha$YearEnd==year & aha$LocationAbbr == loc & aha$Stratification1==class1 & aha$Stratifica

  if(rlang::is_empty(row)){
    return(0)
  }
  else{
  dementvalue<-aha$Data_Value[row]
  dementvalue<-mean(dementvalue)
  return(dementvalue)
  }
}

for(i in 1:nrow(aha)){
  aha[i, "DementiaPercent"]<-searchIndex1(aha[i, "YearEnd"], aha[i, "LocationAbbr"], aha[i, "Stratifica
}
head(aha)
```

```
##    YearEnd LocationAbbr
## 27    2016           US
## 28    2016           MD
## 31    2016          NRE
## 32    2016           WV
## 33    2016         WEST
## 35    2016           NE
##                                                       Question
## 27 Percentage of older adults getting sufficient sleep (>6 hours)
## 28 Percentage of older adults getting sufficient sleep (>6 hours)
## 31 Percentage of older adults getting sufficient sleep (>6 hours)
## 32 Percentage of older adults getting sufficient sleep (>6 hours)
## 33 Percentage of older adults getting sufficient sleep (>6 hours)
## 35 Percentage of older adults getting sufficient sleep (>6 hours)
##    Data_Value_Type Data_Value StratificationCategory1   Stratification1
## 27      Percentage       64.1               Age Group       50-64 years
## 28      Percentage       66.6               Age Group           Overall
## 31      Percentage       71.1               Age Group 65 years or older
## 32      Percentage       55.2               Age Group       50-64 years
## 33      Percentage       64.0               Age Group           Overall
## 35      Percentage       70.6               Age Group       50-64 years
##    StratificationCategoryID2 StratificationID2 DementiaPercent
## 27                    GENDER              MALE            11.2
## 28                    GENDER              MALE             0.0
## 31                    GENDER              MALE            10.7
## 32                    GENDER            FEMALE             0.0
## 33                      RACE               NAA            18.2
## 35                    GENDER              MALE             0.0
```

All set here now on the question bank. We will likely need to do several different subsets of this data depending on factor we're nailing into.

B. What does the final data set look like? There are essentially 3 final data sets. AHA is made up of 9 different questions broken into several categories based on race, sex, and age. MVA is a simpler dataset that is made up of simply subject ID, classification, sex, and eTIV value. Lastly, sdcd is broken down to year, sex, and race data.

C. Questions for future steps? As I move forward, the biggest questions likely revolve around the aha dataset. How I want to carve this up depending on what factor I'm examing may take some extra thought since almost all data values are based on percentages.

For the other two datasets, really thinking about what modeling solution I'd venture with to start would be key.The first one to jump out at me is the MRI data, where it might make the most sense to do a logistical regression or k-nearest-neighbor model.

D. What information is not self-evident? I'm not sure any information is not self-evident right off the bat, I'd say the way that the aha datsets is constructed is just, from onset, a crassly manufactured piece. Once I iron out the kinks there in my analysis, should be pretty set. Again, deciding models for each of these might be more intricate as well.

E. What are different ways you could look at this data? Rather than using each of these datasets to attack each question with different data, I could specify that each question uses a specific dataset, i.e. to answer questions on race & sex implications on dementia, I only use the sdcd dataset, instead of how I've presented it, where I will currently attack that question with all 3 datasets.

F. How do you plan to slice and dice the data?

I will likely break each of these into subsets of the master dataframes I've constructed here, which will hone in on factors presented. Before doing so, I may run a correlation test to ensure the independent variables I'm using here are not affecting one another as well. From there I'll break each factor into a test and training set for model analysis and accuracy/p-value examination

G. How could you summarize your data to answer key questions? The aha dataset, in its current form, would be difficult to interpret through easy summarizations. For sdcd, we could simply create a graphic depicting the percentage for sex, and for different races in two separate visualizations through histograms to interpret them. For the mvd dataset, perhaps a similar visualization could work depicting bins of eTIV values with number of positive dementia classifications

H. What types of plots and tables will help you to illustrate the findings to your questions? Certainly histograms and barcharts are key to the majority of summarizations before any modeling. Once modeling were to kick in, I'll likely need to show some regression, classification, and possibly a k-nearest-neighbor visualization.

I. Do you plan on incorporating any machine learning techniques to answer your research questions? I do plan on using machine learning here. For almost every one of the questions my end goal is to identify if any of these factors, or combinations of them, can help in prediction of dementia classification. As such, I intend to use logistic regression where possible, otherwise I will attempt to see if either k related techniques do better at predicting these case values.

J. Questions for future steps? I suppose the biggest question is whether I should proceed as described or adapt my datasets as laid out in question E, and address how questions will be answered by a specific dataset, rather than each dataset individually.

Part 3. Answering each question via model analysis.

1. Is dementia & Alzheimer's on the rise? Per my earlier discussion regarding different ways of looking at this dataset, I will follow the route of using specific sets to answer these questions rather than cross-mapping each dataset to multiple questions. As such, I will be using the SDCD dataset here to answer this question.

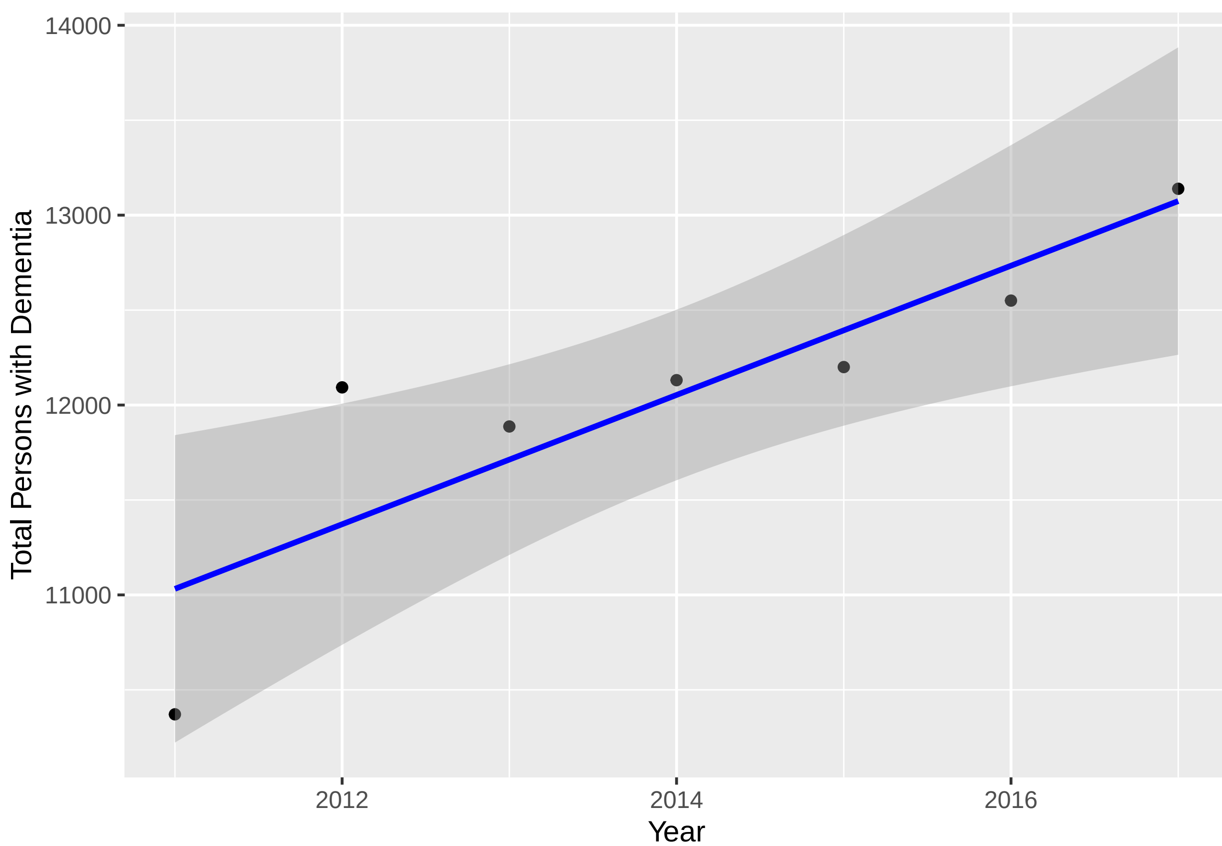
```r
head(sdcd)
```

```
##   Year Total_Male Total_Female White_Total Black_Total Hispanic_Total API_Total
```

```
## 1 2017        534          684          917          42            164          76
## 2 2017         55           69           71          16             23          11
## 3 2017         24           25           34           0              6           0
## 4 2017         16           26           31           0              0           0
## 5 2017         18           24           12          10             14           5
## 6 2017         94          148          195          12             28           0
##   AIAN_Total Other_notAIAN_Total Total
## 1          0                   9  1218
## 2          0                   0   124
## 3          0                   0    49
## 4          0                   0    42
## 5          0                   0    42
## 6          0                   0   242
```

```
sdcd$Total<-as.numeric(gsub(",", "", sdcd$Total))
sdcd[is.na(sdcd)] <-0
sdcd<-ddply(sdcd, "Year", numcolwise(sum))
ggplot(sdcd, aes(Year, Total))+geom_point()+geom_smooth(method="lm", colour="Blue") + labs(x="Year", y =
```

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




```
timedata<-lm(Total ~ Year, data=sdcd)
summary(timedata)
```

```
##
```

```
## Call:
## lm(formula = Total ~ Year, data = sdcd)
##
## Residuals:
##       1       2       3       4       5       6       7
## -660.82  720.79  174.39   78.00 -193.39 -183.79   64.82
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -673498.21  175913.97  -3.829   0.0123 *
## Year            340.39      87.35   3.897   0.0114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 462.2 on 5 degrees of freedom
## Multiple R-squared:  0.7523, Adjusted R-squared:  0.7028
## F-statistic: 15.19 on 1 and 5 DF,  p-value: 0.01144
```

```r
coef_lmbeta<-lm.beta(timedata)
coef_lmbeta
```

```
##      Year
## 0.8673633
```

From this simple linear regression, we can see that the standard deviation of the regression model's errors is about 15% the size of the standard deviation of errors from a simple modeling. Based solely on this data, we can say that the overall number of those affected (or diagnosed) with dementia has increased over time. This does not take into effect the idea that population itself increased, thus effecting the overall percentage of dementia patients.

2. Does diet have an effect on having dementia? This would be answered from the aha dataset that contains info on diet.

```r
ahadiet<-subset(aha, Question == "Percentage of older adults who are eating 2 or more fruits daily" | Qu

head(ahadiet)
```

```
##     YearEnd LocationAbbr
## 85     2017           IA
## 92     2017           PR
## 93     2017           HI
## 99     2017           FL
## 100    2017           NE
## 101    2017           LA
##                                                                 Question
## 85      Percentage of older adults who are eating 2 or more fruits daily
## 92      Percentage of older adults who are eating 2 or more fruits daily
## 93      Percentage of older adults who are eating 2 or more fruits daily
## 99  Percentage of older adults who are eating 3 or more vegetables daily
## 100 Percentage of older adults who are eating 3 or more vegetables daily
## 101 Percentage of older adults who are eating 3 or more vegetables daily
##     Data_Value_Type Data_Value StratificationCategory1   Stratification1
```

```
## 85       Percentage       26.2               Age Group       50-64 years
## 92       Percentage       15.7               Age Group           Overall
## 93       Percentage       41.8               Age Group 65 years or older
## 99       Percentage       15.3               Age Group       50-64 years
## 100      Percentage       15.4               Age Group       50-64 years
## 101      Percentage       11.4               Age Group 65 years or older
##     StratificationCategoryID2 StratificationID2 DementiaPercent
## 85                     GENDER              MALE             0.0
## 92                       RACE               HIS             5.5
## 93                       RACE               WHT             8.0
## 99                       RACE               WHT             0.0
## 100                      RACE               WHT             0.0
## 101                      RACE               WHT             0.0
```

```
nrow(ahadiet)
```

```
## [1] 3780
```

```
set.seed(278613)

ahaddummy<-sample(c(rep(0, 0.8 * nrow(ahadiet)), rep(1, 0.2 * nrow(ahadiet))))

table(ahaddummy)
```

```
## ahaddummy
##    0    1
## 3024  756
```

```
ahadtrain<-ahadiet[ahaddummy==0, ]
ahadtest<-ahadiet[ahaddummy==1, ]

ahadlm<-lm(Data_Value ~ DementiaPercent, data=ahadtrain)
summary(ahadlm)
```

```
##
## Call:
## lm(formula = Data_Value ~ DementiaPercent, data = ahadtrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.829  -8.985  -1.107   8.493  39.793
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     24.40684    0.23758 102.731  < 2e-16 ***
## DementiaPercent -0.13747    0.02799  -4.912 9.51e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.01 on 3022 degrees of freedom
## Multiple R-squared:  0.00792,    Adjusted R-squared:  0.007592
## F-statistic: 24.12 on 1 and 3022 DF,  p-value: 9.509e-07
```

```r
ahadtest$predict<-predict(ahadlm, newdata=ahadtest)

mse<-mean((ahadtest$predict-ahadtest$DementiaPercent)^2)
mse
```

```
## [1] 386.71
```

```r
mae<-mean(abs(ahadtest$predict-ahadtest$DementiaPercent))
mae
```

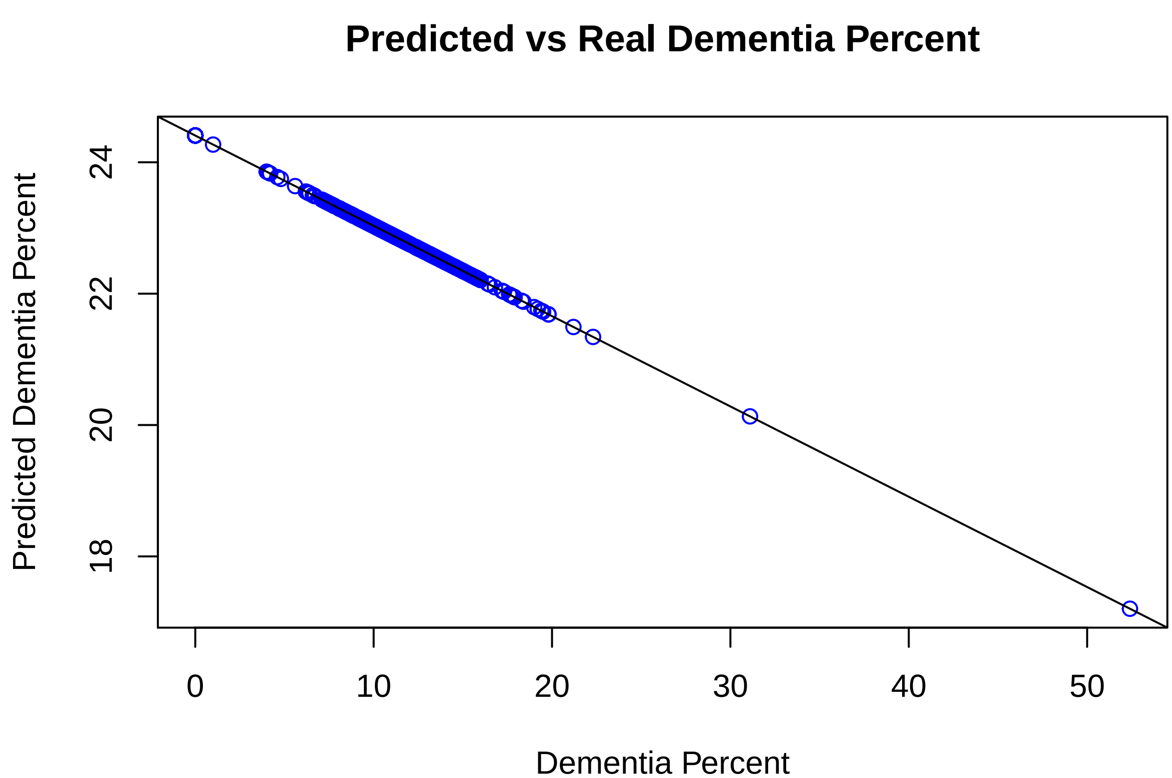
```
## [1] 18.3815
```

```r
rmse<-sqrt(mse)
rmse
```

```
## [1] 19.66494
```

```r
plot(ahadtest$DementiaPercent, ahadtest$predict, col="blue", main = "Predicted vs Real Dementia Percent

abline(ahadlm)
```



Judging from our quite large mean squared error, we're looking at an average of 19% off from actual values, so diet does not seem to be a significant indicator of dementia.

3. Similarly does exercise correlate to cognitive decline/memory loss? THIS WOULD REQUIRE EXERCISE INFORMATION, WHICH IS FOUND ONLY IN AHA. We'll run a similar analysis with a slightly different subset here as we did in question 2.

```
ahaex<-subset(aha, Question=="Percentage of older adults who have not had any leisure time physical acti

head(ahaex)
```

```
##     YearEnd LocationAbbr
## 192    2016          MDW
## 479    2015           NM
## 535    2016          NRE
## 743    2015           TN
## 754    2016           NY
## 955    2017           CT
##                                                                                               Question
## 192 Percentage of older adults who have not had any leisure time physical activity in the past month
## 479 Percentage of older adults who have not had any leisure time physical activity in the past month
## 535 Percentage of older adults who have not had any leisure time physical activity in the past month
## 743 Percentage of older adults who have not had any leisure time physical activity in the past month
## 754 Percentage of older adults who have not had any leisure time physical activity in the past month
## 955 Percentage of older adults who have not had any leisure time physical activity in the past month
##     Data_Value_Type Data_Value StratificationCategory1   Stratification1
## 192      Percentage       32.7               Age Group       50-64 years
## 479      Percentage       24.5               Age Group           Overall
## 535      Percentage       26.6               Age Group       50-64 years
## 743      Percentage       31.4               Age Group       50-64 years
## 754      Percentage       49.9               Age Group 65 years or older
## 955      Percentage       28.7               Age Group 65 years or older
##     StratificationCategoryID2 StratificationID2 DementiaPercent
## 192                      RACE               BLK            10.9
## 479                    GENDER              MALE             0.0
## 535                   OVERALL           OVERALL            10.2
## 743                      RACE               WHT            14.7
## 754                      RACE               NAA             0.0
## 955                   OVERALL           OVERALL             0.0
```

```
nrow(ahaex)
```

```
## [1] 3994
```

```
set.seed(278613)

ahaexdummy<-sample(c(rep(0, 0.8 * nrow(ahaex)), rep(1, 0.2 * nrow(ahaex))))

table(ahaexdummy)
```

```
## ahaexdummy
##    0    1
## 3195  798
```

```r
ahaextrain<-ahaex[ahaexdummy==0, ]
ahaextest<-ahaex[ahaexdummy==1, ]

ahaexlm<-lm(Data_Value ~ DementiaPercent, data=ahaextrain)
summary(ahaexlm)
```

```
##
## Call:
## lm(formula = Data_Value ~ DementiaPercent, data = ahaextrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.746  -5.023  -0.846   4.370  39.133
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     31.44576    0.18073 173.992   <2e-16 ***
## DementiaPercent -0.01910    0.02002  -0.954     0.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.393 on 3194 degrees of freedom
## Multiple R-squared:  0.0002847,  Adjusted R-squared:  -2.83e-05
## F-statistic: 0.9096 on 1 and 3194 DF,  p-value: 0.3403
```

```r
ahaextest$predict<-predict(ahaexlm, newdata=ahaextest)

mse<-mean((ahaextest$predict-ahaextest$DementiaPercent)^2)
mse
```

```
## [1] 666.5719
```

```r
mae<-mean(abs(ahaextest$predict-ahaextest$DementiaPercent))
mae
```

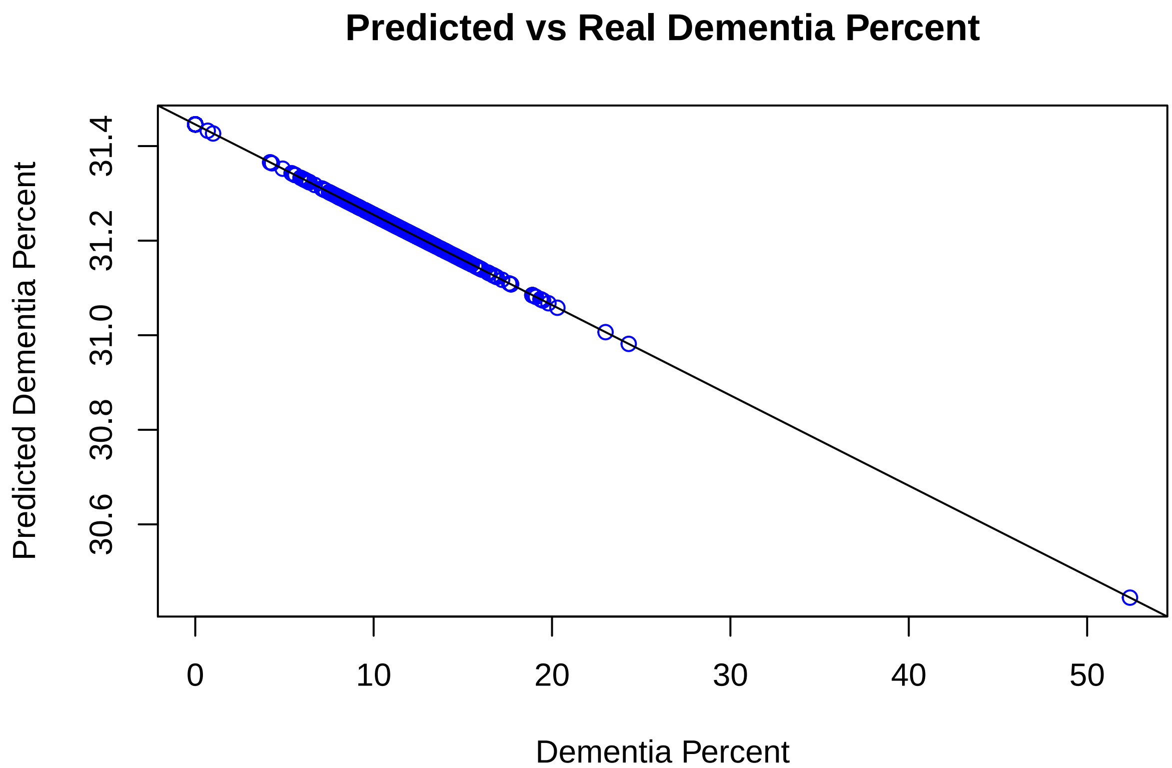
```
## [1] 25.04312
```

```r
rmse<-sqrt(mse)
rmse
```

```
## [1] 25.81805
```

```r
plot(ahaextest$DementiaPercent, ahaextest$predict, col="blue", main = "Predicted vs Real Dementia Perce

abline(ahaexlm)
```



This piece is a bit of a reflection of the last in that we're showing lack of exercise as not being a primary significant component of an indication of dementia. Our p-value here is much higher than as it relates to diet, so we'd say that lack of exercise is an even worse indicator of likelihood to show signs of dementia than diet, which is backed up by the much larger F Stat.

4. Does lack of sleep correlate to cognitive decline / memory loss? ONCE AGAIN, ONLY FOUND IN AHA. Similar to our last model this will do a similar analysis:

```
ahaslp<-subset(aha, Question=="Percentage of older adults getting sufficient sleep (>6 hours)")

head(ahaslp)
```

```
##    YearEnd LocationAbbr
## 27    2016           US
## 28    2016           MD
## 31    2016          NRE
## 32    2016           WV
## 33    2016         WEST
## 35    2016           NE
##                                                        Question
## 27 Percentage of older adults getting sufficient sleep (>6 hours)
## 28 Percentage of older adults getting sufficient sleep (>6 hours)
## 31 Percentage of older adults getting sufficient sleep (>6 hours)
## 32 Percentage of older adults getting sufficient sleep (>6 hours)
## 33 Percentage of older adults getting sufficient sleep (>6 hours)
```

```
## 35 Percentage of older adults getting sufficient sleep (>6 hours)
##    Data_Value_Type Data_Value StratificationCategory1   Stratification1
## 27      Percentage       64.1               Age Group       50-64 years
## 28      Percentage       66.6               Age Group           Overall
## 31      Percentage       71.1               Age Group 65 years or older
## 32      Percentage       55.2               Age Group       50-64 years
## 33      Percentage       64.0               Age Group           Overall
## 35      Percentage       70.6               Age Group       50-64 years
##    StratificationCategoryID2 StratificationID2 DementiaPercent
## 27                    GENDER              MALE            11.2
## 28                    GENDER              MALE             0.0
## 31                    GENDER              MALE            10.7
## 32                    GENDER            FEMALE             0.0
## 33                      RACE               NAA            18.2
## 35                    GENDER              MALE             0.0
```

```
nrow(ahaslp)
```

```
## [1] 2066
```

```
set.seed(278613)

ahaslpdummy<-sample(c(rep(0, 0.8 * nrow(ahaslp)), rep(1, 0.2 * nrow(ahaslp))))

table(ahaslpdummy)
```

```
## ahaslpdummy
##    0    1
## 1652  413
```

```
ahaslptrain<-ahaslp[ahaslpdummy==0, ]
ahaslptest<-ahaslp[ahaslpdummy==1, ]

ahaslplm<-lm(Data_Value ~ DementiaPercent, data=ahaslptrain)
summary(ahaslplm)
```

```
##
## Call:
## lm(formula = Data_Value ~ DementiaPercent, data = ahaslptrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -31.0995  -4.7853   0.8619   5.9783  19.8005
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     65.09952    0.29639  219.64  < 2e-16 ***
## DementiaPercent  0.12731    0.03167    4.02 6.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.996 on 1651 degrees of freedom
```

```
## Multiple R-squared:  0.009695,   Adjusted R-squared:  0.009095
## F-statistic: 16.16 on 1 and 1651 DF,  p-value: 6.076e-05
```

```r
ahaslptest$predict<-predict(ahaslplm, newdata=ahaslptest)

mse<-mean((ahaslptest$predict-ahaslptest$DementiaPercent)^2)
mse
```

```
## [1] 3502.947
```

```r
mae<-mean(abs(ahaslptest$predict-ahaslptest$DementiaPercent))
mae
```

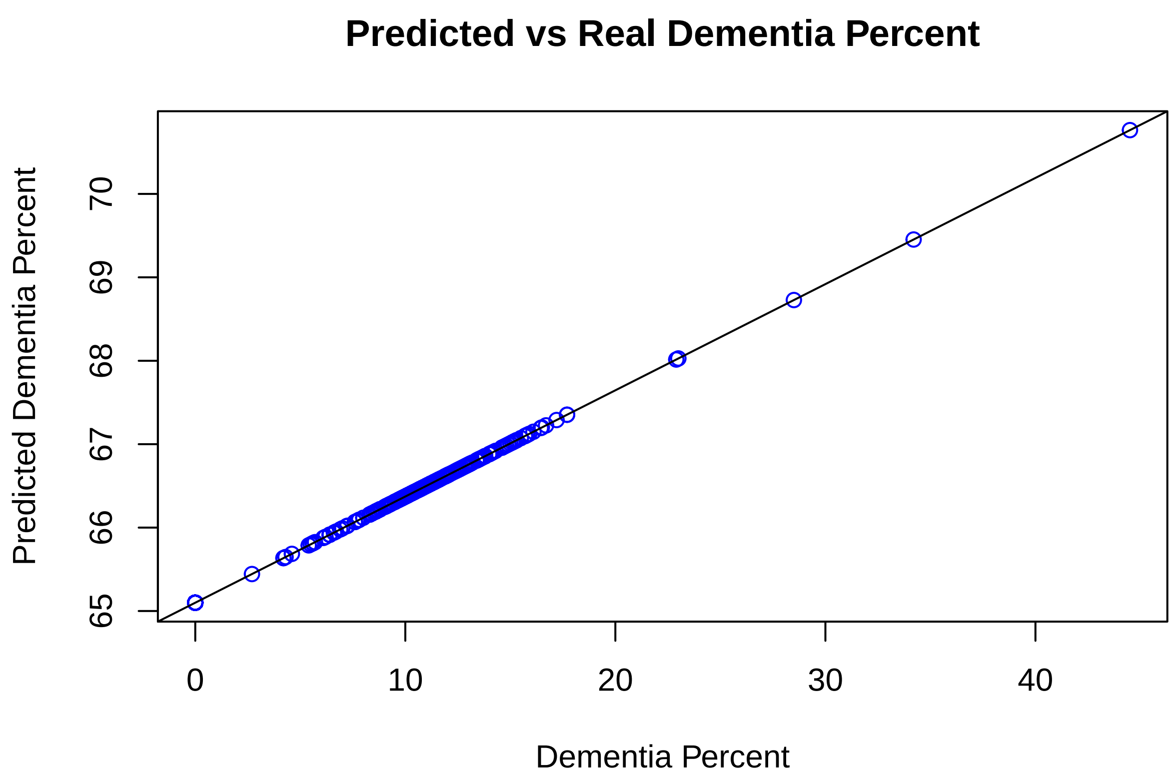
```
## [1] 58.9255
```

```r
rmse<-sqrt(mse)
rmse
```

```
## [1] 59.1857
```

```r
plot(ahaslptest$DementiaPercent, ahaslptest$predict, col="blue", main = "Predicted vs Real Dementia Perc

abline(ahaslplm)
```



Viewing these results, we can see a small p-value and an F-stat falling between the other two samples. We do see a reflection here of our modeling, finally being positive, so we'd say that this is a more significant indicator than exercise, however, not as significant as diet.

5. The individuals I've known to have Alzheimer's and/or dementia have suffered from falls that rapidly increased the speed of mental deterioration. Test this given dataset to see if this rings true. DATA ON FALLS IS FOUND IN AHA.

Another interpretation of this dataset here.

```r
ahainj<-subset(aha, Question =="Percentage of older adults who have fallen and sustained an injury withi

head(ahainj)
```

```
##       YearEnd LocationAbbr
## 2647     2016           US
## 3334     2016          NRE
## 6830     2016           NM
## 7897     2016           US
## 9629     2018          NRE
## 10378    2016           OK
##                                                                                Question
## 2647  Percentage of older adults who have fallen and sustained an injury within last year
## 3334  Percentage of older adults who have fallen and sustained an injury within last year
## 6830  Percentage of older adults who have fallen and sustained an injury within last year
## 7897  Percentage of older adults who have fallen and sustained an injury within last year
## 9629  Percentage of older adults who have fallen and sustained an injury within last year
## 10378 Percentage of older adults who have fallen and sustained an injury within last year
##       Data_Value_Type Data_Value StratificationCategory1   Stratification1
## 2647       Percentage       12.8               Age Group 65 years or older
## 3334       Percentage        8.9               Age Group       50-64 years
## 6830       Percentage       15.8               Age Group           Overall
## 7897       Percentage       11.6               Age Group       50-64 years
## 9629       Percentage        8.9               Age Group           Overall
## 10378      Percentage       13.9               Age Group           Overall
##       StratificationCategoryID2 StratificationID2 DementiaPercent
## 2647                     GENDER            FEMALE            9.60
## 3334                     GENDER              MALE           11.40
## 6830                       RACE               HIS           16.70
## 7897                    OVERALL           OVERALL           10.90
## 9629                       RACE               WHT            9.15
## 10378                      RACE               NAA            0.00
```

```r
nrow(ahainj)
```

```
## [1] 1751
```

```r
set.seed(278613)

ahainjdummy<-sample(c(rep(0, 0.8 * nrow(ahainj)), rep(1, 0.2 * nrow(ahainj))))

table(ahainjdummy)
```

```
## ahainjdummy
##    0    1
## 1400  350
```

```r
ahainjtrain<-ahainj[ahainjdummy==0, ]
ahainjtest<-ahainj[ahainjdummy==1, ]

ahainjlm<-lm(Data_Value ~ DementiaPercent, data=ahainjtrain)
summary(ahainjlm)
```

```
##
## Call:
## lm(formula = Data_Value ~ DementiaPercent, data = ahainjtrain)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.1983 -1.9402 -0.2595  1.4202 19.4960
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     10.67981    0.14025  76.148  < 2e-16 ***
## DementiaPercent  0.04749    0.01415   3.356 0.000812 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.152 on 1399 degrees of freedom
## Multiple R-squared:  0.007986,   Adjusted R-squared:  0.007277
## F-statistic: 11.26 on 1 and 1399 DF,  p-value: 0.0008122
```

```r
ahainjtest$predict<-predict(ahainjlm, newdata=ahainjtest)

mse<-mean((ahainjtest$predict-ahainjtest$DementiaPercent)^2)
mse
```

```
## [1] 38.0886
```

```r
mae<-mean(abs(ahainjtest$predict-ahainjtest$DementiaPercent))
mae
```

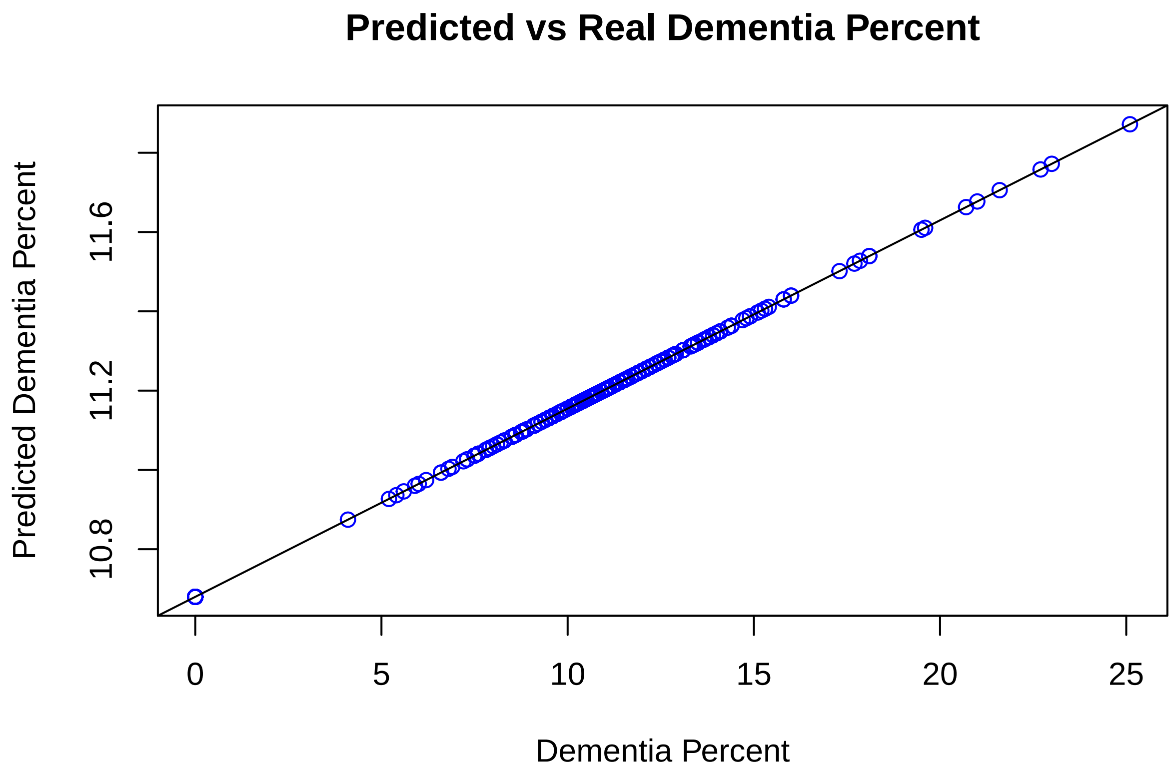
```
## [1] 4.490389
```

```r
rmse<-sqrt(mse)
rmse
```

```
## [1] 6.171596
```

```r
plot(ahainjtest$DementiaPercent, ahainjtest$predict, col="blue", main = "Predicted vs Real Dementia Perc

abline(ahainjlm)
```



One of the most compellingitems here across the board. We see a much smaller mean error value, we have an F stat on par with sleep and diet, and a great p-value to boot. Keep in mind that due to the size of the mses, it's still not a fantastic predictor of dementia, however injuries within the last 12 months seem to be the best correlated to memory issues and dementia so far, which may be biased as it relates to older individuals are more likely to suffer from these falls.

6. Is a certain sex more pre-disposed to having dementia? WE'LL USE MVA DATA HERE

```
smva=mva

split<-sample.split(mva, SplitRatio=0.8)
train<-subset(mva, split == "TRUE")
test <- subset(mva, split == "FALSE")

head(smva)
```

```
##   Subject.ID       Group M.F eTIV
## 1  OAS2_0001 Nondemented   M 1987
## 2  OAS2_0001 Nondemented   M 2004
## 3  OAS2_0002    Demented   M 1678
## 4  OAS2_0002    Demented   M 1738
## 5  OAS2_0002    Demented   M 1698
## 6  OAS2_0004 Nondemented   F 1215
```

```r
model1<-glm(as.factor(Group) ~ M.F, data = train, family = "binomial")
summary(model1)
```

```
##
## Call:
## glm(formula = as.factor(Group) ~ M.F, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2367   0.4136   0.4527   0.4918   0.4918
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.0513     0.2504   8.192 2.57e-16 ***
## M.FM          0.3646     0.4143   0.880    0.379
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 182.05  on 279  degrees of freedom
## Residual deviance: 181.25  on 278  degrees of freedom
## AIC: 185.25
##
## Number of Fisher Scoring iterations: 5
```

```r
predict2<-predict(model1, train, type="response")
confmatrix<-table(Actual_Value=train$Group, Predicted_Value = predict2)
confmatrix
```

```
##              Predicted_Value
## Actual_Value  0.886075949367089 0.918032786885237
##   Converted                  18                10
##   Demented                   45                69
##   Nondemented                95                43
```

```r
(confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)
```

```
## [1] 0.3107143
```

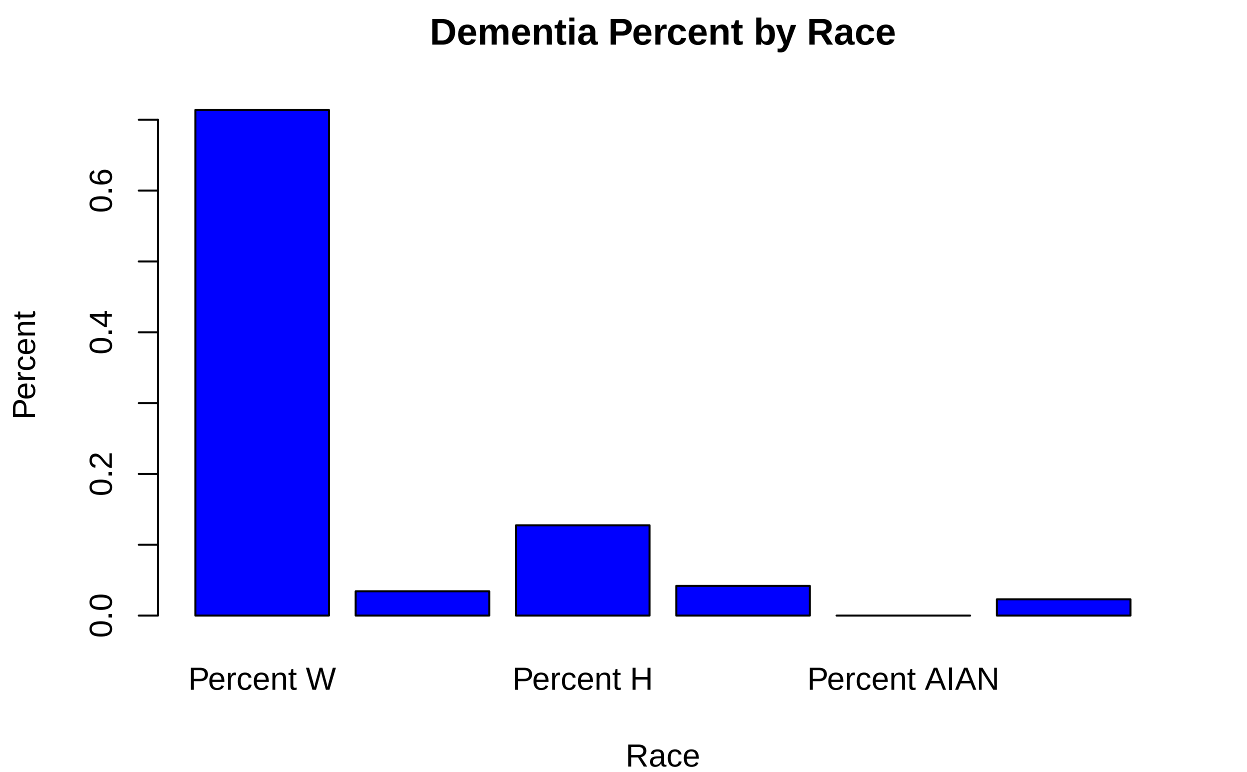
Based on our logistic regression model here, our confidence matrix percent shows an accuracy of a measly 29%. This could be indicative of little correlation and thus predictive power between gender/sex and dementia

7. Does dementia affect certain races more than others? WE'LL USE SDCD DATA HERE, NO PREDICTIVE ELEMENT SINCE WE DO NOT HAVE NON DEMENTED DATA, SO WE'LL NEED TO REPORT ANALYTICS

```r
sdcd2017<-subset(sdcd, sdcd$Year==2017)
percentrow<-c(0,0,0,0,0,0)
labelrow<-c("Percent W", "Percent AA", "Percent H", "Percent API", "Percent AIAN", "Percent O")
for(i in 4:(ncol(sdcd2017)-1))
```

```r
  percentrow[i-3]=(sdcd2017[1,i]/sdcd2017[1,10])


barplot(percentrow, names.arg=labelrow, xlab="Race", ylab="Percent", col="blue", main="Dementia Percent
```




```r
percentrow
```

```
## [1] 0.71390517 0.03432529 0.12748307 0.04193622 0.00000000 0.02298501
```

According to census.gov (https://www.census.gov/quickfacts/fact/table/sandiegocountycalifornia,CA/PST045219), we'd expect roughly 45% to be White, 34% to be Hispanic/Latino, 5% to be African American, and 13% to be Asian/PI.

Based on our output here, the breakdown of those with dementia are made up 71.4% white, 12.7% Hispanic, 3.4% African American, and 4.2% Asian/PI. One thing to keep in mind here, is that even though it shows a pretty heavily biased piece towards white, is that on census data, they broken white into white non-hispanic and white with hispanic. This higher white number shown here might be indicative that these results have not been broken out the same way. However, numbers for that of the asian/pacific islander might show a nice correlation towards less dementia likelihood.

8. Does a specific state have a higher percentage of individuals with dementia? COMPARING STATES ONLY EXISTS IN AHA. Unfortunately we don't have an aggregate number of individuals reporting for the states, so if we were to model the percent of responders who suffered from dementia symptoms for these surveys, it wouldn't tell us anything about whether those states have higher % or even higher numbers of those suffering from dementia. I can't seem to find any studies that report aggregate numbers here, so unfortunately I'll have to throw this question out for now due to lack of resources.

9. Is estimated intracranial volume (eTIV) a solid indicator for predicting dementia? eTIV DATA IS ONLY FOUND IN MVA.

```
mvaeti<-mva

head(mvaeti)
```

```
##   Subject.ID       Group M.F eTIV
## 1  OAS2_0001 Nondemented   M 1987
## 2  OAS2_0001 Nondemented   M 2004
## 3  OAS2_0002    Demented   M 1678
## 4  OAS2_0002    Demented   M 1738
## 5  OAS2_0002    Demented   M 1698
## 6  OAS2_0004 Nondemented   F 1215
```

```
for(i in 1:nrow(mvaeti)){
  if (mvaeti[i, 2]=="Demented" | mvaeti[i,2]=="Converted")
    mvaeti[i,2]=1
  else
    mvaeti[i,2]=0
}
head(mvaeti)
```

```
##   Subject.ID Group M.F eTIV
## 1  OAS2_0001     0   M 1987
## 2  OAS2_0001     0   M 2004
## 3  OAS2_0002     1   M 1678
## 4  OAS2_0002     1   M 1738
## 5  OAS2_0002     1   M 1698
## 6  OAS2_0004     0   F 1215
```

```
split<-sample.split(mvaeti, SplitRatio=0.8)
train<-subset(mvaeti, split == "TRUE")
test <- subset(mvaeti, split == "FALSE")

model1<-glm(as.factor(Group) ~ eTIV, data = train, family = "binomial")
summary(model1)
```

```
##
## Call:
## glm(formula = as.factor(Group) ~ eTIV, family = "binomial", data = train)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.250  -1.169  -1.065   1.183   1.294
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.7606751  1.0236768   0.743    0.457
## eTIV        -0.0005267  0.0006785  -0.776    0.438
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 388.11  on 279  degrees of freedom
## Residual deviance: 387.50  on 278  degrees of freedom
## AIC: 391.5
##
## Number of Fisher Scoring iterations: 3
```

```
predict2<-predict(model1, test, type="response")
confmatrix<-table(Actual_Value=test$Group, Predicted_Value = predict2>0.5)
confmatrix
```

```
##             Predicted_Value
## Actual_Value FALSE TRUE
##            0    23   25
##            1    23   22
```

```
(confmatrix[[1,1]] + confmatrix[[2,2]]) / sum(confmatrix)
```

```
## [1] 0.483871
```

Looking like a 47% accuracy level here with the old fashioned logistic regression. As a result, eTIV looks like a relatively significant measure of dementia symptoms being exhibited as well.