PH 245 Final Project - Flu Absenteeism

In [1]:

```
library(data.table)
library(boot)

prefix = "../absentee/Combined-data/"
filenames = c("absentee_all.csv","absentee-flu.csv", "absentee-nonflu.csv", "ILIData_CA_201101_201739.csv",
              "absentee.RData"
            )
```

In [2]:

```
# Loading Data (using high-speed data.tables)
absenteeData = fread( file=paste(prefix, filenames[1], sep=""), stringsAsFactors=TRUE )
```

```
Read 42797568 rows and 9 (of 9) columns from 2.816 GB file in 00:00:23
```

In [3]:

```
head(absenteeData)
colnames(absenteeData)

# Creating a smaller sample for use until final analysis
smallSampleSize = 1000000
#absenteeData = absenteeData[sample(.N, 1000000)]
nrow(absenteeData)
```

schoolyr	date	grade	race	absent_nonill	absent_ill	dist	school	matchid
2011-12	29aug2011	0	African American	0	0	OUSD	ACORN Woodland Elementary	0
2011-12	29aug2011	0	African American	0	0	OUSD	ACORN Woodland Elementary	0
2011-12	29aug2011	0	African American	0	0	OUSD	ACORN Woodland Elementary	0
2011-12	29aug2011	0	Asian	0	0	OUSD	ACORN Woodland Elementary	0
2011-12	29aug2011	0	Latino	0	0	OUSD	ACORN Woodland Elementary	0
2011-12	29aug2011	0	Latino	0	0	OUSD	ACORN Woodland Elementary	0

'schoolyr' 'date' 'grade' 'race' 'absent_nonill' 'absent_ill' 'dist' 'school' 'matchid'

42797568

In [4]:

```
# Cleaning data and adding more useful variables

absenteeData=absenteeData[,date:=as.Date(absenteeData$date, "%d%b%Y")]
absenteeData=absenteeData[,month:=as.numeric(format(absenteeData$date, "%m"))]
absenteeData=absenteeData[,week:=week(date)]
absenteeData=absenteeData[,yr:=year(date)]

absenteeData$fluseasCDC = ifelse(absenteeData$month <= 4 | absenteeData$month >= 10, 1, 0)

absenteeData$dist.n = ifelse(absenteeData$dist == "OUSD", 1, 0)

absenteeData$grade = as.factor(absenteeData$grade)

absenteeData$race <- factor(absenteeData$race, levels = c("White","African American",
      "Asian","Latino","Multiple Ethnicity","Native American","Not Reported",
      "Pacific Islander"))

# Since WCCUSD has different labeling and fewer races reported that OUSD,
# reduce all races to subset for uniformity
absenteeData = absenteeData[race %in% c("Native American", "Multiple Ethnicity", "Not Reported"),
                            race := "Don't know Other"]

# The sum of any row will be 0 if there was no absence
# or 1 if there was an absence for any reason
absenteeData$absence = absenteeData$absent_nonill + absenteeData$absent_ill

# End result
head(absenteeData)
```

schoolyr	date	grade	race	absent_nonill	absent_ill	dist	school	matchid	month	week	yr	fluseasCDC	dist.n	absence
2011-12	2011-08-29	0	African American	0	0	OUSD	ACORN Woodland Elementary	0	8	35	2011	0	1	0
2011-12	2011-08-29	0	African American	0	0	OUSD	ACORN Woodland Elementary	0	8	35	2011	0	1	0
2011-12	2011-08-29	0	African American	0	0	OUSD	ACORN Woodland Elementary	0	8	35	2011	0	1	0
2011-12	2011-08-29	0	Asian	0	0	OUSD	ACORN Woodland Elementary	0	8	35	2011	0	1	0
2011-12	2011-08-29	0	Latino	0	0	OUSD	ACORN Woodland Elementary	0	8	35	2011	0	1	0
2011-12	2011-08-29	0	Latino	0	0	OUSD	ACORN Woodland Elementary	0	8	35	2011	0	1	0

Exploratory Data Analysis (EDA)

The first, most important thing to do is examine how many absences ocurred in total. Then, we'll break it down year by year and examine absences.

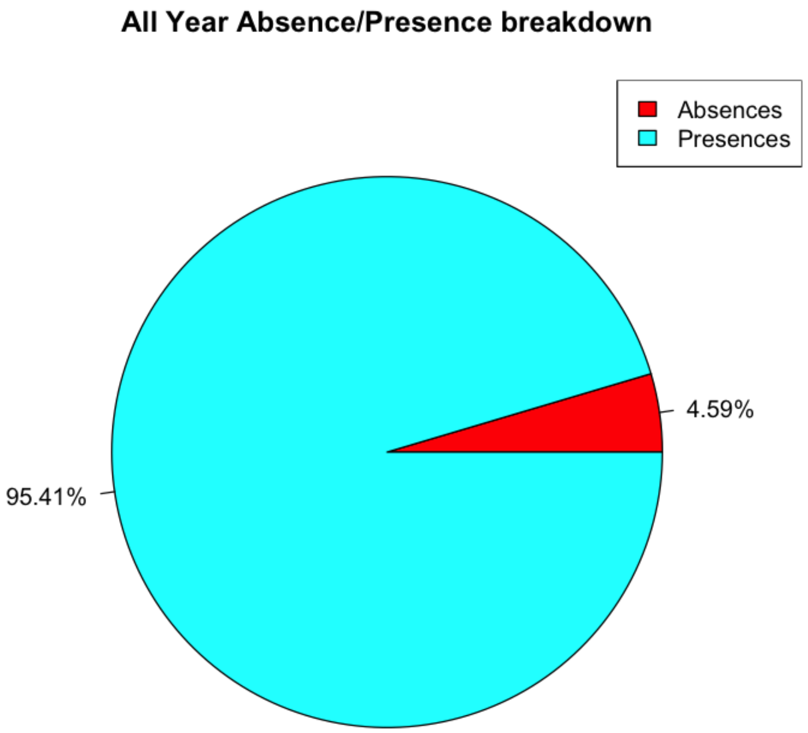
Absences are defined within the absent_nonill and absent_ill columns. Both columns having a 0 means the student was present. A 1 appears in one of the columns if there was an absence.

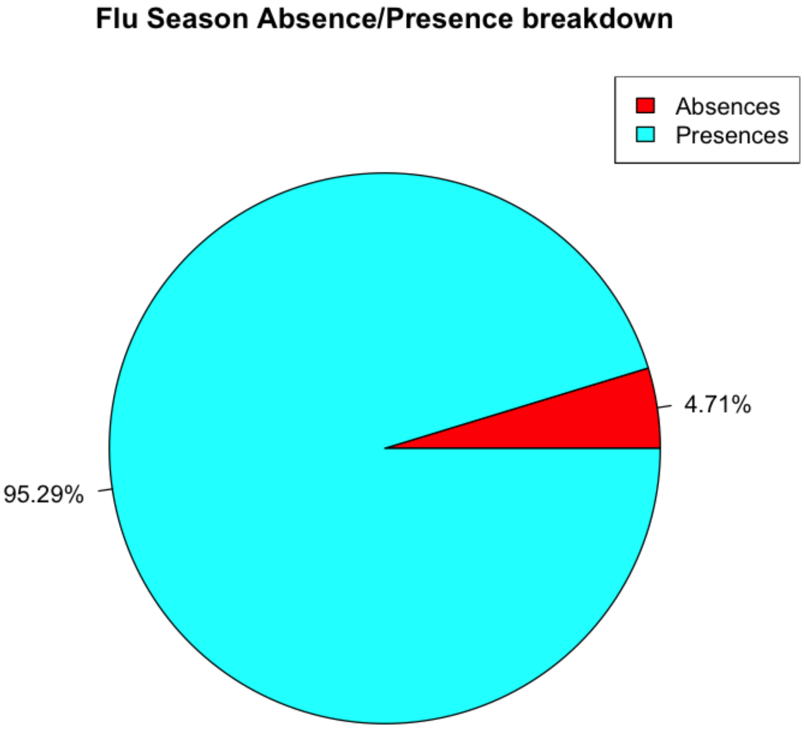
In examining our dataset, some other good things to understand include racial breakdown and grade distribution.

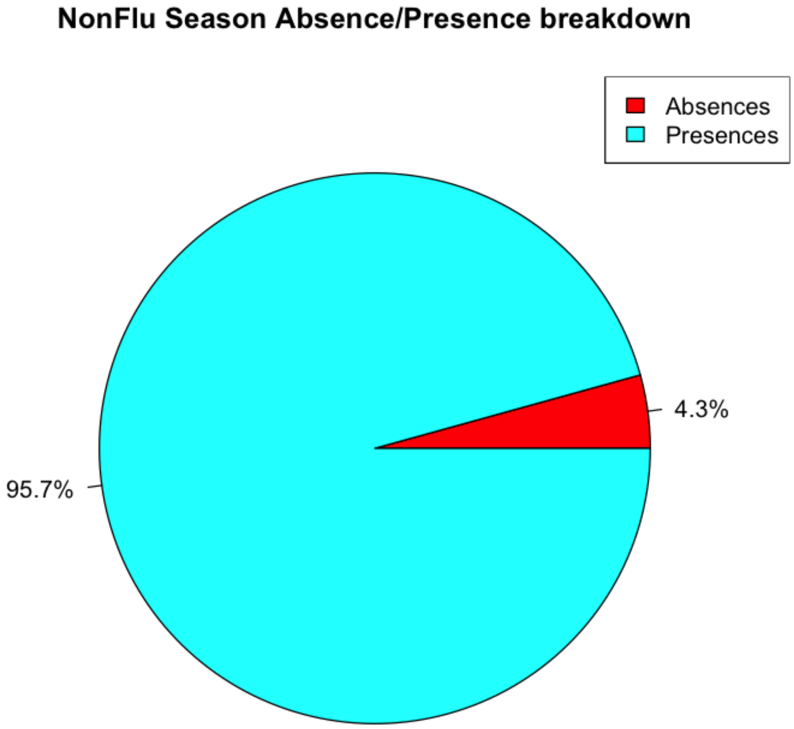
```
In [5]: # Beginning Exploratory Data Analysis
        summary(absenteeData)
```

```
    schoolyr            date               grade                    race         
 2011-12:7210087   Min.   :2011-08-22   0:7358767   Latino          :19605457  
 2012-13:7313735   1st Qu.:2013-01-09   1:6864107   African American: 9528492  
 2013-14:7198778   Median :2014-05-29   2:6746732   Asian           : 6368717  
 2014-15:7193413   Mean   :2014-07-04   3:6616643   White           : 5602174  
 2015-16:7057935   3rd Qu.:2016-01-05   4:6524273   Don't know Other: 1285421  
 2016-17:6823620   Max.   :2017-06-09   5:6254788   Pacific Islander:  407307  
                                        6:2432258   (Other)         :       0  
 absent_nonill       absent_ill           dist         
 Min.   :0.00000   Min.   :0.00000   OUSD  :21764262  
 1st Qu.:0.00000   1st Qu.:0.00000   WCCUSD:21033306  
 Median :0.00000   Median :0.00000                    
 Mean   :0.02254   Mean   :0.02339                    
 3rd Qu.:0.00000   3rd Qu.:0.00000                    
 Max.   :1.00000   Max.   :1.00000                    
                                                      
                 school            matchid          month       
 Lincoln Elementary : 1343324   Min.   : 0.00   Min.   : 1.000  
 Dover Elementary   :  941442   1st Qu.: 3.00   1st Qu.: 3.000  
 Bayview Elementary :  816787   Median :14.00   Median : 5.000  
 Downer Elementary  :  815497   Mean   :14.33   Mean   : 6.297  
 Franklin Elementary:  814717   3rd Qu.:25.00   3rd Qu.:10.000  
 Chavez Elementary  :  799861   Max.   :34.00   Max.   :12.000  
 (Other)            :37265940                                   
      week             yr          fluseasCDC         dist.n      
 Min.   : 1.00   Min.   :2011   Min.   :0.0000   Min.   :0.0000  
 1st Qu.:11.00   1st Qu.:2013   1st Qu.:0.0000   1st Qu.:0.0000  
 Median :21.00   Median :2014   Median :1.0000   Median :1.0000  
 Mean   :25.72   Mean   :2014   Mean   :0.7071   Mean   :0.5085  
 3rd Qu.:42.00   3rd Qu.:2016   3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :53.00   Max.   :2017   Max.   :1.0000   Max.   :1.0000  
                                                                 
    absence       
 Min.   :0.00000  
 1st Qu.:0.00000  
 Median :0.00000  
 Mean   :0.04593  
 3rd Qu.:0.00000  
 Max.   :1.00000  
```

```
In [6]: pieAbsenceBreakdown = function(data, pieTitle) {
            "Creates a pie chart of the absences and presences in dataset"
            numAbsences = sum(data$absence)
            numPresences = length(data$absence) - numAbsences
            rawBreakdown = c(numAbsences, numPresences)

            piePercent = paste(round(100*rawBreakdown/sum(rawBreakdown), 2), "%", sep="")

            pie(rawBreakdown,
                labels=piePercent,
                col=rainbow(length(rawBreakdown)),
                main=pieTitle
               )

            legend("topright",
                   c("Absences","Presences"),
                   fill=rainbow(length(rawBreakdown))
                  )
        }

        # Examining total absence/presence breakdown
        pieAbsenceBreakdown(data=absenteeData, pieTitle="All Year Absence/Presence breakdown")


        # Examining flu-specific absence/presence breakdown
        fluData = absenteeData[fluseasCDC==1]
        nonFluData = absenteeData[fluseasCDC==0]

        pieAbsenceBreakdown(data=fluData, pieTitle="Flu Season Absence/Presence breakdown")
        pieAbsenceBreakdown(data=nonFluData, pieTitle="NonFlu Season Absence/Presence breakdown")
```



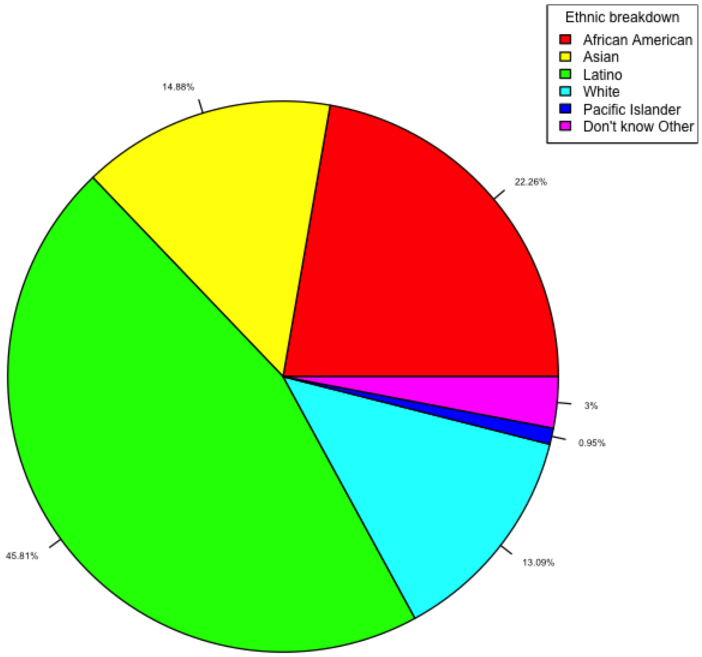

Flu Season Absence/Presence breakdown
🟥 Absences
🟦 Presences
4.71%
95.29%


NonFlu Season Absence/Presence breakdown
Absences
Presences
4.3%
95.7%

In [7]:

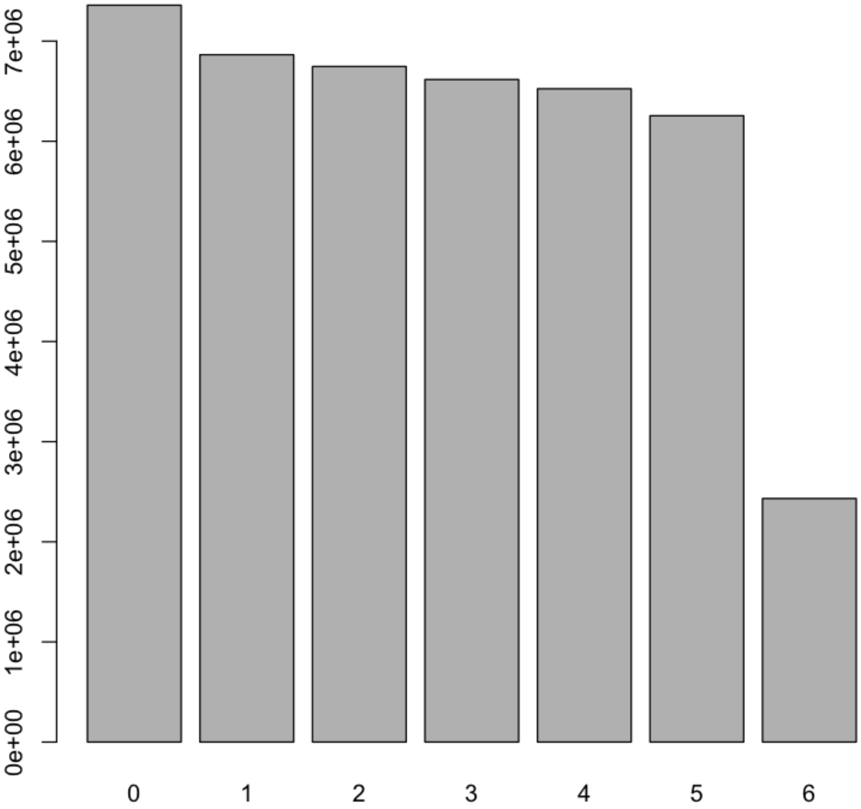
```
# Creating a pie chart of ethnicities

races = absenteeData[,.N,by="race"]
piePercent2 = paste(round(100*races$N/sum(races$N), 2), "%", sep="")

pie(x=races$N, labels=piePercent2, col=rainbow(length(races$race)), cex = 0.4)
legend("topright", legend=races$race, fill=rainbow(length(races$race)), cex = 0.6, title="Ethnic breakdown")
races
```

race	N
African American	9528492
Asian	6368717
Latino	19605457
White	5602174
Pacific Islander	407307
Don't know Other	1285421



In [8]:

```
# Examining overall grade distribution
grades = absenteeData[,.N,by="grade"][order(grade)]

barplot(grades$N, names.arg=grades$grade)
```



In [9]:

```
# Sixth graders are all from one district - drop all sixth graders
sixthGraders = absenteeData[grade==6]
unique(sixthGraders$dist)

head(sixthGraders)

fullNumRows = nrow(absenteeData)
absenteeData = absenteeData[grade != 6]
print(paste("Lost", (fullNumRows-nrow(absenteeData)), "rows in eliminating sixth graders.",
            nrow(absenteeData), "rows remain")
     )
```

WCCUSD

schoolyr	date	grade	race	absent_nonill	absent_ill	dist	school	matchid	month	week	yr	fluseasCDC	dist.n	absence
2011-12	2011-08-22	6	African American	0	0	WCCUSD	Bayview Elementary	34	8	34	2011	0	0	0
2011-12	2011-08-22	6	African American	0	0	WCCUSD	Bayview Elementary	34	8	34	2011	0	0	0
2011-12	2011-08-22	6	African American	0	0	WCCUSD	Bayview Elementary	34	8	34	2011	0	0	0
2011-12	2011-08-22	6	African American	0	0	WCCUSD	Bayview Elementary	34	8	34	2011	0	0	0
2011-12	2011-08-22	6	African American	0	0	WCCUSD	Bayview Elementary	34	8	34	2011	0	0	0
2011-12	2011-08-22	6	African American	0	0	WCCUSD	Bayview Elementary	34	8	34	2011	0	0	0

```
[1] "Lost 2432258 rows in eliminating sixth graders. 40365310 rows remain"
```

Interpreting Our EDA Results

So, we see that we have a relatively small number of absences in our overall dataset (this is good!). Since we have a huge sample size, we'll have plenty of absences to examine.

The first thing we did is examine overall number of absences during flu season versus during the nonflu season. As one would expect, flu season had slightly a slightly greater percentage of students absent.

In the rest of our EDA, we explored the ethnic breakdown and grade distributions of our dataset. One thing to note is that our subject population is quite different in terms of ethnic breakdown from the entire United States, so our projects extensibility to other populations with different breakdowns is a bit less certain.

One thing to note is that our 6th grade population is so small because only one of the two school districts contributed data to that bin, so for this analysis, we'll proceed analyzing only grades K-5.

Analyzing Absenteeism Variation among Matched Schools

To continue, let's try to understand how much variation in absenteeism there was between matched schools during the nonflu season. This will be important as a baseline for analyzing the variance between the same matched schools during flu season when the intervention took place. Schools that were matched have matchid's that are *not* 0.

In [10]:

```
# Calculating the average percentage of absences per school
# For now, we'll only include the intervention time period
nonFluDataInterventionTime = nonFluData[nonFluData$yr > 2014 | nonFluData$schoolyr == "2014-15"]

nonFluAbsenceAverages = nonFluDataInterventionTime[,.(absenceAverage=mean(absence)),by=c("matchid", "dist", "school")][order(matchid, c
head(nonFluAbsenceAverages)
tail(nonFluAbsenceAverages)
```

matchid	dist	school	absenceAverage
0	OUSD	ACORN Woodland Elementary	0.03588439
0	OUSD	Esperanza Elementary	0.04137591
0	OUSD	Futures Elementary	0.07901656
0	OUSD	Greenleaf Elementary	0.03681576
0	OUSD	Hillcrest School (K-8)	0.01849695
0	OUSD	Hoover Elementary	0.06010090

matchid	dist	school	absenceAverage
32	OUSD	Parker Elementary	0.06488845
32	WCCUSD	Lincoln Elementary	0.06025072
33	OUSD	Bridges Academy	0.04812210
33	WCCUSD	Chavez Elementary	0.04812621
34	OUSD	Manzanita Community School	0.06330087
34	WCCUSD	Bayview Elementary	0.05490917

In [11]:

```
# Drop schools that were not matched by the matching algorithm and group by matchid
nonFluMatchedAbsenceAverages = nonFluAbsenceAverages[matchid != 0][order(matchid, dist)]
head(nonFluMatchedAbsenceAverages)
```

matchid	dist	school	absenceAverage
1	OUSD	Horace Mann Elementary	0.06545300
1	WCCUSD	Sheldon Elementary	0.04343917
2	OUSD	Emerson Elementary	0.05102712
2	WCCUSD	Shannon Elementary	0.05075521
3	OUSD	Laurel Elementary	0.04668948
3	WCCUSD	Tara Hills Elementary	0.05095789

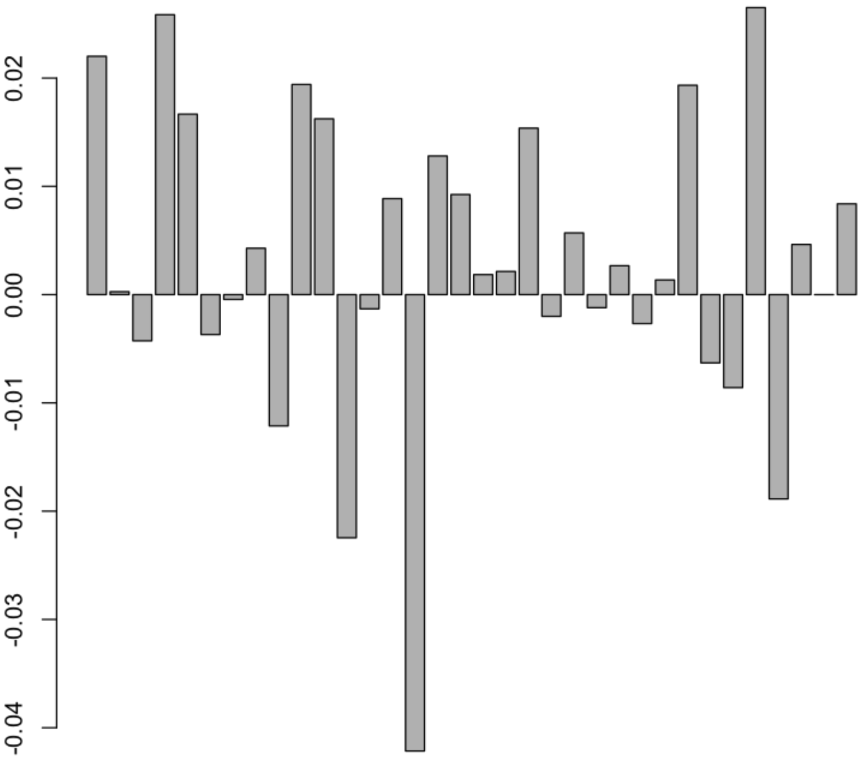
In [12]:

```
# Let's find the baseline difference between the two groups for each matched school

OUSDNonFlu = nonFluMatchedAbsenceAverages[dist=="OUSD"][order(matchid)]
WCCUSDNonFlu = nonFluMatchedAbsenceAverages[dist=="WCCUSD"][order(matchid)]

differenceNonFlu = OUSDNonFlu[,difference:=(OUSDNonFlu$absenceAverage - WCCUSDNonFlu$absenceAverage)][,c("matchid", "difference")]
head(differenceNonFlu)
barplot(differenceNonFlu$difference)

print("Mean difference in percentage of absences between matched pairs of schools during nonflu season")
mean(differenceNonFlu$difference)
```

matchid	difference
1	0.0220138316
2	0.0002719033
3	-0.0042684101
4	0.0258548299
5	0.0166706288
6	-0.0036856728

```
[1] "Mean difference in percentage of absences between matched pairs of schools during nonflu season"
```

0.00286923396948978



In [13]:

```
# Now, let's repeat the same set of steps to analyze whether the intervention seemed to have any effect.
# We would expect OUSD, which had the intervention, to have absenteeism less impacted by illness.
# On the other hand WCCUSD, which did not have any intervention
# would have greater absenteeism as flu became more prevalent during flu season.
# Thus, we would expect a downward shift in the barplot
fluDataInterventionTime = fluData[fluData$yr > 2014 | fluData$schoolyr == "2014-15"]


fluAbsenceAverages = fluDataInterventionTime[,.(absenceAverage=mean(absence)),by=c("matchid", "dist", "school")][order(matchid, dist)]
fluMatchedAbsenceAverages = fluAbsenceAverages[matchid != 0][order(matchid, dist)]
OUSDFlu = fluMatchedAbsenceAverages[dist=="OUSD"][order(matchid)]
WCCUSDFlu = fluMatchedAbsenceAverages[dist=="WCCUSD"][order(matchid)]

differenceFlu = OUSDFlu[,difference:=(OUSDFlu$absenceAverage - WCCUSDFlu$absenceAverage)][,c("matchid", "difference")]
head(differenceFlu)
barplot(differenceFlu$difference, col="black")

print("Mean difference in percentage of absences between matched pairs of schools during flu season")
mean(differenceFlu$difference)

# Calculate the percentage of schools where expected "downward shift" during flu season occurred
print("Percentage of matched pairs with expected downward shift:")
sum(differenceFlu$difference < differenceNonFlu$difference)/length(differenceFlu$difference)
```

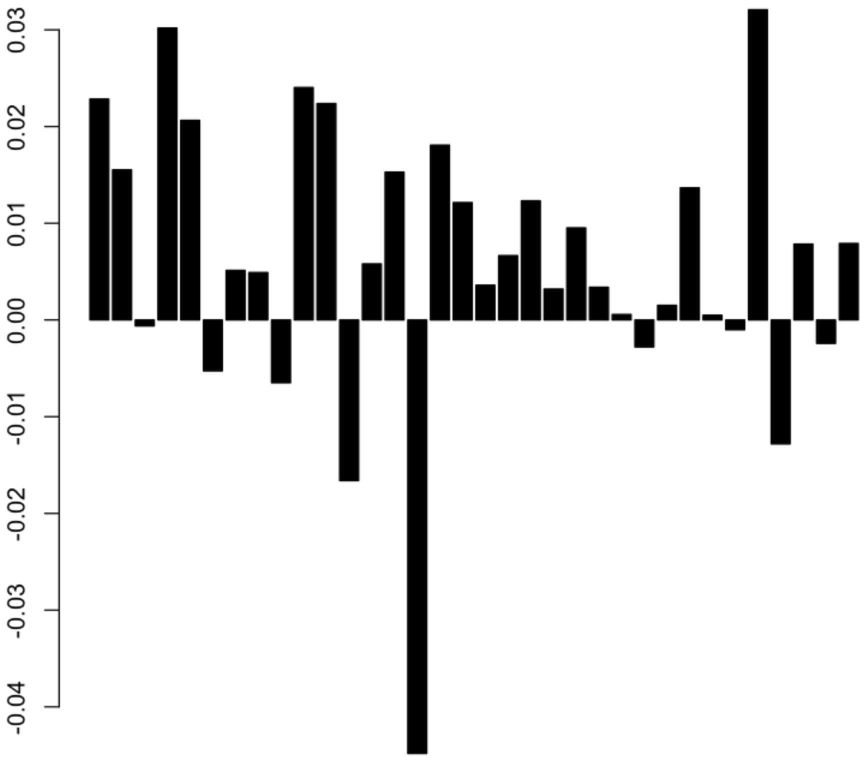
matchid	difference
1	0.0228454306
2	0.0155312832
3	-0.0006223461
4	0.0301872645
5	0.0206435702
6	-0.0052682861

```
[1] "Mean difference in percentage of absences between matched pairs of schools during flu season"
```

0.00608590693643027

```
[1] "Percentage of matched pairs with expected downward shift:"
```

0.235294117647059



Interpreting the result

This is... mildly worrying, if I'm interpreting the data correctly, though the test we ran was rather informal and intended to understand whether the data would fit to our intuitions. However, it seems as if schools receiving the intervention actually had a larger increase in absenteeism during the flu season vs rest of the year compared to the matched control group which did not receive the intervention. While our analysis did not look at illness specific data (which is pretty important to making an actual conclusion), the trends in the data are very counterintuitive.

Moving Forward

Nevertheless, we'll move on to fitting statistical models for linear and logistic regression in an attempt to be able to predict how certain factors affect all-cause and illness specific absenteeism.

In [14]:

```
# Since we're generating predictions with regression, need to bring in other school-specific variables to fit on

getSchoolData = function(aggregationData, dropColumns, aggregationColumns) {
    oldw <- getOption("warn")
    options(warn = -1)

    cleanAggregationData = aggregationData[,(dropColumns):=NULL]
    groupedSchoolData = cleanAggregationData[,head(.SD, 1),by=aggregationColumns]

    options(warn = oldw)

    print(paste("Data collected for", nrow(groupedSchoolData), "schools"))

    return(groupedSchoolData)
}

# Dropping irrelevant columns (for specific schools) from aggregation data
dropColumns = c("V1", "schoolyr", "date", "grade", "race", "absent_nonill", "absent_ill",
                "matchid", "month", "flusesn", "absent_all", "weekending", "peakwk", "week", "yr",
                "fluseasCDPH", "fluseasCDC"
               )

aggregationColumns = c("dist", "school", "enrolled") # Unique identifying key for a school

#load(file = paste(prefix, filenames[5], sep=""))
attach(paste(prefix, filenames[5], sep=""));
flu = flu;
detach()

schoolData = getSchoolData(aggregationData=flu, dropColumns=dropColumns, aggregationColumns=aggregationColumns)
head(schoolData)
colnames(schoolData)
```

```
[1] "Data collected for 68 schools"
```

dist	school	enrolled	mn.class.size	per.not_hsg	per.hsg	per.some_col	per.col_grad	per.grad_sch	per.englearn	per.freelunch	API13	API12	mean.cst.ela	per.adv.ela	pe
OUSD	Allendale Elementary	425	26.56250	27	33	27	12	3	41.17647	79.91	663	725	329.625	8.75	
WCCUSD	Bayview Elementary	685	28.54167	31	46	17	6	1	53.57664	73.37	675	681	321.000	9.80	
OUSD	Bella Vista Elementary	525	21.87500	24	31	24	15	5	42.28571	75.33	813	849	369.825	29.75	
OUSD	Bridges Academy	381	19.05000	55	30	11	3	1	79.26509	77.00	678	715	320.050	9.75	
OUSD	Brookfield Village Elementary	367	16.68182	41	33	15	8	2	58.03815	66.21	687	738	329.675	8.25	
OUSD	Burckhalter Elementary	298	22.92308	12	22	40	20	6	11.74497	71.81	769	808	358.950	22.50	

```
'dist'  'school'  'enrolled'  'mn.class.size'  'per.not_hsg'  'per.hsg'  'per.some_col'  'per.col_grad'  'per.grad_sch'  'per.englearn'  'per.freelunch'  'API13'  'API12'
'mean.cst.ela'  'per.adv.ela'  'per.basic.ela'  'mean.cst.m'  'per.adv.m'  'per.basic.m'  'dist.n'
```

In [15]:
```
# Merging school level data into our set of patients
combinedFluDataInterventionTime = merge(x=fluDataInterventionTime[matchid!=0,!c("schoolyr", "date", "absence")],
                                        y=schoolData,
                                        by=c("dist", "school", "dist.n")
                                       )
head(combinedFluDataInterventionTime)
colnames(combinedFluDataInterventionTime)
```

dist	school	dist.n	grade	race	absent_nonill	absent_ill	matchid	month	week	⋯	per.englearn	per.freelunch	API13	API12	mean.cst.ela	per.adv.ela	per.basic.ela	r
OUSD	Allendale Elementary	1	0	African American	0	0	14	10	40	⋯	41.17647	79.91	663	725	329.625	8.75	38.5	
OUSD	Allendale Elementary	1	0	African American	0	0	14	10	40	⋯	41.17647	79.91	663	725	329.625	8.75	38.5	
OUSD	Allendale Elementary	1	0	African American	0	0	14	10	40	⋯	41.17647	79.91	663	725	329.625	8.75	38.5	
OUSD	Allendale Elementary	1	0	African American	0	0	14	10	40	⋯	41.17647	79.91	663	725	329.625	8.75	38.5	
OUSD	Allendale Elementary	1	0	African American	0	0	14	10	40	⋯	41.17647	79.91	663	725	329.625	8.75	38.5	
OUSD	Allendale Elementary	1	0	African American	0	0	14	10	40	⋯	41.17647	79.91	663	725	329.625	8.75	38.5	

```
'dist'   'school'   'dist.n'   'grade'   'race'   'absent_nonill'   'absent_ill'   'matchid'   'month'   'week'   'yr'   'fluseasCDC'   'enrolled'   'mn.class.size'   'per.not_hsg'   'per.hsg'
'per.some_col'   'per.col_grad'   'per.grad_sch'   'per.englearn'   'per.freelunch'   'API13'   'API12'   'mean.cst.ela'   'per.adv.ela'   'per.basic.ela'   'mean.cst.m'   'per.adv.m'
'per.basic.m'
```

In [16]:
```
# Fitting logistic regression for illness-specific absenteeism and nonspecific absenteeism

glm.log.ill = glm(absent_ill~., data=combinedFluDataInterventionTime[,!c("dist", "school", "absent_nonill", "matchid")])
glm.log.nonill = glm(absent_nonill~., data=combinedFluDataInterventionTime[,!c("dist", "school", "absent_ill", "matchid")])

summary(glm.log.ill)
summary(glm.log.nonill)
```

```
Call:
glm(formula = absent_ill ~ ., data = combinedFluDataInterventionTime[,
    !c("dist", "school", "absent_nonill", "matchid")])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.05454  -0.03151  -0.02558  -0.01983   0.99889

Coefficients: (1 not defined because of singularities)
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.231e-01  1.138e-01   3.718 0.000201 ***
dist.n                1.100e-03  2.212e-04   4.974 6.55e-07 ***
grade1               -6.267e-03  1.625e-04 -38.559  < 2e-16 ***
grade2               -8.792e-03  1.622e-04 -54.202  < 2e-16 ***
grade3               -1.110e-02  1.620e-04 -68.508  < 2e-16 ***
grade4               -1.188e-02  1.625e-04 -73.128  < 2e-16 ***
grade5               -1.241e-02  1.640e-04 -75.691  < 2e-16 ***
grade6               -1.221e-02  2.135e-04 -57.200  < 2e-16 ***
raceAfrican American  3.874e-03  1.849e-04  20.953  < 2e-16 ***
raceAsian            -6.234e-03  1.908e-04 -32.675  < 2e-16 ***
raceLatino            1.995e-03  1.750e-04  11.399  < 2e-16 ***
racePacific Islander  1.776e-03  4.721e-04   3.762 0.000169 ***
raceDon't know Other  2.188e-03  3.139e-04   6.970 3.17e-12 ***
month                 3.144e-03  1.576e-04  19.953  < 2e-16 ***
week                 -9.273e-04  3.624e-05 -25.592  < 2e-16 ***
yr                   -1.456e-04  5.629e-05  -2.586 0.009703 **
fluseasCDC                   NA         NA      NA       NA
enrolled             -1.409e-05  5.221e-07 -26.994  < 2e-16 ***
mn.class.size        -1.960e-05  2.507e-05  -0.782 0.434300
per.not_hsg          -1.424e-03  7.895e-05 -18.036  < 2e-16 ***
per.hsg              -1.245e-03  7.777e-05 -16.008  < 2e-16 ***
per.some_col         -1.373e-03  7.984e-05 -17.195  < 2e-16 ***
per.col_grad         -9.076e-04  7.814e-05 -11.615  < 2e-16 ***
per.grad_sch         -1.194e-03  7.587e-05 -15.740  < 2e-16 ***
per.englearn          2.976e-06  7.935e-06   0.375 0.707639
per.freelunch         1.675e-04  7.640e-06  21.926  < 2e-16 ***
API13                 1.229e-05  2.494e-06   4.926 8.39e-07 ***
API12                -2.774e-05  3.311e-06  -8.379  < 2e-16 ***
mean.cst.ela          3.677e-04  1.608e-05  22.865  < 2e-16 ***
per.adv.ela          -8.760e-04  2.545e-05 -34.425  < 2e-16 ***
per.basic.ela        -4.865e-04  1.708e-05 -28.475  < 2e-16 ***
mean.cst.m           -1.941e-04  1.165e-05 -16.664  < 2e-16 ***
per.adv.m             4.086e-04  2.677e-05  15.262  < 2e-16 ***
per.basic.m           3.416e-04  1.736e-05  19.671  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.02536359)

    Null deviance: 303319  on 11927280  degrees of freedom
Residual deviance: 302518  on 11927248  degrees of freedom
AIC: -9977892

Number of Fisher Scoring iterations: 2


Call:
glm(formula = absent_nonill ~ ., data = combinedFluDataInterventionTime[,
    !c("dist", "school", "absent_ill", "matchid")])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.05385  -0.02859  -0.02224  -0.01425   1.00480

Coefficients: (1 not defined because of singularities)
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)          -2.643e+00  1.052e-01 -25.117  < 2e-16 ***
dist.n                1.900e-03  2.046e-04   9.288  < 2e-16 ***
grade1               -5.700e-03  1.503e-04 -37.929  < 2e-16 ***
grade2               -7.395e-03  1.500e-04 -49.299  < 2e-16 ***
grade3               -8.038e-03  1.498e-04 -53.656  < 2e-16 ***
grade4               -8.148e-03  1.503e-04 -54.223  < 2e-16 ***
grade5               -7.733e-03  1.517e-04 -50.985  < 2e-16 ***
grade6               -6.651e-03  1.974e-04 -33.692  < 2e-16 ***
raceAfrican American  1.136e-02  1.710e-04  66.435  < 2e-16 ***
raceAsian            -5.145e-03  1.764e-04 -29.159  < 2e-16 ***
raceLatino            6.676e-04  1.618e-04   4.125 3.71e-05 ***
racePacific Islander  8.244e-03  4.366e-04  18.881  < 2e-16 ***
raceDon't know Other  4.134e-03  2.903e-04  14.241  < 2e-16 ***
month                 1.453e-04  1.457e-04   0.997   0.3186
week                 -7.538e-05  3.351e-05  -2.250   0.0245 *
yr                    1.220e-03  5.206e-05  23.433  < 2e-16 ***
fluseasCDC                   NA         NA      NA       NA
enrolled              1.246e-05  4.828e-07  25.802  < 2e-16 ***
mn.class.size        -1.719e-04  2.319e-05  -7.414 1.23e-13 ***
per.not_hsg           1.858e-03  7.302e-05  25.451  < 2e-16 ***
per.hsg               1.671e-03  7.192e-05  23.232  < 2e-16 ***
per.some_col          1.601e-03  7.383e-05  21.690  < 2e-16 ***
per.col_grad          1.594e-03  7.227e-05  22.055  < 2e-16 ***
per.grad_sch          1.499e-03  7.016e-05  21.363  < 2e-16 ***
per.englearn         -2.260e-04  7.338e-06 -30.797  < 2e-16 ***
per.freelunch        -8.037e-05  7.066e-06 -11.375  < 2e-16 ***
API13                -7.108e-05  2.307e-06 -30.817  < 2e-16 ***
API12                -7.146e-05  3.062e-06 -23.341  < 2e-16 ***
mean.cst.ela          1.553e-04  1.487e-05  10.440  < 2e-16 ***
per.adv.ela           1.609e-05  2.353e-05   0.684   0.4942    
per.basic.ela         3.794e-04  1.580e-05  24.013  < 2e-16 ***
mean.cst.m            3.098e-04  1.077e-05  28.767  < 2e-16 ***
per.adv.m            -4.359e-04  2.476e-05 -17.605  < 2e-16 ***
per.basic.m          -1.476e-04  1.606e-05  -9.191  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.02169131)

    Null deviance: 259858  on 11927280  degrees of freedom
Residual deviance: 258718  on 11927248  degrees of freedom
AIC: -11843356

Number of Fisher Scoring iterations: 2
```

In [17]:

```
# Using Cross-Validation to estimate prediction error of our two models

oldw <- getOption("warn")
options(warn = -1)

cv.log.ill.predError = cv.glm(data=combinedFluDataInterventionTime[,!c("dist", "school", "absent_nonill", "matchid")],
                              glmfit = glm.log.ill,
                              K=2
                             )$delta

cv.log.nonill.predError = cv.glm(data=combinedFluDataInterventionTime[,!c("dist", "school", "absent_ill", "matchid")],
                              glmfit = glm.log.nonill,
                              K=2
                             )$delta

options(warn = oldw)

cv.log.ill.predError
cv.log.nonill.predError
```

```
0.0253637016069678   0.0253636420812935

0.0216913966369064   0.021691347238908
```

Logistic Regression Interpretation

Though our prediction accuracies are actually very good, its important to recognize how biased our data was to begin with. We started with a dataset composed of < 5% absences, so simply guessing "present" every time, a naive model could still get a 95%+ accuracy. This model, thus, is able to pick up on some of the variables which are important to the classification but it has a biased view of which variables are extremely important because of how skewed the data is to one class. That said, dist.n *is* thankfully one of the significant predictors, though that should be taken with a grain of salt due to the above.

To further explore whether Shoo-the-flu had an impact:

Multiple Linear Regression on All-Cause and Illness-Specific School-level Absenteeism

In [18]:
```
# Having fit a logistic regression model, a regularized multiple linear regression model may now help us discern
# effects of many of these variables on absenteeism percentage by school

# These GLM models took wayyyyy too much RAM (115gb+). My computer couldn't handle it

granularSchoolAbsenceAverages = absenteeData[sample(.N, smallSampleSize),.(absenceAverage=mean(absence)*100, yr=yr,
                                                illnessAbsenceAverage=mean(absent_ill)*100),
                                              by=c("matchid", "dist", "school", "schoolyr", "fluseasCDC")][order(matchid, dist)]
head(granularSchoolAbsenceAverages)
tail(granularSchoolAbsenceAverages)
```

matchid	dist	school	schoolyr	fluseasCDC	absenceAverage	yr	illnessAbsenceAverage
0	OUSD	Esperanza Elementary	2016-17	1	4.037267	2016	2.380952
0	OUSD	Esperanza Elementary	2016-17	1	4.037267	2017	2.380952
0	OUSD	Esperanza Elementary	2016-17	1	4.037267	2016	2.380952
0	OUSD	Esperanza Elementary	2016-17	1	4.037267	2016	2.380952
0	OUSD	Esperanza Elementary	2016-17	1	4.037267	2017	2.380952
0	OUSD	Esperanza Elementary	2016-17	1	4.037267	2016	2.380952

matchid	dist	school	schoolyr	fluseasCDC	absenceAverage	yr	illnessAbsenceAverage
34	WCCUSD	Bayview Elementary	2015-16	0	5.647383	2016	2.203857
34	WCCUSD	Bayview Elementary	2015-16	0	5.647383	2016	2.203857
34	WCCUSD	Bayview Elementary	2015-16	0	5.647383	2015	2.203857
34	WCCUSD	Bayview Elementary	2015-16	0	5.647383	2015	2.203857
34	WCCUSD	Bayview Elementary	2015-16	0	5.647383	2015	2.203857
34	WCCUSD	Bayview Elementary	2015-16	0	5.647383	2016	2.203857

In [19]:
```
# Merging school level data into our set of all-cause absenteeism
combinedGranularSchoolAbsenceAverages = merge(x=granularSchoolAbsenceAverages,
                                              y=schoolData,
                                              by=c("dist", "school")
                                             )

head(combinedGranularSchoolAbsenceAverages)
tail(combinedGranularSchoolAbsenceAverages)
colnames(combinedGranularSchoolAbsenceAverages)
```

dist	school	matchid	schoolyr	fluseasCDC	absenceAverage	yr	illnessAbsenceAverage	enrolled	mn.class.size	⋯	per.freelunch	API13	API12	mean.cst.ela	per.adv.ela
OUSD	Allendale Elementary	14	2013-14	1	3.811102	2013	2.485501	425	26.5625	⋯	79.91	663	725	329.625	8.75
OUSD	Allendale Elementary	14	2013-14	1	3.811102	2014	2.485501	425	26.5625	⋯	79.91	663	725	329.625	8.75
OUSD	Allendale Elementary	14	2013-14	1	3.811102	2014	2.485501	425	26.5625	⋯	79.91	663	725	329.625	8.75
OUSD	Allendale Elementary	14	2013-14	1	3.811102	2014	2.485501	425	26.5625	⋯	79.91	663	725	329.625	8.75
OUSD	Allendale Elementary	14	2013-14	1	3.811102	2013	2.485501	425	26.5625	⋯	79.91	663	725	329.625	8.75
OUSD	Allendale Elementary	14	2013-14	1	3.811102	2014	2.485501	425	26.5625	⋯	79.91	663	725	329.625	8.75

dist	school	matchid	schoolyr	fluseasCDC	absenceAverage	yr	illnessAbsenceAverage	enrolled	mn.class.size	⋯	per.freelunch	API13	API12	mean.cst.ela	per.adv.
WCCUSD	Wilson Elementary	13	2015-16	0	4.293381	2016	1.788909	538	26.9	⋯	71.56	745	778	344.98	1
WCCUSD	Wilson Elementary	13	2015-16	0	4.293381	2016	1.788909	538	26.9	⋯	71.56	745	778	344.98	1
WCCUSD	Wilson Elementary	13	2015-16	0	4.293381	2015	1.788909	538	26.9	⋯	71.56	745	778	344.98	1
WCCUSD	Wilson Elementary	13	2015-16	0	4.293381	2015	1.788909	538	26.9	⋯	71.56	745	778	344.98	1
WCCUSD	Wilson Elementary	13	2015-16	0	4.293381	2015	1.788909	538	26.9	⋯	71.56	745	778	344.98	1
WCCUSD	Wilson Elementary	13	2015-16	0	4.293381	2016	1.788909	538	26.9	⋯	71.56	745	778	344.98	1

```
'dist'  'school'  'matchid'  'schoolyr'  'fluseasCDC'  'absenceAverage'  'yr'  'illnessAbsenceAverage'  'enrolled'  'mn.class.size'  'per.not_hsg'  'per.hsg'  'per.some_col'
'per.col_grad'  'per.grad_sch'  'per.englearn'  'per.freelunch'  'API13'  'API12'  'mean.cst.ela'  'per.adv.ela'  'per.basic.ela'  'mean.cst.m'  'per.adv.m'  'per.basic.m'
'dist.n'
```

In [20]:
```
 Marking rows that schools were under intervention - the hope is of course that intervention contributes significantly to each type of 

ombinedGranularSchoolAbsenceAverages = combinedGranularSchoolAbsenceAverages[
   ,"intervention":= ifelse( (yr>2014|schoolyr=="2014-2015"), dist.n, 0)]

rint("Percentage of all rows under intervention: ")
ean(combinedGranularSchoolAbsenceAverages$intervention)
```

```
[1] "Percentage of all rows under intervention: "
```

0.189042141997257

In [21]:

```
:()
m.linReg.absenceAverage = glm(absenceAverage~., data=combinedGranularSchoolAbsenceAverages[,!c("dist", "school", "matchid", "illnessAbs
```

	used	(Mb)	gc trigger	(Mb)	max used	(Mb)
Ncells	12490667	667.1	20885653	1115.5	13458772	718.8
Vcells	4112849467	31378.6	9044353458	69003.0	9043430446	68995.9

In [22]:

```
gc()
glm.linReg.illnessAbsenceAverage = glm(illnessAbsenceAverage~.,
                                        data=combinedGranularSchoolAbsenceAverages[,!c("dist", "school", "matchid", "absenceAverage")])
```

	used	(Mb)	gc trigger	(Mb)	max used	(Mb)
Ncells	12491114	667.1	20885653	1115.5	13458772	718.8
Vcells	4175877774	31859.5	9044353458	69003.0	9043430446	68995.9

```
In [23]: gc()
         summary(glm.linReg.absenceAverage)
         summary(glm.linReg.illnessAbsenceAverage)
```

	used	(Mb)	gc trigger	(Mb)	max used	(Mb)
Ncells	12491397	667.2	20885653	1115.5	13458772	718.8
Vcells	4238945955	32340.6	9044353458	69003.0	9043430446	68995.9

```
Call:
glm(formula = absenceAverage ~ ., data = combinedGranularSchoolAbsenceAverages[, 
    !c("dist", "school", "matchid", "illnessAbsenceAverage")])

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.6494  -0.5973  -0.0252   0.5309   5.8252  

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)    
(Intercept)     -1.580e+01  4.565e+00   -3.461 0.000539 ***
schoolyr2012-13 -1.319e-01  4.412e-03  -29.910  < 2e-16 ***
schoolyr2013-14 -1.699e-01  5.916e-03  -28.722  < 2e-16 ***
schoolyr2014-15  2.820e-01  7.721e-03   36.521  < 2e-16 ***
schoolyr2015-16  2.817e-01  9.779e-03   28.808  < 2e-16 ***
schoolyr2016-17  4.839e-01  1.188e-02   40.742  < 2e-16 ***
fluseasCDC       3.771e-01  2.454e-03  153.659  < 2e-16 ***
yr               8.458e-03  2.267e-03    3.731 0.000191 ***
enrolled        -1.106e-03  1.246e-05  -88.786  < 2e-16 ***
mn.class.size   -1.562e-02  6.130e-04  -25.487  < 2e-16 ***
per.not_hsg     -3.176e-02  1.881e-03  -16.884  < 2e-16 ***
per.hsg         -4.229e-02  1.852e-03  -22.829  < 2e-16 ***
per.some_col    -6.396e-02  1.899e-03  -33.684  < 2e-16 ***
per.col_grad    -4.428e-02  1.860e-03  -23.806  < 2e-16 ***
per.grad_sch    -6.359e-02  1.808e-03  -35.179  < 2e-16 ***
per.englearn    -4.343e-02  1.895e-04 -229.186  < 2e-16 ***
per.freelunch    1.354e-02  1.837e-04   73.741  < 2e-16 ***
API13           -4.957e-03  5.892e-05  -84.135  < 2e-16 ***
API12           -1.057e-02  7.888e-05 -133.995  < 2e-16 ***
mean.cst.ela     6.139e-02  3.898e-04  157.478  < 2e-16 ***
per.adv.ela     -7.851e-02  6.142e-04 -127.825  < 2e-16 ***
per.basic.ela   -2.222e-02  4.052e-04  -54.848  < 2e-16 ***
mean.cst.m       5.686e-03  2.847e-04   19.975  < 2e-16 ***
per.adv.m       -1.171e-02  6.470e-04  -18.092  < 2e-16 ***
per.basic.m      1.550e-02  4.263e-04   36.363  < 2e-16 ***
dist.n           3.666e-01  5.622e-03   65.210  < 2e-16 ***
intervention    -5.416e-02  4.286e-03  -12.636  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.97373)

    Null deviance: 1598085  on 788310  degrees of freedom
Residual deviance:  767576  on 788284  degrees of freedom
AIC: 2216173

Number of Fisher Scoring iterations: 2


Call:
glm(formula = illnessAbsenceAverage ~ ., data = combinedGranularSchoolAbsenceAverages[, 
    !c("dist", "school", "matchid", "absenceAverage")])

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-3.1651  -0.5179  -0.0602   0.4354   4.6297  

Coefficients:
                  Estimate Std. Error  t value Pr(>|t|)    
(Intercept)     -2.435e+01  3.747e+00   -6.499 8.10e-11 ***
schoolyr2012-13 -2.446e-02  3.621e-03   -6.756 1.42e-11 ***
schoolyr2013-14 -1.660e-01  4.856e-03  -34.188  < 2e-16 ***
schoolyr2014-15  3.210e-01  6.337e-03   50.654  < 2e-16 ***
schoolyr2015-16  2.888e-01  8.026e-03   35.980  < 2e-16 ***
schoolyr2016-17  3.068e-01  9.749e-03   31.473  < 2e-16 ***
fluseasCDC       6.553e-01  2.014e-03  325.366  < 2e-16 ***
yr               1.419e-02  1.861e-03    7.624 2.45e-14 ***
enrolled        -1.711e-03  1.023e-05 -167.307  < 2e-16 ***
mn.class.size   -2.224e-02  5.031e-04  -44.202  < 2e-16 ***
per.not_hsg     -7.459e-02  1.544e-03  -48.314  < 2e-16 ***
per.hsg         -6.653e-02  1.520e-03  -43.762  < 2e-16 ***
per.some_col    -5.207e-02  1.558e-03  -33.411  < 2e-16 ***
per.col_grad    -4.499e-02  1.526e-03  -29.476  < 2e-16 ***
per.grad_sch    -4.360e-02  1.484e-03  -29.388  < 2e-16 ***
per.englearn     3.975e-03  1.555e-04   25.560  < 2e-16 ***
per.freelunch    1.957e-02  1.507e-04  129.856  < 2e-16 ***
API13           -6.216e-04  4.836e-05  -12.855  < 2e-16 ***
API12           -2.116e-03  6.474e-05  -32.681  < 2e-16 ***
mean.cst.ela     4.113e-02  3.199e-04  128.566  < 2e-16 ***
per.adv.ela     -6.779e-02  5.041e-04 -134.487  < 2e-16 ***
per.basic.ela   -4.551e-02  3.325e-04 -136.866  < 2e-16 ***
mean.cst.m      -1.738e-02  2.336e-04  -74.413  < 2e-16 ***
per.adv.m        1.533e-02  5.310e-04   28.873  < 2e-16 ***
per.basic.m      9.110e-03  3.499e-04   26.037  < 2e-16 ***
dist.n          -9.993e-02  4.614e-03  -21.660  < 2e-16 ***
```

```
intervention    -2.022e-01  3.518e-03  -57.474  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.6558767)

    Null deviance: 766616  on 788310  degrees of freedom
Residual deviance: 517017  on 788284  degrees of freedom
AIC: 1904663

Number of Fisher Scoring iterations: 2
```

In [24]:

```
gc()
print("Cross Validation Linear Regression Prediction Error for all cause absenteeism:")
cv.linReg.absenceAverage.predError = cv.glm(data=combinedGranularSchoolAbsenceAverages[,!c("dist", "school", "matchid", "illnessAbsence
                               glmfit = glm.linReg.absenceAverage,
                               K=2
                              )$delta

cv.linReg.absenceAverage.predError[1]

print("Compare to the mean proportionof all-cause absenteeism across schools:")
mean(combinedGranularSchoolAbsenceAverages$absenceAverage)
```

	used	(Mb)	gc trigger	(Mb)	max used	(Mb)
Ncells	12491505	667.2	20885653	1115.5	13458772	718.8
Vcells	4239734619	32346.7	9044353458	69003.0	9043430446	68995.9

```
[1] "Cross Validation Linear Regression Prediction Error for all cause absenteeism:"
```

0.973837511960211

```
[1] "Compare to the mean proportionof all-cause absenteeism across schools:"
```

4.55175685738243

In [25]:

```
gc()
print("Cross Validation Linear Regression Prediction Error for illness-specific absenteeism:")
cv.linReg.absenceAverage.predError = cv.glm(data=combinedGranularSchoolAbsenceAverages[,!c("dist", "school", "matchid", "absenceAverage
                               glmfit = glm.linReg.illnessAbsenceAverage,
                               K=2
                              )$delta

cv.linReg.absenceAverage.predError[1]

print("Compare to the mean proportion of illness-specific absenteeism across schools:")
mean(combinedGranularSchoolAbsenceAverages$illnessAbsenceAverage)
```

	used	(Mb)	gc trigger	(Mb)	max used	(Mb)
Ncells	12491497	667.2	20885653	1115.5	13458772	718.8
Vcells	4238946782	32340.6	9044353458	69003.0	9043430446	68995.9

```
[1] "Cross Validation Linear Regression Prediction Error for illness-specific absenteeism:"
```

0.655892760769659

```
[1] "Compare to the mean proportion of illness-specific absenteeism across schools:"
```

2.28894433795799

Interpreting our linear regression

So, in this case, based on our cross validation predictions, our linear regression model isn't awful, but it isn't great either at using these school level variables to detect either type of absenteeism, with significant residuals. Unfortunately, we are no closer to discovering how important our intervention variable really is, and can only note that it also was a significant contributor to the regression combination, but since every other variable was as well... that doesn't say much. Our regression does, however, allow us to predict (albeit with a very large margin of error) average absenteeism over any given time period at the school level. This, of course, has the potential to highlight schools in areas that require