Leverage and big data

Aimee Barciauskas

November 21, 2015

1. SimDat1

Simulate a dataset from a linear regression model as follows. Generate 100K replications of 30 inputs using a 30-dim Gaussian with mean 0 and covariance the identity matrix. Generate the output as a linear regression on the inputs (without intercept) with coefficients randomly drawn from a multivariate Gaussian with mean 0 and variance 3. This will be SimDat1.

```r
# install.packages('mvtnorm')
# install.packages('scales')
library(mvtnorm)
library(scales)

# Generate 100k replications of 30 inputs
m <- 30
# n <- 100000
n <- 10000
sigma <- diag(m)
simdat1.xs <- rmvnorm(n, mean = rep(0, m), sigma = sigma)
# How can this be a multivariate gaussian?
simdat1.betas <- rnorm(m, mean = 0, sd = 1.5)

# SimDat1
simdat1.ys <- simdat1.xs %*% simdat1.betas
# plot(density(simdat1.ys))
# summary(simdat1.ys)
# plot(simdat1.ys, col = 'blue', ylim = c(-50,50))
```

2. Revise and briefly describe what is a mixture of Gaussian

distributions (see e.g. Bishop) and describe a procedure for simulating from one.

A mixture of Gaussian distributions is a linear combination of multiple Gaussian Normal distributions, each having a mixing coefficient where the sum of mixture coefficients is equal to 1. Intuitively, the mixture coefficient k is the probability of an observation having been drawn from distribution k.

To simulate from a mixture of k Gaussians, one could randomly sample from a multinomial distribution where the kth Gaussian has the kth mixing coefficient (probability). This observation would be drawn from distribution k.

3. SimDat2

Simulate another linear regression dataset as follows. Generate 100K replications of 30 inputs from a mixture of 2 30-dim Gaussians; the first with mean 0 and covariance the identity matrix; the second with mean 0 and covariance 10 times the identity. The first component has weight 0.95 and the second 0.05. Generate the output as a linear regression on the inputs (without intercept) with coefficients randomly drawn from a multivariate Gaussian with mean 0 and variance 3. This will be SimDat2.

Question: Why do we generate the output?

```
# Assign observations to mixtures
mixture_assignment <- rmultinom(n = n, size = 1, prob = c(0.95,0.05))
# Check we didn't do anything silly
# sum(mixture_assignment[1,])/n
# sum(mixture_assignment[2,])/n
assignments <- apply(mixture_assignment, 2, function(col) { match(1,col)})

sigma.gaussian1 <- diag(m)
sigma.gaussian2 <- diag(m)*10

rdraw.gaussian1 <- function() {
  rmvnorm(1, mean = rep(0, m), sigma = sigma.gaussian1)
}

rdraw.gaussian2 <- function() {
  rmvnorm(1, mean = rep(0, m), sigma = sigma.gaussian2)
}

simdat2.xs <- matrix(0, nrow = n, ncol = m)
# draw from 2 different gaussians
# waiting ...
for (idx in 1:length(assignments)) {
  assignment <- assignments[idx]
  if (assignment == 1) {
    simdat2.xs[idx,] <- rdraw.gaussian1()[1,]
  } else {
    simdat2.xs[idx,] <- rdraw.gaussian2()[1,]
  }
}

# SimDat2
simdat2.betas <- rnorm(m, mean = 0, sd = 1.5)
simdat2.ys <- simdat2.xs %*% simdat2.betas
# plot(density(simdat2.ys))
# summary(simdat2.ys)
```

4. Describe briefly what is leverage for linear regression, how it is computed and propose a linear transformation of the leverage that makes it numerically comparable across datasets with different number of inputs and different number of replications.

Leverage of the ith observation is the value $H_{i,i}$ of the hat matrix H:

$H = \phi(\phi^T \phi)^{-1} \phi^T$

So all the values of leverage would be the diagonal entries of the hatmatrix.

The average leverage will be $(m + 1)/n$, where m is the dimension of the inputs and n is the number of observations.

To compare leverages across different data sets, we will find leverages which are greater than 99% of the distribution of leverages (which we expect to follow a chi-squared distribution). So the leverages which are outside of C_α the 99% confidence region of the chi-squared distribution with m+1 degrees of freedom.

Classify as high leverage those who fail the following test:

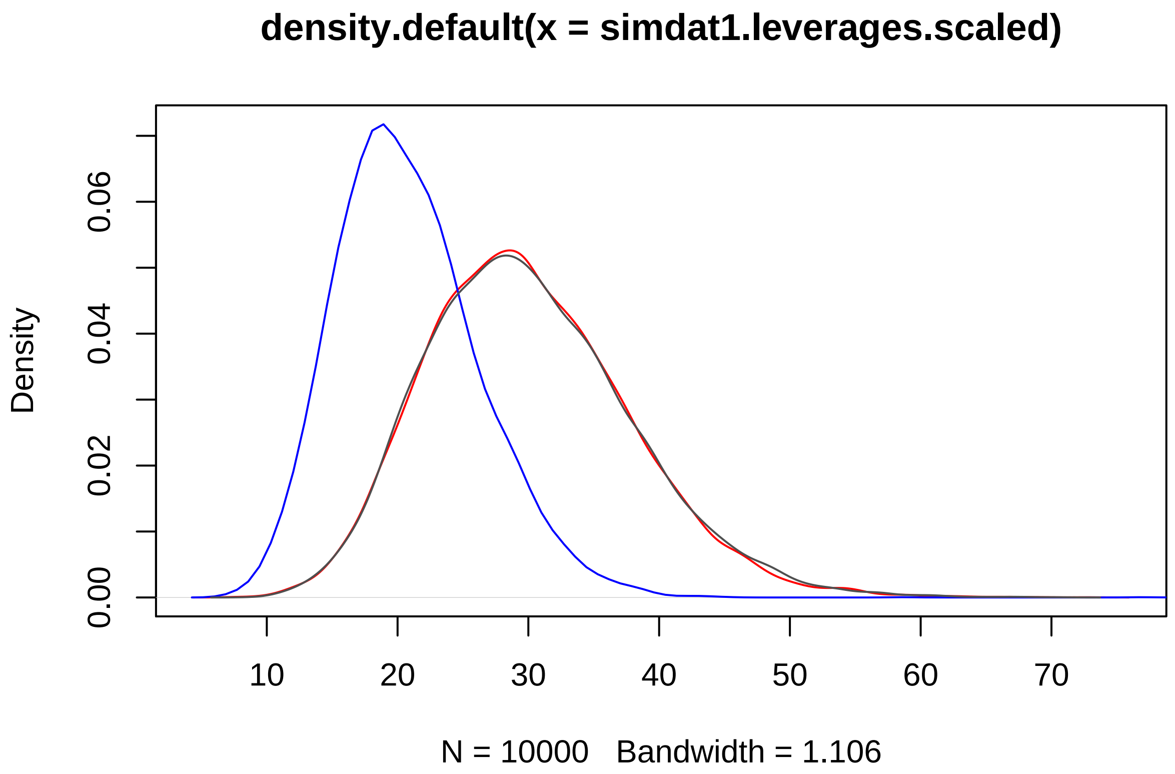
$C_\alpha = \left\{H_{i,i} \leq \chi^2_{1-\alpha}(d)\right\}$

5. Find from ML (or otherwise) data repositories (e.g. UCI) a number of big datasets (maximum 10) with continuous output and continuous inputs (maximum work with 50 inputs). For each of these datasets compute the leverages.

Question Why does it matter if the outputs are continuous?

6. The main output of this project is a visualization of the leverage distributions, one for each of the ML datasets as well as SimDat1 and SimDat2, that allows direct comparison between them. The aim is to to uncover structure that might be common in big data sets amenable to regression analysis.

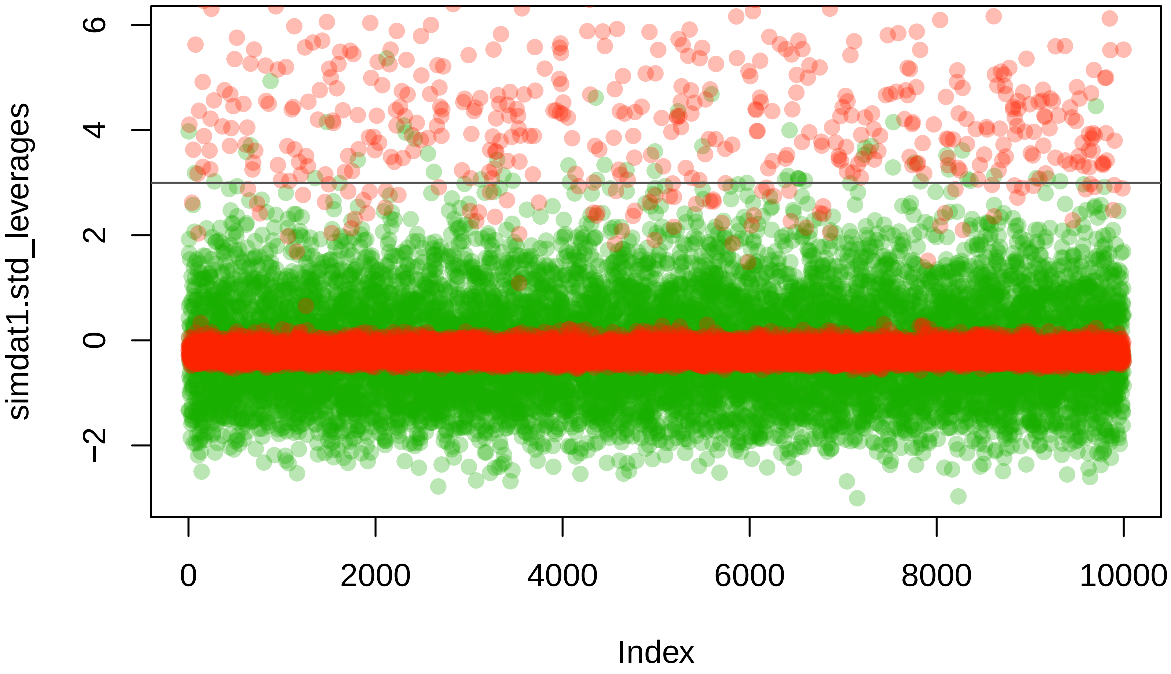
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9.10   24.40   29.39   30.09   34.99   70.36
```



Using z-scores…

```
### OLD
stdzd_leverages <- function(leverages, m, n) (leverages - (m+1)/n) / sd(leverages)

simdat1.leverages <- leverages(simdat1.xs, scaled = FALSE)
simdat2.leverages <- leverages(simdat2.xs, scaled = FALSE)
simdat1.std_leverages <- stdzd_leverages(simdat1.leverages, m, n)
simdat2.std_leverages <- stdzd_leverages(simdat2.leverages, m, n)
```

```r
# These summaries don't look right
library(scales)
heat.cols <- heat.colors(m, alpha = 0.3)
ter.cols <- terrain.colors(m, alpha = 0.3)
plot(simdat1.std_leverages,
     col = ter.cols[3],
     pch = 19,
     ylim = c(-3,6))
points(x = simdat2.std_leverages,
     col = heat.cols[4],
     pch = 19)
abline(a = 3, b = 0, col = 'grey31')
```



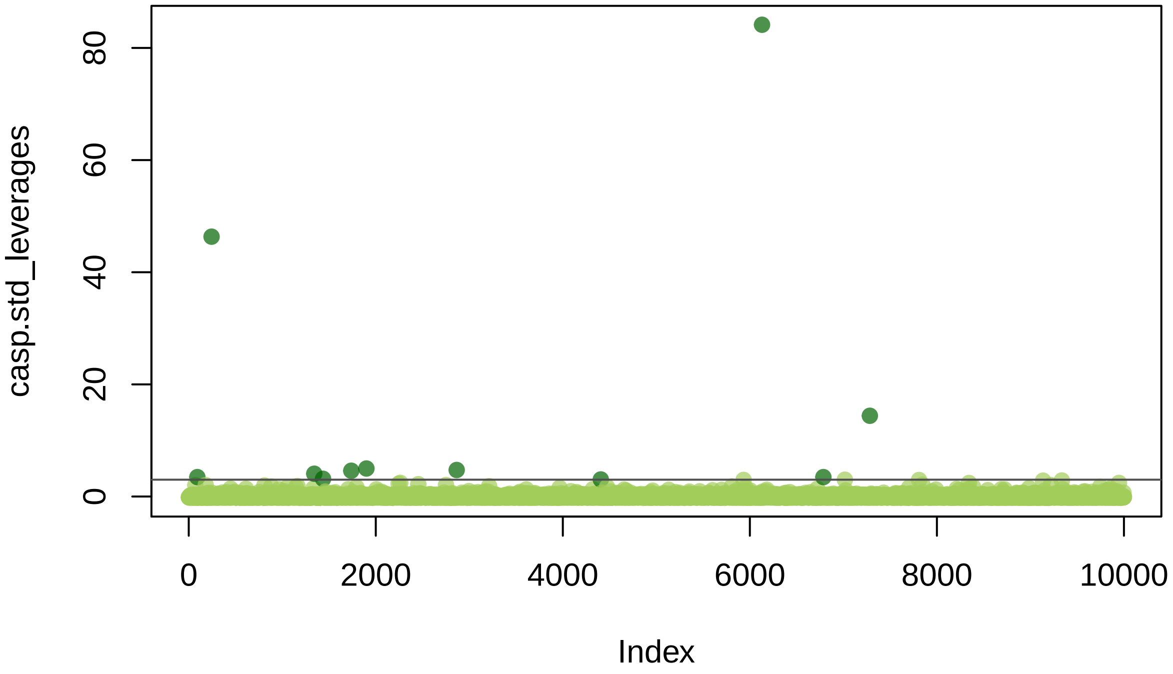
5. Continued

CASP Data: Physicochemical Properties of Protein Tertiary Structure

Dataset Homepage

```r
casp.data <- read.csv('~/Projects/data_files/CASP.csv')
casp.features <- casp.data[1:10000,2:ncol(casp.data)]

casp.leverages <- leverages(as.matrix(casp.features))
casp.std_leverages <- stdzd_leverages(casp.leverages, ncol(casp.features), nrow(casp.features))
# summary(casp.std_leverages)
```

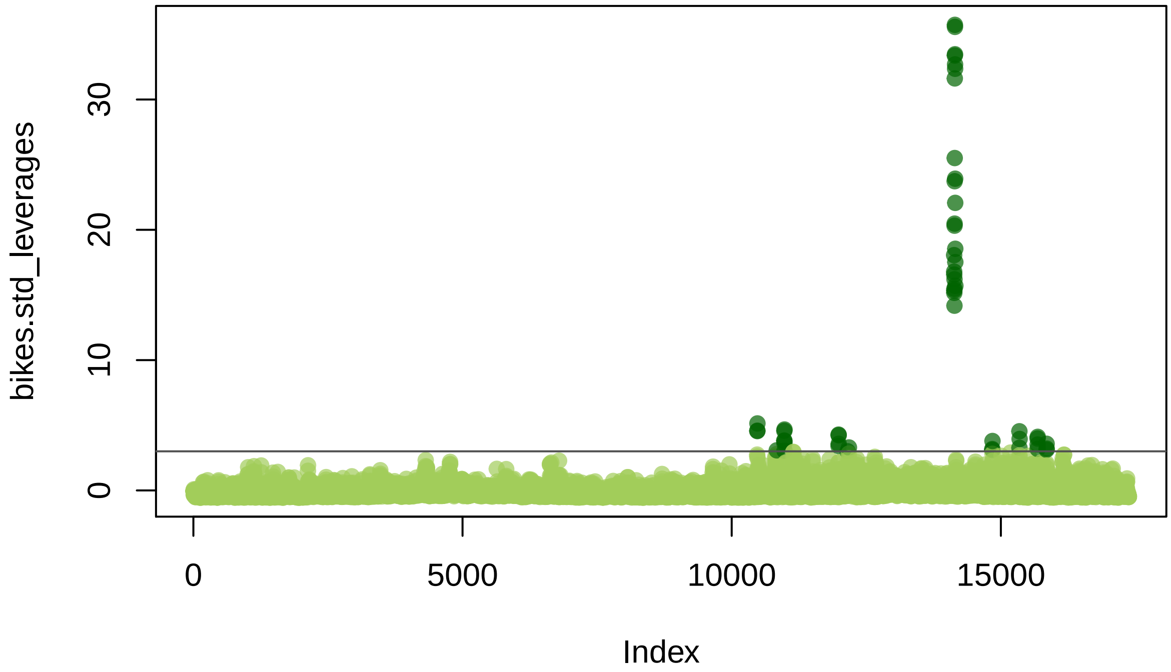
```r
plot(x = casp.std_leverages,
     col = alpha(ifelse(abs(casp.std_leverages) < 3,'darkolivegreen3','darkgreen'), 0.7),
     pch = 19)
abline(a = 3, b = 0, col = 'grey31')
```



Bike Sharing Dataset

Dataset Homepage

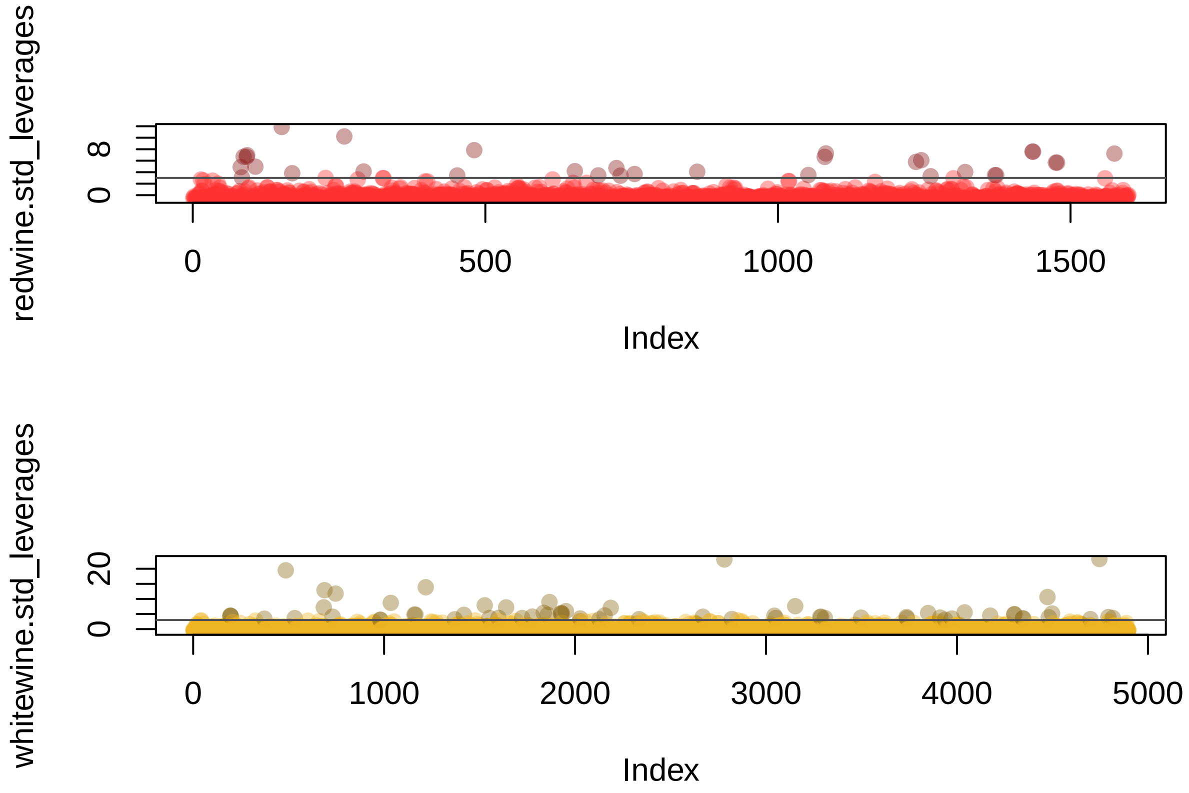
```r
bikes.data <- read.csv('~/Projects/data_files/Bike-Sharing-Dataset/hour.csv')
bikes.features <- bikes.data[,c('temp','atemp','hum','windspeed','casual', 'registered')]

bikes.leverages <- leverages(as.matrix(bikes.features))
bikes.std_leverages <- stdzd_leverages(bikes.leverages, ncol(bikes.features), nrow(bikes.features))

plot(x = bikes.std_leverages,
     col = alpha(ifelse(abs(bikes.std_leverages) < 3,'darkolivegreen3','darkgreen'), 0.7),
     pch = 19)
abline(a = 3, b = 0, col = 'grey31')
```



Wine Quality Dataset

Dataset Homepage

```
redwine.data <- read.delim('~/Projects/data_files/wine/winequality-red.csv', sep=';')
redwine.features <- redwine.data[,setdiff(colnames(redwine.data), 'quality')]

redwine.leverages <- leverages(as.matrix(redwine.features))
redwine.std_leverages <- stdzd_leverages(redwine.leverages, ncol(redwine.features), nrow(redwine.featur

whitewine.data <- read.delim('~/Projects/data_files/wine/winequality-white.csv', sep=';')
whitewine.features <- whitewine.data[,setdiff(colnames(whitewine.data), 'quality')]

whitewine.leverages <- leverages(as.matrix(whitewine.features))
whitewine.std_leverages <- stdzd_leverages(whitewine.leverages, ncol(whitewine.features), nrow(whitewin

par(mfrow = c(2,1))

plot(x = redwine.std_leverages,
     col = alpha(ifelse(abs(redwine.std_leverages) < 3,'firebrick1','firebrick4'), 0.4),
     pch = 19)
abline(a = 3, b = 0, col = 'grey31')

plot(x = whitewine.std_leverages,
     col = alpha(ifelse(abs(whitewine.std_leverages) < 3,'goldenrod2','goldenrod4'), 0.4),
     pch = 19)
abline(a = 3, b = 0, col = 'grey31')
```

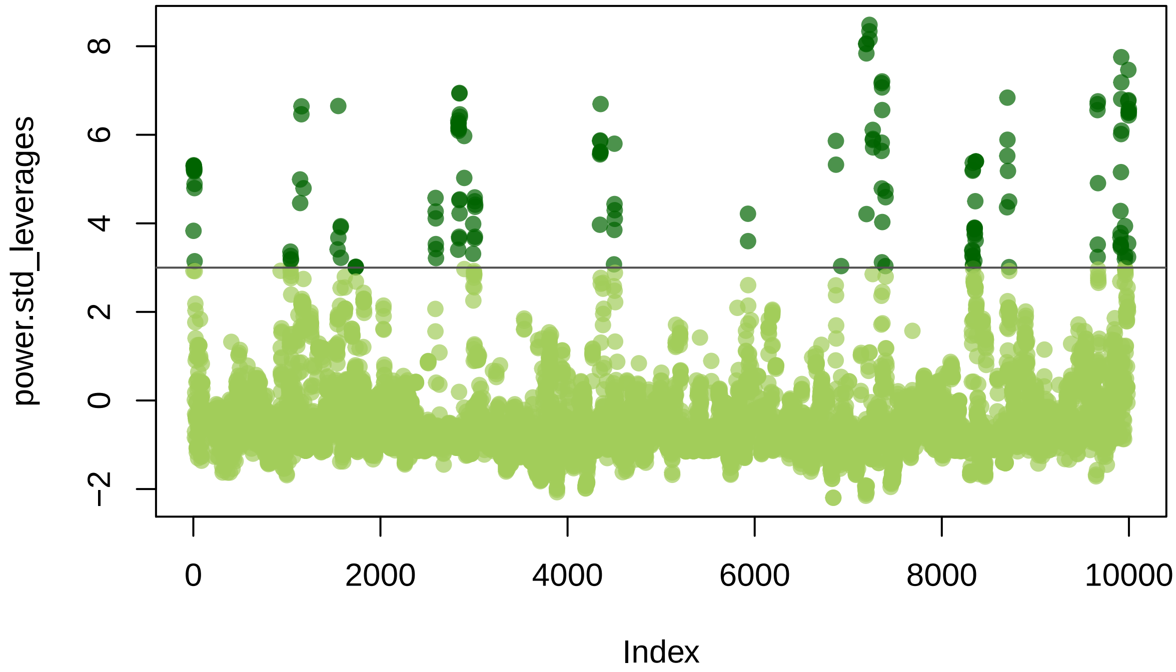


Individual household electric power consumption Data Set

Dataset Homepage

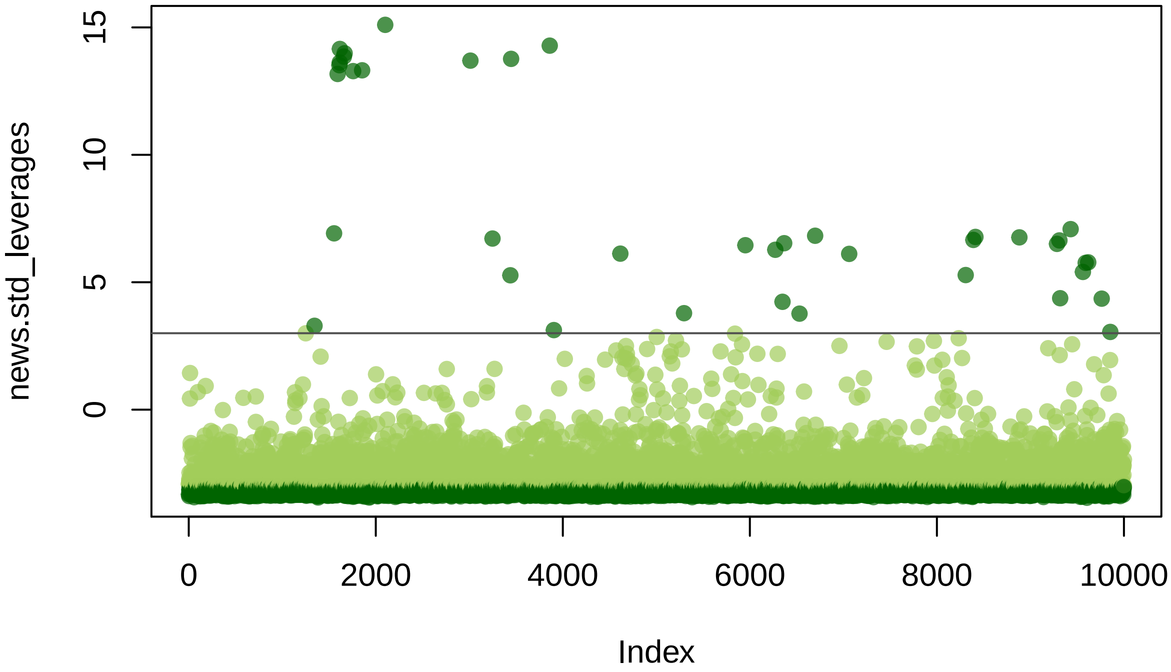
Note: Sequential nature of data clumps high leverage points

```
power.data <- read.delim('~/Projects/data_files/household_power_consumption.txt', nrow=10000, sep = ';'
power.features <- power.data[,c('Global_active_power','Global_reactive_power','Voltage','Global_intensi
power.features <- data.matrix(power.features)

power.leverages <- leverages(power.features)
power.std_leverages <- stdzd_leverages(power.leverages, ncol(power.features), nrow(power.features))

par(mfrow = c(1,1))
plot(x = power.std_leverages,
     col = alpha(ifelse(abs(power.std_leverages) < 3,'darkolivegreen3','darkgreen'), 0.7),
     pch = 19)
abline(a = 3, b = 0, col = 'grey31')
```



Online News Popularity

```
news.data <- read.csv('~/Projects/data_files/OnlineNewsPopularity/OnlineNewsPopularity.csv', nrow=10000
news.features <- news.data[,c('n_tokens_title', 'n_tokens_content','num_hrefs', 'num_imgs', 'num_videos

news.leverages <- leverages(as.matrix(news.features))
news.std_leverages <- stdzd_leverages(news.leverages, ncol(news.data), nrow(news.data))


plot(x = news.std_leverages,
     col = alpha(ifelse(abs(news.std_leverages) < 3,'darkolivegreen3','darkgreen'), 0.7),
     pch = 19)
abline(a = 3, b = 0, col = 'grey31')
```



Final Thoughts

In data which is sequential, often high leverage points are near eachother in the order of rows.