Midterm Project

Achyut and Nathaniel_Bowerman

4/5/2018

```
options(warn = -1)
```

data prep

```
# libraries
library(fBasics)
```

```
## Loading required package: timeDate
## Loading required package: timeSeries
##
## Rmetrics Package fBasics
## Analysing Markets and calculating Basic Statistics
## Copyright (C) 2005-2014 Rmetrics Association Zurich
## Educational Software for Financial Engineering and Computational Science
## Rmetrics is free software and comes with ABSOLUTELY NO WARRANTY.
## https://www.rmetrics.org --- Mail to: info@rmetrics.org
```

```
library(Hotelling)
```

```
## Loading required package: corpcor
```

```
library(ICSNP)
```

```
## Loading required package: mvtnorm
## Loading required package: ICS
```

```
# loading the data
car_data = read.table('CarBodyAssembly.dat')

car_dt1 <- car_data[1:30, ]
car_dt2 <- car_data[31:50 , ]
```

a)

```
#Multivariate T-Test against mean of zero vector
HotellingsT2(car_data)
```

```
##
##  Hotelling's one sample T2-test
##
```

```
## data:  car_data
## T.2 = 74.917, df1 = 6, df2 = 44, p-value < 2.2e-16
## alternative hypothesis: true location is not equal to c(0,0,0,0,0,0)
```

```
################
#p-value=2.2e-16, thus we have evidence that the mean vector is different than [0,0,0,0,0,0]
################
```

b)

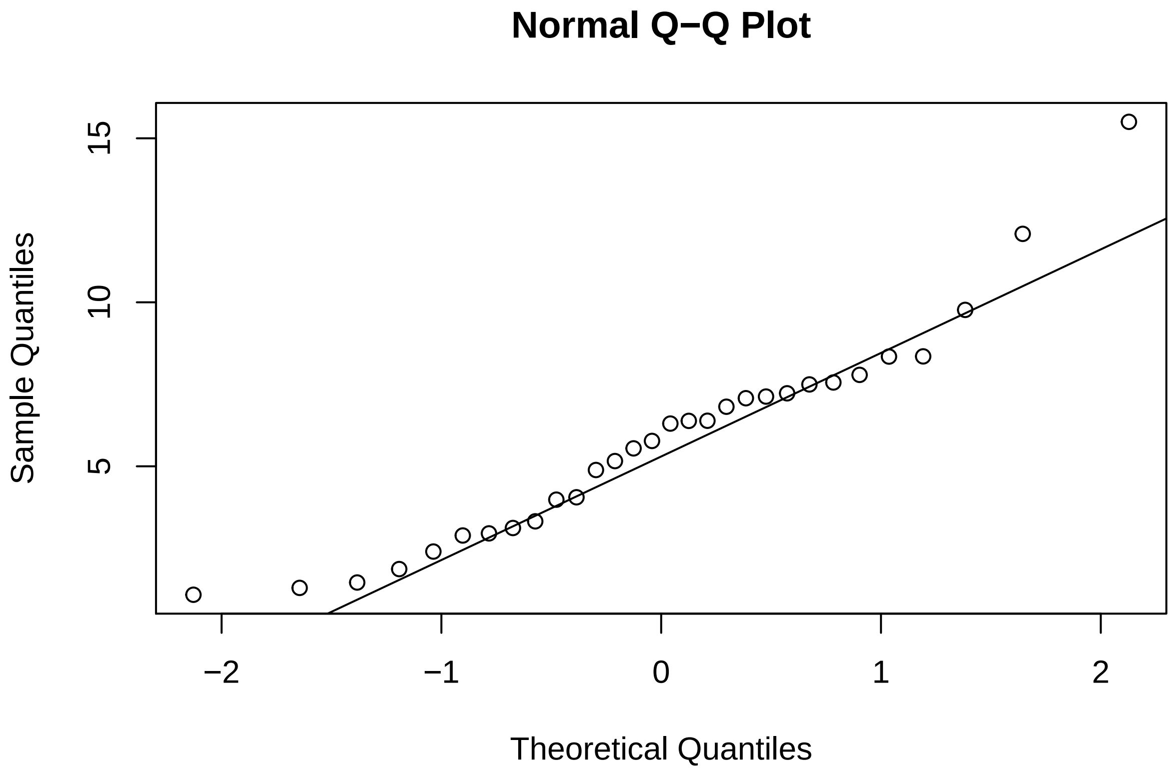
```
#test of difference in means between first 30 and last 20
HotellingsT2(car_dt1,car_dt2)
```

```
##
##  Hotelling's two sample T2-test
##
## data:  car_dt1 and car_dt2
## T.2 = 1.8563, df1 = 6, df2 = 43, p-value = 0.1107
## alternative hypothesis: true location difference is not equal to c(0,0,0,0,0,0)
```

```
##############
#p-value is 0.1107, thus we have no evidence that there is a difference
#between the mean vectors of the first 30 and last 20 observations
##############
```

c)

```
#Examine QQ plots of first 30 observations to determine MVN
mah <- mahalanobis(car_dt1,colMeans(car_dt1),var(car_dt1))

qqnorm(mah)
qqline(mah)
```



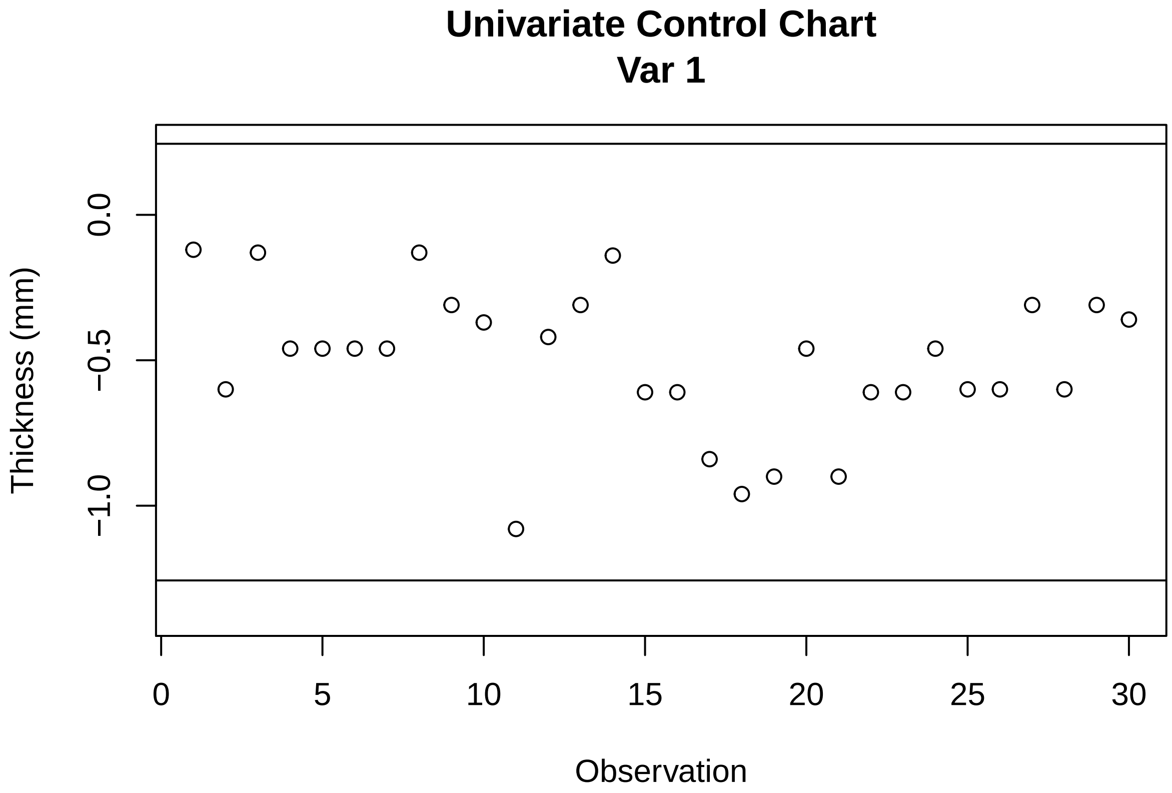
```
#Shapiro test for MVN
shapiro.test(qnorm(pchisq(mah,6)))
```

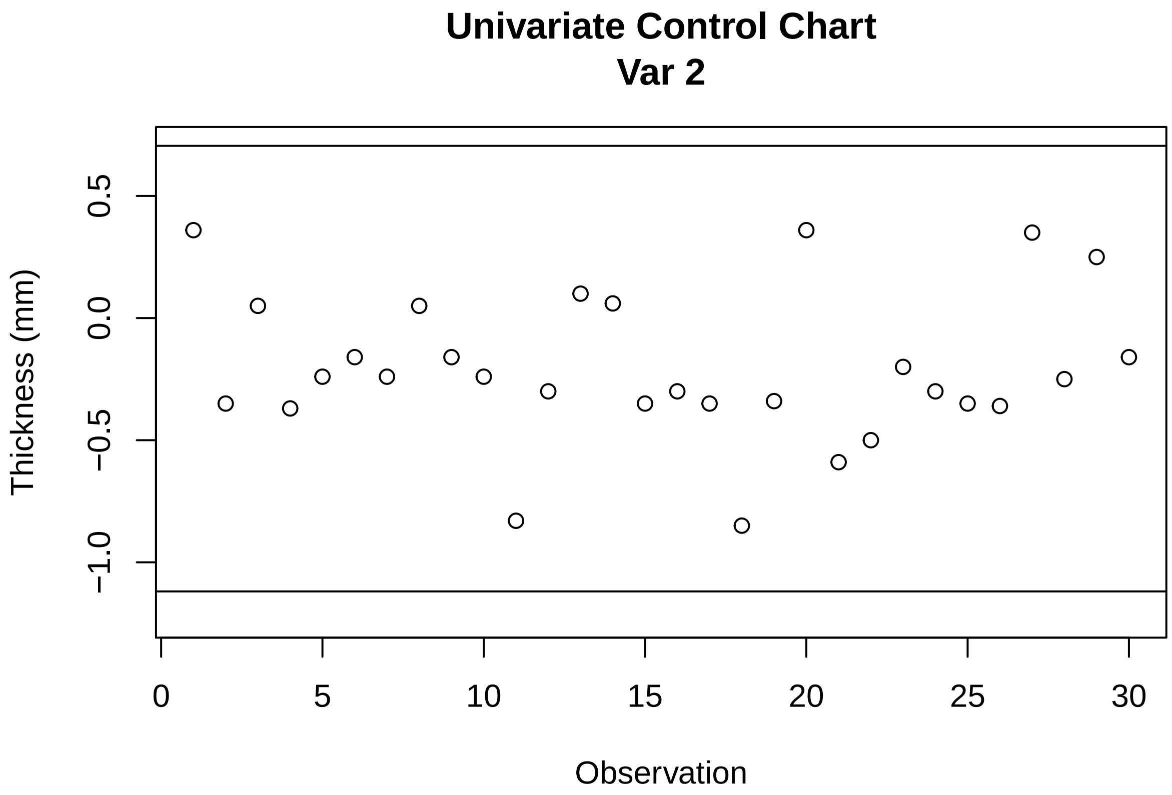
```
##
##  Shapiro-Wilk normality test
##
## data:  qnorm(pchisq(mah, 6))
## W = 0.96728, p-value = 0.4678
```

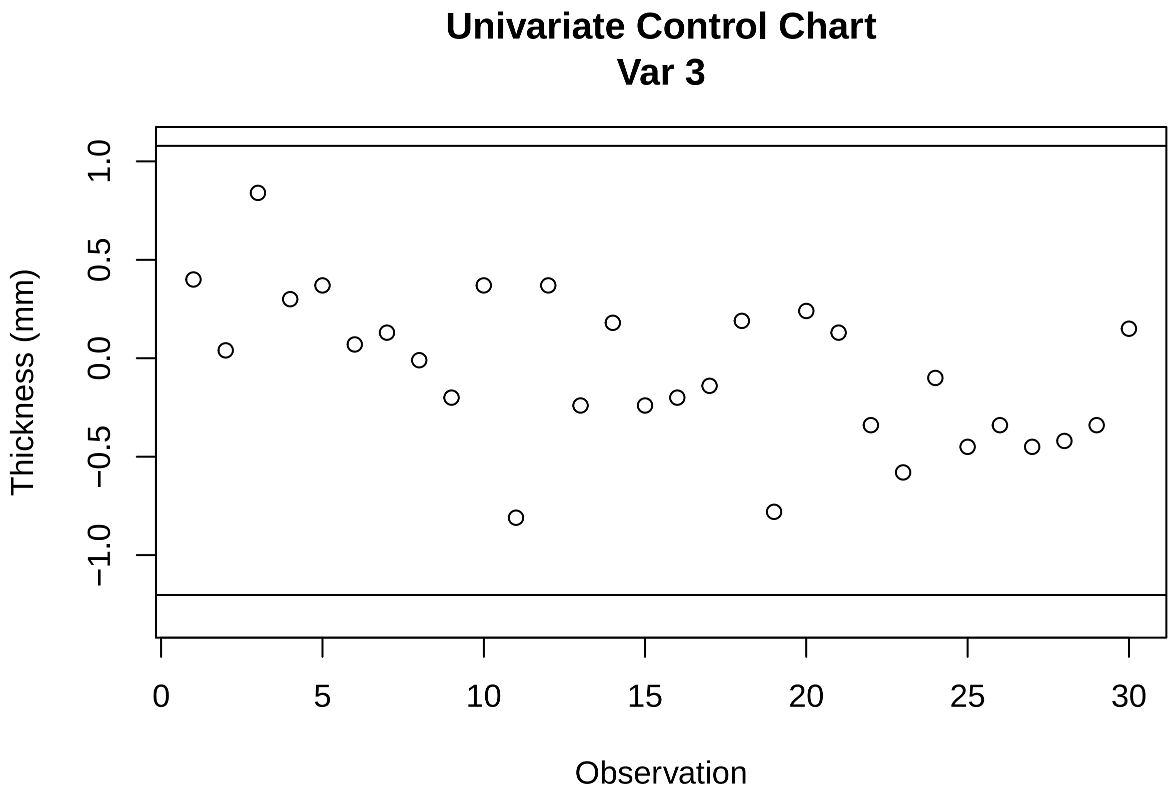
```
##################
#p-value = 0.4678 and the QQ plot only has a couple points on the ends that are far from the expected
#line is the data was multivariate normal.  Based on this, it appears that the data are normal for
#the first 30 observations.
#################
```

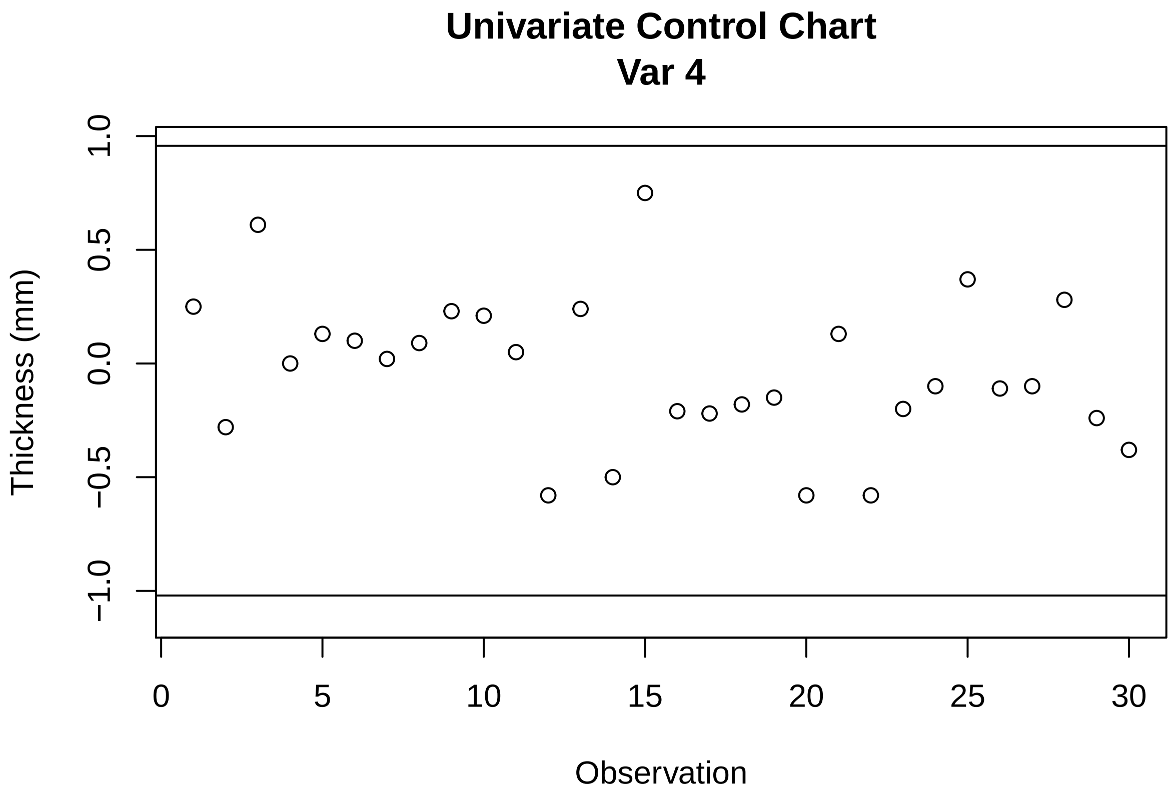
d)

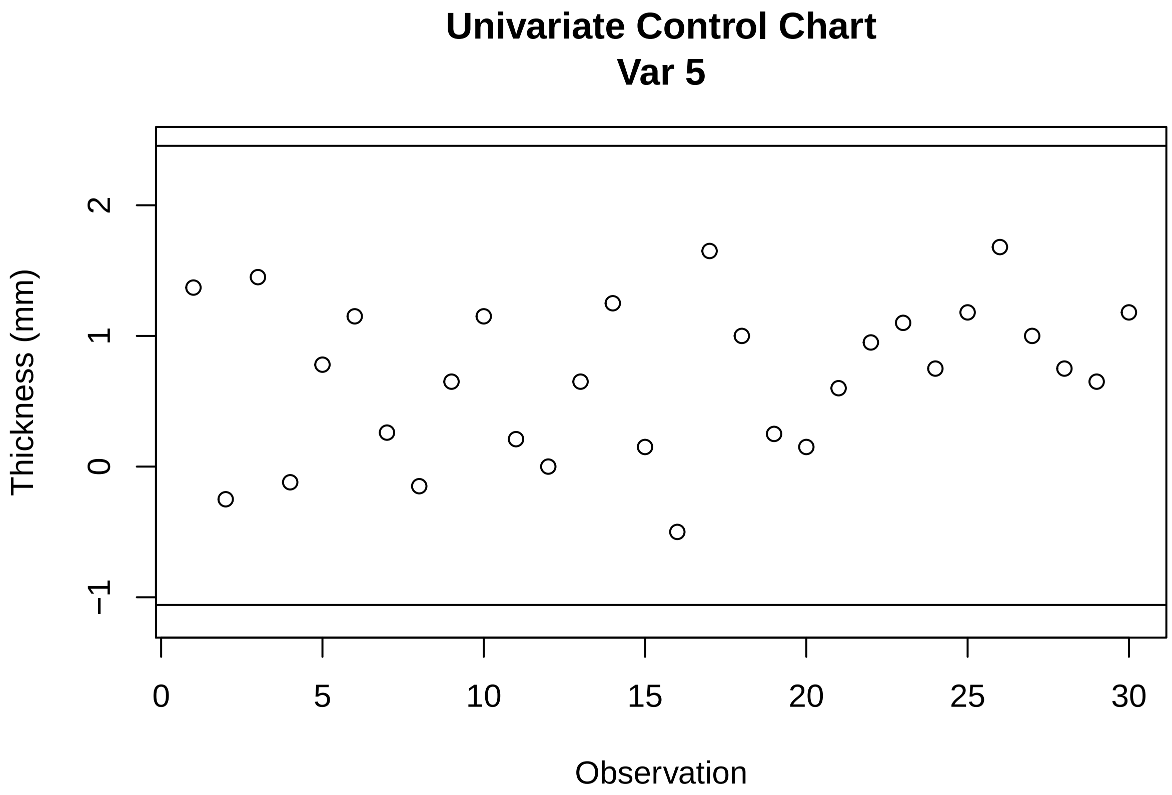
```
#calculate means and var for first 30 observations
my_means <- colMeans(car_dt1)
my_var <- diag(var(car_dt1))


#plot the control charts for all 6 variables for the first 30 observations
for (i in 1:6){
  plot(1:30,car_dt1[,i],ylim=c(1.1*min(my_means[i]-3*sqrt(my_var[i]),car_dt1[,i]),max
                               (my_means[i]+3*sqrt(my_var[i]),car_dt1[,i])),xlab="Observation",
       ylab=paste("Thickness (mm)"))
  title(main=paste("Univariate Control Chart\nVar",i))
  abline(h=c(my_means[i]-3*sqrt(my_var[i]),my_means[i]+3*sqrt(my_var[i])))
}
```

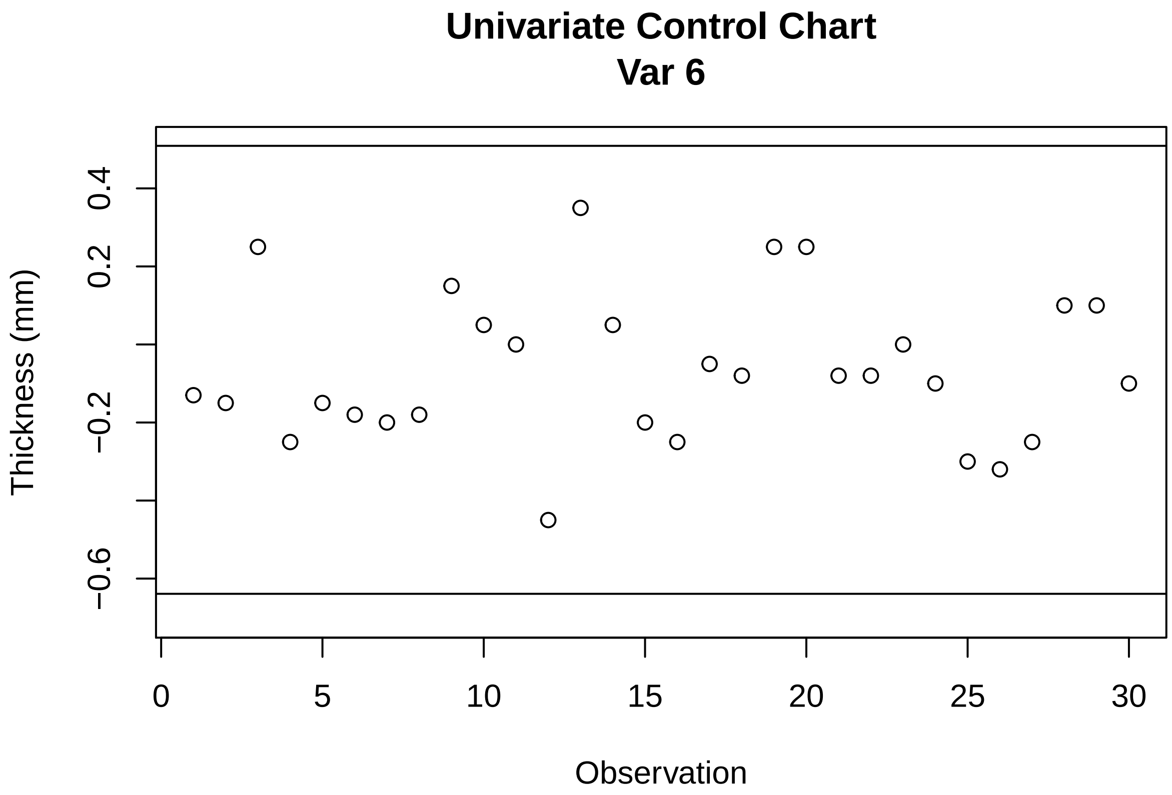



Univariate Control Chart
Var 2
Thickness (mm)
0.5
0.0
−0.5
−1.0
0
5
10
15
20
25
30
Observation


Univariate Control Chart
Var 3
Thickness (mm)
1.0
0.5
0.0
−0.5
−1.0
0
5
10
15
20
25
30
Observation


Univariate Control Chart
Var 4
Thickness (mm)
1.0
0.5
0.0
−0.5
−1.0
0
5
10
15
20
25
30
Observation

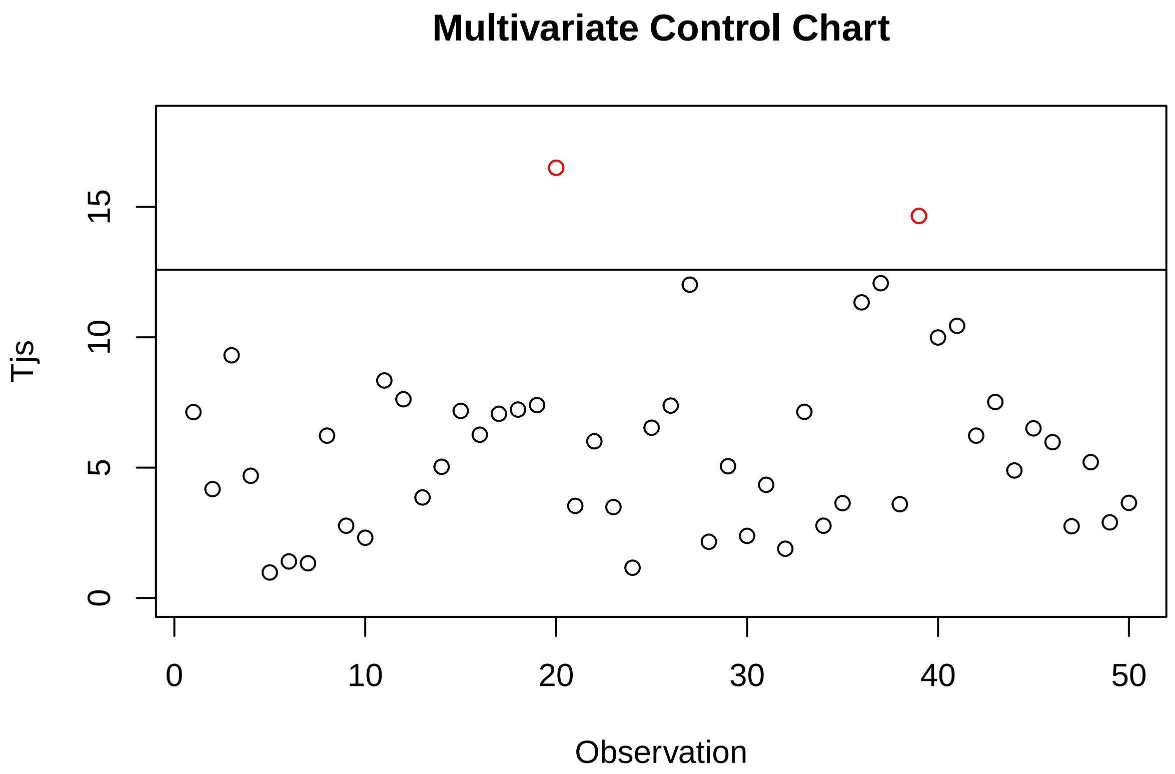

Univariate Control Chart
Var 5
Thickness (mm)
2
1
0
−1
0
5
10
15
20
25
30
Observation



e)

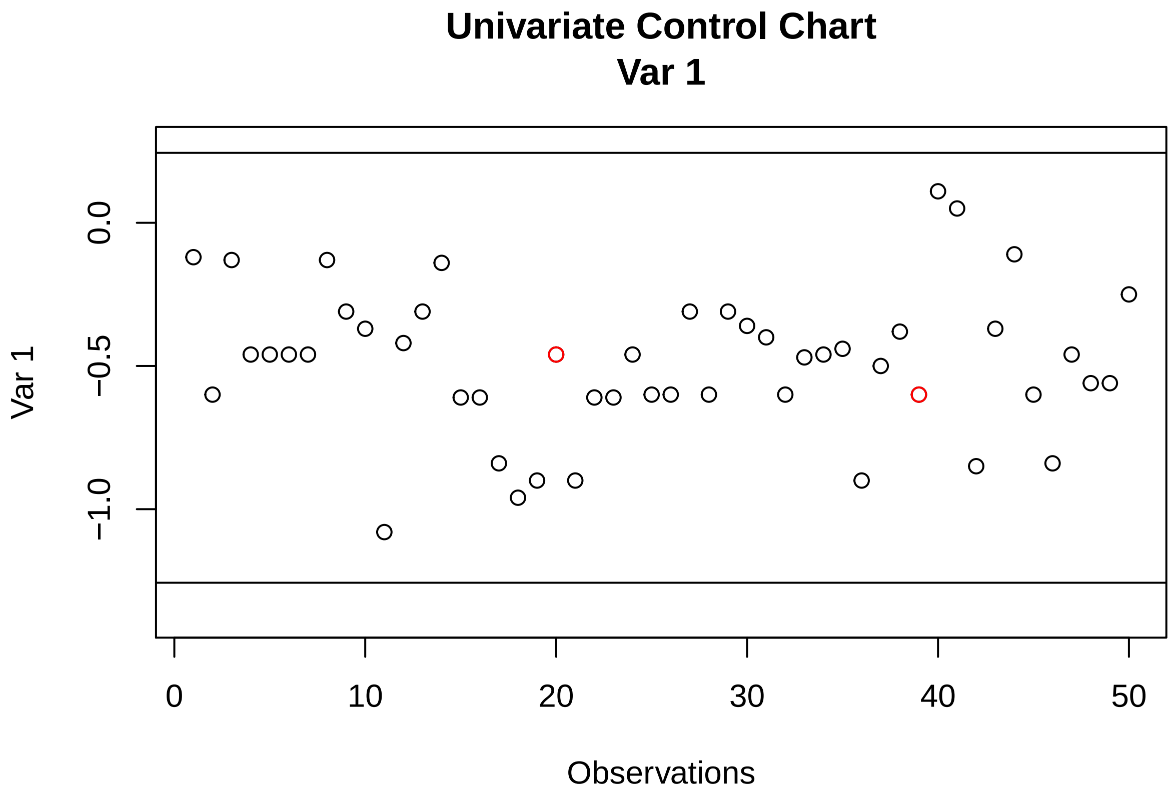
```
#assign dataset and create components for the calculations
data <- unname(unlist(data.matrix(car_data)))
S <- var(data)
x_bar <- colMeans(data)
S_inv<-solve(S)

#calculate Tj’s for all observations
Tjs<-c()
for (i in 1:50) {
  diff<-x_bar-data[i,]
  Tj<-t(diff)%*%S_inv%*%diff
  Tjs<-c(Tjs,Tj)
}

#determine upper control limit
UCL<-qchisq(0.95,6)

#plot multivariate control chart and highlight points over UCL as red
plot(1:50,Tjs,ylim=c(0,1.1*max(Tjs,UCL)),xlab=’Observation’)
title(main="Multivariate Control Chart")
over<-which(Tjs>UCL)
points(over, Tjs[over], col = "red")
```

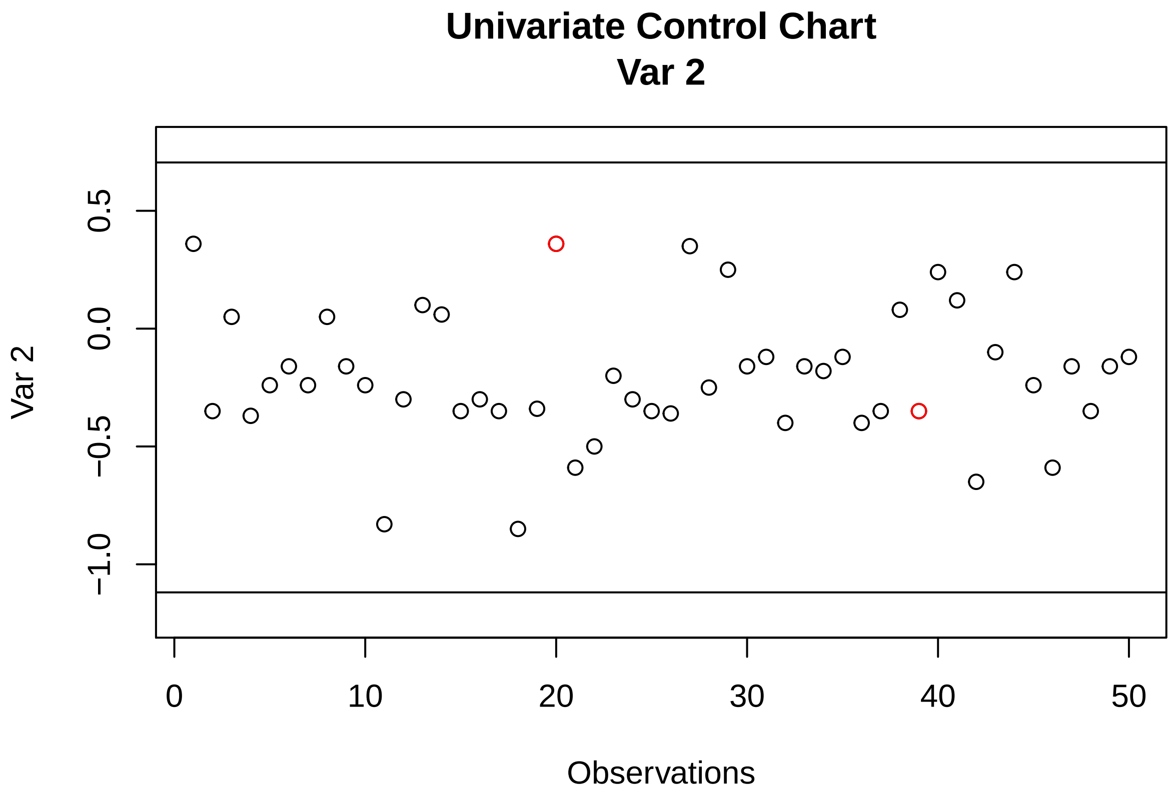
```
abline(h=UCL)
```

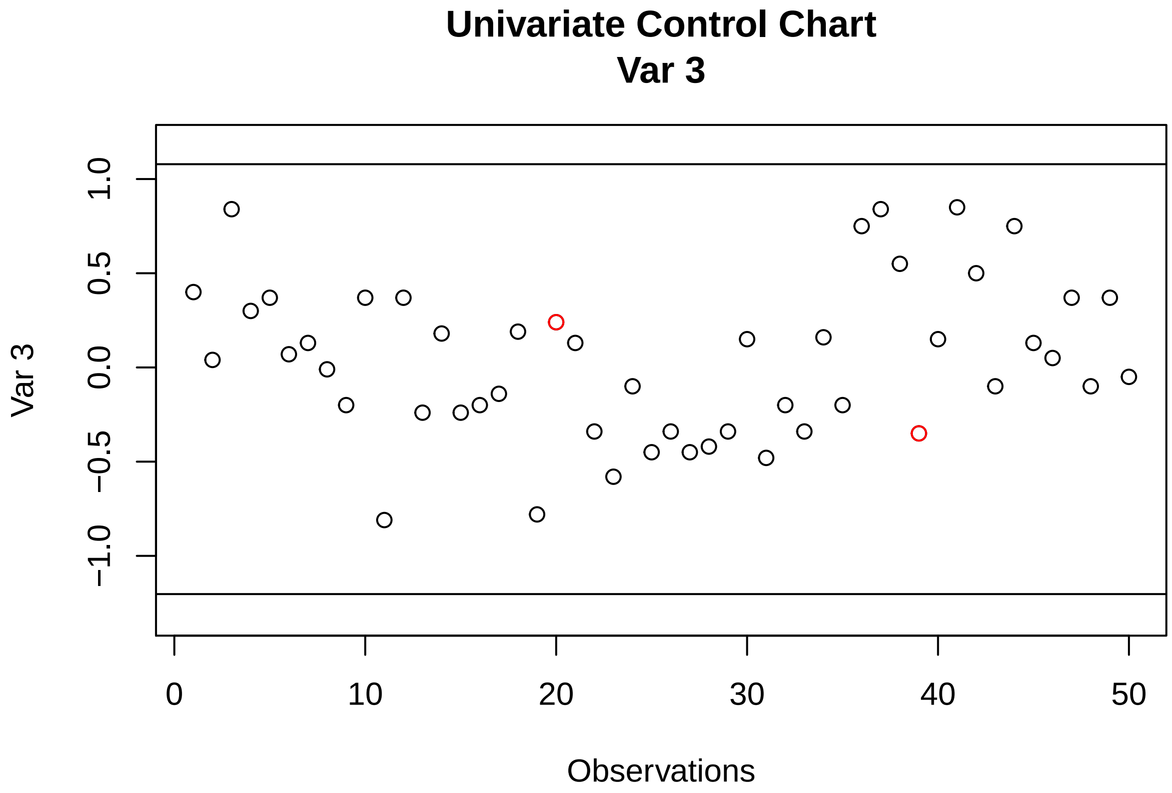


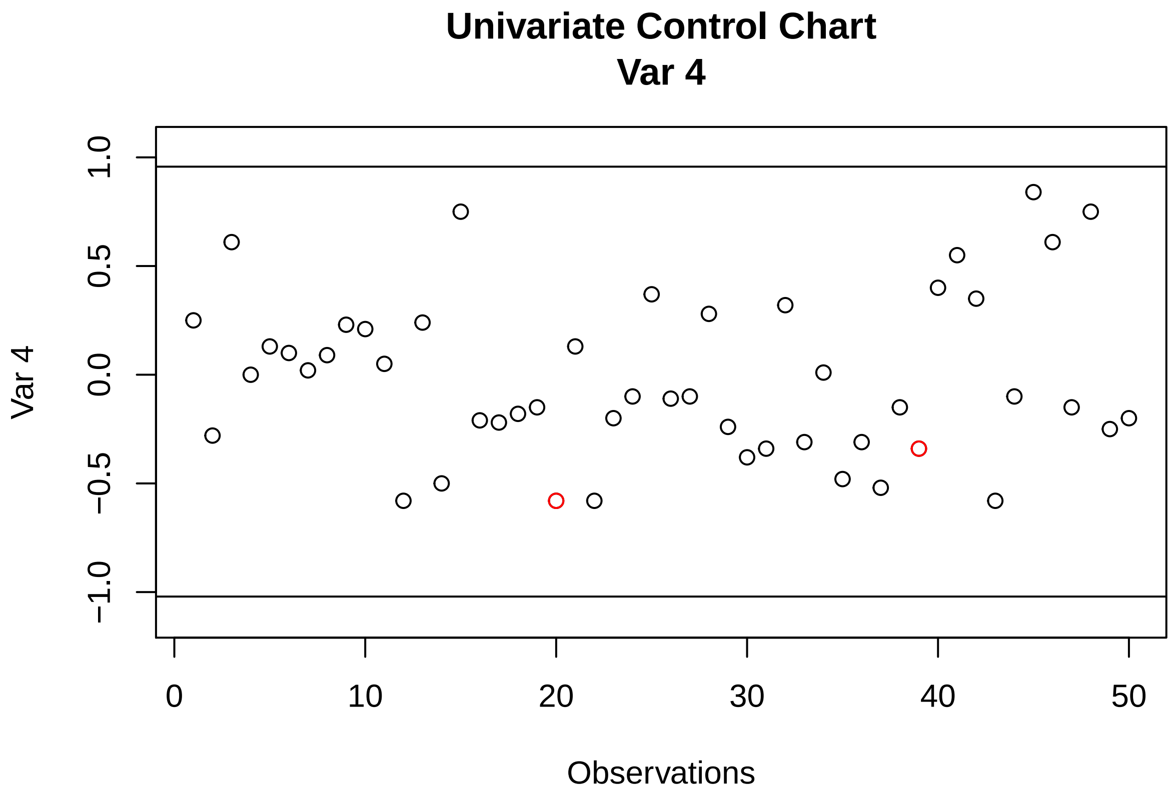
f)

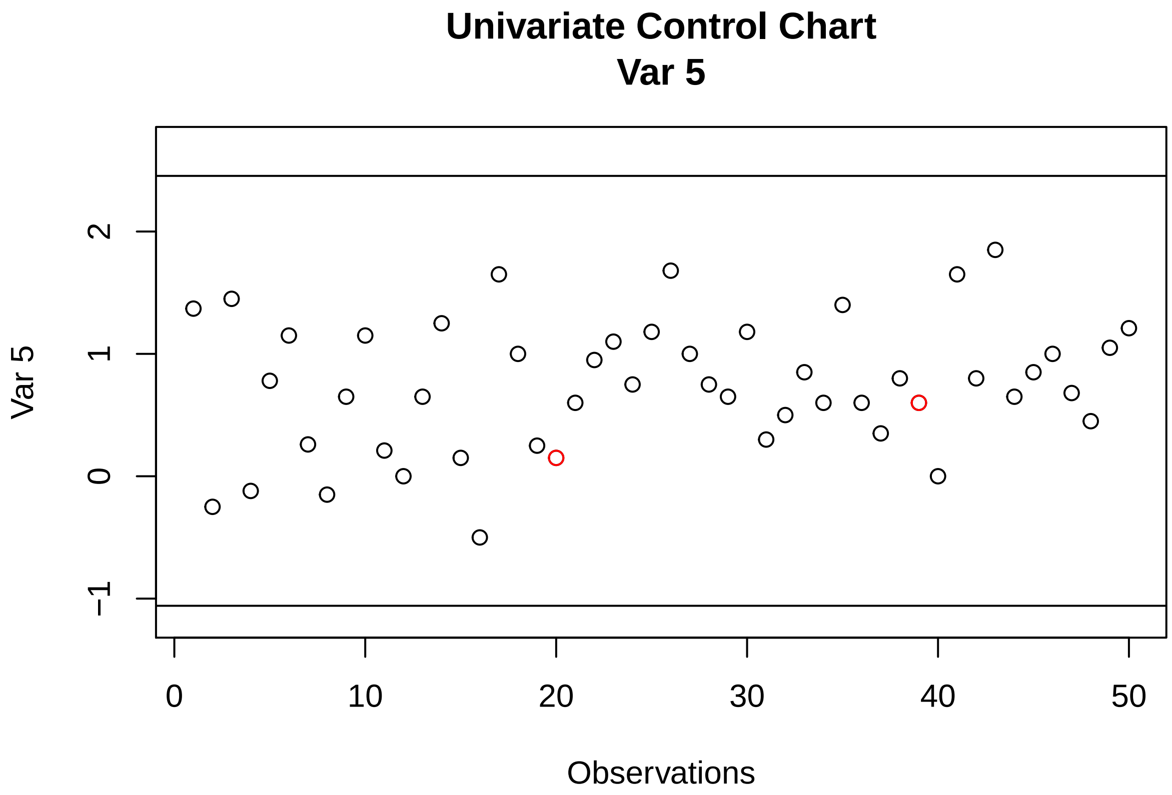
```
#plot Univariate control plots for all 50 observations and highlight points over
#MV control limit as red
for (i in c(1,2,3,4,5,6)){
  plot(1:50,data[,i],ylim=c(1.1*min(data[,i],my_means[i]-3*sqrt(my_var[i])),1.1*max(
    data[,i],my_means[i]+3*sqrt(my_var[i]))),xlab="Observations",ylab=paste("Var",i))
  title(main=paste("Univariate Control Chart\nVar",i))
  points(over, data[over,i], col = "red")
  abline(h=c(my_means[i]-3*sqrt(my_var[i]),my_means[i]+3*sqrt(my_var[i])))
}
```

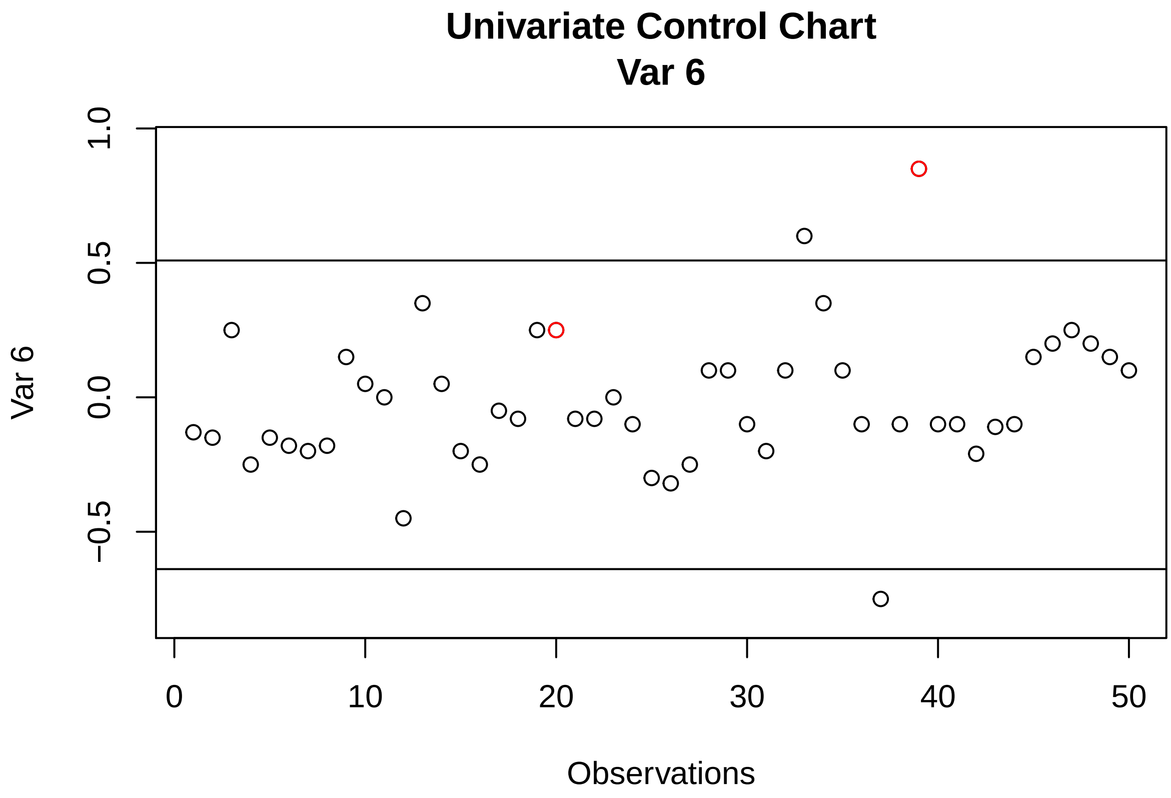

Univariate Control Chart
Var 1
Var 1
0.0
−0.5
−1.0
0
10
20
30
40
50
Observations


Univariate Control Chart
Var 2
Var 2
0.5
0.0
−0.5
−1.0
0
10
20
30
40
50
Observations


Univariate Control Chart
Var 3
Var 3
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
40
50
Observations


Univariate Control Chart
Var 4
Var 4
1.0
0.5
0.0
−0.5
−1.0
0
10
20
30
40
50
Observations


Univariate Control Chart
Var 5
Var 5
2
1
0
−1
0
10
20
30
40
50
Observations



```r
cor(data)
```

```
##              [,1]        [,2]        [,3]        [,4]       [,5]
## [1,]  1.00000000  0.83567296  0.33940689  0.04802316 0.12355479
## [2,]  0.83567296  1.00000000  0.20433646 -0.06522073 0.11744953
## [3,]  0.33940689  0.20433646  1.00000000  0.07918343 0.09678582
## [4,]  0.04802316 -0.06522073  0.07918343  1.00000000 0.01516116
## [5,]  0.12355479  0.11744953  0.09678582  0.01516116 1.00000000
## [6,] -0.03145274  0.08941637 -0.24156171  0.09727123 0.11471950
##              [,6]
## [1,] -0.03145274
## [2,]  0.08941637
## [3,] -0.24156171
## [4,]  0.09727123
## [5,]  0.11471950
## [6,]  1.00000000
```

```r
##############################
#Observations 20 and 39 are over the MV control limit.  Observation 39 seems to mostly be driven
#by variable 6 being outside the univariate control limit.
#No variables have observation 20 outside of the the univariate control limit, but observation 20
#is close to the control limit of variable 2, 4, and 6.  The UCL for multivariate used a less
#strict cutoff of .95, rather than the 6sigma cutoff, which is .999.  Also, observation 20 has a
#high Var2, but the Var1 is close to the mean.  When examining the correlation matrix, var 1 and 2
#have by far the highest correlation of 0.8356, so the values of var1 and var2 for observation 20
#would be unexpected.  The combination of all of these could be driving the Tj being outside of the
#control limit for observation 20.
```

```
################################
```

g)

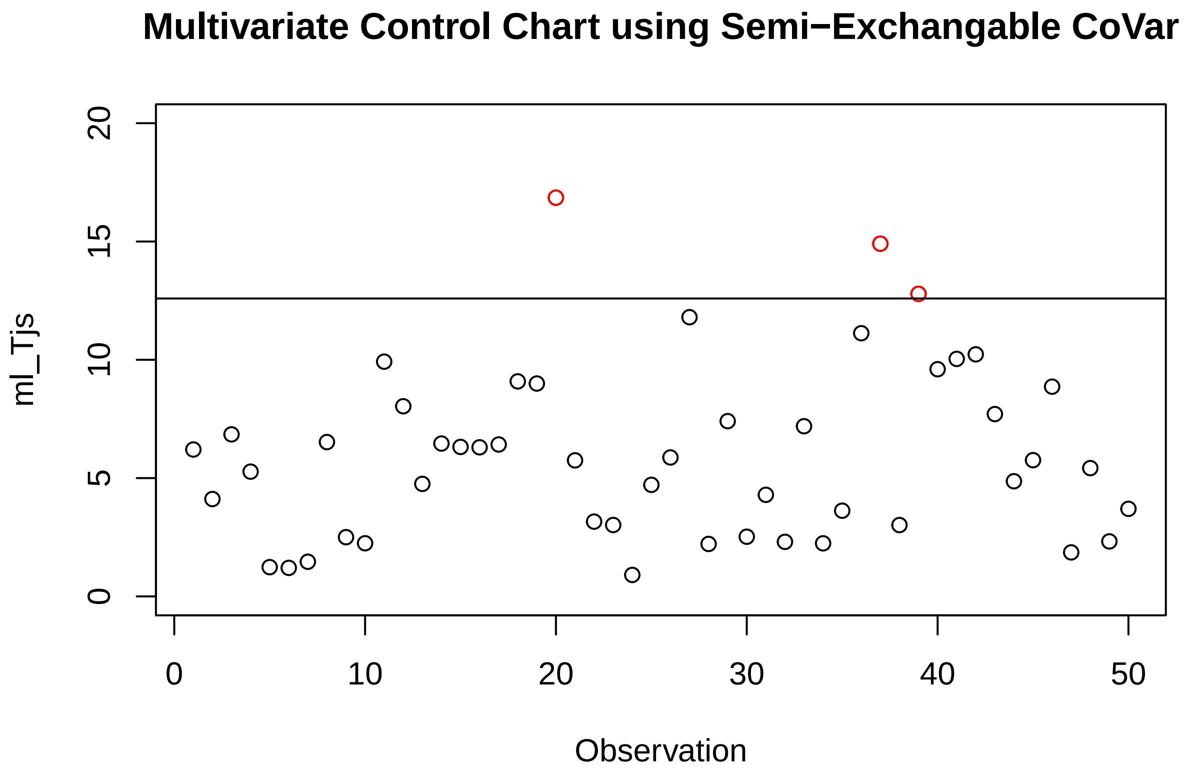
```
#create the new model
mvnll <- function(parms){
  n<-50
  p<-6
  mu <- parms[1:6]
  rho <- parms[7:9]
  sigma <- parms[10:15]

  resid <- car_data - t(matrix(rep(mu,n),p,n))

  cov1 <- matrix(rho[1],4,4)-diag(rep(rho[1],4))+diag(sigma[1:4])
  cov2 <- matrix(rho[2],4,2)
  cov3 <- matrix(rho[3],2,2)-diag(rep(rho[3],2))+diag(sigma[5:6])
  cov <- rbind(cbind(cov1,cov2),cbind(t(cov2),cov3))

  mvnll <- sum(dmvnorm(resid,sigma =cov,log = TRUE))

  -mvnll
}

#perform maximum likelihood estimation of parameters for new model
colMeans(data)
```

```
## [1] -0.4876 -0.1996  0.0358 -0.0170  0.7426 -0.0134
```

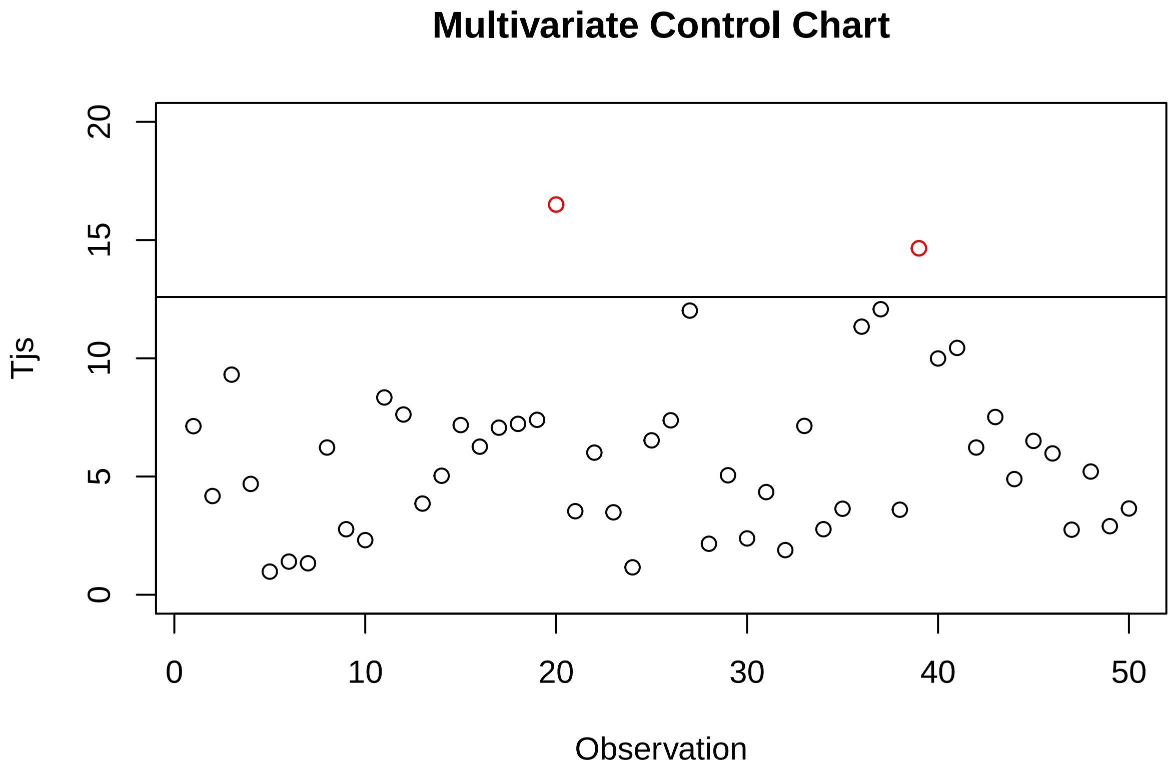
```
diag(S)
```

```
## [1] 0.06494922 0.07731820 0.16863302 0.14106633 0.28596657 0.06756167
```

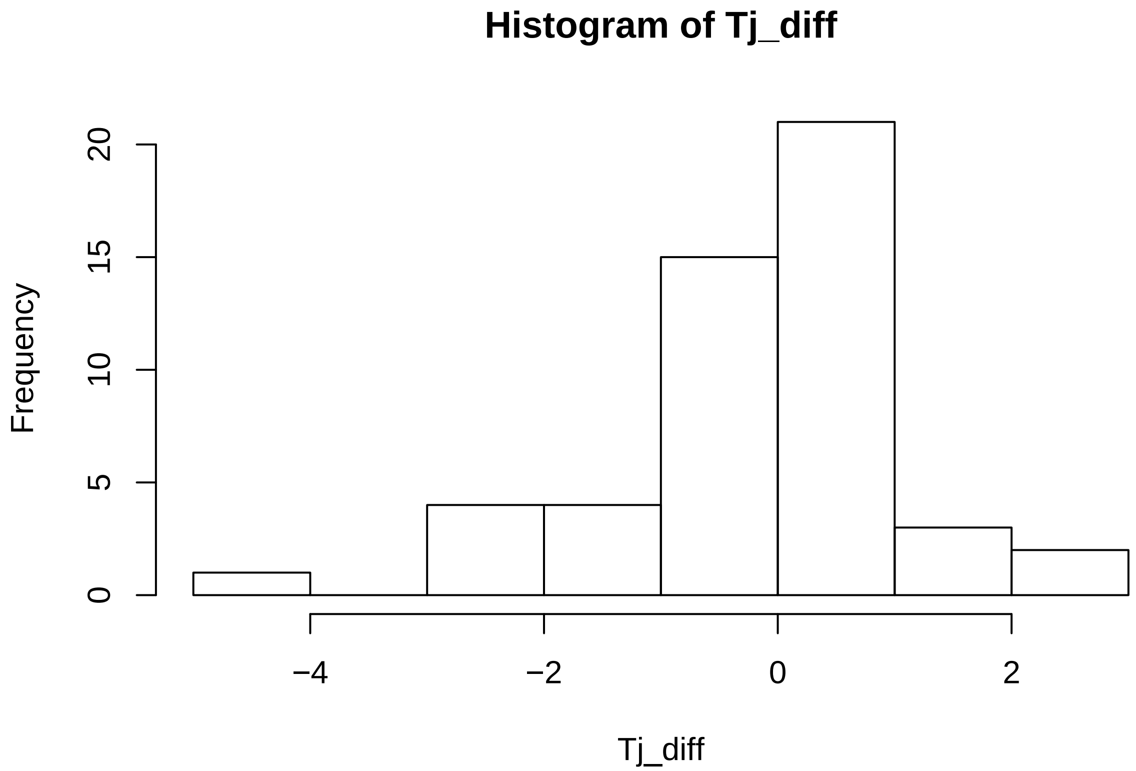
```
nlm.out <-nlm(mvnll,c(rep(0,6),rep(0.5,3),rep(1,6)),hessian = TRUE)

#create the new S matrix for calculation of Tjs
ml_means <- nlm.out$estimate[1:6]
ml_rho <- nlm.out$estimate[7:9]
ml_sigma <- nlm.out$estimate[10:15]

ml_cov1 <- matrix(ml_rho[1],4,4)-diag(rep(ml_rho[1],4))+diag(ml_sigma[1:4])
ml_cov2 <- matrix(ml_rho[2],4,2)
ml_cov3 <- matrix(ml_rho[3],2,2)-diag(rep(ml_rho[3],2))+diag(ml_sigma[5:6])
ml_cov <- rbind(cbind(ml_cov1,ml_cov2),cbind(t(ml_cov2),ml_cov3))

#assign dataset and create components for the calculations for new model
data <- unname(unlist(data.matrix(car_data)))
ml_S <- ml_cov
ml_S_inv<-solve(ml_S)

#calculate Tj's for all observations for new model
ml_Tjs<-c()
for (i in 1:50) {
  diff<-ml_means-data[i,]
  ml_Tj<-t(diff)%*%ml_S_inv%*%diff
  ml_Tjs<-c(ml_Tjs,ml_Tj)
}

#determine upper control limit
UCL<-qchisq(0.95,6)

#plot multivariate control chart and highlight points over UCL as red for new model
plot(1:50,ml_Tjs,ylim=c(0,20),xlab='Observation')
title(main="Multivariate Control Chart using Semi-Exchangable CoVar")
ml_over<-which(ml_Tjs>UCL)
points(ml_over, ml_Tjs[ml_over], col = "red")
abline(h=UCL)
```



```
#reprint multivariate control chart for original model for comparison
plot(1:50,Tjs,ylim=c(0,20),xlab='Observation')
title(main="Multivariate Control Chart")
over<-which(Tjs>UCL)
points(over, Tjs[over], col = "red")
abline(h=UCL)
```

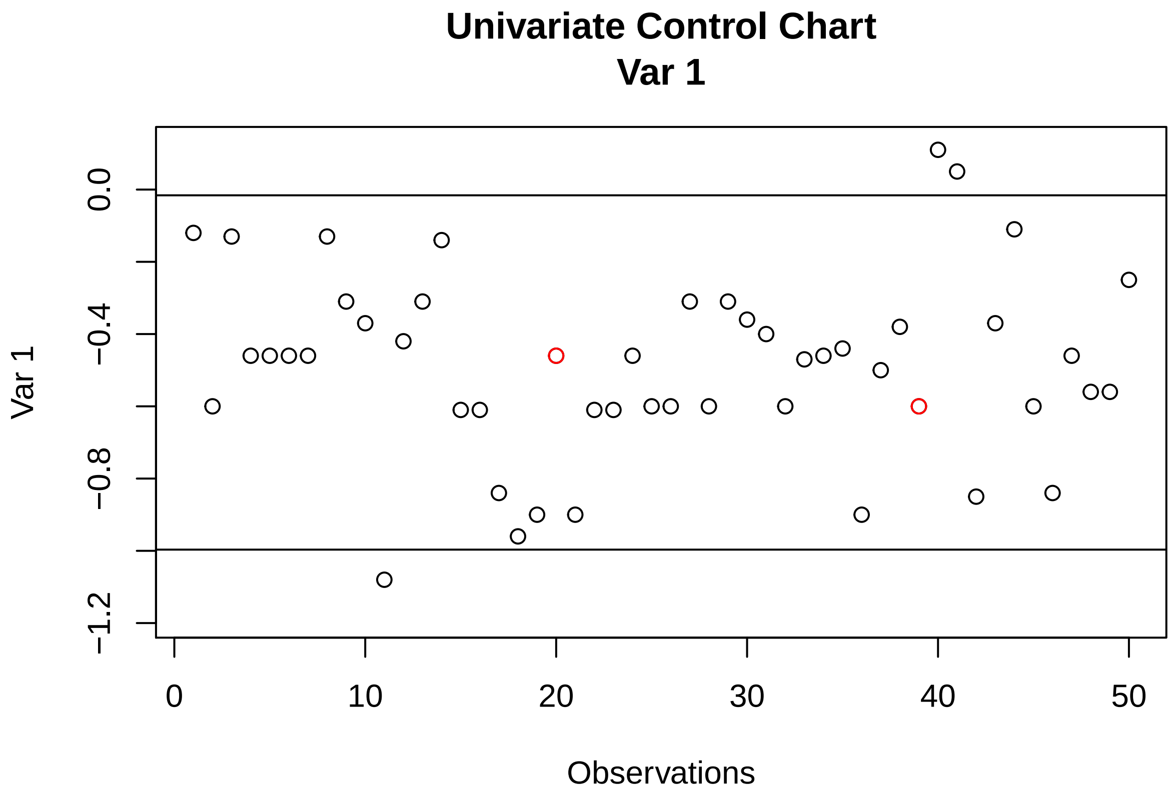


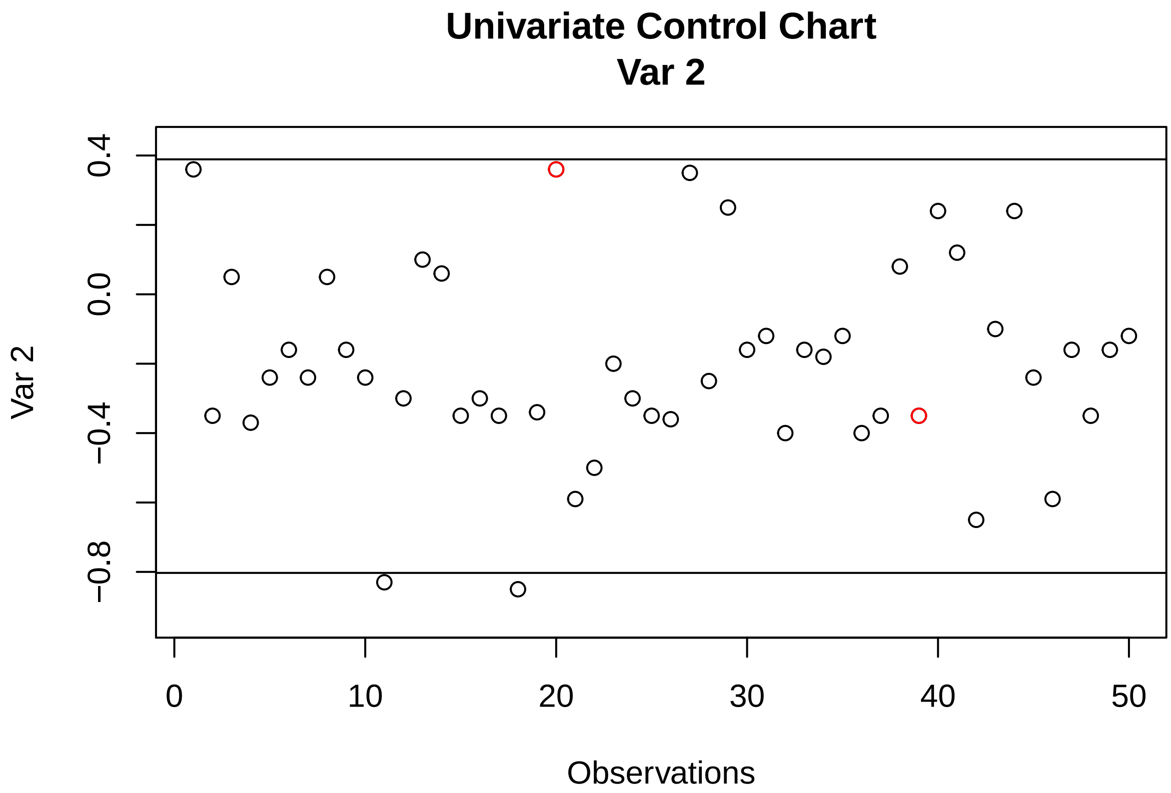
```
#calculate the differences in Tjs and plot using a histogram
Tj_diff<-Tjs-ml_Tjs
hist(Tj_diff)
```

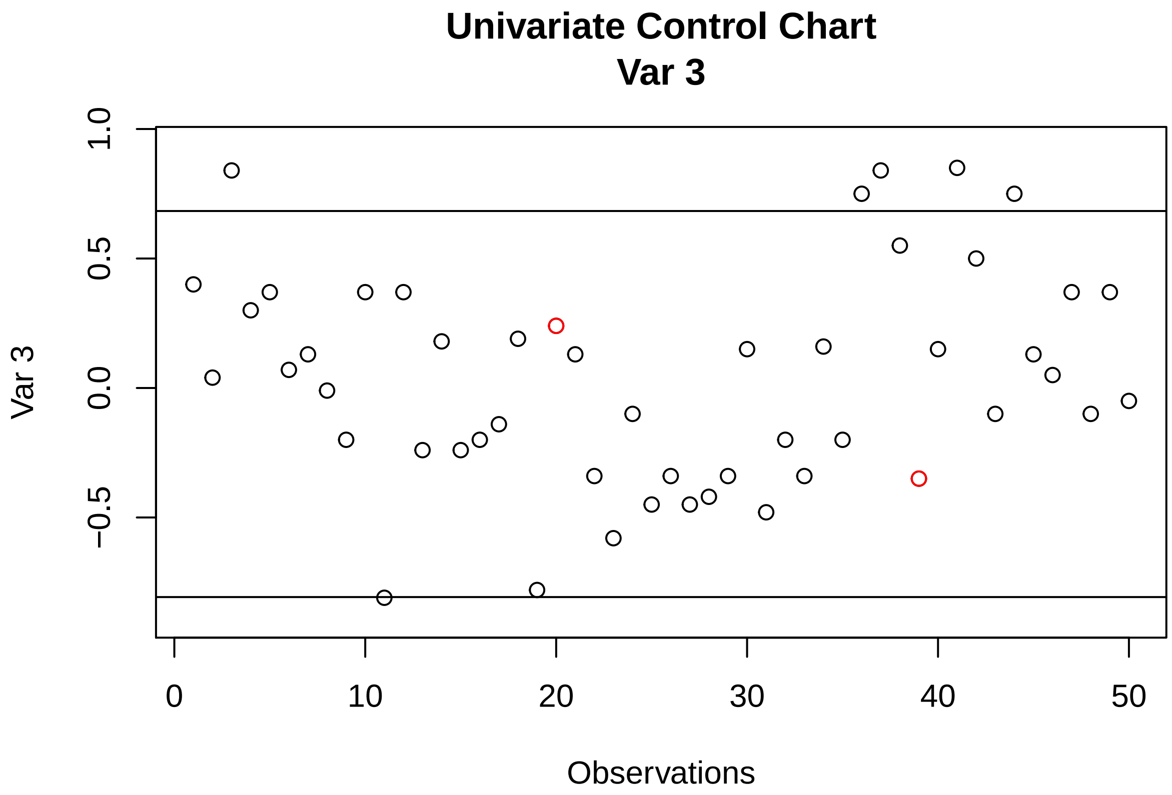


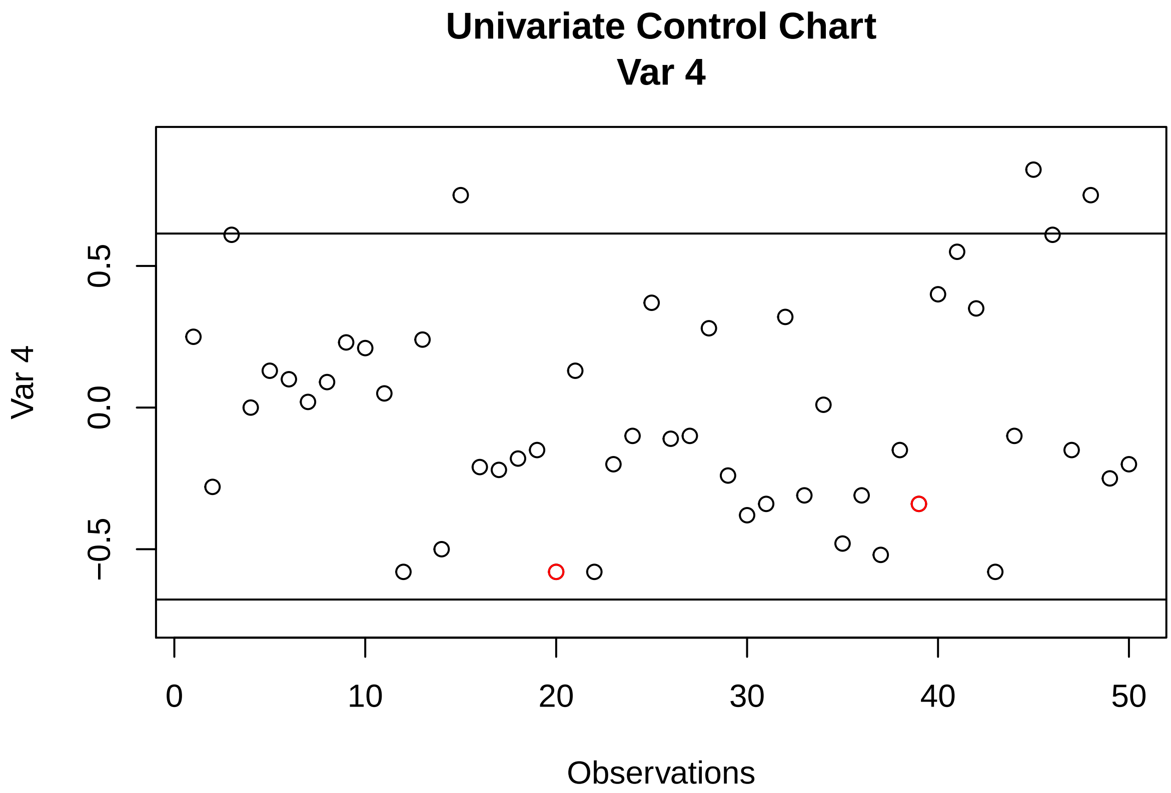
```
###############
#Comparing the two plots and the histogram of the differences in the Tj's, there was a
#significant difference between the Tj's calculated.  Seven of the 50 observations
#changed by over 2, which is a significant portion of the UCL of 12.59.  This resulted
#in one new observation, 37, being labeled as out of control.  Observations 20 and 39
#were also labeled as out of control using this new model.
###############
```

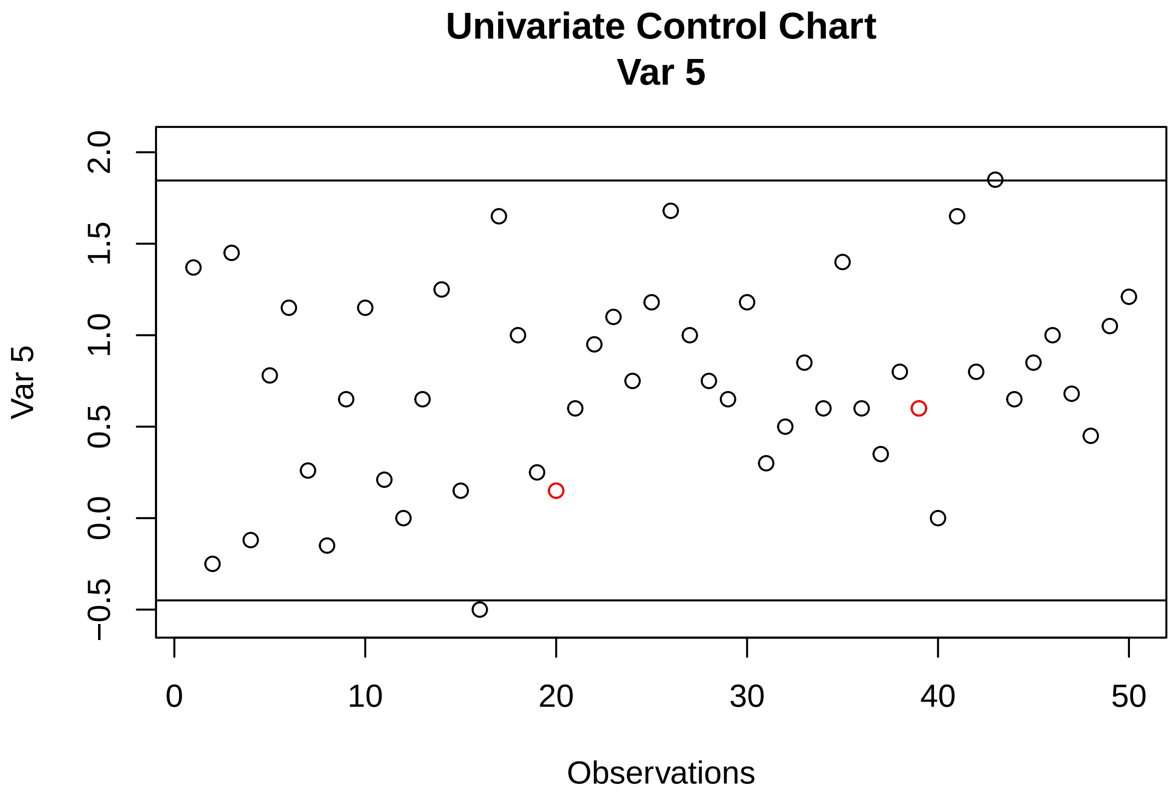
```
#EXPLORATORY STUFF

#plot Univariate control plots for all 50 observations and highlight points over MV control limit as
#red, change cutoff so 95% of data within control limits
for (i in c(1,2,3,4,5,6)){
  plot(1:50,data[,i],ylim=c(1.1*min(data[,i],my_means[i]-1.959964*sqrt(my_var[i])),
                            1.1*max(data[,i],my_means[i]+1.959964*sqrt(my_var[i]))),xlab="Observations"
       ylab=paste("Var",i))
  title(main=paste("Univariate Control Chart\nVar",i))
  points(over, data[over,i], col = "red")
  abline(h=c(my_means[i]-1.959964*sqrt(my_var[i]),my_means[i]+1.959964*sqrt(my_var[i])))
}
```

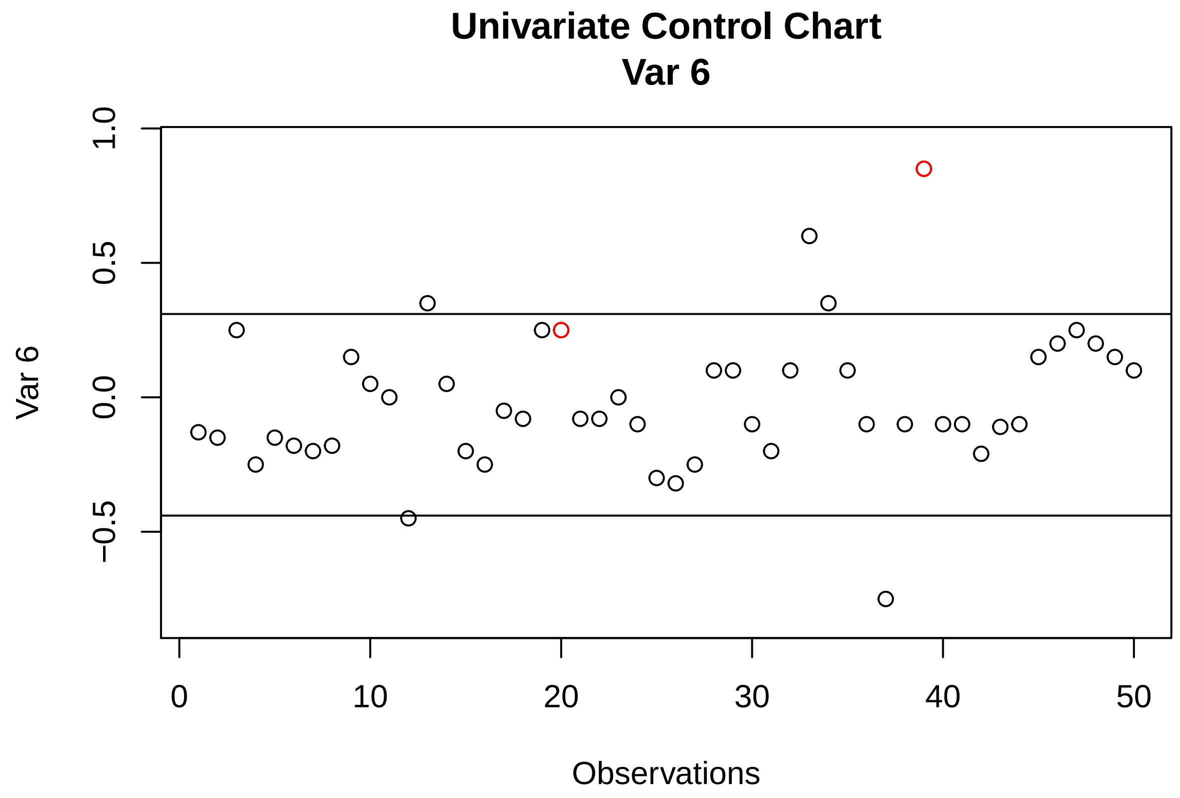

Univariate Control Chart
Var 1
Var 1
0.0
−0.4
−0.8
−1.2
0
10
20
30
40
50
Observations


Univariate Control Chart
Var 2
Var 2
0.4
0.0
−0.4
−0.8
0
10
20
30
40
50
Observations

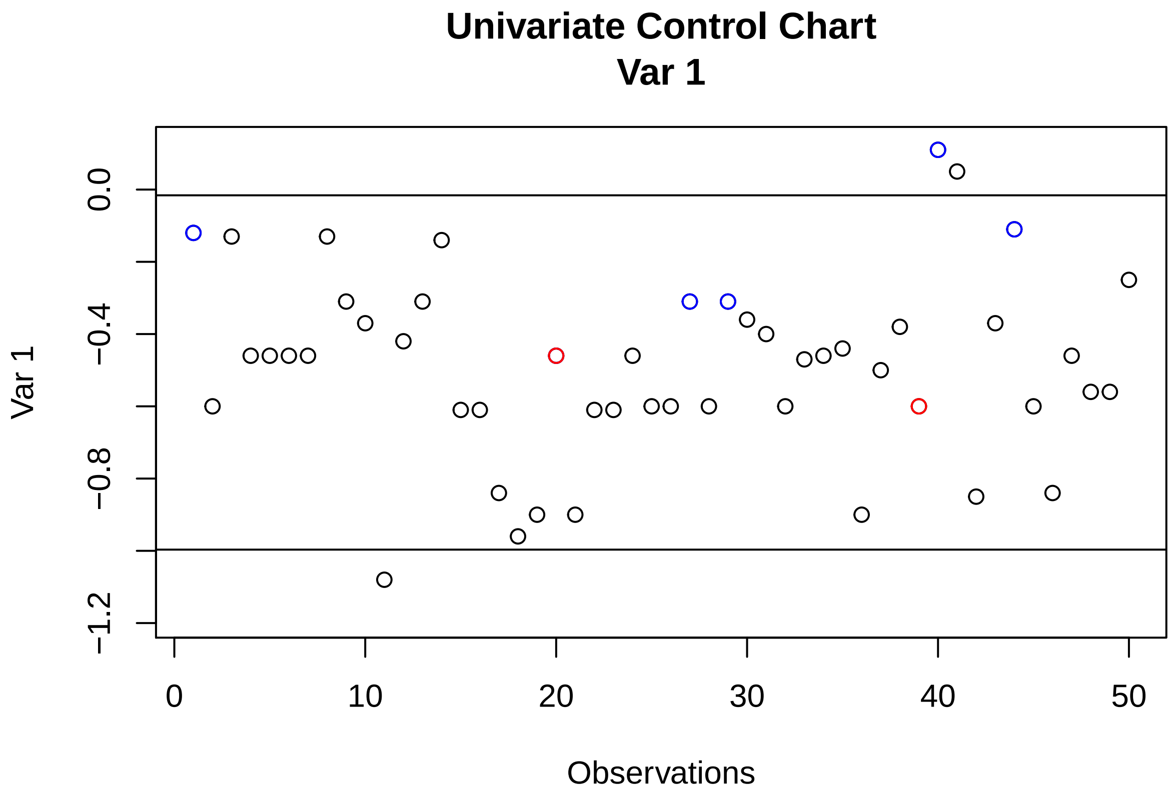

Univariate Control Chart
Var 3
Var 3
1.0
0.5
0.0
−0.5
0
10
20
30
40
50
Observations


Univariate Control Chart
Var 4
Var 4
0.5
0.0
−0.5
0
10
20
30
40
50
Observations


Univariate Control Chart
Var 5
Var 5
−0.5
0.0
0.5
1.0
1.5
2.0
0
10
20
30
40
50
Observations



```
#plot var1 again, highlight values with high var2
over2<-which(data[,2]>.2)
i=1
plot(1:50,data[,i],ylim=c(1.1*min(data[,i],my_means[i]-1.959964*sqrt(my_var[i])),1.1*max(
  data[,i],my_means[i]+1.959964*sqrt(my_var[i]))),xlab="Observations",ylab=paste("Var",i))
  title(main=paste("Univariate Control Chart\nVar",i))
  points(over2, data[over2,i], col = "blue")
  points(over, data[over,i], col = "red")
  abline(h=c(my_means[i]-1.959964*sqrt(my_var[i]),my_means[i]+1.959964*sqrt(my_var[i])))
```


Univariate Control Chart
Var 1
Var 1
0.0
−0.4
−0.8
−1.2
0
10
20
30
40
50
Observations