cusum_revised.R

zhuoxun.yang001

2022-10-09

```
####分为模型建设与结论两个阶段。
###模型建设(R语言)
########initialize##############
##### coding by yang zx
##### import data and transform to time series
# clear & reset environment
rm(list = ls())
# import qcc
library(qcc)
# import data(广东中支)
library(readxl)
df <- read_excel("C:/Users/zhuoxun.yang001/Desktop/广东中支.xlsx", sheet = "Sheet4")
```

```
New names:
• `` -> `...1`
```

```
# check head
head(df)
```

```
# A tibble: 6 × 19
  ...1   A4400-广…¹ A4404-…² A4405…³ A4406…⁴ A4407…⁵ A4408…⁶ A4409…⁷ A4412…⁸ A4413…⁹ A4414…ˣ A4415…ˣ A4416…ˣ A4418…ˣ A4419…ˣ A4420…ˣ A4451…ˣ
  <chr>       <dbl>    <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
1 2019.1      131.      16.1   27.8    261.     4.08   102.    89.5    25.6     74.2   190.    102.   27.0     73.1     47.5    76.2   151.
2 2019.2       54.6     58.2   13.6     67.7    2.28    45.0   50.2    16.4     38.3   102.     82.5  11.7     42.7     11.8    92.8    99.2
3 2019.3      310.      66.0   22.8    368.    35.6    331.   169.     64.7    152.    361.    212.   72.8    101.      92.2   231.    311.
4 2019.4       28.1     25.6    4.23    38.7   -3.94    39.7    1.66    5.61    16.2    26.2    47.8   0.259    77.63    10.7    37.7    47.4
5 2019.5      131.      35.2   35.4    137.    10.7    163.   108.     47.0     62.0   230.     99.8  31.4     41.      4
```

```
3.0     40.7   131.    245.
6 2019.6      108.      54.6   34.8    102.     6.27   144.   106.     36.9     79.7   143.     90.8  49.0     3
9.3     26.7   136.    261.
# … with 2 more variables: `A4452-广东分公司-揭阳中心支公司` <dbl>, `A4456-广东分公司-云浮中心支公司` <dbl>, and abbrevia
ted variable names
#   ¹`A4400-广东分公司-广州中心支公司`, ²`A4404-广东分公司-珠海中心支公司`, ³`A4405-广东分公司-汕头中心支公司`,
#   ⁴`A4406-广东分公司-佛山中心支公司`, ⁵`A4407-广东分公司-江门中心支公司`, ⁶`A4408-广东分公司-湛江中心支公司`,
#   ⁷`A4409-广东分公司-茂名中心支公司`, ⁸`A4412-广东分公司-肇庆中心支公司`, ⁹`A4413-广东分公司-惠州中心支公司`,
#   ˣ`A4414-广东分公司-梅州中心支公司`, ˣ`A4415-广东分公司-汕尾中心支公司`, ˣ`A4416-广东分公司-河源中心支公司`,
#   ˣ`A4418-广东分公司-清远中心支公司`, ˣ`A4419-广东分公司-东莞中心支公司`, ˣ`A4420-广东分公司-中山中心支公司`,
#   ˣ`A4451-广东分公司-潮州中心支公司`
```

```r
# check statistic summary
summary(df)
```

```
    ...1           A4400-广东分公司-广州中心支公司 A4404-广东分公司-珠海中心支公司 A4405-广东分公司-汕头中心支公司
 Length:36          Min.   : 13.32                 Min.   :-0.8667               Min.   :  1.926
 Class :character   1st Qu.: 44.51                 1st Qu.: 4.1878               1st Qu.:  9.116
 Mode  :character   Median : 67.20                 Median : 9.2830               Median : 18.493
                    Mean   :105.74                 Mean   :15.4656               Mean   : 24.128
                    3rd Qu.:120.72                 3rd Qu.:16.9775               3rd Qu.: 32.273
                    Max.   :442.74                 Max.   :66.0368               Max.   :108.360
 A4406-广东分公司-佛山中心支公司 A4407-广东分公司-江门中心支公司 A4408-广东分公司-湛江中心支公司 A4409-广东分公司-茂名中心支公司
 Min.   :  9.318                Min.   : -3.939                Min.   : 13.50                 Min.   :  1.661
 1st Qu.: 50.481                1st Qu.:  1.269                1st Qu.: 39.55                 1st Qu.: 13.734
 Median : 83.262                Median :  6.441                Median : 49.07                 Median : 24.649
 Mean   :128.843                Mean   : 19.089                Mean   : 96.81                 Mean   : 41.478
 3rd Qu.:184.743                3rd Qu.: 19.178                3rd Qu.:127.14                 3rd Qu.: 50.558
 Max.   :409.566                Max.   :117.287                Max.   :417.62                 Max.   :168.848
 A4412-广东分公司-肇庆中心支公司 A4413-广东分公司-惠州中心支公司 A4414-广东分公司-梅州中心支公司 A4415-广东分公司-汕尾中心支公司
 Min.   : 0.8281                Min.   : -12.55                Min.   :-23.31                 Min.   :  3.57
 1st Qu.: 5.0420                1st Qu.:  19.99                1st Qu.: 33.59                 1st Qu.: 15.64
 Median :12.1404                Median :  91.02                Median :108.14                 Median : 39.29
 Mean   :17.6070                Mean   : 295.65                Mean   :166.51                 Mean   : 54.67
 3rd Qu.:24.6906                3rd Qu.: 476.02                3rd Qu.:178.56                 3rd Qu.: 81.05
 Max.   :64.6636                Max.   :1313.80                Max.   :779.54                 Max.   :219.71
 A4416-广东分公司-河源中心支公司 A4418-广东分公司-清远中心支公司 A4419-广东分公司-东莞中心支公司 A4420-广东分公司-中山中心支公司
 Min.   :  0.1609               Min.   :  0.462                Min.   :  0.6687               Min.   : 12.45
 1st Qu.:  9.0494               1st Qu.:  5.956                1st Qu.: 11.5137               1st Qu.: 47.85
 Median : 16.4542               Median : 13.896                Median : 25.0935               Median : 73.01
 Mean   : 24.0360               Mean   : 23.672                Mean   : 35.3524               Mean   :104.31
```

```
 3rd Qu.: 30.5444                3rd Qu.: 34.528                3rd Qu.: 48.0409                3rd Qu.:131.37
 Max.   :100.2474                Max.   :100.972                Max.   :157.1021                Max.   :447.96
 A4451-广东分公司-潮州中心支公司 A4452-广东分公司-揭阳中心支公司 A4456-广东分公司-云浮中心支公司
 Min.   :  2.85                  Min.   :  5.695                Min.   : 10.24
 1st Qu.: 48.62                  1st Qu.: 20.302                1st Qu.: 40.83
 Median :100.56                  Median : 49.119                Median : 92.42
 Mean   :146.46                  Mean   : 60.375                Mean   :119.06
 3rd Qu.:231.54                  3rd Qu.: 78.731                3rd Qu.:162.70
 Max.   :507.13                  Max.   :238.633                Max.   :497.92
```

```
# check structure of dataframe
str(df)
```

```
tibble [36 × 19] (S3: tbl_df/tbl/data.frame)
 $ ...1                          : chr [1:36] "2019.1" "2019.2" "2019.3" "2019.4" ...
 $ A4400-广东分公司-广州中心支公司: num [1:36] 131.3 54.6 309.9 28.1 130.6 ...
 $ A4404-广东分公司-珠海中心支公司: num [1:36] 16.1 58.2 66 25.6 35.2 ...
 $ A4405-广东分公司-汕头中心支公司: num [1:36] 27.84 13.6 22.79 4.23 35.4 ...
 $ A4406-广东分公司-佛山中心支公司: num [1:36] 261 67.7 367.6 38.7 137 ...
 $ A4407-广东分公司-江门中心支公司: num [1:36] 4.08 2.28 35.61 -3.94 10.66 ...
 $ A4408-广东分公司-湛江中心支公司: num [1:36] 102.3 45 330.7 39.7 162.7 ...
 $ A4409-广东分公司-茂名中心支公司: num [1:36] 89.53 50.25 168.85 1.66 107.84 ...
 $ A4412-广东分公司-肇庆中心支公司: num [1:36] 25.64 16.44 64.66 5.61 47.02 ...
 $ A4413-广东分公司-惠州中心支公司: num [1:36] 74.2 38.3 151.7 16.2 62 ...
 $ A4414-广东分公司-梅州中心支公司: num [1:36] 189.9 101.6 361 26.2 230.1 ...
 $ A4415-广东分公司-汕尾中心支公司: num [1:36] 101.8 82.5 211.8 47.8 99.8 ...
 $ A4416-广东分公司-河源中心支公司: num [1:36] 26.99 11.653 72.83 0.259 31.432 ...
 $ A4418-广东分公司-清远中心支公司: num [1:36] 73.14 42.75 100.97 7.63 42.97 ...
 $ A4419-广东分公司-东莞中心支公司: num [1:36] 47.5 11.8 92.2 10.7 40.7 ...
 $ A4420-广东分公司-中山中心支公司: num [1:36] 76.2 92.8 230.8 37.7 131.1 ...
 $ A4451-广东分公司-潮州中心支公司: num [1:36] 151.1 99.2 311.5 47.4 245 ...
 $ A4452-广东分公司-揭阳中心支公司: num [1:36] 85.5 76.2 192.5 29 107.1 ...
 $ A4456-广东分公司-云浮中心支公司: num [1:36] 125.7 88.5 317.2 29.5 181.5 ...
```

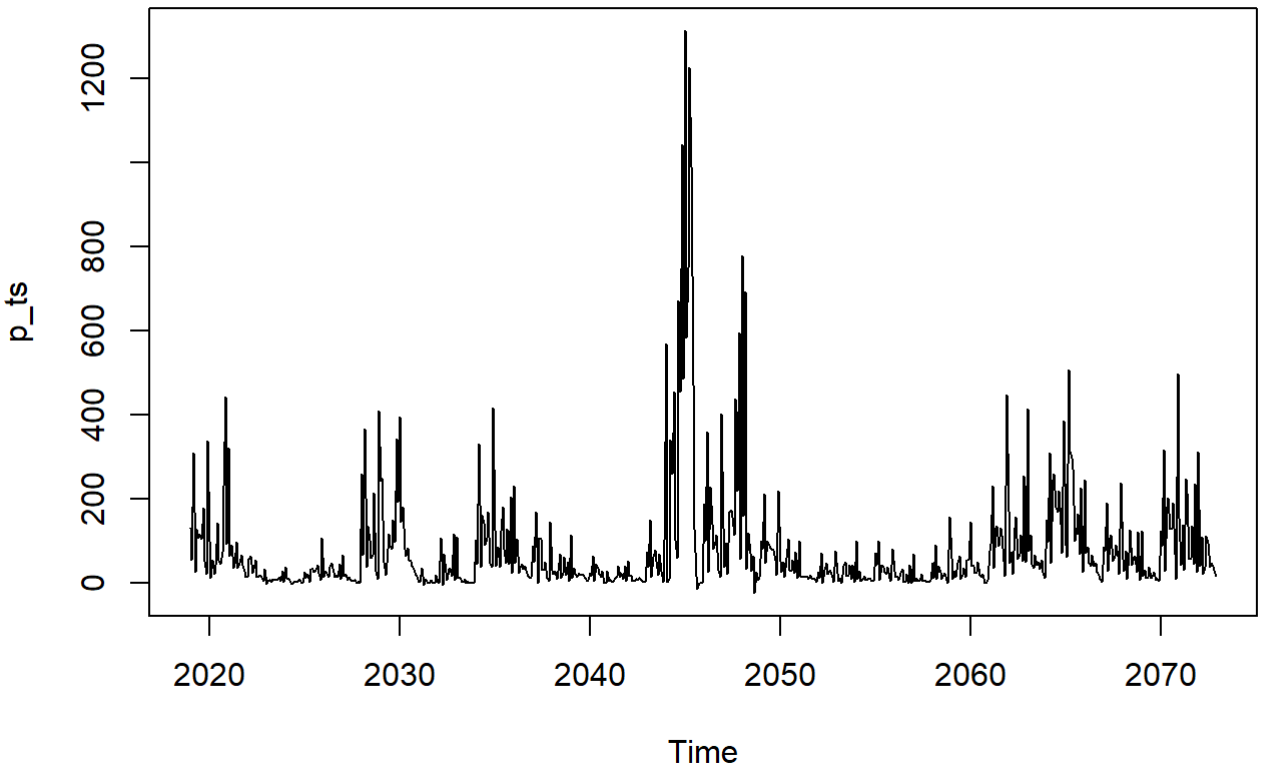
```
# Create matrix for easier processing and referencing
premium = matrix(df[,2:ncol(df)])
premium_vec <- as.vector(unlist(premium))
# check the shape of premium
dim(premium)
```

```
[1] 18  1
```

```
# check the structure of premium
str(premium)
```

```
List of 18
 $ : num [1:36] 131.3 54.6 309.9 28.1 130.6 ...
 $ : num [1:36] 16.1 58.2 66 25.6 35.2 ...
 $ : num [1:36] 27.84 13.6 22.79 4.23 35.4 ...
 $ : num [1:36] 261 67.7 367.6 38.7 137 ...
 $ : num [1:36] 4.08 2.28 35.61 -3.94 10.66 ...
 $ : num [1:36] 102.3 45 330.7 39.7 162.7 ...
 $ : num [1:36] 89.53 50.25 168.85 1.66 107.84 ...
 $ : num [1:36] 25.64 16.44 64.66 5.61 47.02 ...
 $ : num [1:36] 74.2 38.3 151.7 16.2 62 ...
 $ : num [1:36] 189.9 101.6 361 26.2 230.1 ...
 $ : num [1:36] 101.8 82.5 211.8 47.8 99.8 ...
 $ : num [1:36] 26.99 11.653 72.83 0.259 31.432 ...
 $ : num [1:36] 73.14 42.75 100.97 7.63 42.97 ...
 $ : num [1:36] 47.5 11.8 92.2 10.7 40.7 ...
 $ : num [1:36] 76.2 92.8 230.8 37.7 131.1 ...
 $ : num [1:36] 151.1 99.2 311.5 47.4 245 ...
 $ : num [1:36] 85.5 76.2 192.5 29 107.1 ...
 $ : num [1:36] 125.7 88.5 317.2 29.5 181.5 ...
 - attr(*, "dim")= int [1:2] 18 1
```

```
# plot time series
p_ts <- ts(premium_vec, start = 2019, frequency = 12)
plot(p_ts)
```




```
# replicate avg_premium of all premium vector
avg_premium <- rep(0,nrow(premium))
sd_premium <- rep(0,nrow(premium))
avg_premium
```

```
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
sd_premium
```

```
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
# write loop to loop through all the data and assign value to avg & std
for (i in 1:nrow(premium)){
```

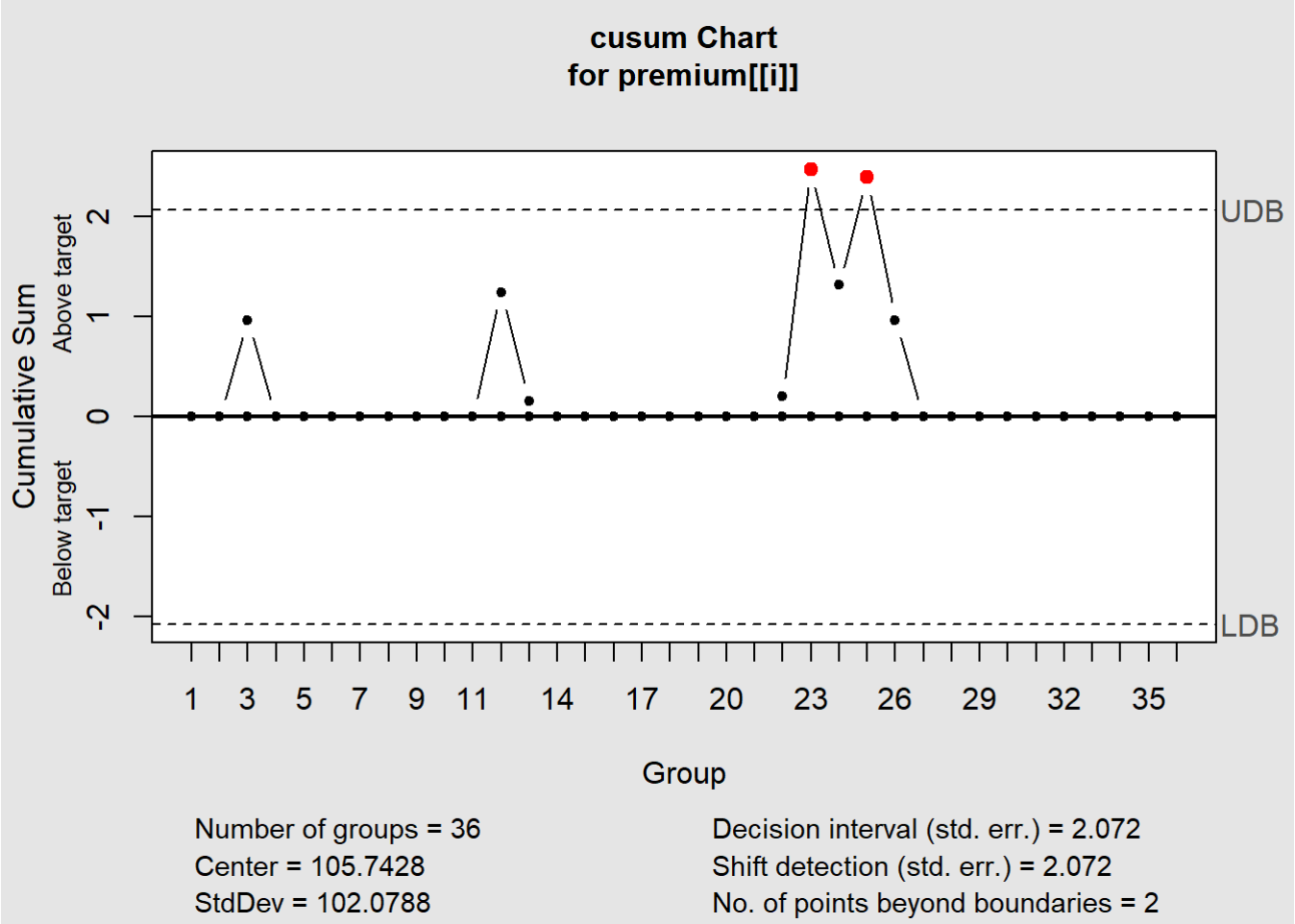
```
  avg_premium[i] <- mean(premium[[i]][1:36])
  sd_premium[i] <- sd(premium[[i]][1:36])}
## print avg & std
avg_premium
```

```
 [1] 105.74282  15.46557  24.12774 128.84287  19.08867  96.81201  41.47769  17.60703 295.65157 166.50599  54.6725
3  24.03601  23.67219
[14]  35.35237 104.31223 146.45808  60.37516 119.06094
```

```
sd_premium
```

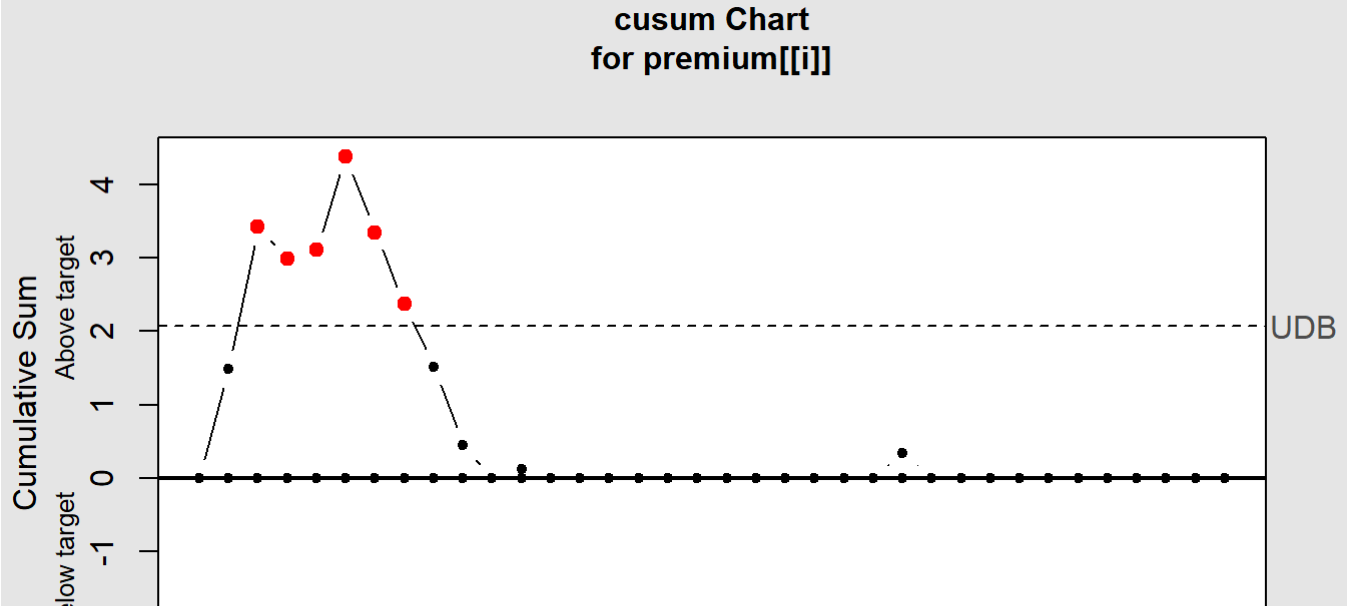
```
 [1] 102.07881  16.97702  20.96781 113.18912  32.10424  89.97145  41.37637  15.89647 377.51614 193.67195  51.2578
8  23.33728  24.96922
[14]  35.53090  98.19291 120.17026  53.22775 106.16691
```

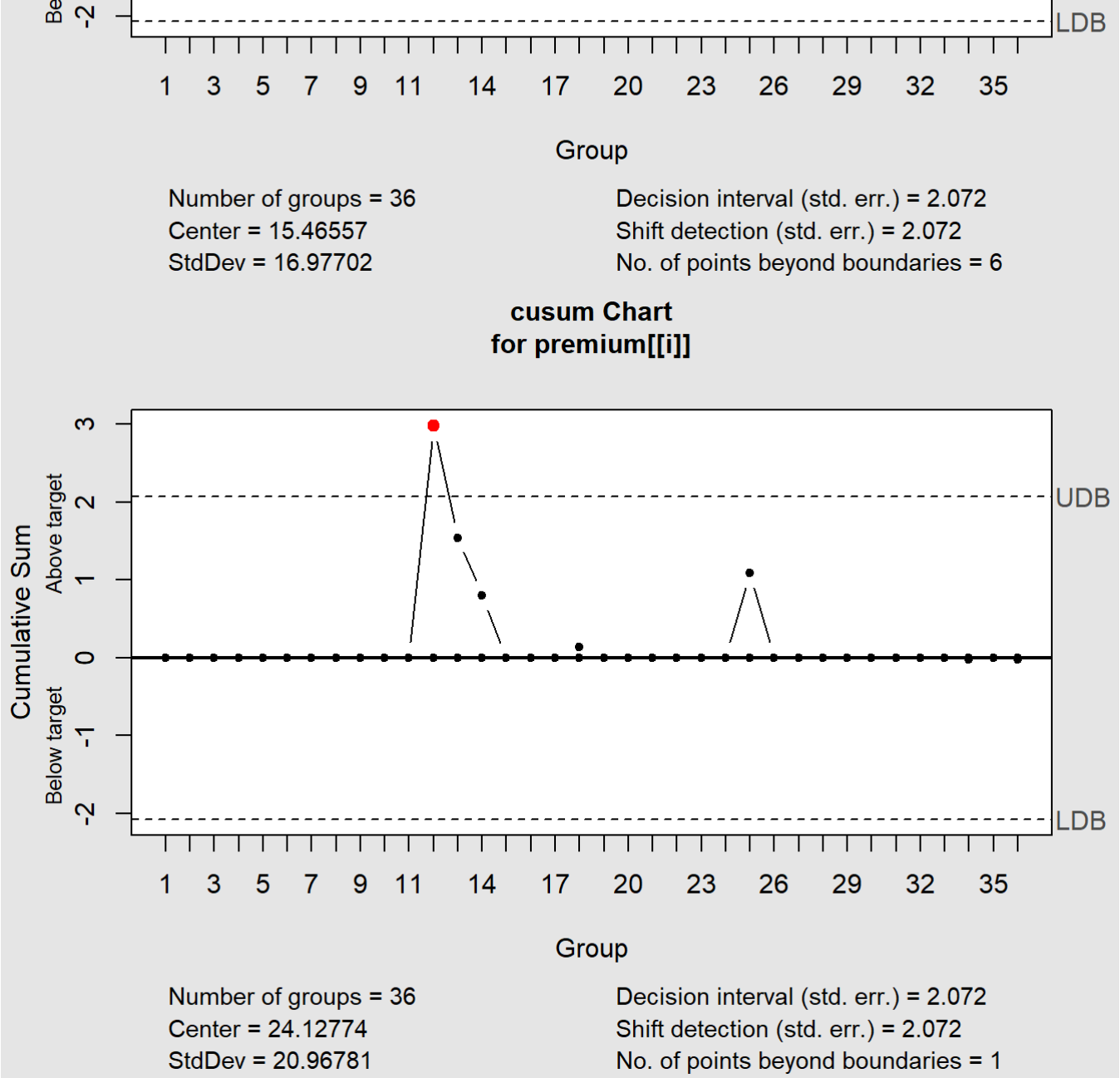
```
######Applying CUSUM (change of detection)
#####
###
CUSUMmodels <- vector(mode="list", length=nrow(premium))
CUSUMviolations <- vector(mode="list", length=nrow(premium))
####
ma_pro <- movavg(avg_premium, n=17, type='e')
di <- 2.072
ss <- 2.072
for (i in 1:nrow(premium)){
  CUSUMmodels[[i]] <- cusum(premium[[i]], center=avg_premium[i], std.dev = sd_premium[i], decision.interval=di, s
e.shift=ss, plot = TRUE)
  CUSUMviolations[[i]] <- CUSUMmodels[[i]]$violations}
```

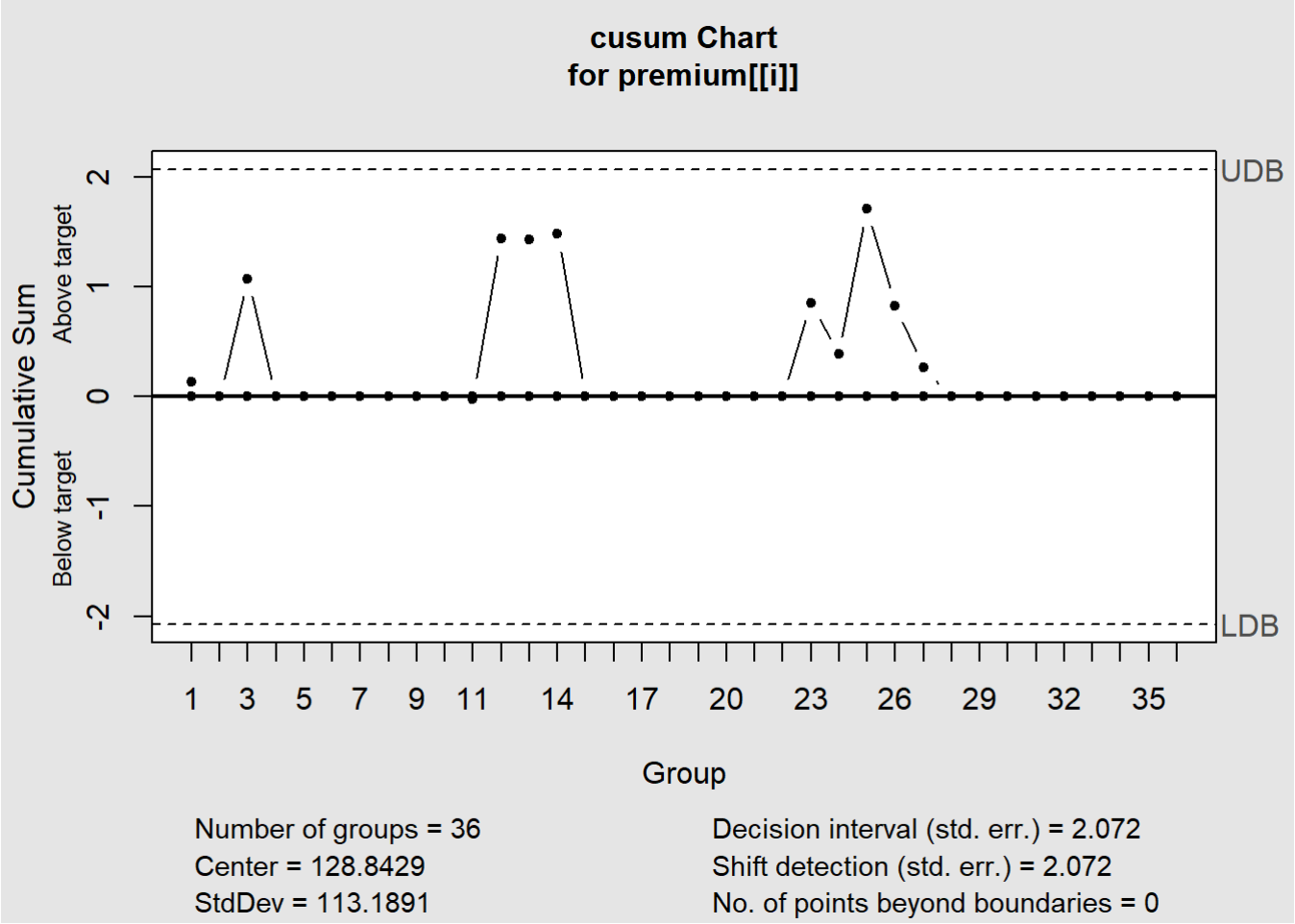


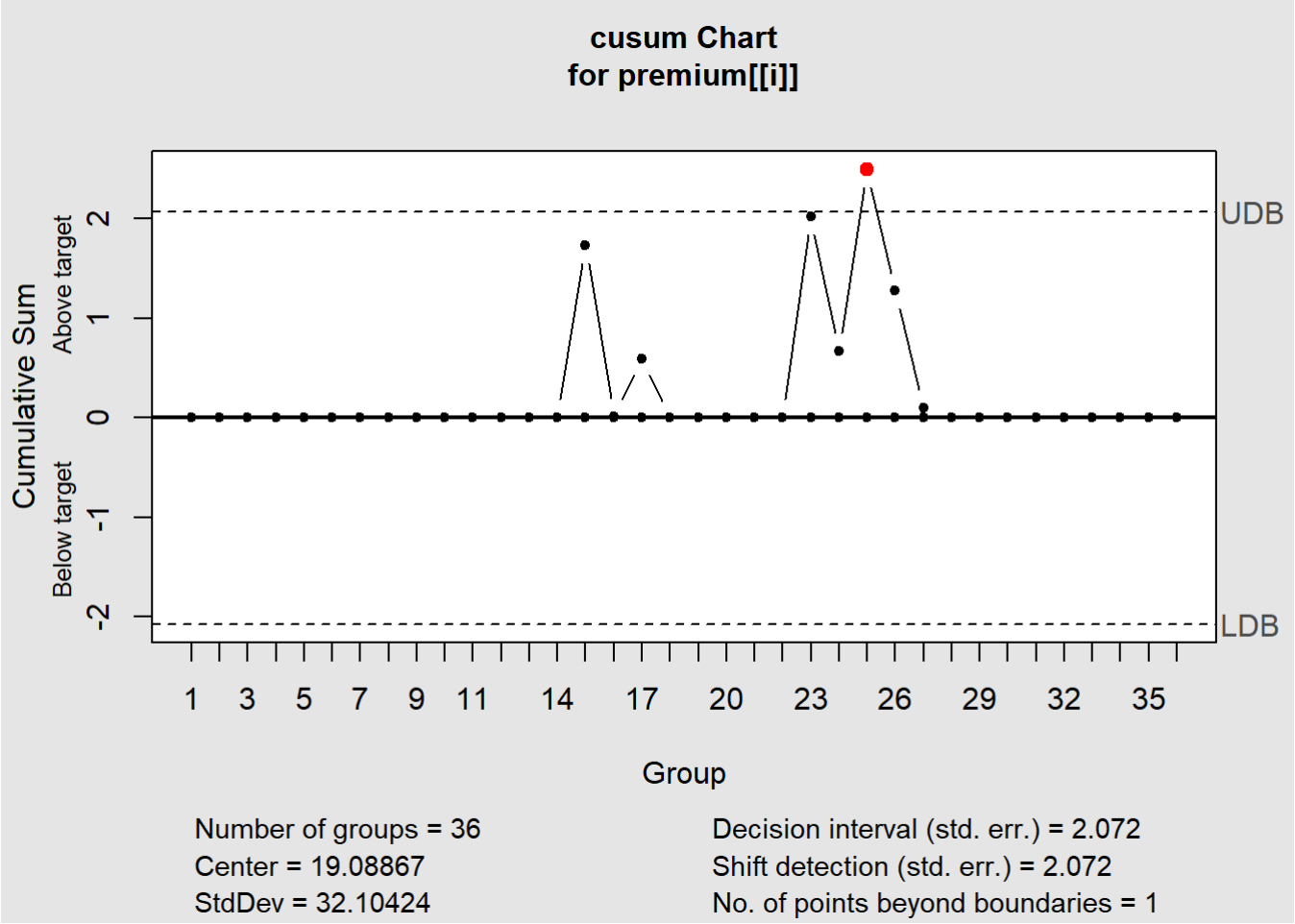
Number of groups = 36
Center = 105.7428
StdDev = 102.0788

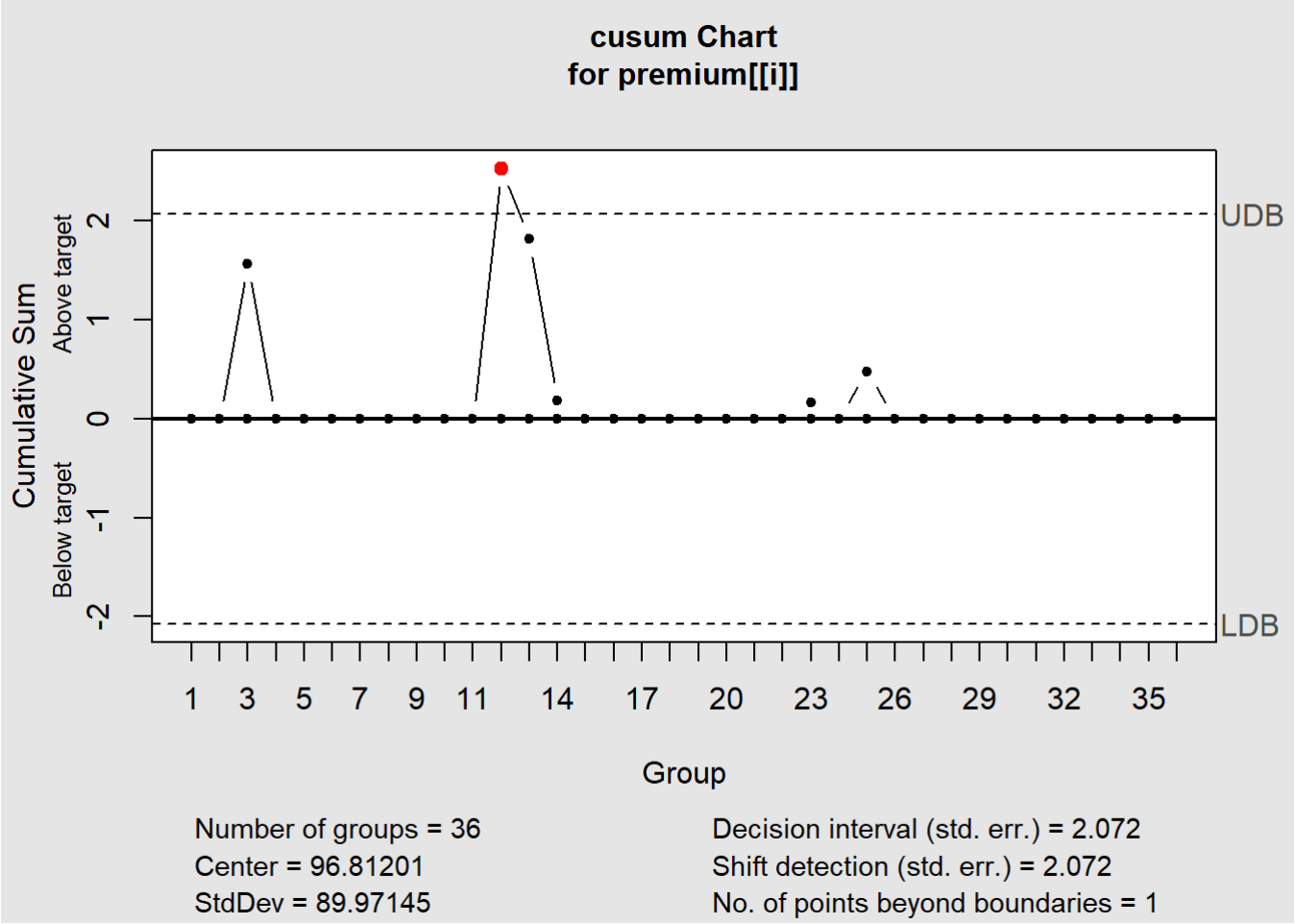
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 2

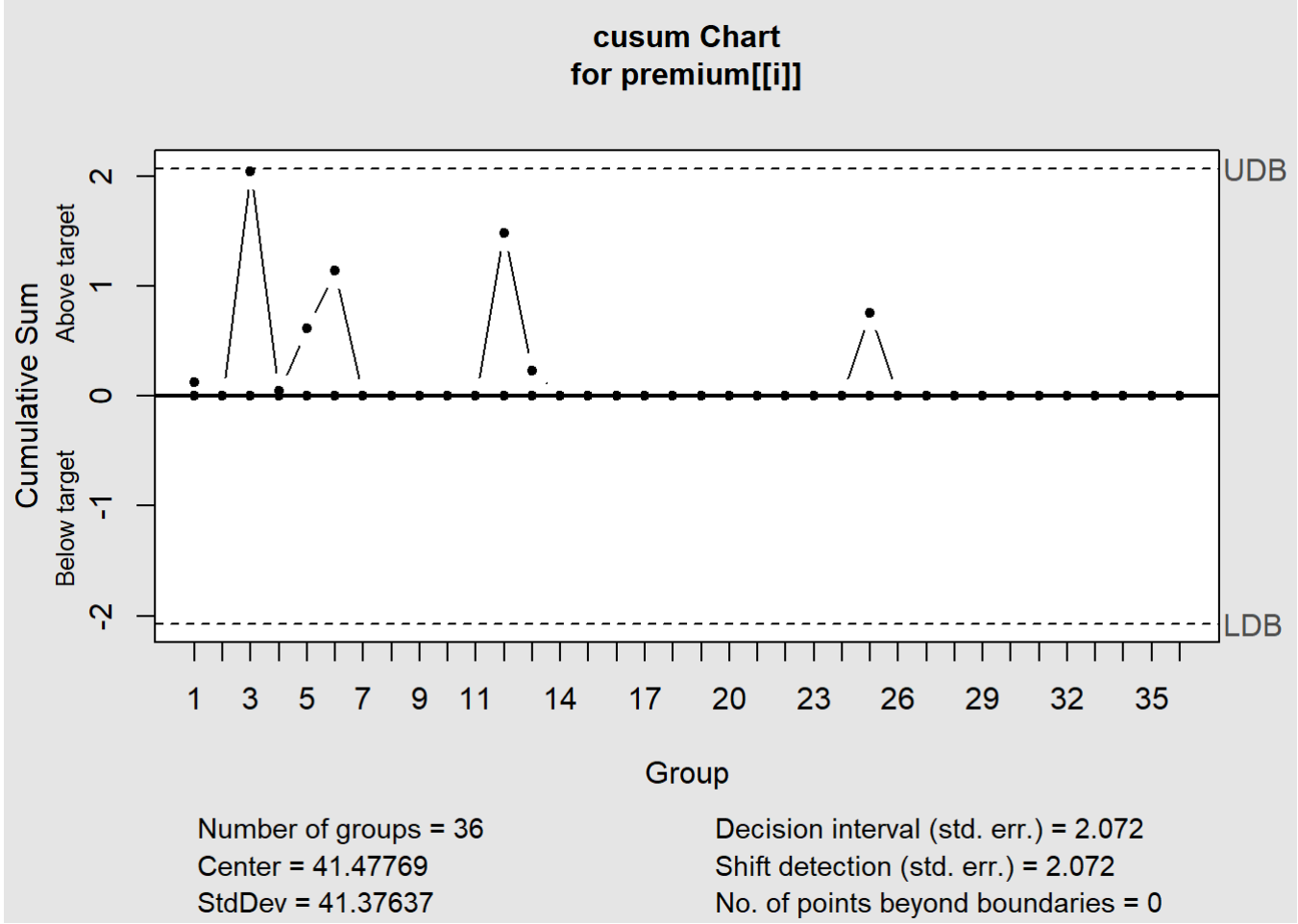


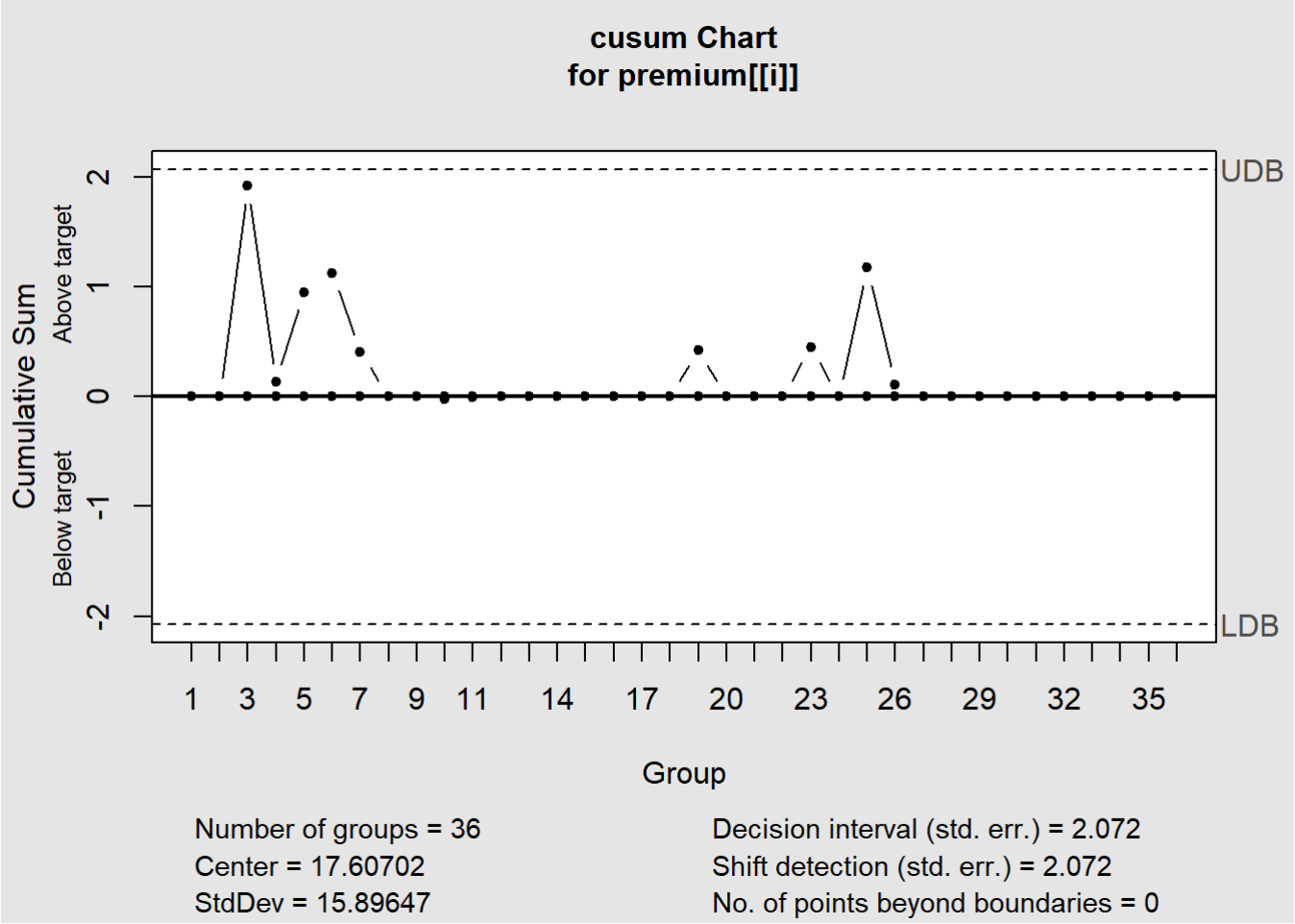

-2
LDB
1 3 5 7 9 11 14 17 20 23 26 29 32 35
Group
Number of groups = 36
Center = 15.46557
StdDev = 16.97702
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 6
cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
3
2
1
0
-1
-2
UDB
LDB
1 3 5 7 9 11 14 17 20 23 26 29 32 35
Group
Number of groups = 36
Center = 24.12774
StdDev = 20.96781
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 1

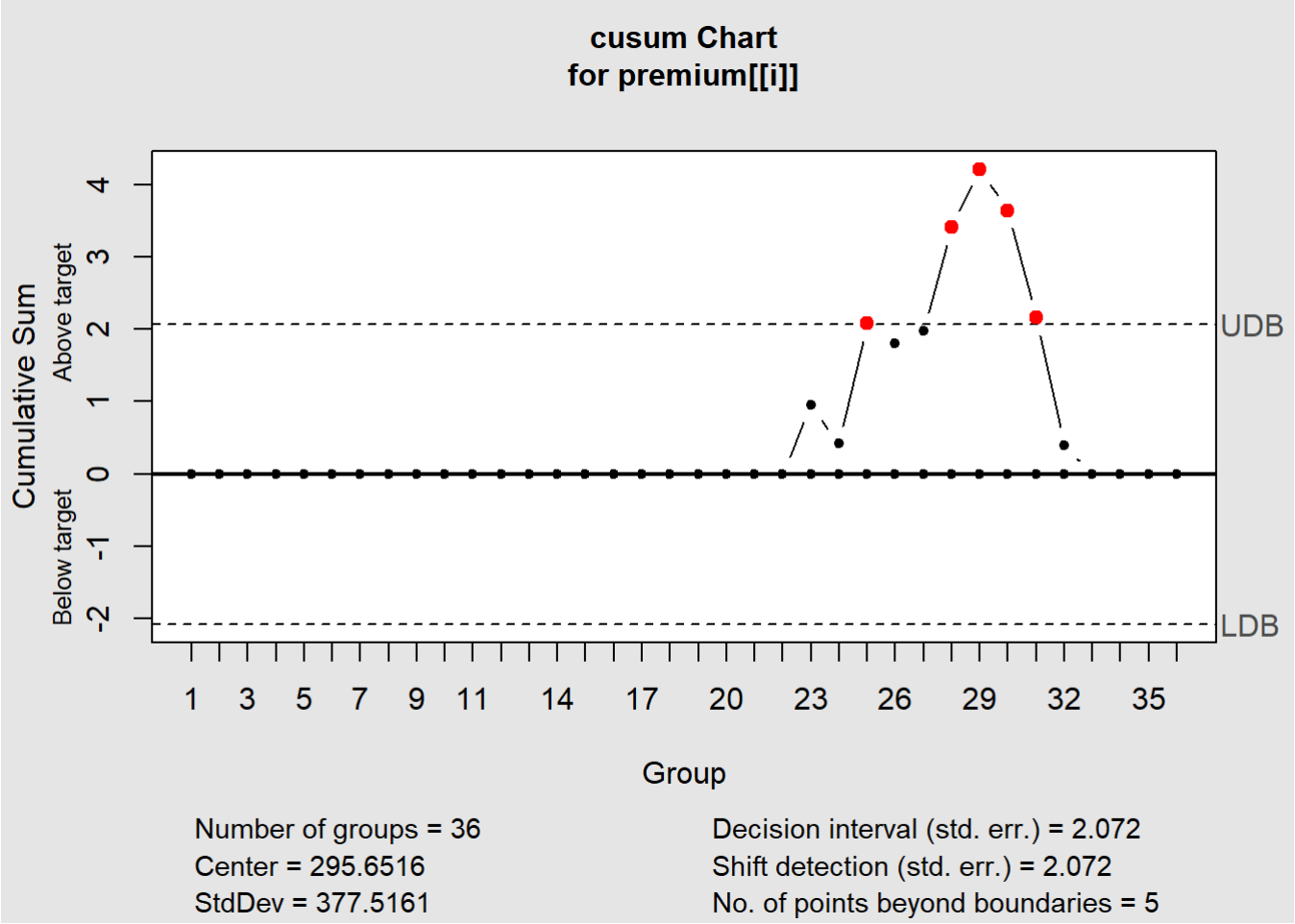

cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 128.8429
StdDev = 113.1891
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 0

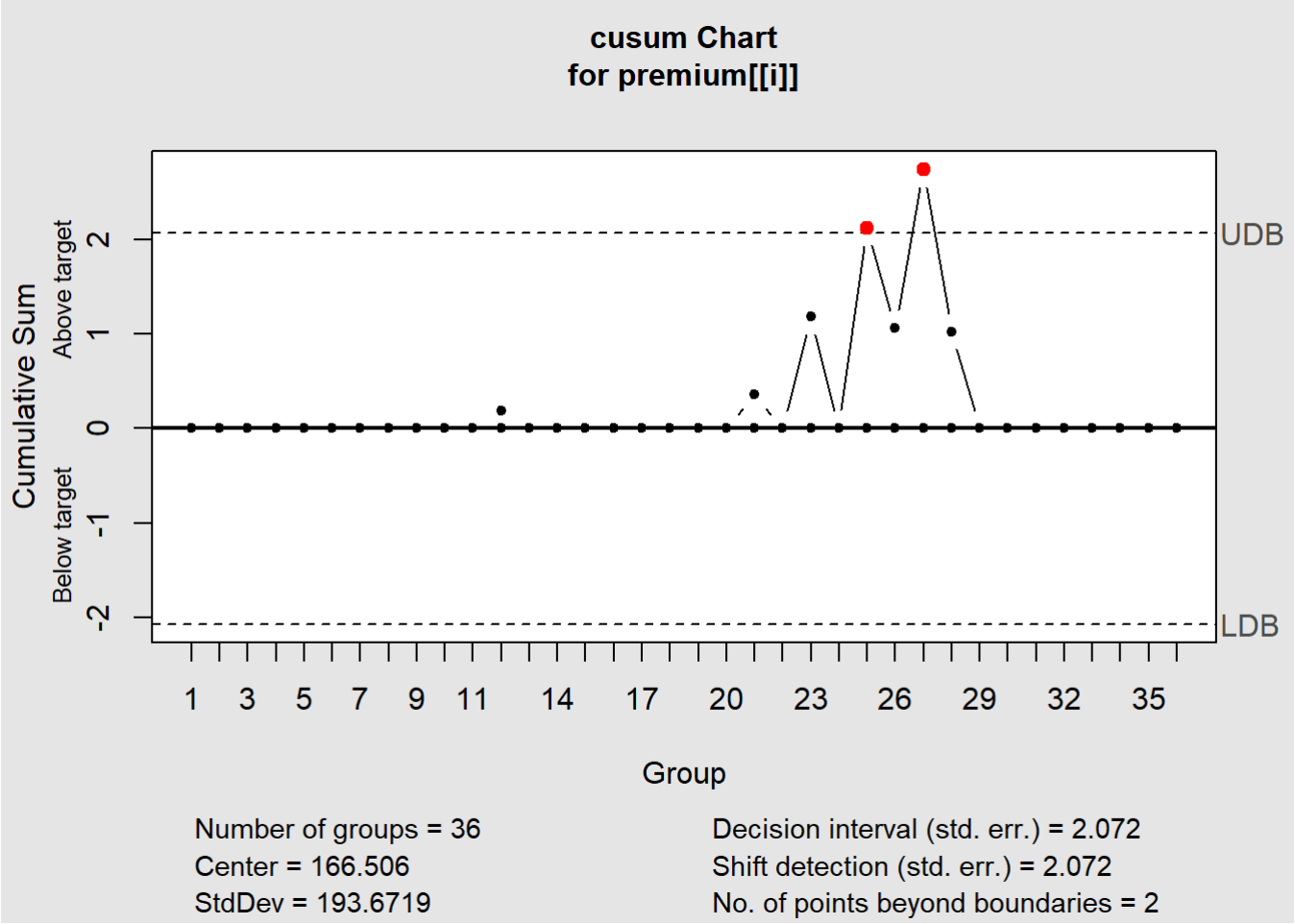

cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 19.08867
StdDev = 32.10424
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 1

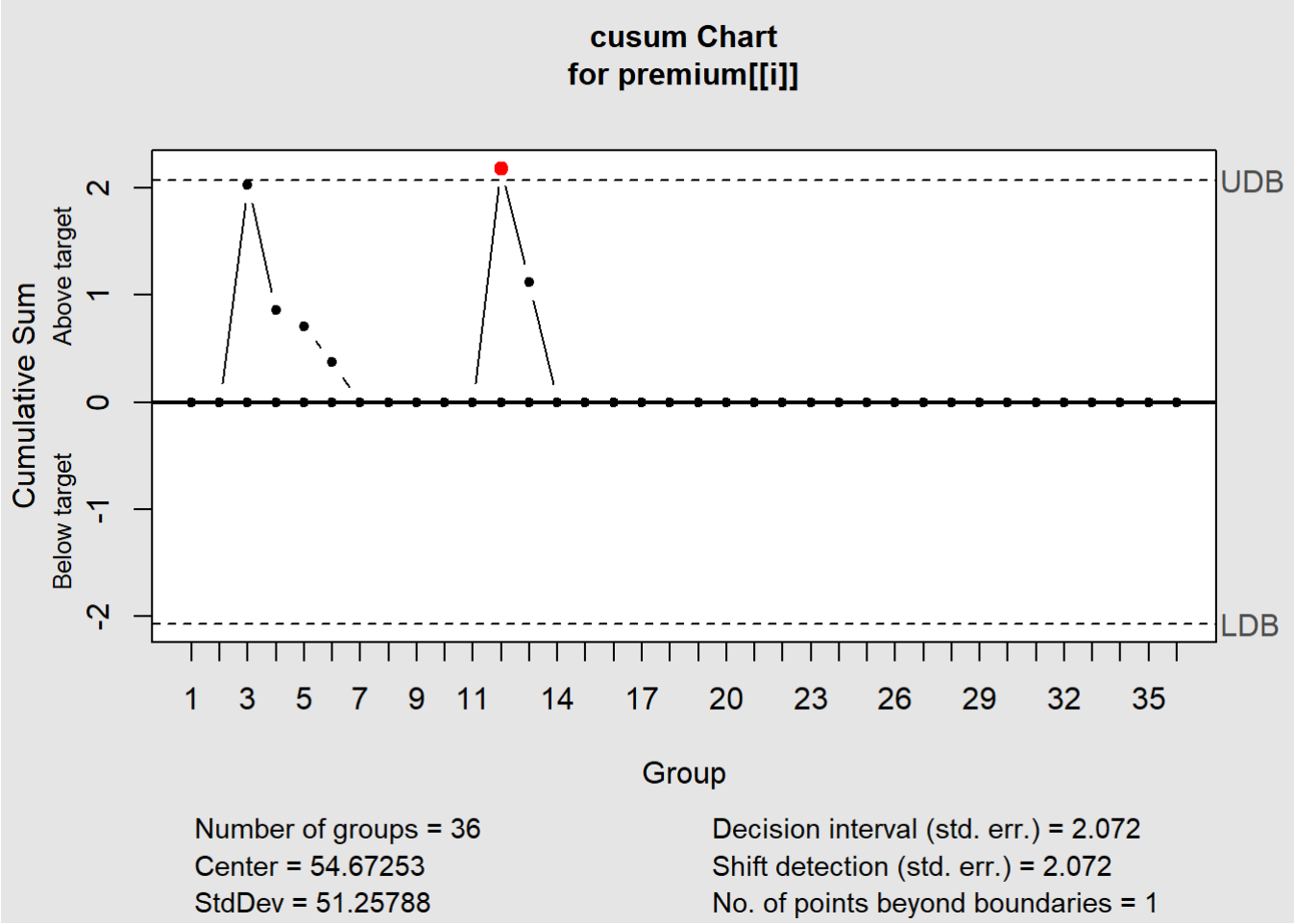

cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 96.81201
StdDev = 89.97145
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 1

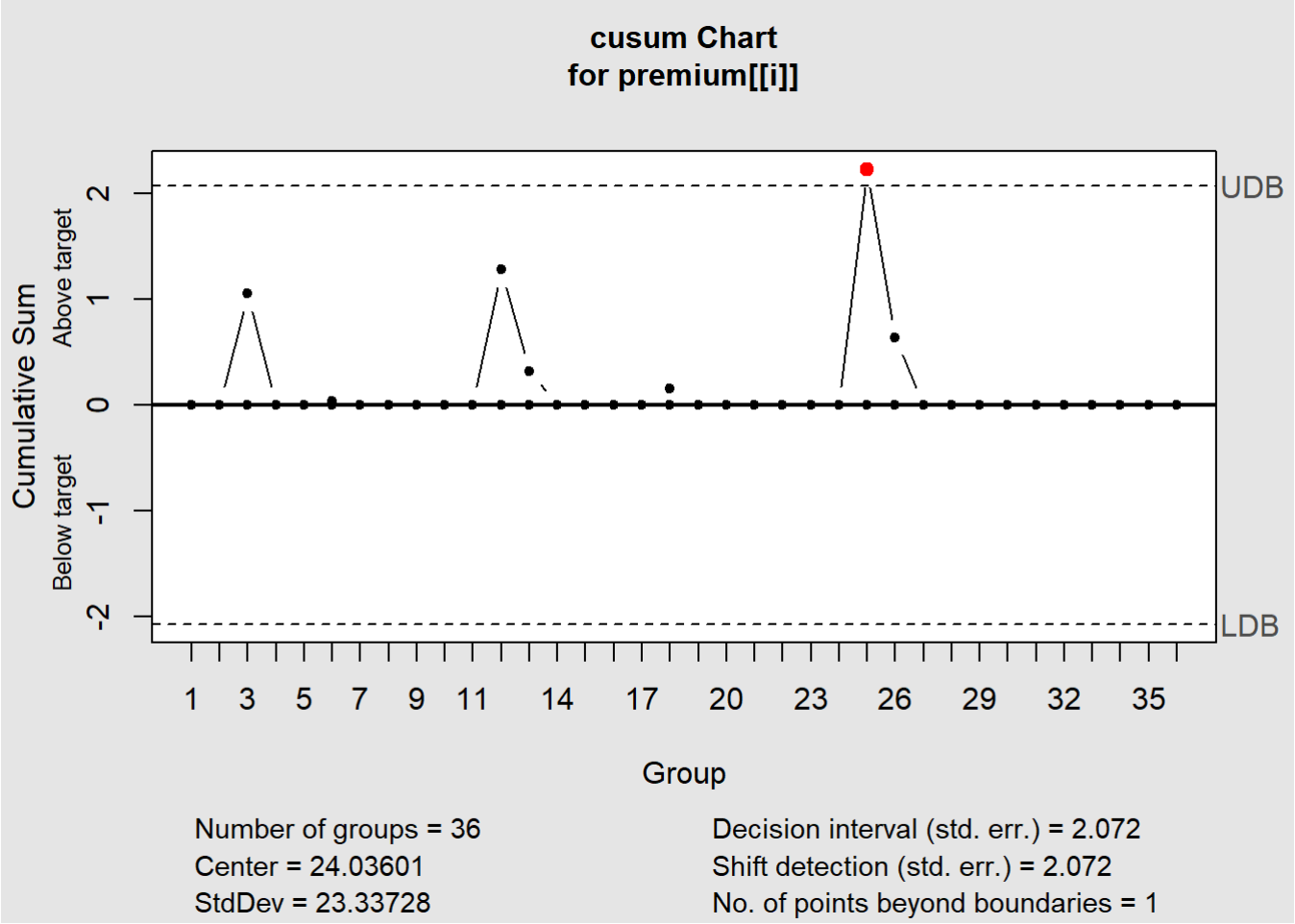

cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 41.47769
StdDev = 41.37637
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 0

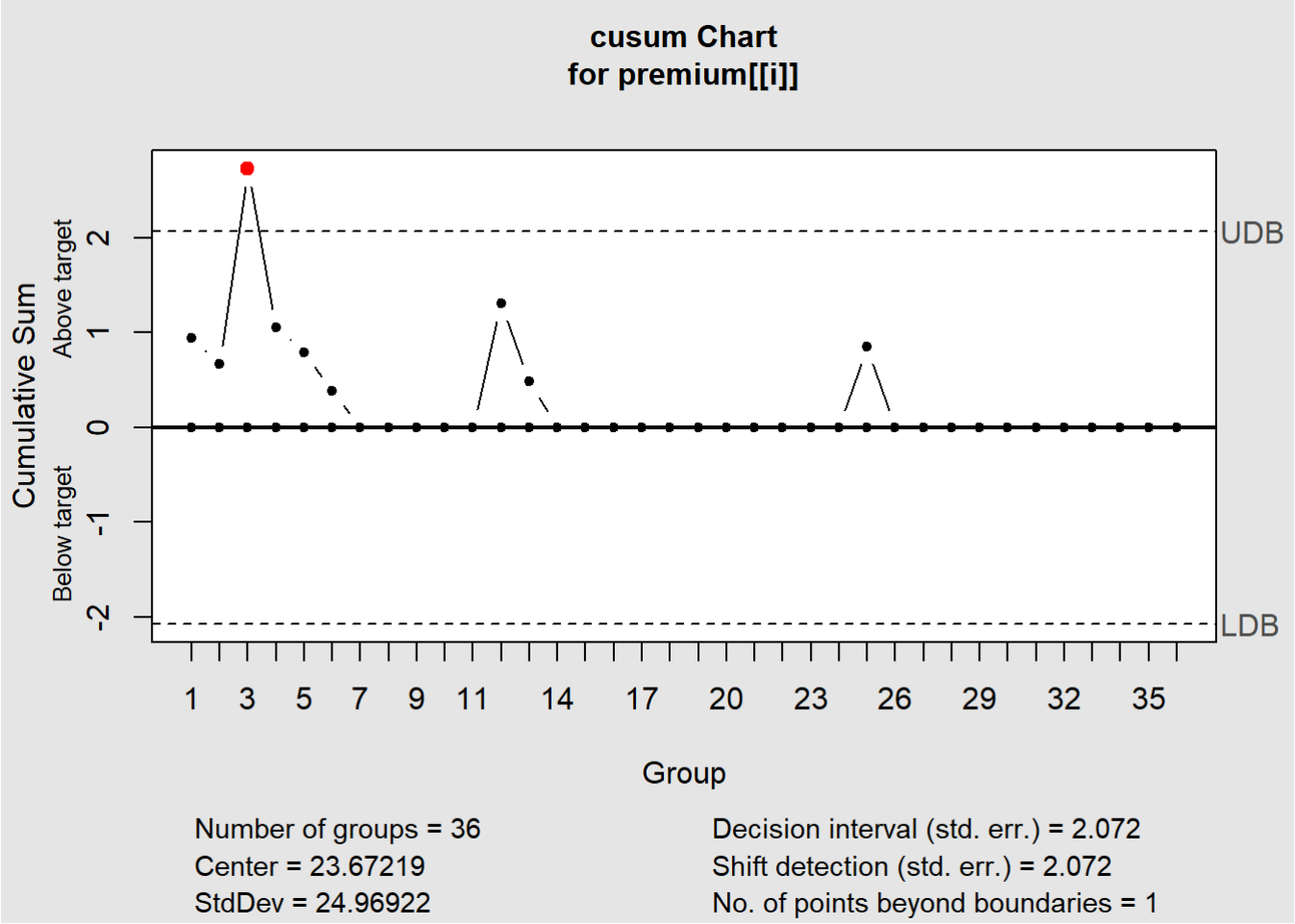

cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 17.60702
StdDev = 15.89647
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 0

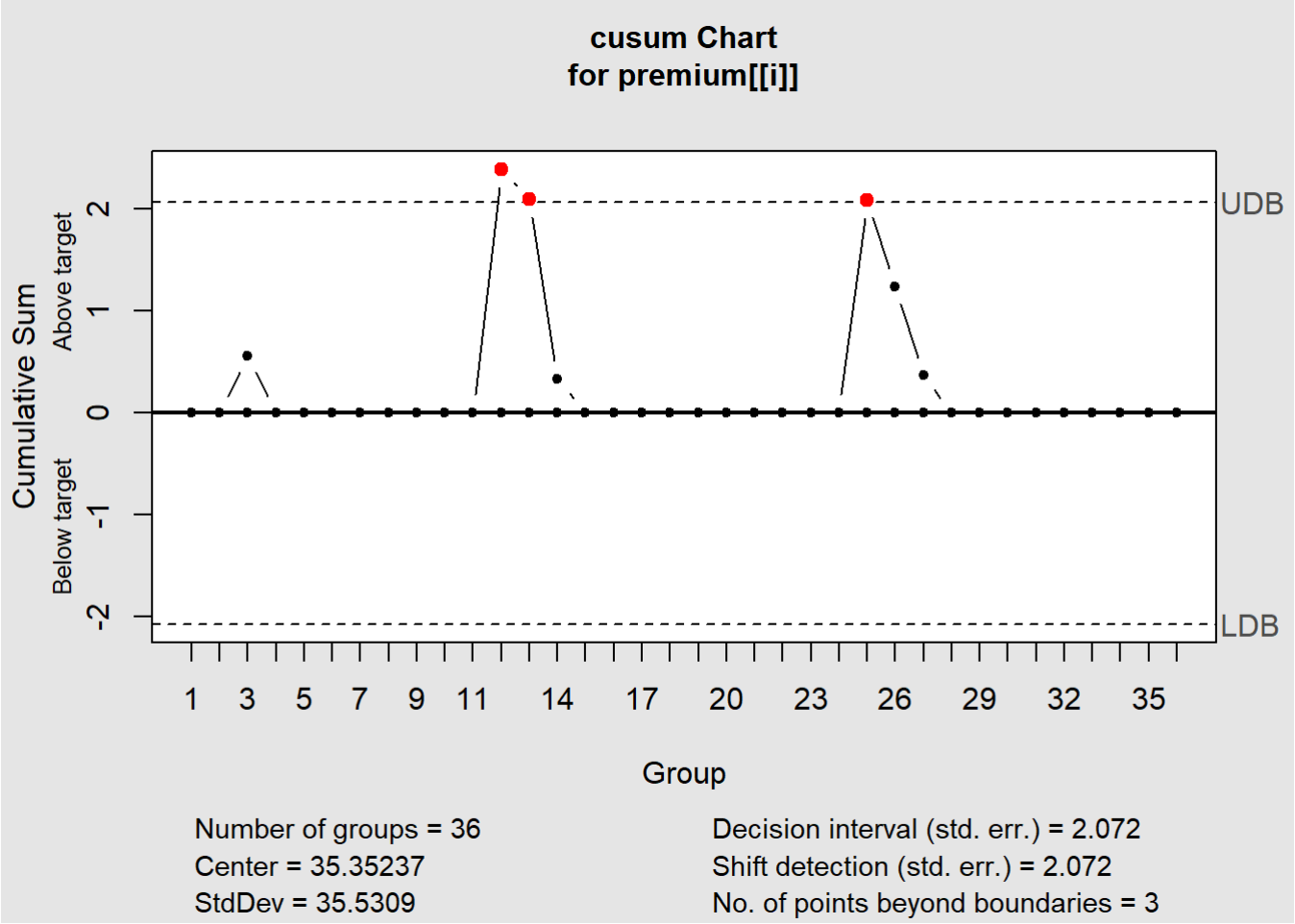

cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
4
3
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 295.6516
StdDev = 377.5161
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 5

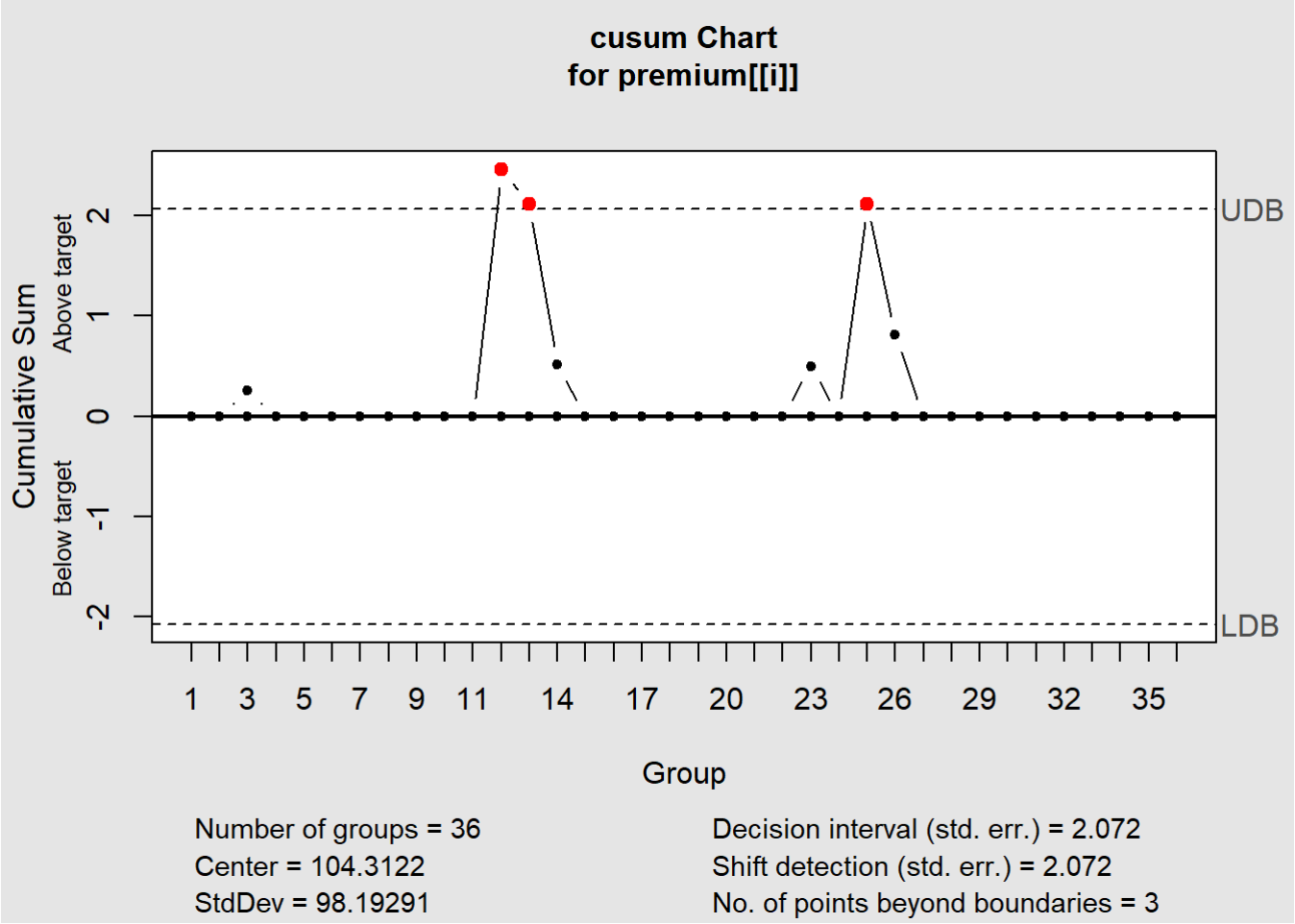

cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 166.506
StdDev = 193.6719
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 2

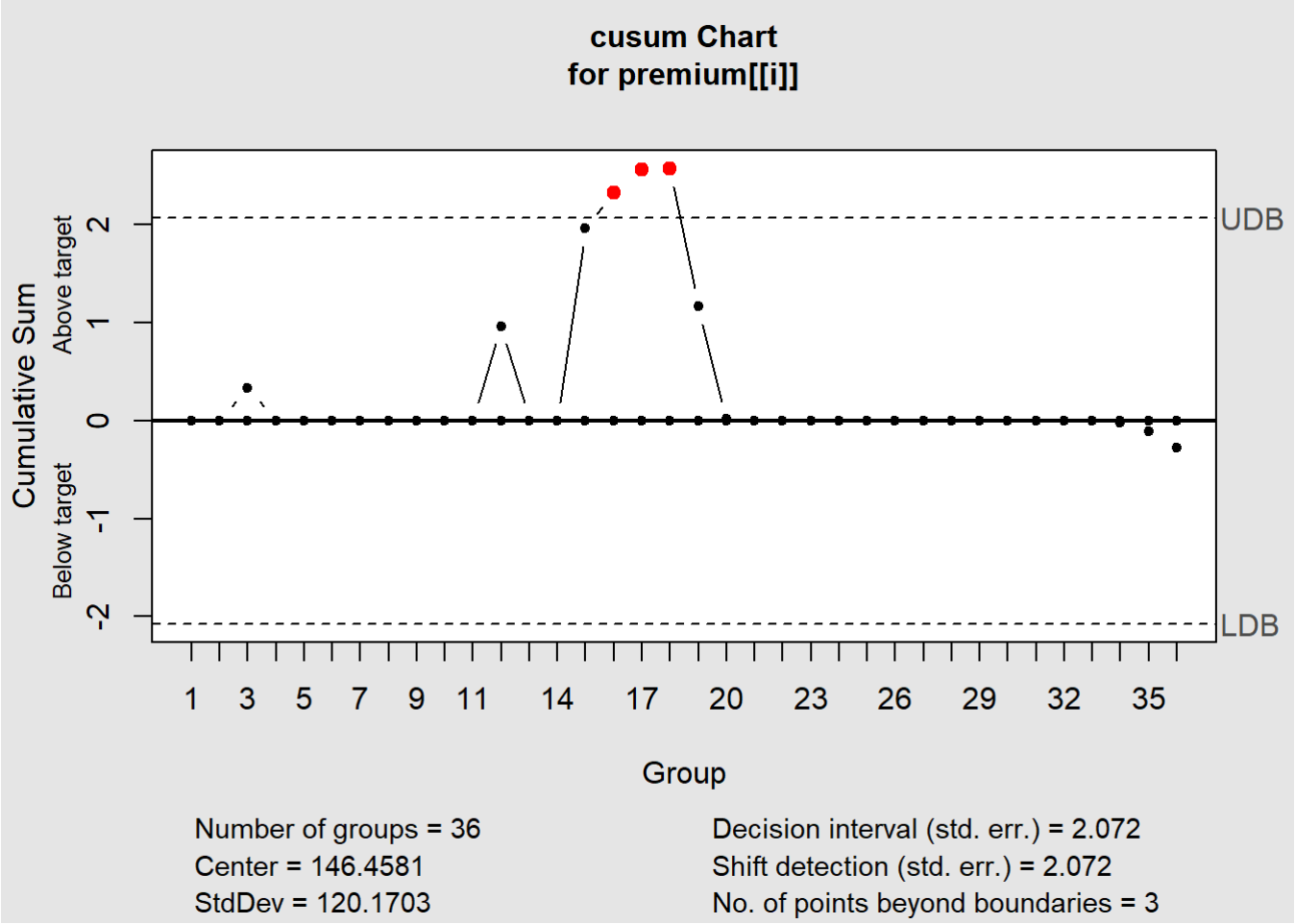

cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 54.67253
StdDev = 51.25788
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 1

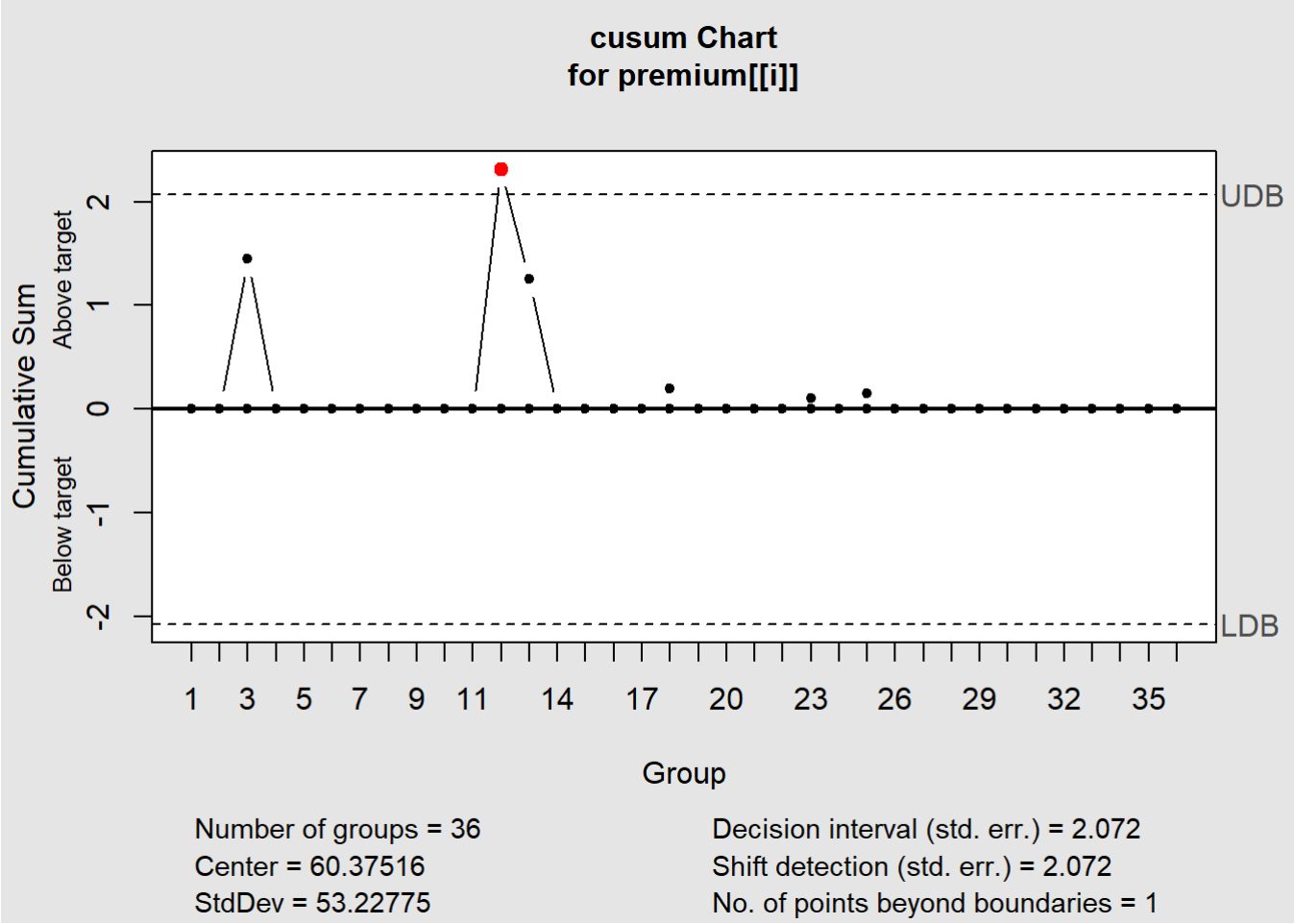

cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 24.03601
StdDev = 23.33728
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 1

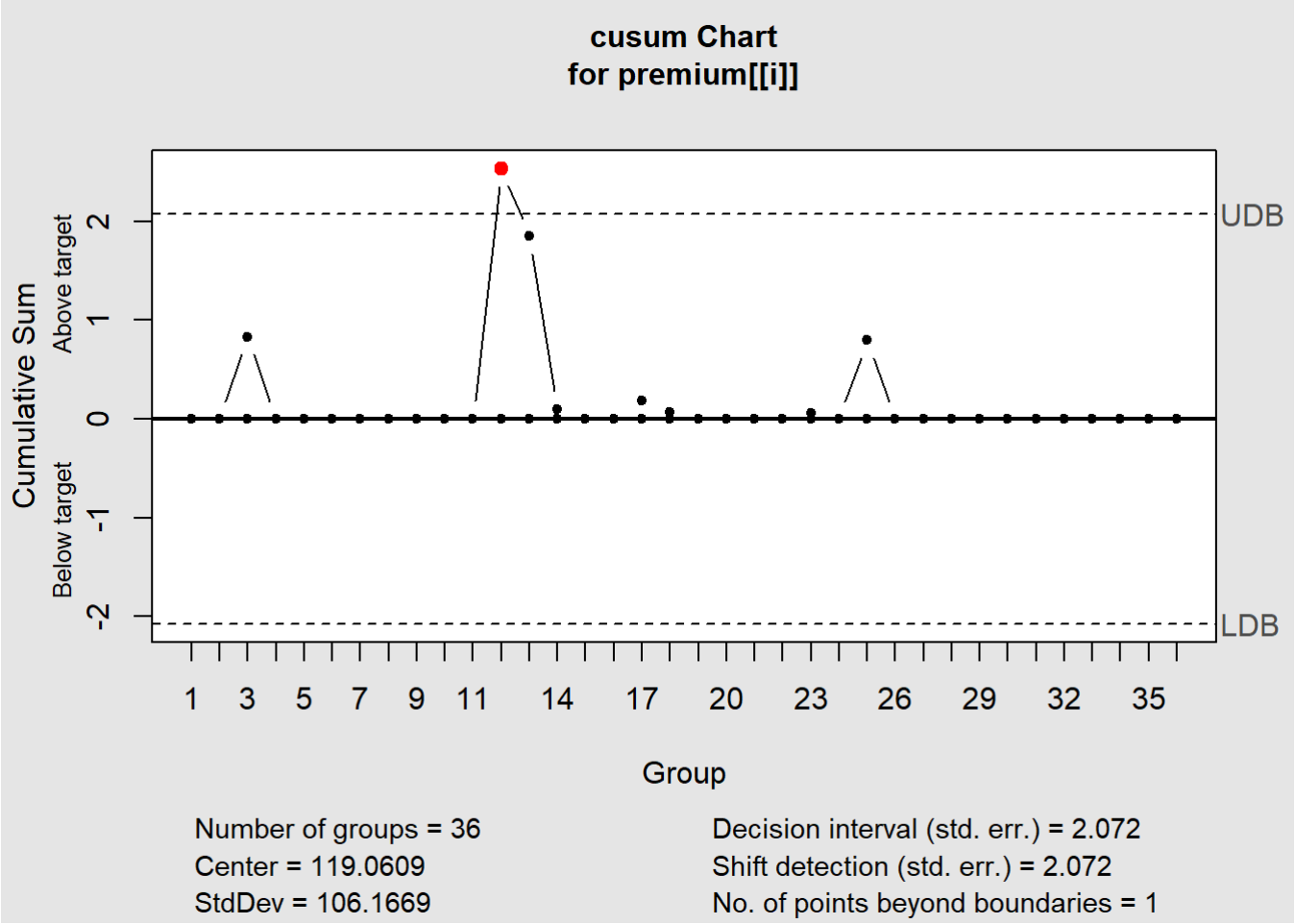

cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 23.67219
StdDev = 24.96922
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 1


cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 35.35237
StdDev = 35.5309
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 3


cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 104.3122
StdDev = 98.19291
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 3


cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 146.4581
StdDev = 120.1703
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 3


cusum Chart
for premium[[i]]
Cumulative Sum
Above target
Below target
2
1
0
-1
-2
UDB
LDB
1
3
5
7
9
11
14
17
20
23
26
29
32
35
Group
Number of groups = 36
Center = 60.37516
StdDev = 53.22775
Decision interval (std. err.) = 2.072
Shift detection (std. err.) = 2.072
No. of points beyond boundaries = 1




```
#####
### end
##### 此次采取r语言使用统计学方法（CUSUM）检测单个值或者群组值的偏移是否目标既定值。目标既定值可以理解为样本的均值，可以为一个数值，
#或者可以为一个群体的测量值。数据样本来自于研究小组发的中支数据，此处仅以广东中支为例。
#具体步骤涉及清理引入相关数据包，读取数据，查验数据结构，建立数据矩阵和矢量，转换成时间序列数据，这里不一一赘述。
#结论以图表方式呈现。以图一为例子。
#cusum模型将整个数据源分成了36个组。36个组是基于36个观察值。其中这组观察值的中心是119.0609.此处我将decision interval和shift det
ection设置
#成同一值为了方便计算。现实可以根据对业务的理解自主设置值从而得到更精确的模型输出。这幅图最重要的部分就是高于udb的值。其有可能是异常
值。
#需要去查看对应该值的时间节点发生了什么导致保费异常。以图一为例，23和25是两个高于上层边界值，cusum模型中把其当作异常值。对应分组后的序
列，可以看到
#23和25分别是广州中支2020.11保费443，和广州中支2021.1的保费321。

####写到最后
```

```
#CUSUM是检测偏移自目标值最好的方法之一。在检测较大偏移时cusum本身
#算法检测较慢（可能需要额外数据量）。该模型仍有改进的空间。同时关于di（decision.interval）和ss（se.shift)的取值较大程度上决定了模型预测的结果
#结果预测的准确程度会随着对业务理解而提升。
```