Initial Data Summaries

Lee Panter

Description

This script will produce numerical and graphical summaries of relevant models considered in each of the model developed as described in the ReadMe.

Begin Script

Exploratory Data Analysis

Following:

- Quantitative summary tables
- Histogram plots of Predictor and Outcome
- Scatter Plot of Predictor v Outcome
- Numerical five number summaries of Predictor and Outcome

Quantitative Data Summaries

```
# variables to include
#  - Subject count
#  - Observation count
#  - Min, Max, Mean, Median, Mode observations
#  -

mdataFilterQ=mdataFilter

seqFilterQ=seqFilter
seqFilterQ=t(seqFilterQ)
seqFilterQ=data.frame(seqFilterQ)
seqFilterQ$subject.no=subject.no

mdataCountCols=
    mdataFilterQ %>%
      select(subject.no, well, Perc.Mt) %>%
      group_by(subject.no) %>%
      count(subject.no)

colnames(mdataCountCols)=c("subjectNO", "quantity_byPerctMT")

mdataSummaryCols=
  mdataFilterQ %>%
    select(subject.no, well, Perc.Mt) %>%
    group_by(subject.no) %>%
```

```
    summarise(group_minPerct.Mt=min(Perc.Mt),
              group_maxPerct.Mt=max(Perc.Mt),
              group_avgPerct.Mt=mean(Perc.Mt),
              group_medPerct.Mt=median(Perc.Mt))

colnames(mdataSummaryCols)=c("subjectNO",
                             "group_minPerct.Mt",
                             "group_maxPerct.Mt",
                             "group_avgPerct.Mt",
                             "group_medPerct.Mt")


#seqCountCols=data.frame(subject.no, cd19, mala)
seqCountCols=
    seqFilterQ %>%
      select(subject.no, CD19, MALAT1) %>%
      group_by(subject.no) %>%
      count(subject.no)

colnames(seqCountCols)=c("subject.no", "quantity_bySeq")
seqCountCols
```

subject.no	quantity_bySeq
5	58
6	86
7	32
9	31
10	21
11	107
13	107
14	100
15	25
17	122
19	127
20	75
22	87
24	79
26	53

```
seqCD19.SummaryCols=
  seqFilterQ %>%
  select(subject.no, CD19, MALAT1) %>%
  group_by(subject.no) %>%
  summarise(group_minCD19 =    min(CD19),
            group_maxCD19 =    max(CD19),
            group_avgCD19 =   mean(CD19),
            group_medCD19 = median(CD19))
seqCD19.SummaryCols
```

subject.no	group_minCD19	group_maxCD19	group_avgCD19	group_medCD19
5	0	678	36.67241	0.0

subject.no	group_minCD19	group_maxCD19	group_avgCD19	group_medCD19
6	0	299	36.68605	7.5
7	0	10	2.12500	1.0
9	0	1052	89.41935	3.0
10	0	158	37.57143	2.0
11	0	339	28.31776	1.0
13	0	629	56.08411	18.0
14	0	251	42.26000	19.0
15	0	148	26.60000	0.0
17	0	982	112.37705	16.0
19	0	665	59.33858	5.0
20	0	287	40.12000	23.0
22	0	380	43.44828	1.0
24	0	282	55.01266	27.0
26	0	1624	268.41509	110.0

```
seqMALA.SummaryCols=
  seqFilterQ %>%
  select(subject.no, CD19, MALAT1) %>%
  group_by(subject.no) %>%
  summarise(group_minMALA =    min(MALAT1),
            group_maxMALA =    max(MALAT1),
            group_avgMALA =   mean(MALAT1),
            group_medMALA = median(MALAT1))
seqMALA.SummaryCols
```

subject.no	group_minMALA	group_maxMALA	group_avgMALA	group_medMALA
5	67	40812	10206.362	9195.0
6	757	30774	11568.279	11689.0
7	441	17916	6868.000	4039.5
9	311	18239	5703.935	5983.0
10	1875	17160	6638.571	6190.0
11	349	34082	9716.028	8826.0
13	99	25572	5867.944	4895.0
14	355	15740	6154.150	5720.5
15	157	11923	3839.080	3467.0
17	337	8342	2960.254	2692.0
19	227	91961	13959.984	10125.0
20	379	21736	7301.413	6417.0
22	161	28429	6881.747	5068.0
24	240	42792	6448.823	5955.0
26	1114	32426	8463.170	6426.0

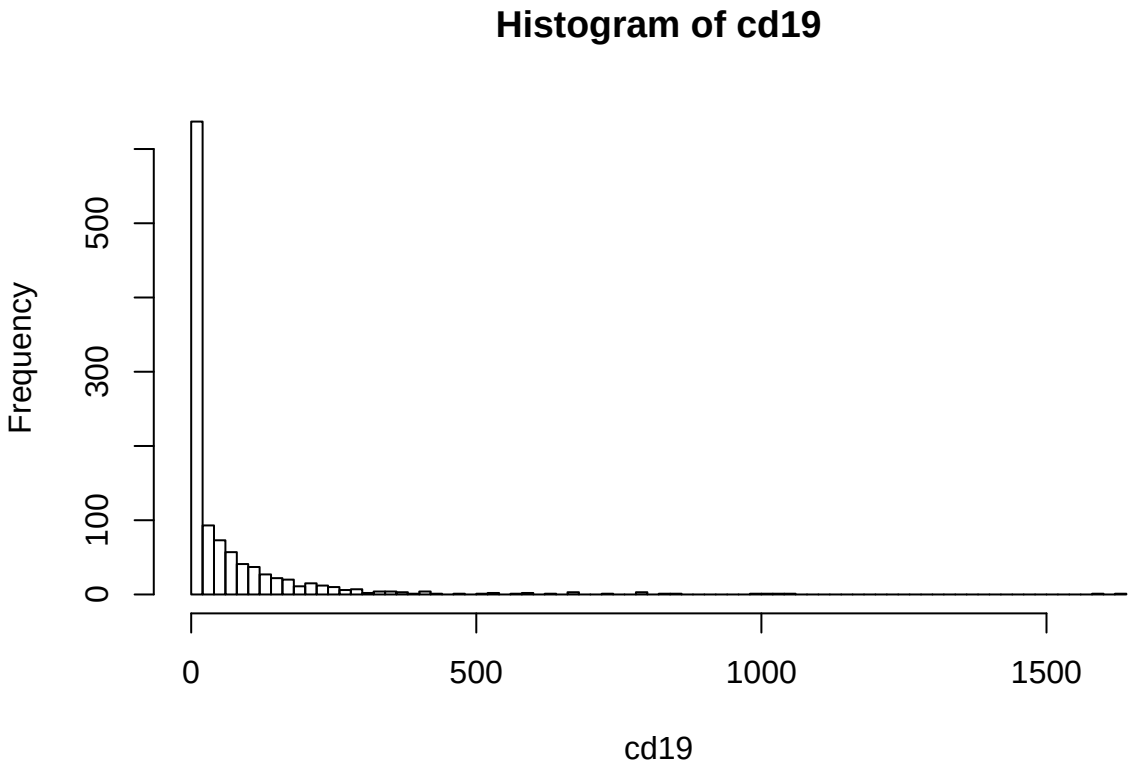
Predictor Summaries

```
# FIVE NUMBER SUMMARY
summary(dat$cd19)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00    0.00    7.00   62.56   70.00 1624.00
```

```
# HISTOGRAM
# p1=ggplot(dat, aes(x=cd19,fill=subject.no, color=subject.no))+
#   geom_histogram(alpha=0.5, position = "dodge", binwidth = 30)+
#   theme(legend.position = "right")
# p1

hist(cd19, xlab = "cd19", breaks= 100 )
```



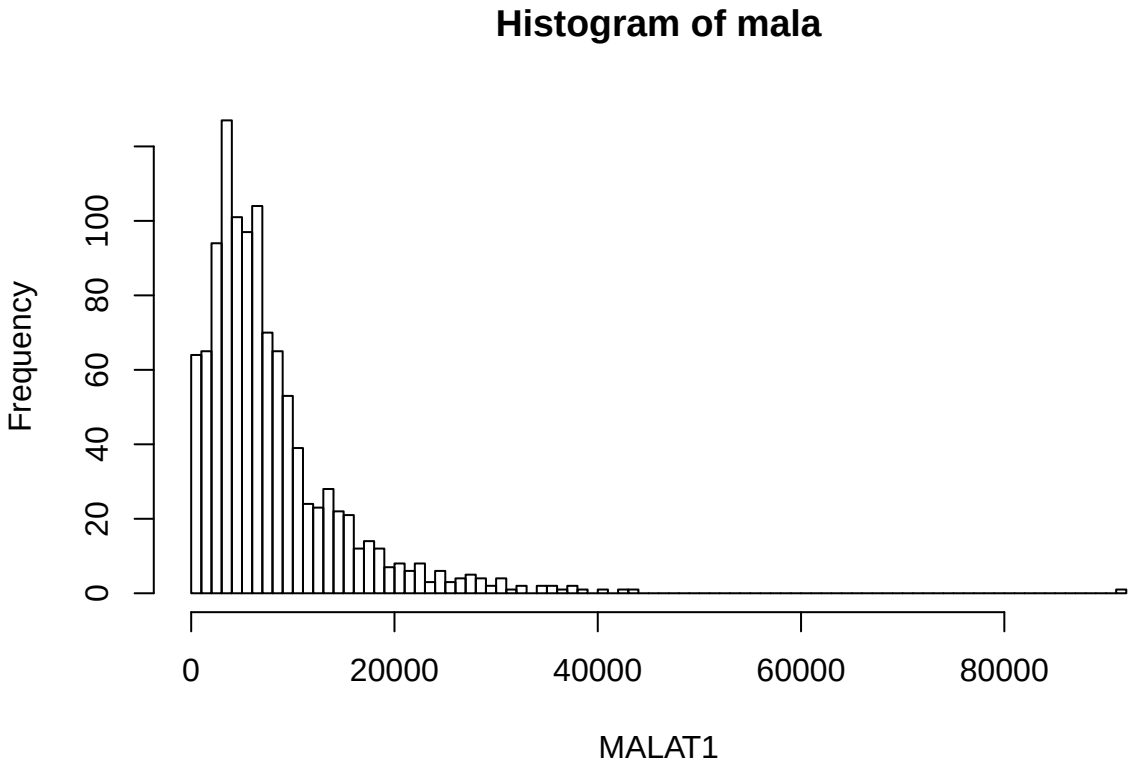
Outcome Summaries

```
# FIVE NUMBER SUMMARY
summary(dat$mala)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      67    3349    6098    7874    9774   91961
```

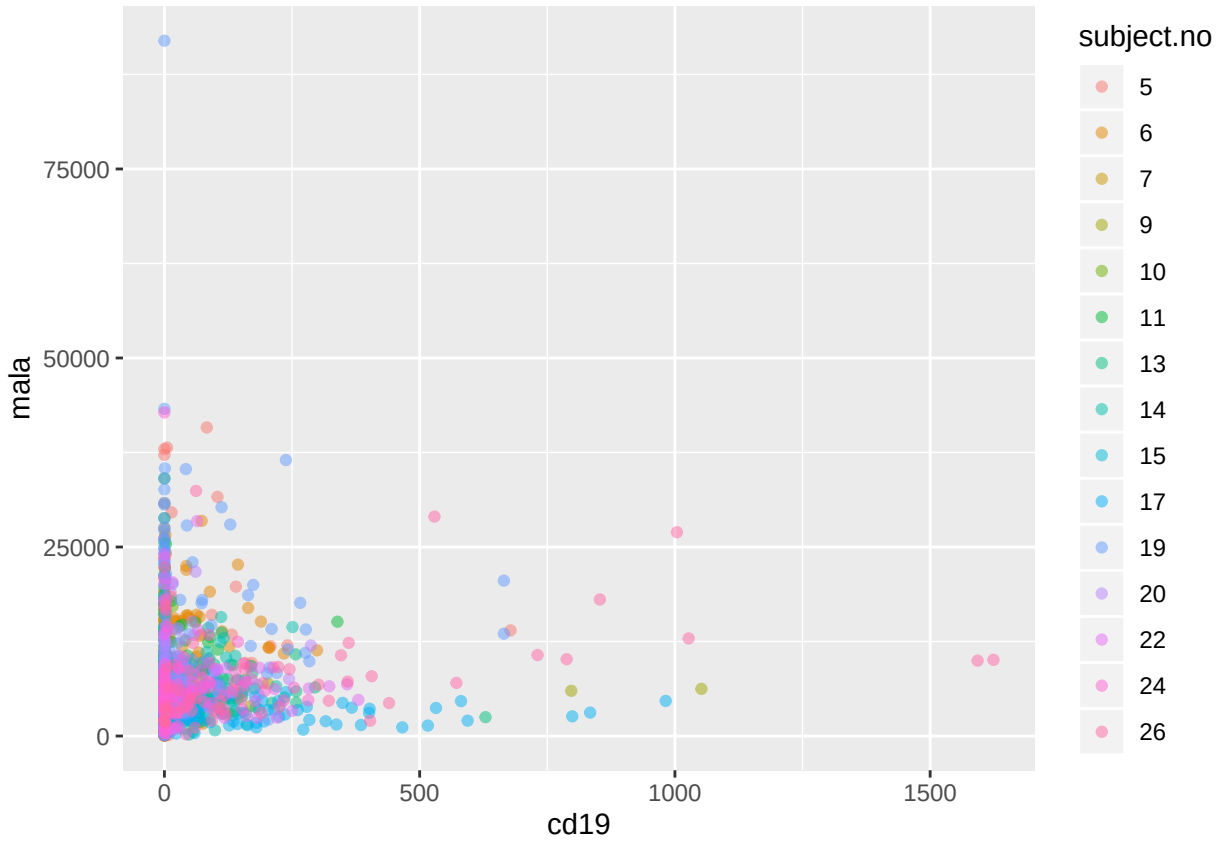
```
# HISTOGRAM
# p2=ggplot(dat, aes(x=mala,fill=subject.no, color=subject.no))+
#   geom_histogram(alpha=0.5, position = "dodge", binwidth = 30)+
#   theme(legend.position = "right")
# p2

hist(mala, xlab = "MALAT1", breaks = 100)
```



Scatter Plot Outcome ~ Predictor

```
p3=ggplot(dat, aes(x=cd19, y=mala, color=subject.no))+
  geom_point(alpha=0.5)
p3
```



Transformed Variables (log transformations)

We will apply the transformation $Y = log(x + 1)$ to the outcome and response variables to create new-transformed variables.

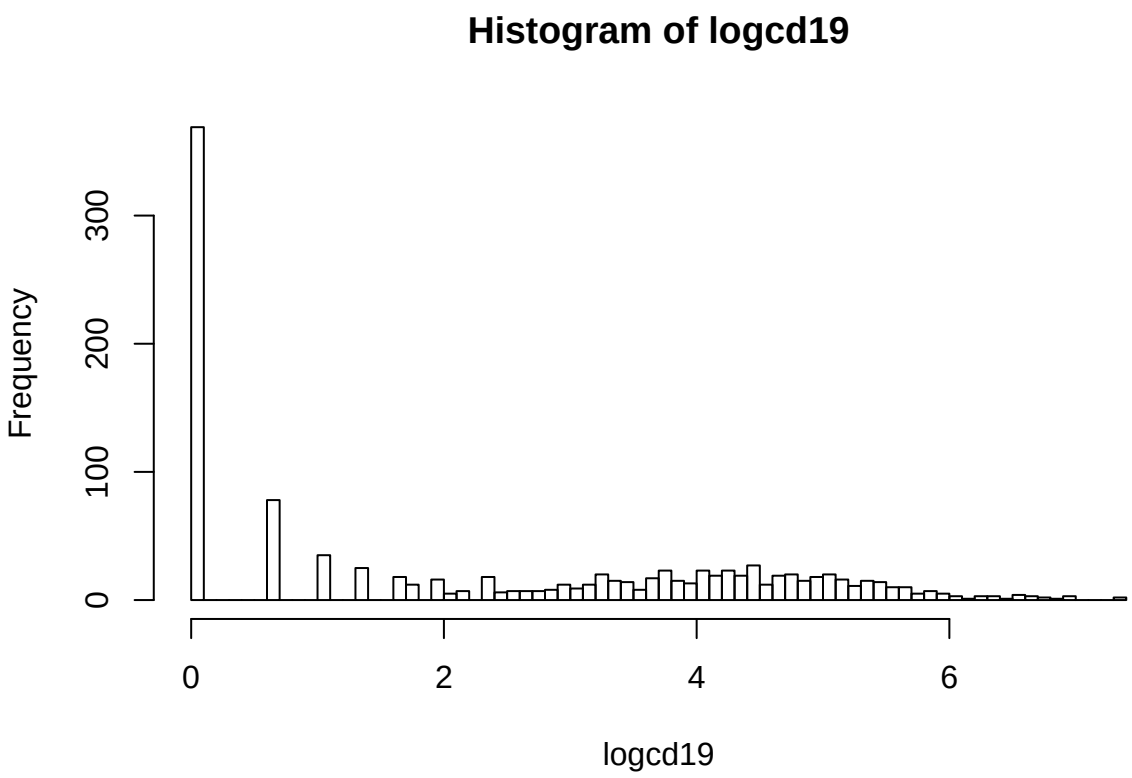
```
dat$logcd19=log(cd19+1, base = exp(1))
dat$logmala=log(mala+1, base = exp(1))
logcd19=dat$logcd19
logmala=dat$logmala

# FIVE NUMBER SUMMARY
summary(dat$logcd19)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   2.079   2.318   4.263   7.393
```

```
# HISTOGRAM
# p1=ggplot(dat, aes(x=logcd19,fill=subject.no, color=subject.no))+
#   geom_histogram(alpha=0.5, position = "dodge", binwidth = 1)+
#   theme(legend.position = "right")
# p1

hist(logcd19, xlab = "logcd19", breaks = 100)
```



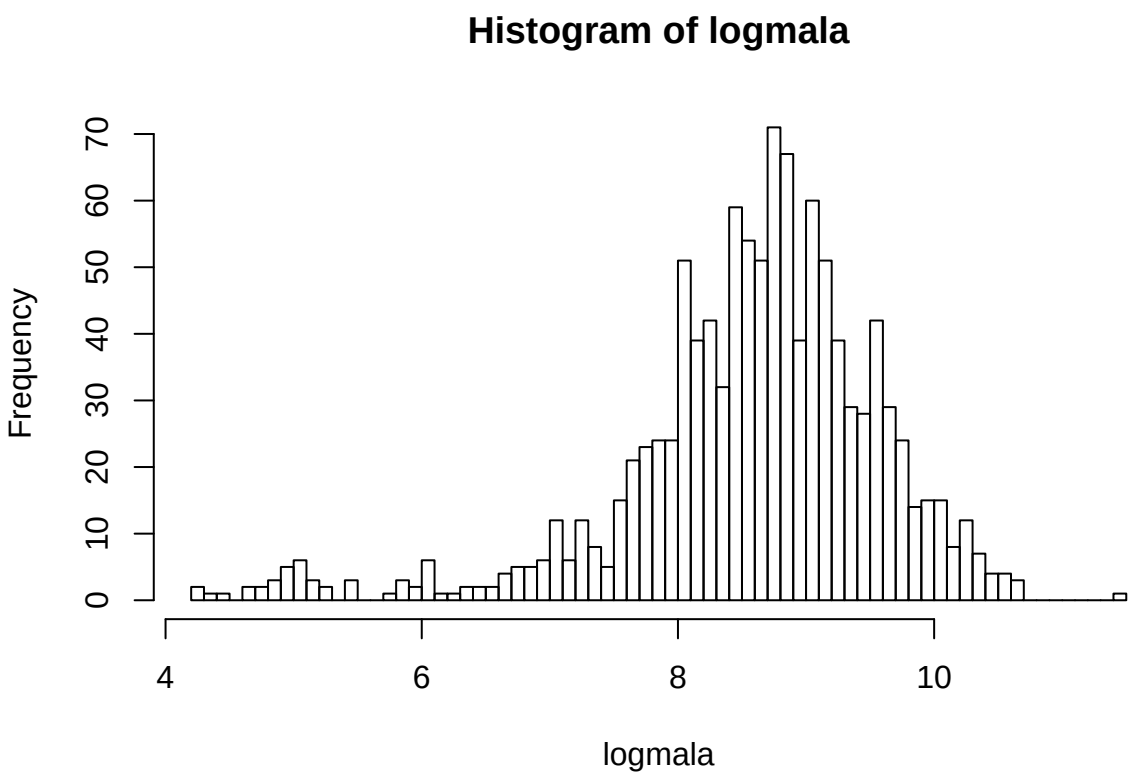

```
# FIVE NUMBER SUMMARY
summary(dat$logmala)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.220   8.117   8.716   8.576   9.188  11.429
```

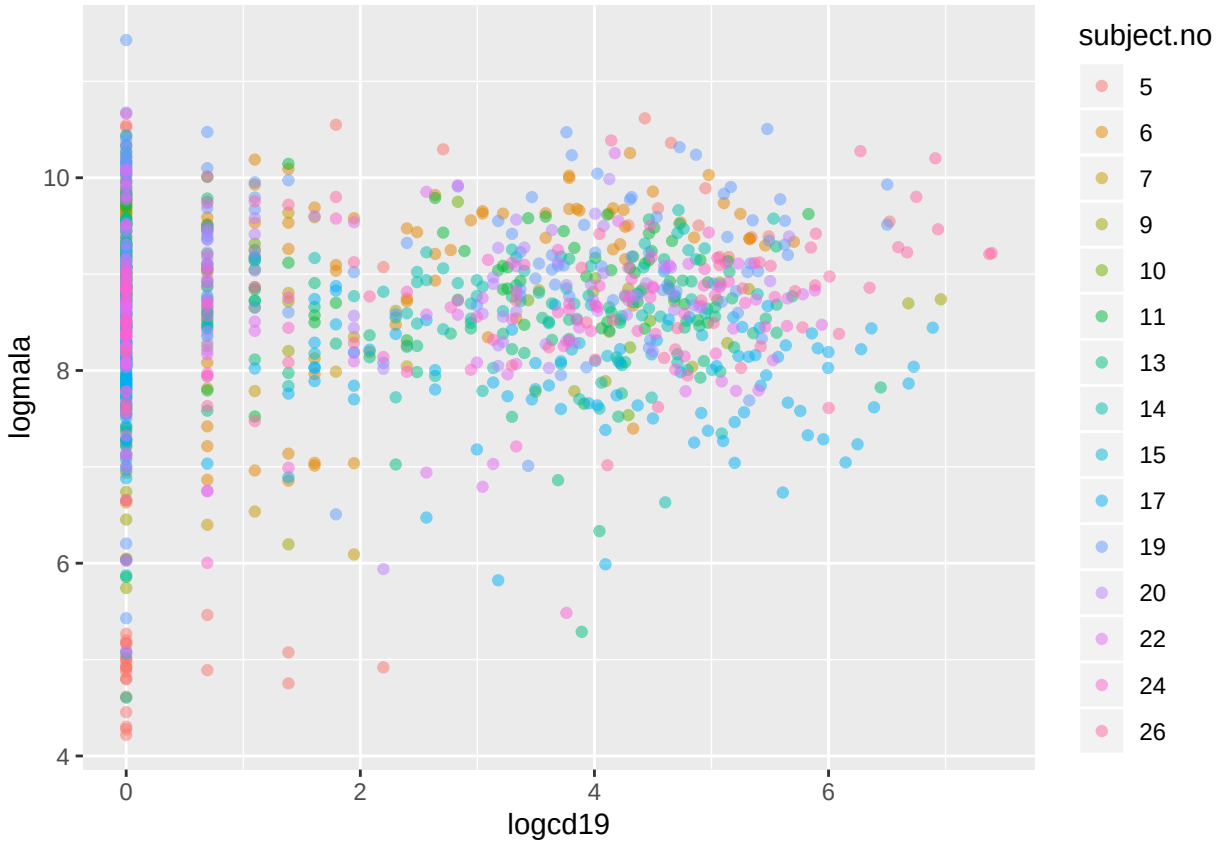
```
# HISTOGRAM
# p1=ggplot(dat, aes(x=logmala,fill=subject.no, color=subject.no))+
#   geom_histogram(alpha=0.5, position = "dodge", binwidth = 1)+
```

```
#   theme(legend.position = "right")
# p1

hist(logmala, xlab = "logmala", breaks = 100)
```




```
p3=ggplot(dat, aes(x=logcd19, y=logmala, color=subject.no))+
  geom_point(alpha=0.5)
p3
```


logmala
logcd19
10
8
6
4
0
2
4
6
subject.no
5
6
7
9
10
11
13
14
15
17
19
20
22
24
26

End Script