Final project

Miza Syafiqah Mohamad Shanudin

12/23/2020

```
rm(list = ls())
library(ggplot2)
library(ggmosaic)
library(dplyr)
library(lmerTest)
library(lme4)
library(lmeresampler)
library(chi)
library(VGAM)
```

```
library(jsonlite)
df <- fromJSON("https://query.data.world/s/2o4phxd36r6xkjug2gcnnrgzbr4bla")
```

1. The data is found at data.world (https://data.world/durhamnc/north-carolina-school-performance-data/workspace/file?filename=north-carolina-school-performance-data_1.csv) and it is about North Carolina school performance. There are 8 variables with 2425 observation chosen (each observation is a school). For this project I will ignore the type of school:

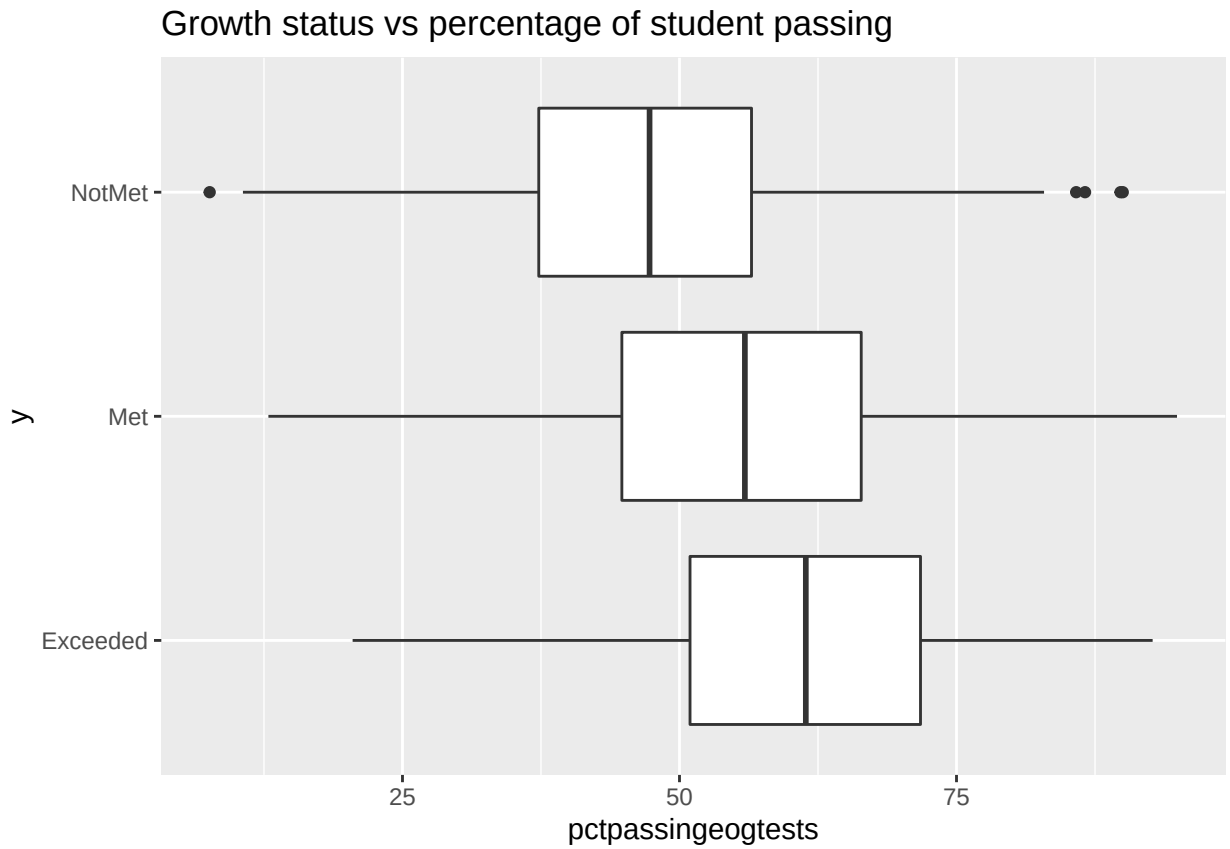
district: the district of the school
school: the name of the school
growthstatus: the academic growth status of the school (NotMet, Met,Exceeded)
pctpassingseogtest: percentage of students passing EOG Math & English Test
pctfreereduced: percentage of students who receive free or reduced lunch
hasstringprogram: whether the school has a string program (0,1)
schoolscore: school performance grade score
grade: school grade (A-F)

The questions of interest here are does school score, percentage of students passing eog test and percentage of students of get free or reduced lunch significant in determine growth status of the schools in North Carolina and does district and growth status affect school score?

2. EDA
   GLM:

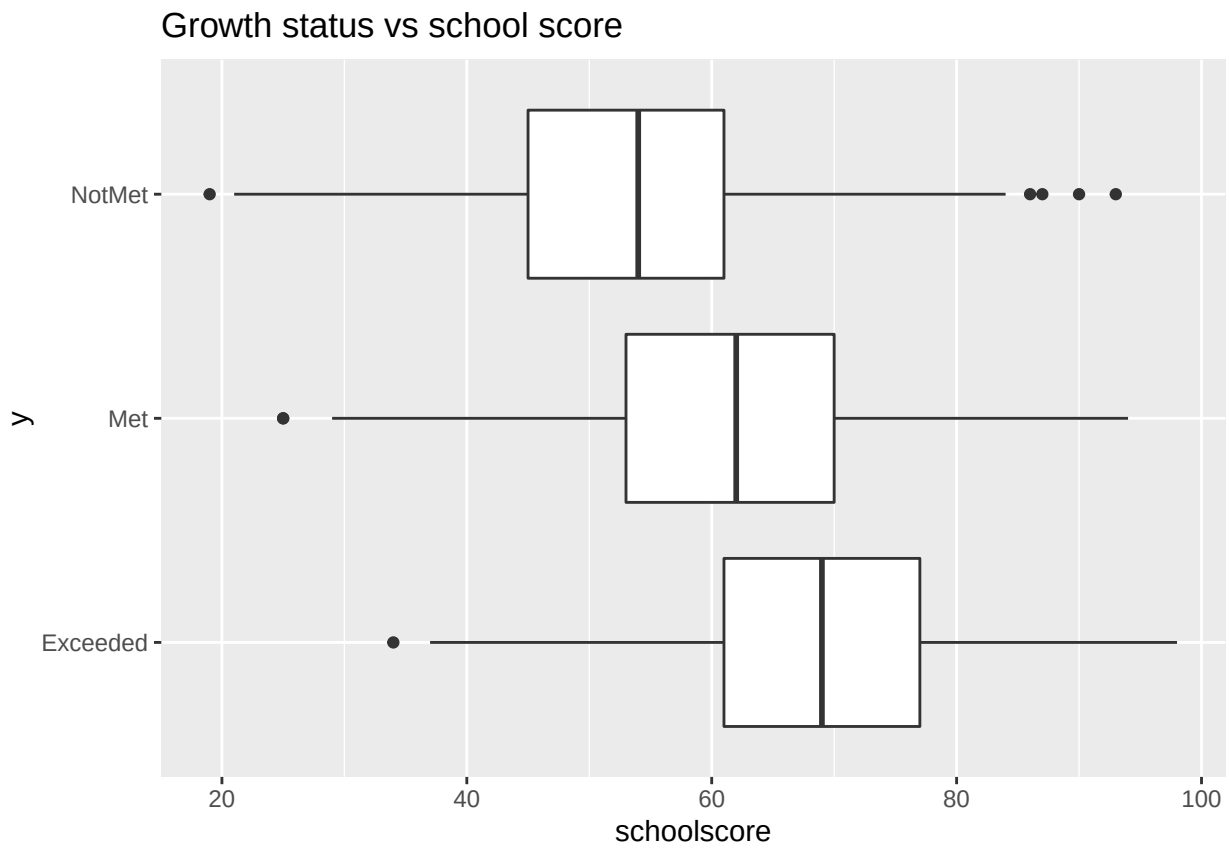
```
df <- df%>% filter(growthstatus !="N/A")%>%
  mutate(y=growthstatus)

ggplot(data=df, aes(x=pctpassingeogtests,y=y, na.rm = TRUE))+
  geom_boxplot(na.rm = TRUE)+ ggtitle("Growth status vs percentage of student passing")
```



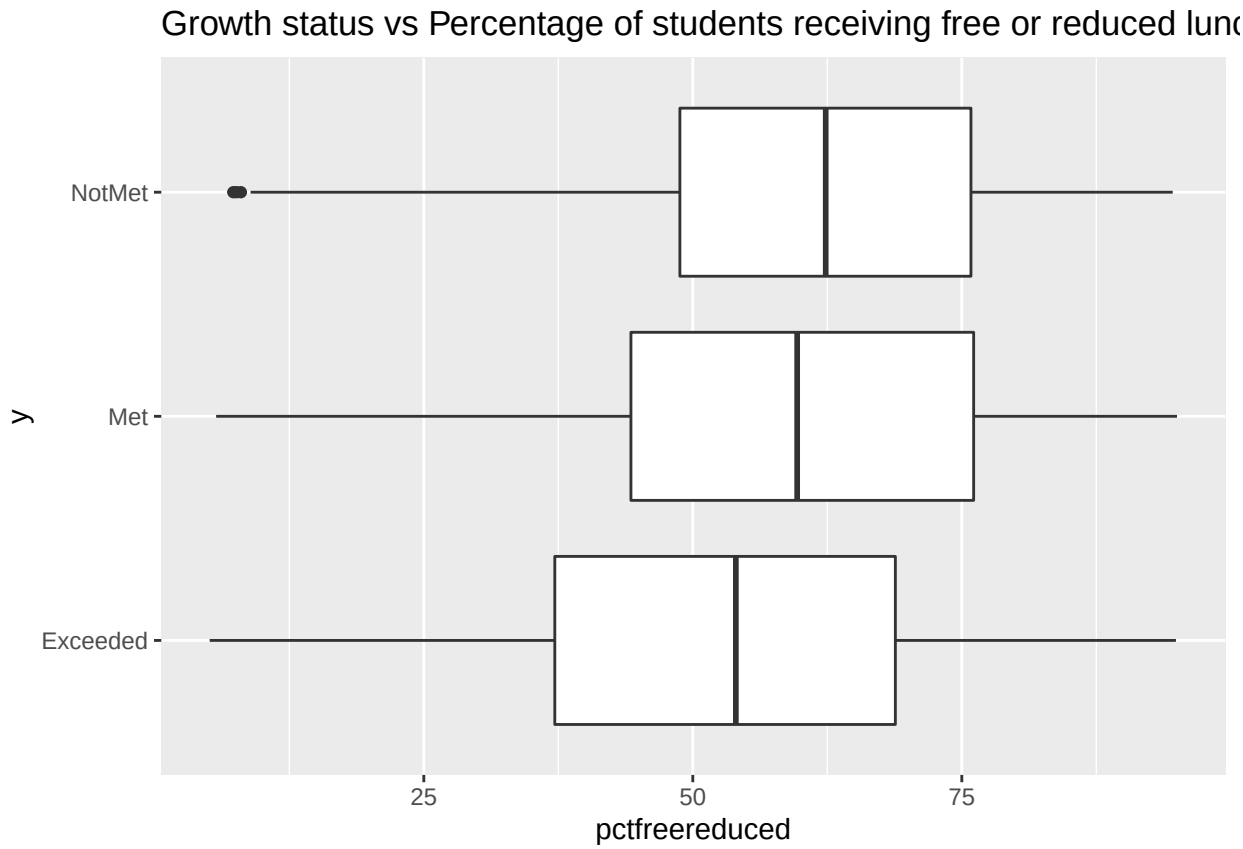
From the plot, it seems like percentage of passing EOG test does not have a significant effect of growth status of school. But, some school at lower percentage do not meet the growth status. But overall, the mean percentage are different for the 3 growth status

```
ggplot(data=df, aes(x=schoolscore,y=y, na.rm = TRUE))+
  geom_boxplot(na.rm = TRUE)+ ggtitle("Growth status vs school score")
```



From the plot, schools that have higher score tend to exceeded the growth status although there are some that don't met. But it does seem like school with lower school score have not yet met the growth status

```
ggplot(data=df, aes(x=pctfreereduced,y=y, na.rm = TRUE))+
  geom_boxplot(na.rm = TRUE)+ ggtitle("Growth status vs Percentage of students receiving free or reduce
```



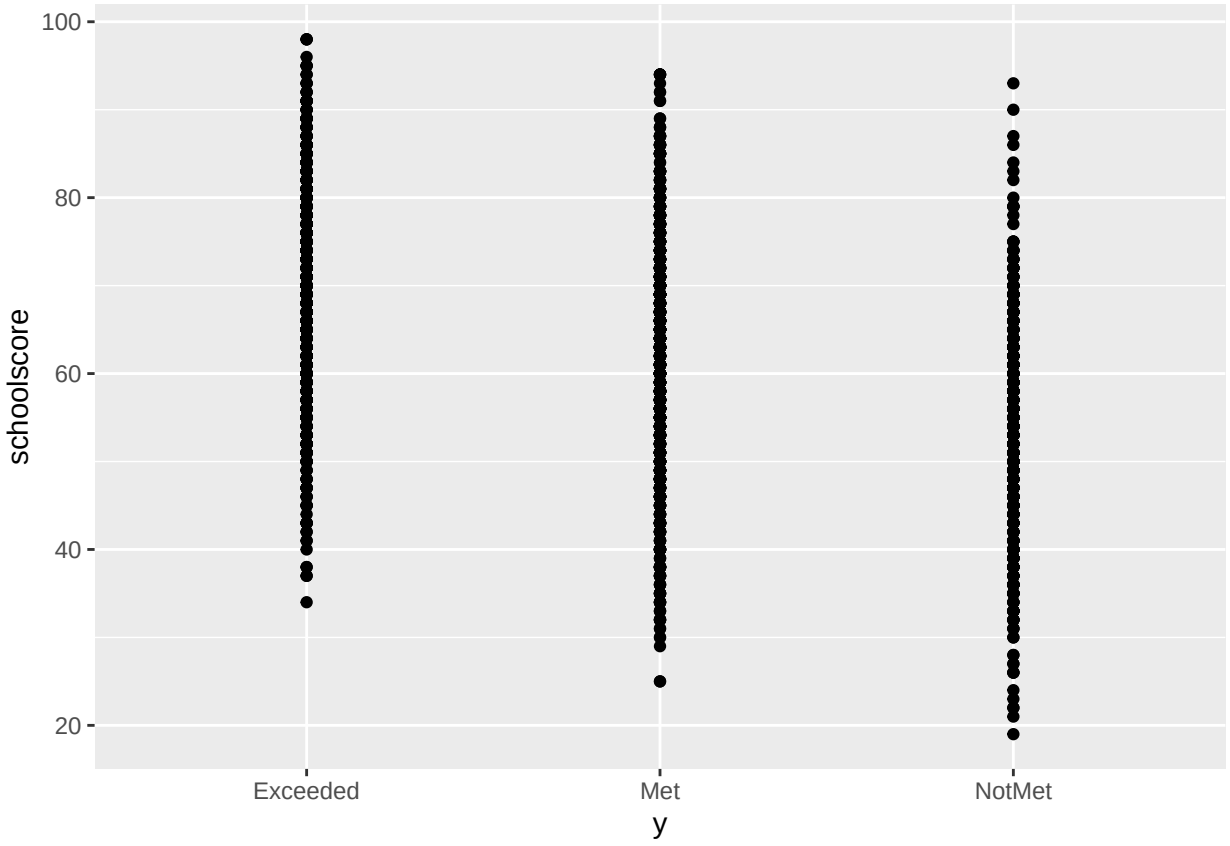
From the plot, although the mean are different, overall it does not seem like percentage of student receiving free or reduced lunch has any effect on growth status.

Multilevel: EDA for multilevel data has to take into account the differently-sized experimental units; in this case, we have districts (level 2 experiment unit) and school (level 1 experimental unit).

Level 1: Since level 1 is school, I assess individual growth status. The plot b

```
ggplot(data=df, aes(x=y,y=schoolscore)) + geom_point(binaxis="y")
```

```
## Warning: Ignoring unknown parameters: binaxis
```



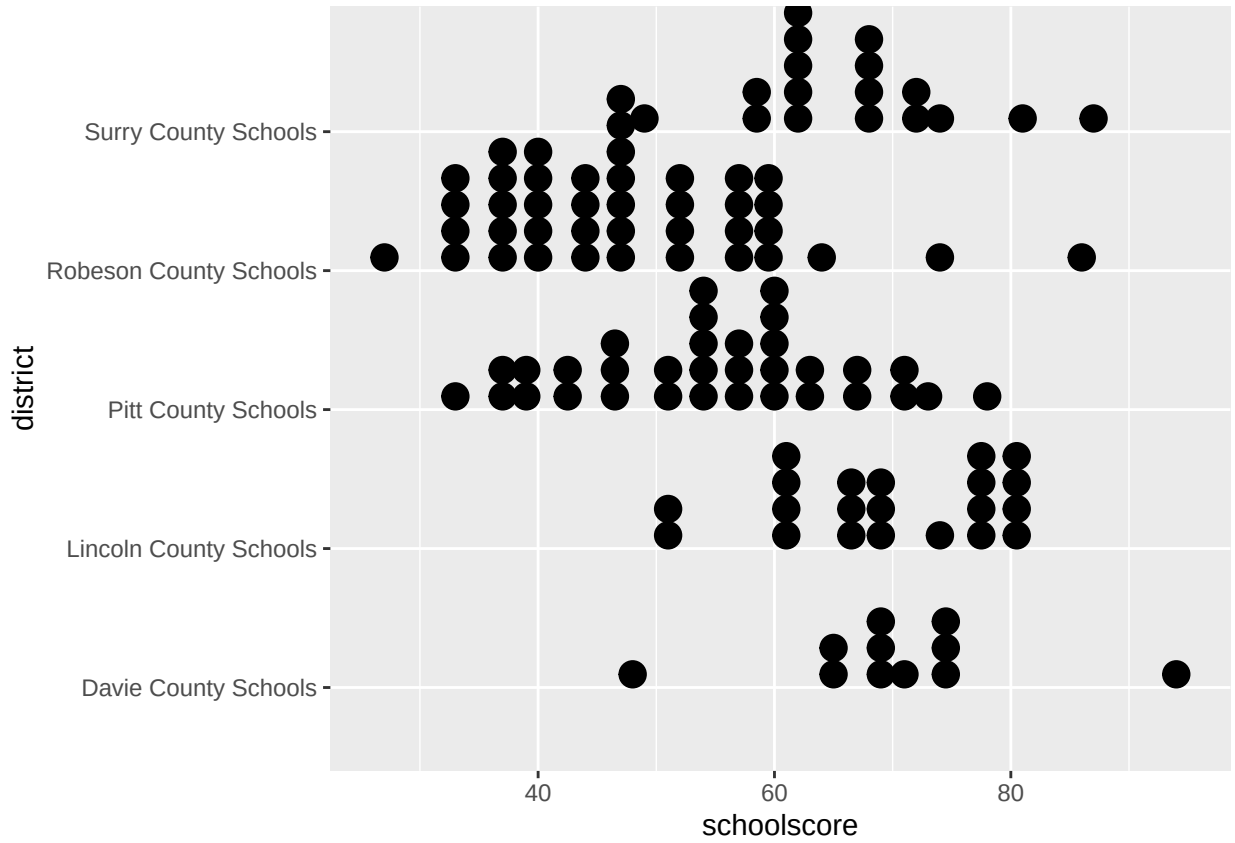
Since level 1 is school, I assess individual growth status. The plot shows that schools that has not yet met the growth status has lower school score and schools that has exceeded the growth status has higher school score.

Level 2:

```
id1 <-df%>% filter(district == "Surry County Schools")
id2 <-df%>% filter(district == "Robeson County Schools")
id3 <-df%>% filter(district == "Davie County Schools")
id4 <-df%>% filter(district == "Pitt County Schools")
id5 <-df%>% filter(district == "Lincoln County Schools")

df2<- bind_rows(id1,id2,id3,id4,id5)
ggplot(data=df2, aes(x=district, y=schoolscore)) + geom_dotplot(binaxis = "y") + coord_flip()
```

```
## `stat_bindot()` using `bins = 30`. Pick better value with `binwidth`.
```



Level 2 is district and since there are 117 districts, I will look at 5 districts. From the plot, it can clearly be seen that every district has different school score. However, the districts have different number of school

3. Analysis and Conclusion

GLM: Since growth status of school has 3 categorical variables, multinomial model is chosen. the proportional-odds model is considered and it has 5 parameters (2 intercepts, 3 slopes. The non proportional-odds model has 8 parameter (2 intercepts, 6 slopes). The degree of freedom is 8-5=3 and the p-value that is calculated is extremely low so we can reject the proportional-odds model. Therefore, school score, percentage passing eog test and percentage of students who receive free or reduced lunch do contribute to school growth status because the variables all have low p-value. But the model chosen would be the non-proportional one.

```
fit<- vglm(y~schoolscore+pctpassingeogtests+pctfreereduced,cumulative(parallel=TRUE), data = df)
fit1<- vglm(y~schoolscore+pctpassingeogtests+pctfreereduced,cumulative(parallel=FALSE), data = df)

L0 <- logLik(fit)
L1 <- logLik(fit1)
1-pchisq(2*(L1-L0),3)
```

```
## [1] 5.541058e-07
```

The interaction between the explanatory variables are not considered in the model because the log-likehood is N/A

```
fit2<- vglm(y~(schoolscore+pctpassingeogtests+pctfreereduced)^2,cumulative(parallel=TRUE), data = df)
fit2
```

```
##
## Call:
## vglm(formula = y ~ (schoolscore + pctpassingeogtests + pctfreereduced)^2,
##     family = cumulative(parallel = TRUE), data = df)
##
##
## Coefficients:
##                       (Intercept):1                       (Intercept):2
##                       -2.923138e+01                       -2.667953e+01
##                         schoolscore                 pctpassingeogtests
##                        4.219643e-01                        1.080006e-01
##                     pctfreereduced    schoolscore:pctpassingeogtests
##                        1.860048e-01                       -2.090057e-03
##        schoolscore:pctfreereduced pctpassingeogtests:pctfreereduced
##                       -9.912245e-04                       -1.278731e-03
##
## Degrees of Freedom: 4614 Total; 4606 Residual
## Residual deviance: 4023.471
## Log-likelihood: -2011.735
```

Multilevel: The p-value here is 0.7219. We fail to reject the null hypothesis. The extra term (district:growth) is unnecessary. The reduced model is preferred

```
model <- lmer(schoolscore~district + growthstatus + district:growthstatus + (1|school), REML = T,data =
```

```
## fixed-effect model matrix is rank deficient so dropping 30 columns / coefficients
```

```
model1<-lmer(schoolscore~district + growthstatus+ (1 |school), REML = T, data = df)
anova(model,model1)
```

```
## refitting model(s) with ML (instead of REML)

## Data: df
## Models:
## model1: schoolscore ~ district + growthstatus + (1 | school)
## model: schoolscore ~ district + growthstatus + district:growthstatus +
## model:     (1 | school)
##        npar   AIC   BIC  logLik deviance  Chisq  Df Pr(>Chisq)
## model1  121 18565 19265 -9161.6    18323
## model   323 18780 20648 -9066.8    18134 189.76 202     0.7219
```

To confirm whether district or growthstatus alone would be a better fit to the model, the following anova test is carried. Both the pvalue for term district alone and growthstatus alone are $< 2.2e\text{-}16$. Which means that we can reject the null hypothesis. model 1 is preferred.

```
model2 <- lmer(schoolscore~district + (1 |school), REML = T, data = df)
```

```
## boundary (singular) fit: see ?isSingular
```

```
anova(model1,model2)
```

```
## refitting model(s) with ML (instead of REML)

## Data: df
## Models:
## model2: schoolscore ~ district + (1 | school)
## model1: schoolscore ~ district + growthstatus + (1 | school)
##        npar   AIC   BIC  logLik deviance  Chisq Df Pr(>Chisq)
## model2  119 19015 19703 -9388.4    18777
## model1  121 18565 19265 -9161.6    18323 453.54  2  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model3 <- lmer(schoolscore~growthstatus + (1 |school), REML = T, data = df)
```

```
## boundary (singular) fit: see ?isSingular
```

```
anova(model1,model3)
```

```
## refitting model(s) with ML (instead of REML)

## Data: df
## Models:
## model3: schoolscore ~ growthstatus + (1 | school)
## model1: schoolscore ~ district + growthstatus + (1 | school)
##        npar   AIC   BIC  logLik deviance  Chisq  Df Pr(>Chisq)
## model3    5 18952 18981 -9471.0    18942
## model1  121 18565 19265 -9161.6    18323 618.66 116  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```