Hao Qin

Final project

3/12/2018

Patient survey (HCAHPS) - Hospital final Project

Abstract:

With the progress of society, although the medical equipment and technology of hospitals are gradually improving, more and more patients are getting sick. So, what I'm interested in is whether the patient's medical experience in the hospital and the use of equipment will affect the patient's recovery. So, I search the website to get some information about that and know there is a national standardized survey which called the HCAHPS. The second step is to download the dataset into R, and clean that in order to get a better dataset.

Then, I will do some EDA about my project and fit the best model to find the relationship between the experience from the patient and variety variables.

Introduction:

I: HCAHPS is a national, standardized survey of hospital patients about their experiences during a recent inpatient hospital stay. As we all know, the facility service and condition environment in the hospital can also affect whether a patient can recover smoothly

II: Before I did the project, I will Choose a dataset that is relevant to my career goals or my personal interest and propose an analysis that includes fitting at least a multilevel model. Then, my professor provided us kinds of example data and some websites with the dataset.

So, I pick the website that I am interested in, which is about the Medical. I spent a lot of times on that web, and finally find a dataset which is about the Patient survey (HCAHPS) – Hospital.

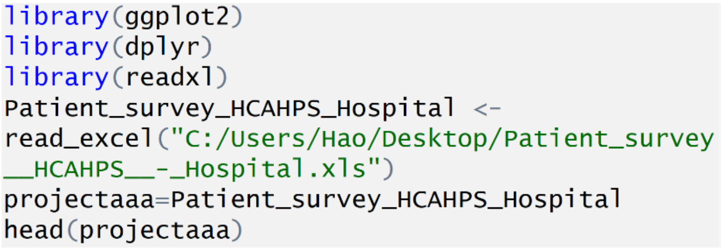
Method:

I: This is my dataset website: https://data.medicare.gov/Hospital-Compare/Patient-survey-HCAHPS-Hospital/dgck-syfz which contains the 246K rows, and 23 columns. The variable called Star will be the outcome, and other variables will be the response.

II: since my dataset has many different variables, I will use the multilevel model to check the how many variables that related to the outcome (Star).

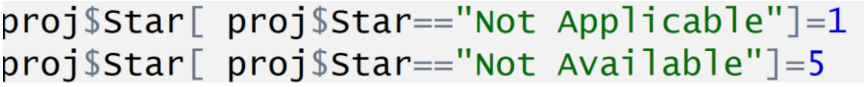
Results: (Those graphs are from the R screen shot)

1, I will plug my first dataset into the R

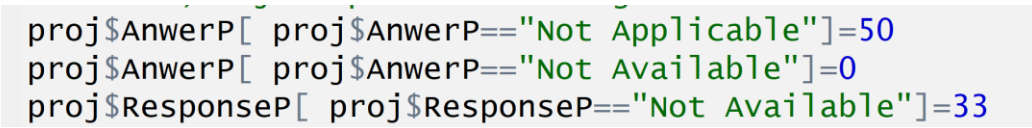


```
library(ggplot2)
library(dplyr)
library(readxl)
Patient_survey_HCAHPS_Hospital <-
read_excel("C:/Users/Hao/Desktop/Patient_survey
__HCAHPS__-_Hospital.xls")
projectaaa=Patient_survey_HCAHPS_Hospital
head(projectaaa)
```

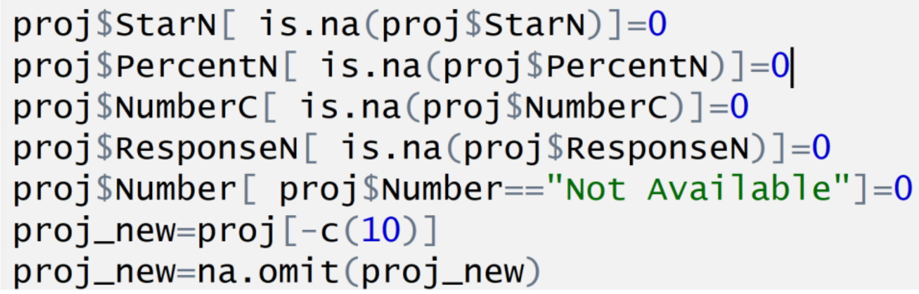
2, After I plug this, I find that there is a lot of print mistake in dataset such as there is a lot of data which contains the NA, the Not Available, the Not Applicable, I search the website, which shows that we can know that in order to make the 'Not Applicable' and 'the Not Available' more easily to understand, I just rank then as one point star and five point star, the same as the percentage of answer the question and the percentage of the response, in order to make the data more reliable, I just pick the average 50 and 33. Also there are too many NAs in that, so I just change the NA into zero, and delete the unrelated variable



```
proj$Star[ proj$Star=="Not Applicable"]=1
proj$Star[ proj$Star=="Not Available"]=5
```

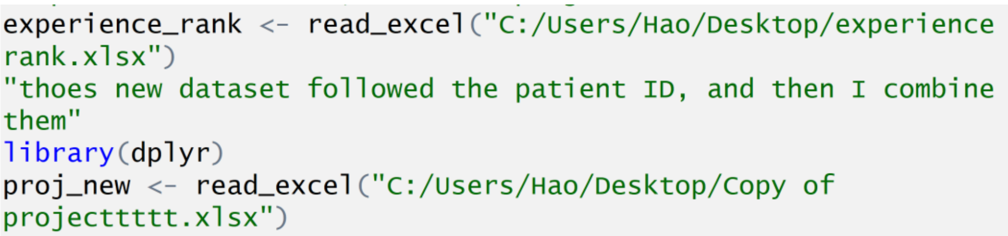


```
proj$AnwerP[ proj$AnwerP=="Not Applicable"]=50
proj$AnwerP[ proj$AnwerP=="Not Available"]=0
proj$ResponseP[ proj$ResponseP=="Not Available"]=33
```



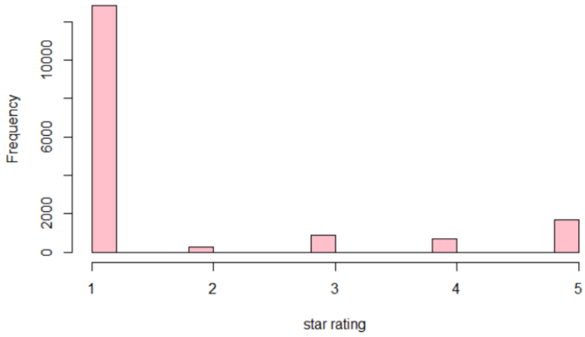
```
proj$StarN[ is.na(proj$StarN)]=0
proj$PercentN[ is.na(proj$PercentN)]=0
proj$NumberC[ is.na(proj$NumberC)]=0
proj$ResponseN[ is.na(proj$ResponseN)]=0
proj$Number[ proj$Number=="Not Available"]=0
proj_new=proj[-c(10)]
proj_new=na.omit(proj_new)
```

3. What's more, after receiving my professor's advice, I got another dataset which is rank by the experience, when the rank range is between 0-20, the star will be 1, and the rank range is between 21-40, the star will be 2, etc... Combine those two datasets



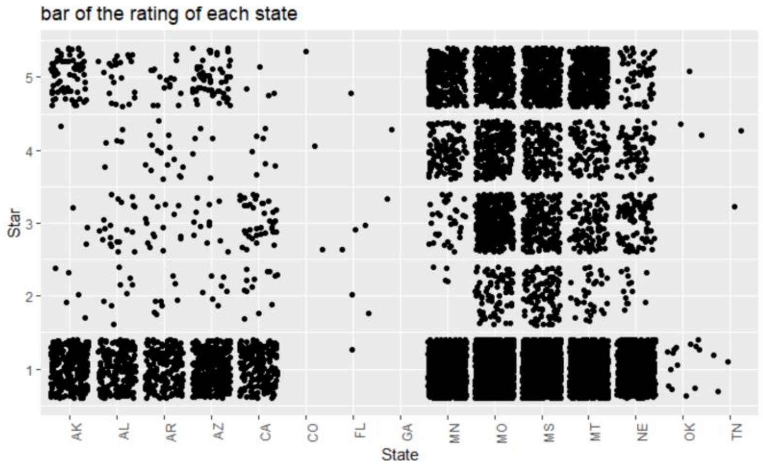
```
experience_rank <- read_excel("C:/Users/Hao/Desktop/experience
rank.xlsx")
"thoes new dataset followed the patient ID, and then I combine
them"
library(dplyr)
proj_new <- read_excel("C:/Users/Hao/Desktop/Copy of
projectttttt.xlsx")
```

4. for the next step, I will do some graphs to get more details about my project, the first graph is about the counts of the star ratings.

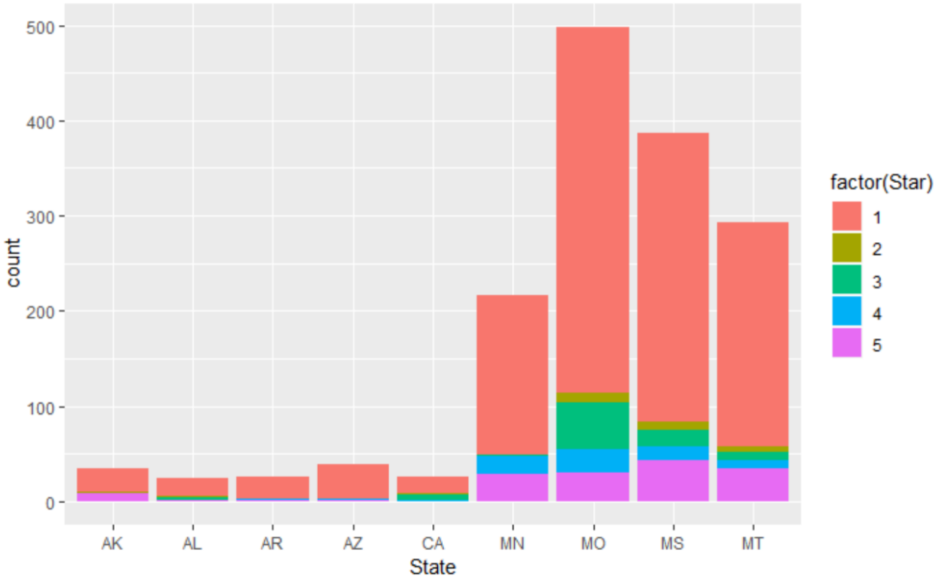


From the outcome, we can get that the star one has the largest proportion in this plot, so we can get a basic result that most of the people for now did not get the satisfaction in hospital. For the other comparison, I will do more graphs about other variables.

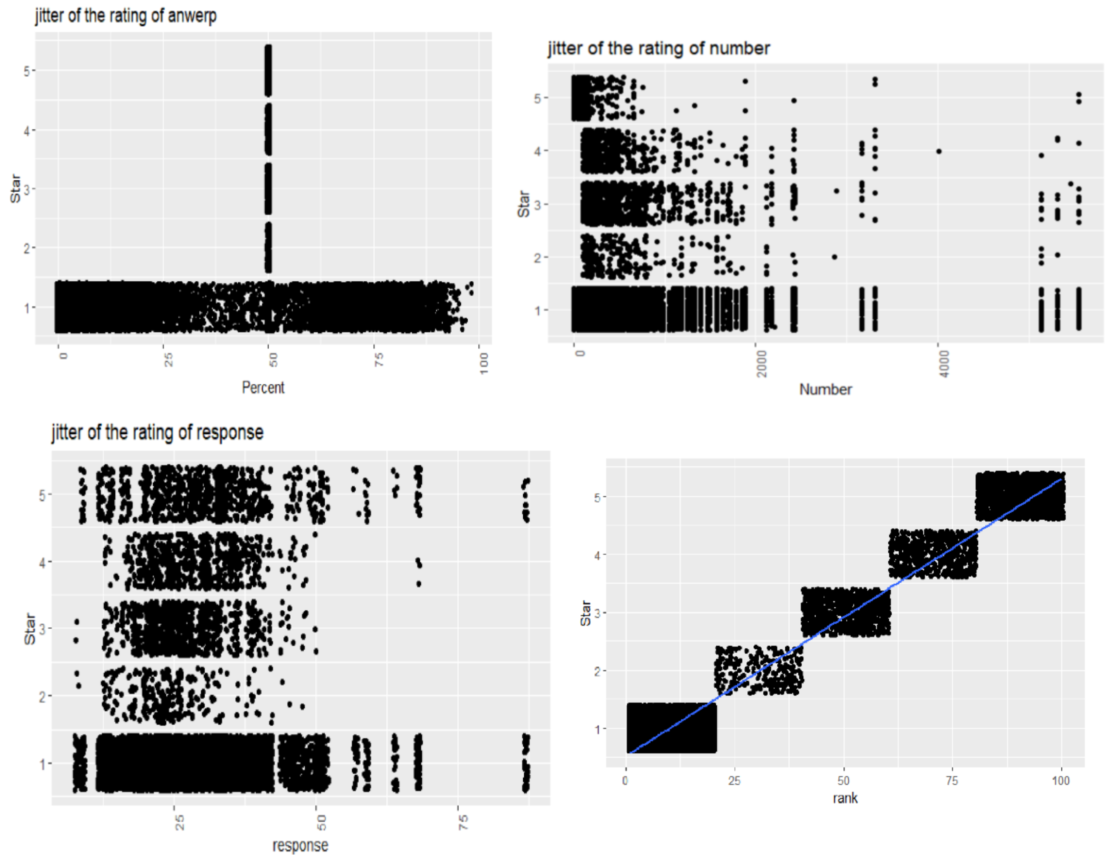
5. since different states has different stars from the patients, so the next step, I will pick the state and star to create the plot.



From the outcome, we can get that most of the points are in AK AL AR AZ CA MN MO MS MT NE, so I will choose those to find more information about this.

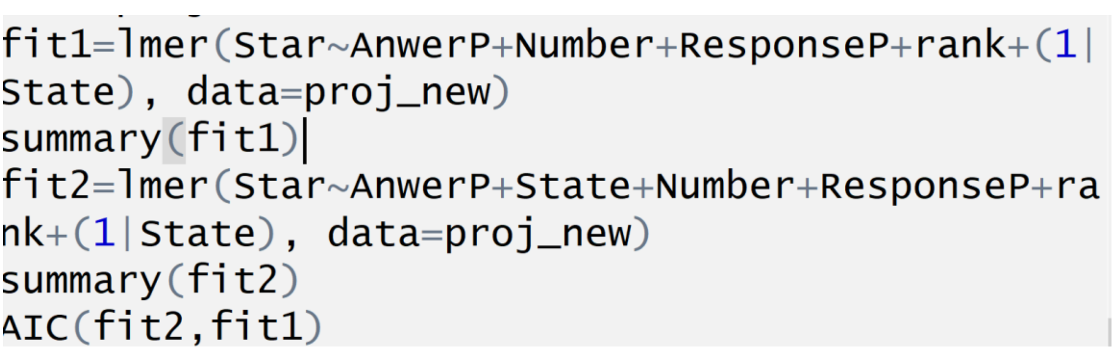


From the plot, we can get that for the star 4, 3 and 1, in MO, it has the largest proportion. Except this variable, I also set the variable answer, response, number and rank as the response, and continue to do some plot. You can see above!

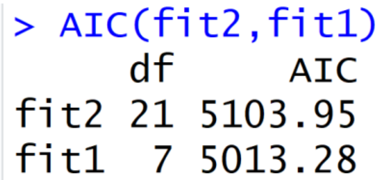


From the first plot, you can see most of the star, which is located in the 1, also the same as the Number and the Response, however in the last plot, you can see there is a linear regression on that, with the rank going up, the star will increase.

The next important step is to fit the model, since the requirement is to let us use the multilevel, so let's see the R outcome

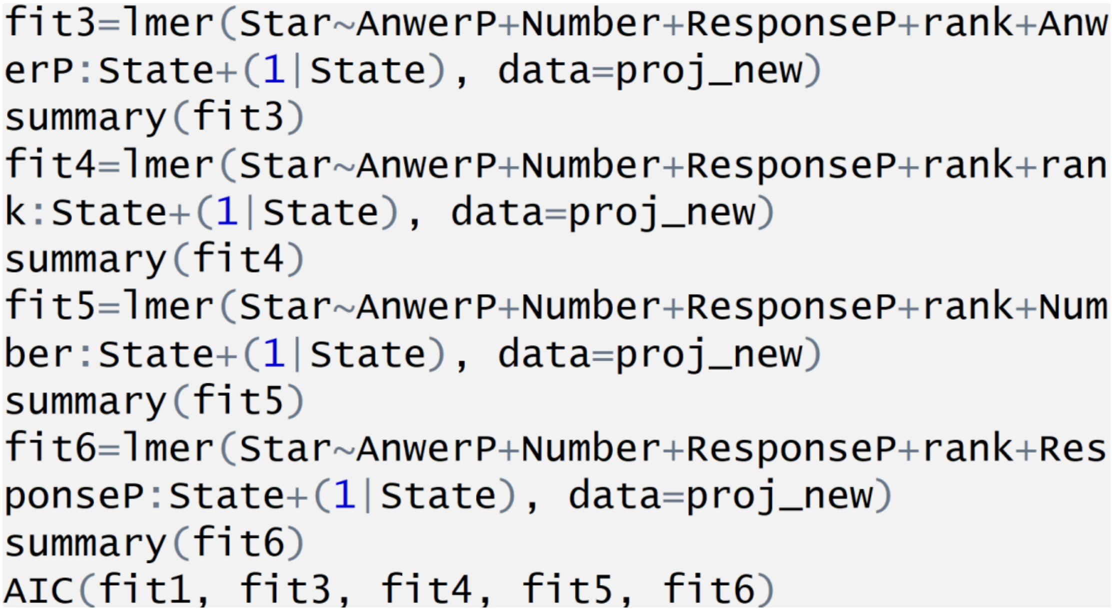


```
fit1=lmer(Star~AnwerP+Number+ResponseP+rank+(1|
State), data=proj_new)
summary(fit1)
fit2=lmer(Star~AnwerP+State+Number+ResponseP+ra
nk+(1|State), data=proj_new)
summary(fit2)
AIC(fit2,fit1)
```

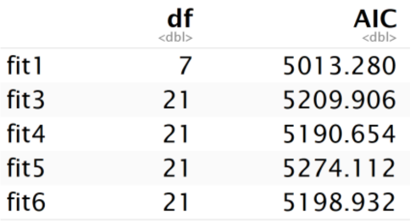


```
> AIC(fit2,fit1)
     df     AIC
fit2 21 5103.95
fit1  7 5013.28
```

from the first two fitted model, I will pick the first one as my best, since it has a lower AIC between them, but for the fit1 and fit2, there is no interaction between, so the next I will add the 4 different interaction to find which one is better.



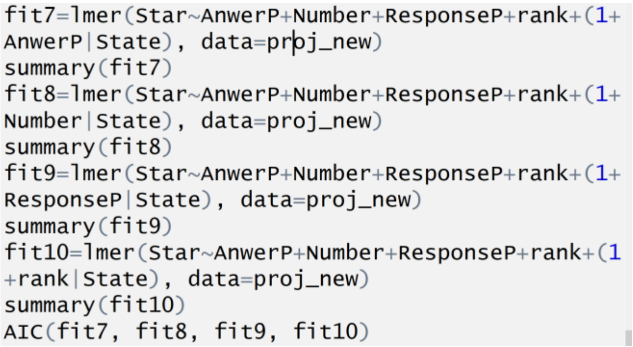
```
fit3=lmer(Star~AnwerP+Number+ResponseP+rank+Anw
erP:State+(1|State), data=proj_new)
summary(fit3)
fit4=lmer(Star~AnwerP+Number+ResponseP+rank+ran
k:State+(1|State), data=proj_new)
summary(fit4)
fit5=lmer(Star~AnwerP+Number+ResponseP+rank+Num
ber:State+(1|State), data=proj_new)
summary(fit5)
fit6=lmer(Star~AnwerP+Number+ResponseP+rank+Res
ponseP:State+(1|State), data=proj_new)
summary(fit6)
AIC(fit1, fit3, fit4, fit5, fit6)
```



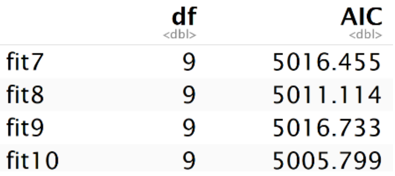
```
        df       AIC
     <dbl>     <dbl>
fit1     7  5013.280
fit3    21  5209.906
fit4    21  5190.654
fit5    21  5274.112
fit6    21  5198.932
```

from the outcome, the fit1 is still better.

However, for the previous models that I fit, there are only random intercept effect. I picked the best model fit and now I will be fitting random slope models.



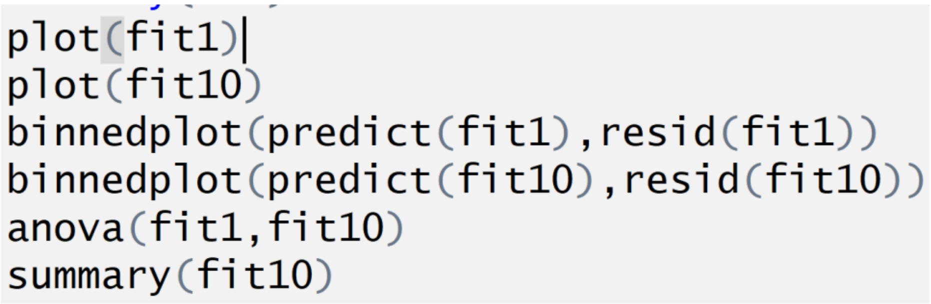
```
fit7=lmer(Star~AnwerP+Number+ResponseP+rank+(1+
AnwerP|State), data=proj_new)
summary(fit7)
fit8=lmer(Star~AnwerP+Number+ResponseP+rank+(1+
Number|State), data=proj_new)
summary(fit8)
fit9=lmer(Star~AnwerP+Number+ResponseP+rank+(1+
ResponseP|State), data=proj_new)
summary(fit9)
fit10=lmer(Star~AnwerP+Number+ResponseP+rank+(1
+rank|State), data=proj_new)
summary(fit10)
AIC(fit7, fit8, fit9, fit10)
```



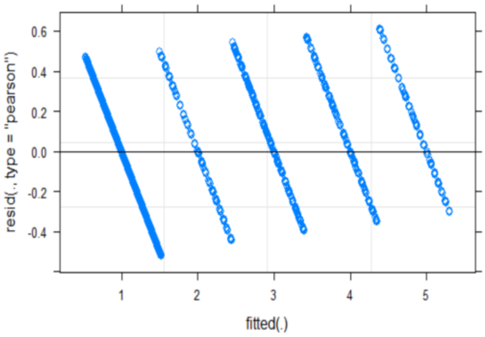
```
         df       AIC
      <dbl>     <dbl>
fit7      9  5016.455
fit8      9  5011.114
fit9      9  5016.733
fit10     9  5005.799
```

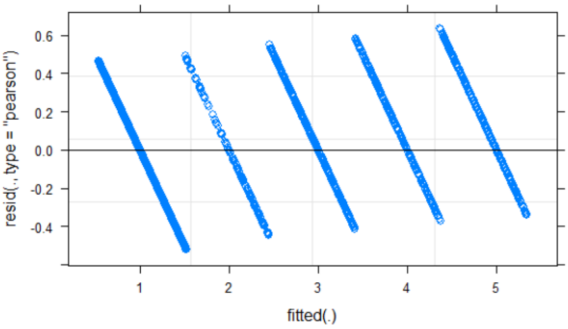
From the outcome, the fit10 has the lowest AIC, so for now, the only two left are fit1 and fit10.

After fit1 model and the fir10 model, I will do the comparison, and find the best one

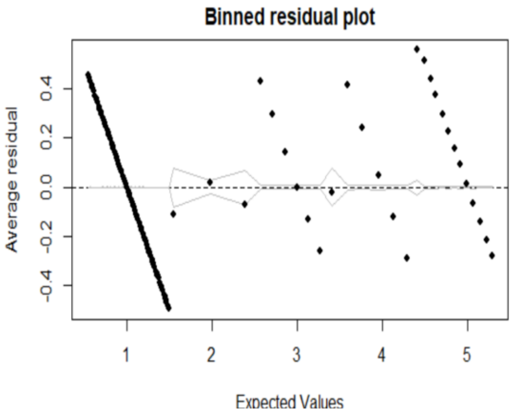


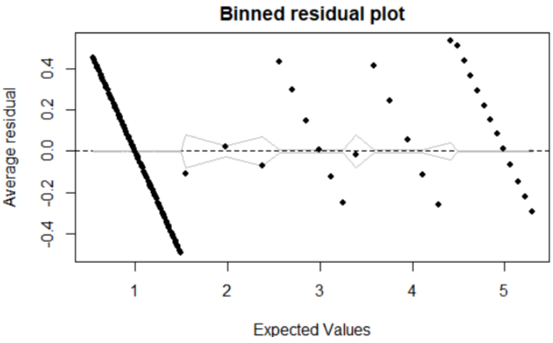
```
plot(fit1)
plot(fit10)
binnedplot(predict(fit1),resid(fit1))
binnedplot(predict(fit10),resid(fit10))
anova(fit1,fit10)
summary(fit10)
```

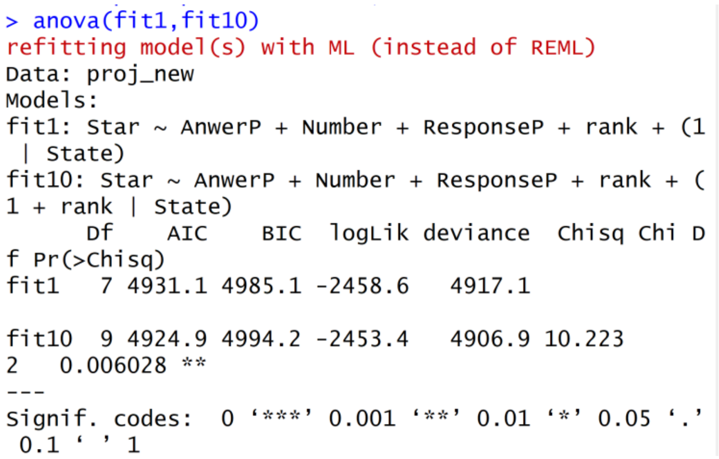




From the plot, we can see that the fit10's plot is more connected, and more cluster.







```
> anova(fit1,fit10)
refitting model(s) with ML (instead of REML)
Data: proj_new
Models:
fit1: Star ~ AnwerP + Number + ResponseP + rank + (1
 | State)
fit10: Star ~ AnwerP + Number + ResponseP + rank + (
1 + rank | State)
      Df    AIC    BIC  logLik deviance  Chisq Chi D
f Pr(>Chisq)
fit1   7 4931.1 4985.1 -2458.6   4917.1

fit10  9 4924.9 4994.2 -2453.4   4906.9 10.223
2   0.006028 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
 0.1 ‘ ’ 1
```

To compare the 2 different fitted model, I will use the anova to show that, you can see it from the outcome, fit1 has a lower AIC, so for now I think the best model is fit1

Overall, I fit lots of random intercept models and random slope models and find the best fit for these two kinds, and then do some model checking and find the best model fit model fit2 which is a random intercept model