Midterm Exam

Minqi Li

11/2/2020

Instruction

This is your midterm exam that you are expected to work on it alone. You may NOT discuss any of the content of your exam with anyone except your instructor. This includes text, chat, email and other online forums. We expect you to respect and follow the GRS Academic and Professional Conduct Code (http://www.bu.edu/cas/files/2017/02/GRS-Academic-Conduct-Code-Final.pdf).

Although you may NOT ask anyone directly, you are allowed to use external resources such as R codes on the Internet. If you do use someone's code, please make sure you clearly cite the origin of the code.

When you finish, please compile and submit the PDF file and the link to the GitHub repository that contains the entire analysis.

Introduction

In this exam, you will act as both the client and the consultant for the data that you collected in the data collection exercise (20pts). Please note that you are not allowed to change the data. The goal of this exam is to demonstrate your ability to perform the statistical analysis that you learned in this class so far. It is important to note that significance of the analysis is not the main goal of this exam but the focus is on the appropriateness of your approaches.

Data Description (10pts)

Please explain what your data is about and what the comparison of interest is. In the process, please make sure to demonstrate that you can load your data properly into R.

```
data<- read.csv(file="Dataset_Minqi_Li.csv", header = T)
data$Age<-factor(data$Age,levels=c("<18","18-24","24-30",">30"))
data$Percentage<-factor(data$Percentage,levels=c("<30%","30%-60%",">60%"))
data$Price..RMB.<-factor(data$Price..RMB.,levels=c("<12","12-18","18-25",">25"))
head(data,n=5)
```

Age <fct>	Sex <chr>	Type <chr>	Frequency <int>	Percentage <fct>	Price..RMB. <fct>	Income..RMB. <chr>
1 18-24	female	fruit tea	2	30%-60%	12-18	0
2 18-24	female	bubble tea	2	<30%	12-18	0
3 24-30	female	bubble tea	1	<30%	<12	0
4 18-24	female	bubble tea	0	<30%	12-18	0
5 18-24	female	cheese tea	5	>60%	12-18	0

5 rows

In China, bubble tea, cheese tea and fruit tea are often sold in the milk tea shop and now they are very popular among people. In this dataset, I wonder which age group has the most number of drinking bubble tea, cheese tea and fruit tea in one week among my close contacts and their friends. Besides, I want to explore the relationship between other variables and the total number of drinking bubble tea, cheese tea and fruit tea in one week.

Variables description:

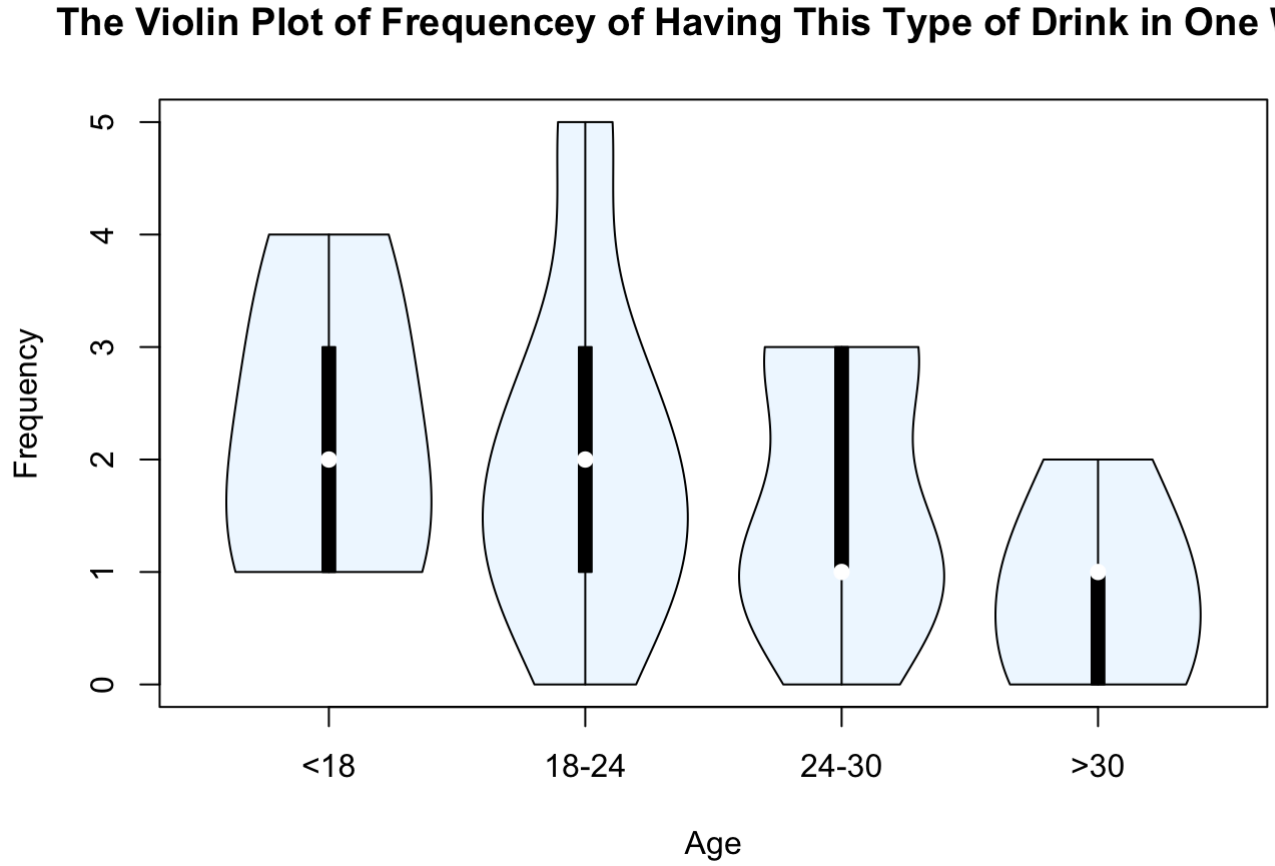
- Type: the favorite type among bubble tea, cheese tea and fruit tea.
- Frequency: the total number of drinking bubble tea, cheese tea and fruit tea in one week.
- Percentage: the percentage of total number of drinking bubble tea, cheese tea and fruit tea among drinks in one week.
- Price(RMB): the price of favorite drink among bubble tea, cheese tea and fruit tea.

EDA (10pts)

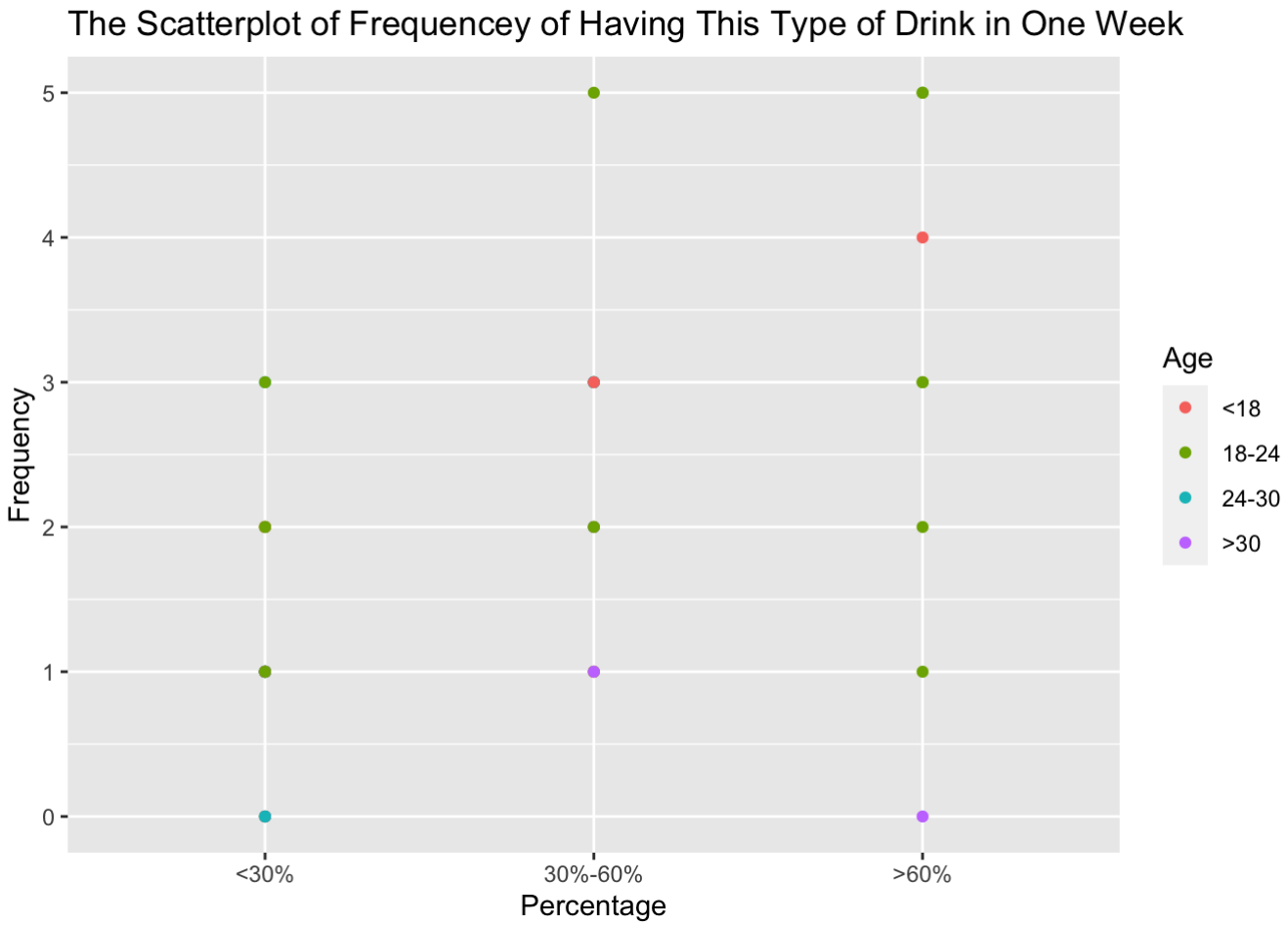
Please create one (maybe two) figure(s) that highlights the contrast of interest. Make sure you think ahead and match your figure with the analysis. For example, if your model requires you to take a log, make sure you take log in the figure as well.

Through viewing the dataset, I think the percentage and age have more influence on the frequency than other variables. Therefore, I use two plot to display their relationships.

```
# violin plot
x1<-data$Frequency[data$Age=="<18"]
x2<-data$Frequency[data$Age=="18-24"]
x3<-data$Frequency[data$Age=="24-30"]
x4<-data$Frequency[data$Age==">30"]
vioplot(x1,x2,x3,x4,
        names = c("<18","18-24","24-30",">30"),
        col = "aliceblue")
title("The Violin Plot of Frequencey of Having This Type of Drink in One Week",
      ylab = "Frequency",
      xlab = "Age")
```




```
#scatterplot
ggplot(data, aes(x = Percentage, y = Frequency, color=Age)) +
  geom_point()+
  labs(title='The Scatterplot of Frequencey of Having This Type of Drink in One Week'
)
```



- Based on the violin plot, we can conclude that people whose ages less than 18 years old have almost the same frequency of having this kind of drink on average in one week as people whose ages between 18 and 24 years old. Besides, people whose ages more than 30 years old have the lowest frequency of having this kind of drink on average in one week. In conclusion, the age has a weak negative influence on the frequency.
- Based on the scatterplot, we can find that the percentage has a positive influence on the frequency among most ages group.

Power Analysis (10pts)

Please perform power analysis on the project. Use 80% power, the sample size you used and infer the level of effect size you will be able to detect. Discuss whether your sample size was enough for the problem at hand. Please note that method of power analysis should match the analysis. Also, please clearly state why you should NOT use the effect size from the fitted model.

- Based on the dataset, I will build GLM model with Age and Percentage as parameters. Thus, I use pwr.f2.test() from pwr package to perform power analysis. Because u=2, v=41-2-1=38. Besides, confidence interval is set to be 95% and power is set to be 80%. I can get that effect size is 0.25. The following is the code:

```
pwr.f2.test(u=2,v=38,f2=NULL,sig.level=0.05,power=0.8)
```

```
##
##      Multiple regression power calculation
##
##               u = 2
##               v = 38
##              f2 = 0.254537
##       sig.level = 0.05
##           power = 0.8
```

- When confidence interval is set to be 95% and power is set to be 80%, I use medium effect size from Cohen's f2 which is 0.15. I can get that sample size is 65+2+1=68. Therefore, my sample size is not enough for the problem. The following is the code:

```
pwr.f2.test(u=2,v=NULL,f2=0.15,sig.level=0.05,power=0.8)
```

```
##
##      Multiple regression power calculation
##
##               u = 2
##               v = 64.31932
##              f2 = 0.15
##       sig.level = 0.05
##           power = 0.8
```

- Why we should not use the effect size from the fitted model.

1. Effect size is generally overestimated if we have an underpowered study.
2. Post power analysis may lead to a false direction for next study.

Modeling (10pts)

Please pick a regression model that best fits your data and fit your model. Please make sure you describe why you decide to choose the model. Also, if you are using GLM, make sure you explain your choice of link function as well.

The outcome of my model is Frequency which is the total number of drinking bubble tea, cheese tea and fruit tea in one week. Because Poisson regression is used when the outcome is positive discrete counts, I use Poisson regression to fit my model.

```
fit <- stan_glm(Frequency~Age+Percentage, family=poisson,data=data,refresh=0)
print(fit)
```

```
## stan_glm
##  family:       poisson [log]
##  formula:      Frequency ~ Age + Percentage
##  observations: 41
##  predictors:   6
## ------
##                    Median MAD_SD
## (Intercept)         0.2    0.4
## Age18-24            0.0    0.3
## Age24-30            0.0    0.4
## Age>30             -1.1    0.6
## Percentage30%-60%   0.8    0.3
## Percentage>60%      0.9    0.3
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

```
round(coef(fit),2)
```

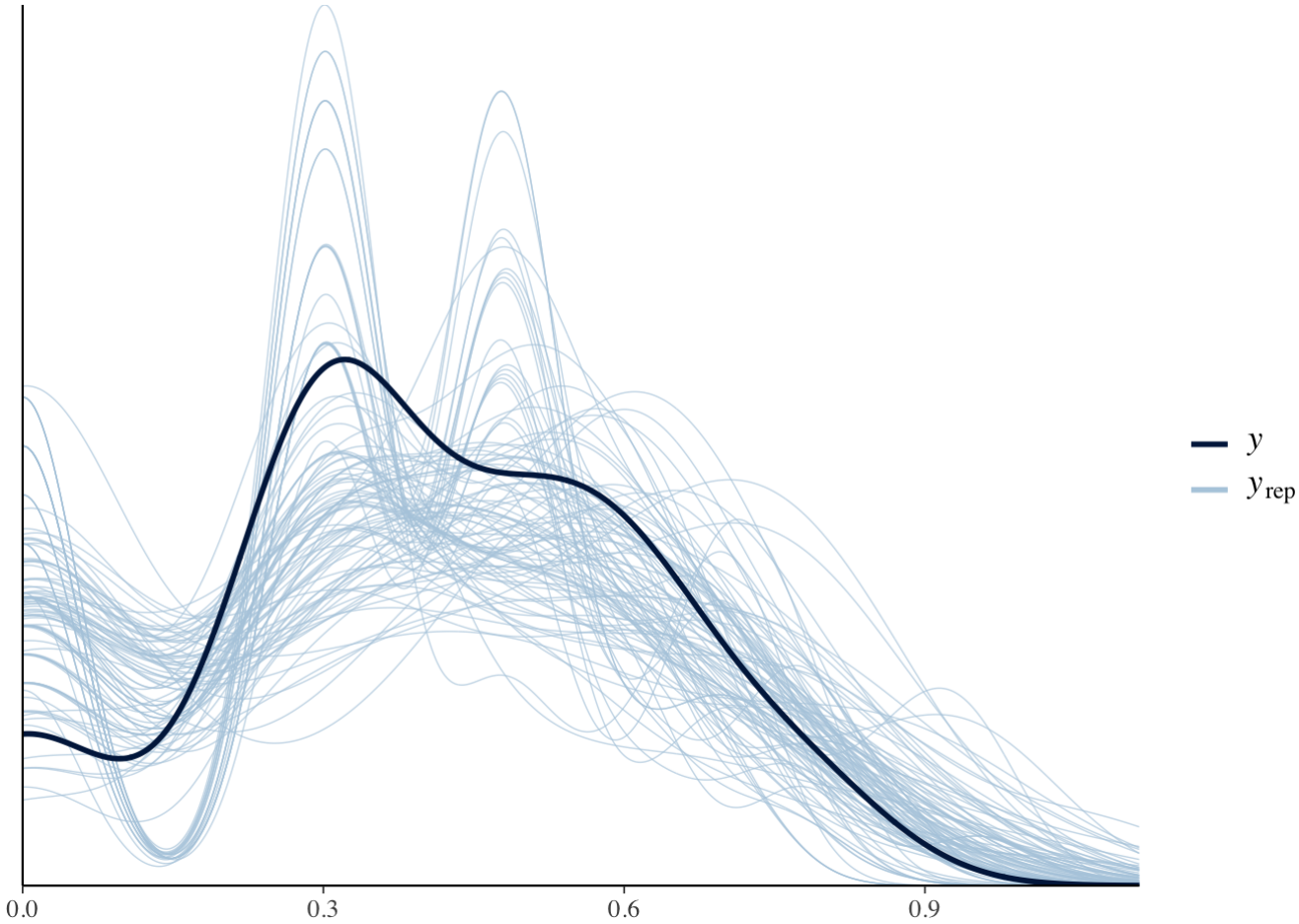
```
##       (Intercept)          Age18-24          Age24-30            Age>30
##              0.18              0.02              0.03             -1.06
## Percentage30%-60%    Percentage>60%
##              0.76              0.92
```

Validation (10pts)

Please perform a necessary validation and argue why your choice of the model is appropriate.

Posterior predictive checks

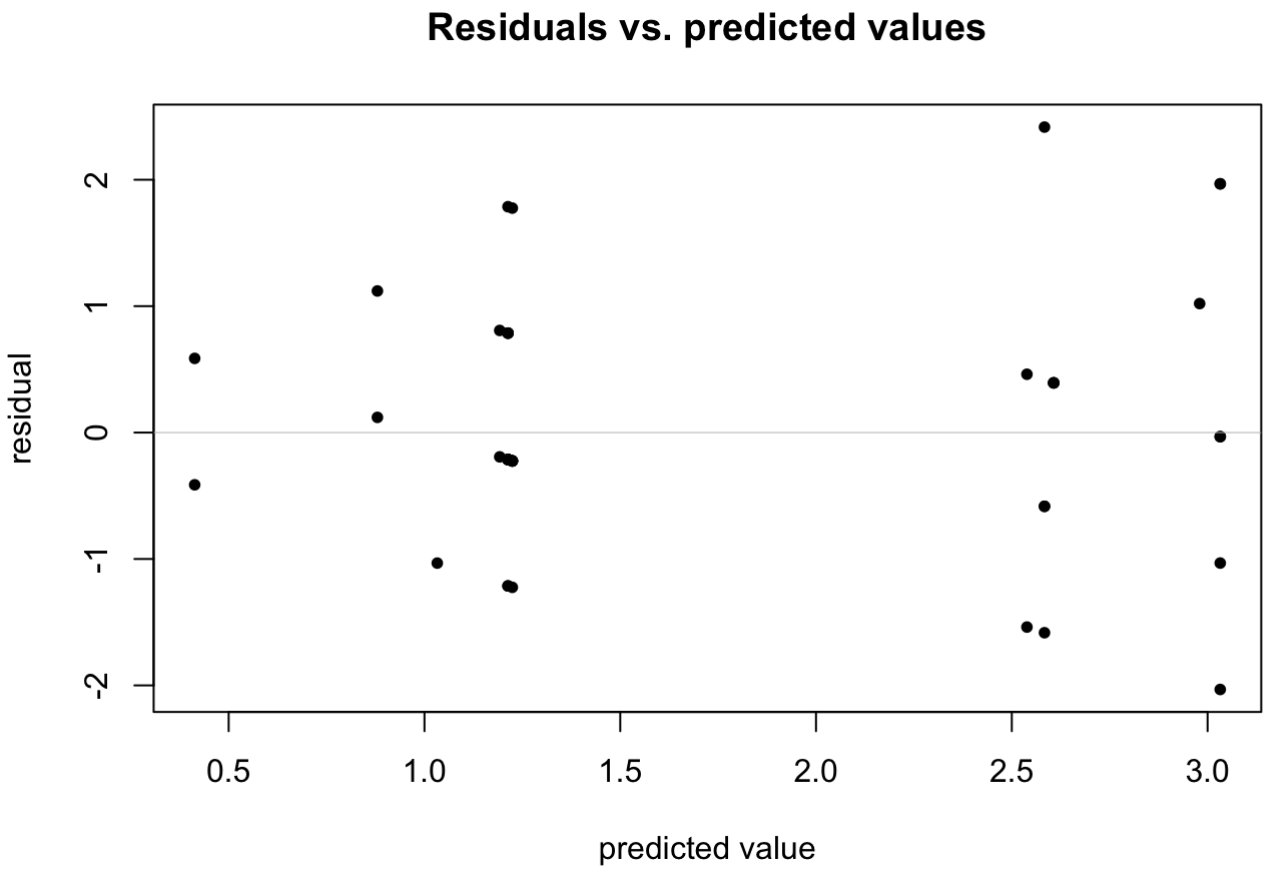
```
# Posterior predictive checks
y_rep<-posterior_predict(fit)
n_sims<-nrow(y_rep)
subset<-sample(n_sims,100)
check<-ppc_dens_overlay(log10(data$Frequency+1),log10(y_rep[subset,]+1))
check
```



Posterior predictive checks suggest some discrepancy with the fitted model and the data.

Residual plot

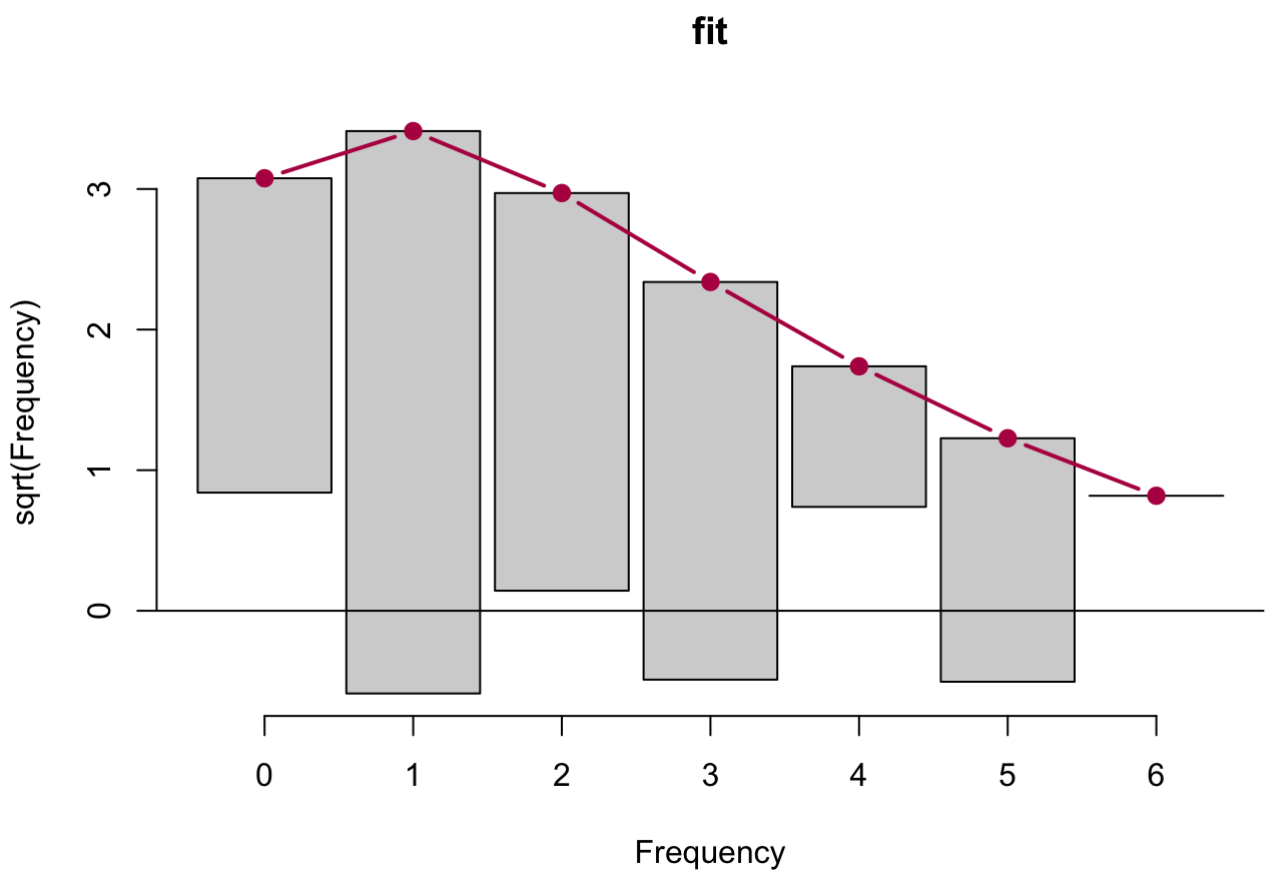
```
pred<-predict(fit, type="response")
plot(pred, resid(fit), xlab="predicted value", ylab="residual",
     main="Residuals vs.\ predicted values",  pch=20, yaxt="n")
axis(2, seq(-2,2,1))
abline(0, 0, col="gray", lwd=.5)
```



Based on the residual plot, most points are symmetrically scattered on both sides of the horizontal line. There are few outliers.

Check overdispersion

```
rootogram(fit)
```




```
# calculate overdispersion factor for poisson
z <- (data$Frequency-pred)/sqrt(pred)
n=length(data$Frequency)
k=length(data)-1
cat ("overdispersion ratio is ", sum(z^2)/(n-k), "\n")
```

```
## overdispersion ratio is  0.6967742
```

According to the plot and overdispersion ratio, data is not in the situation of overdispersion.

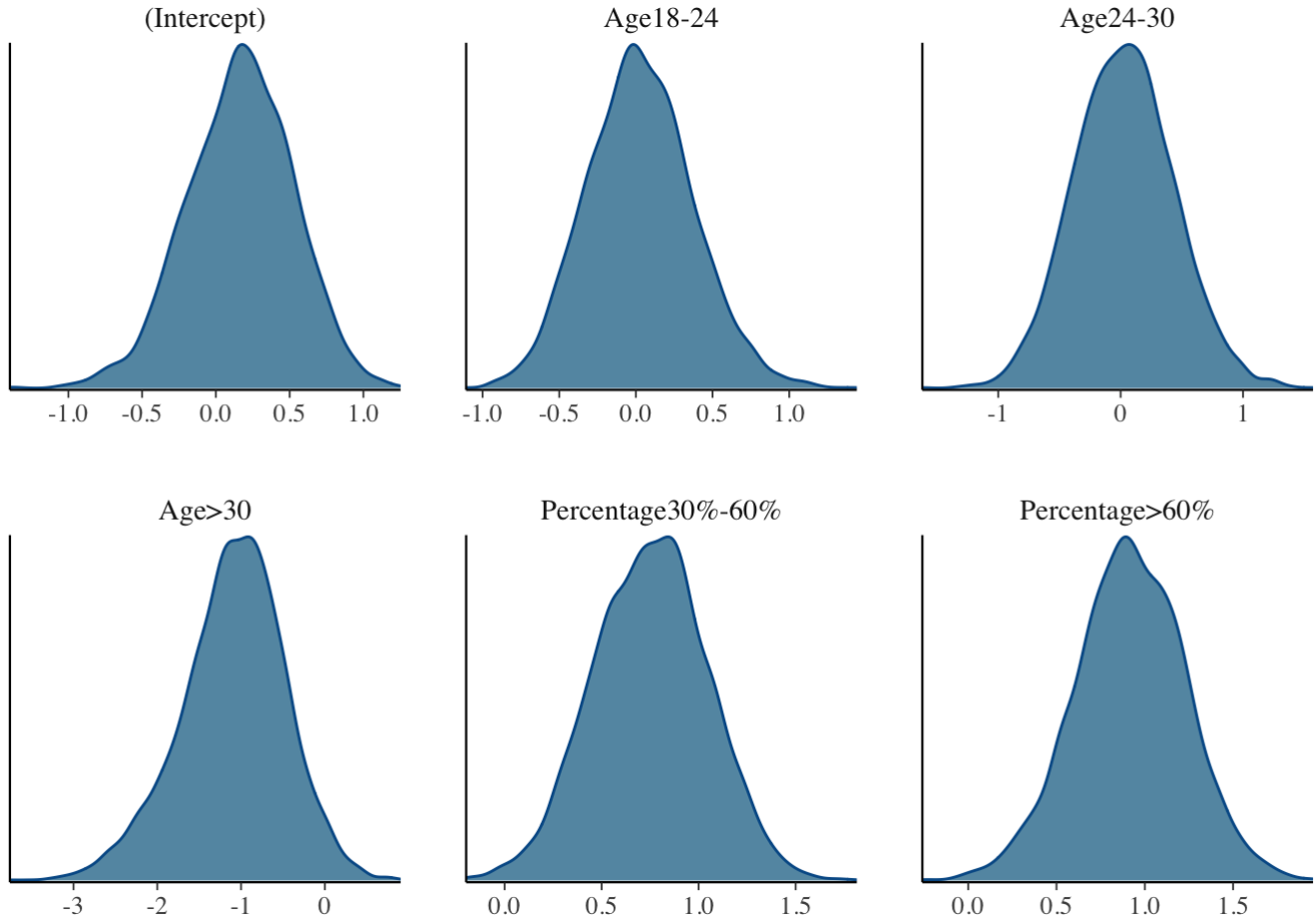
To sum up, the model is relatively appropriate from the residual plot and overdispersion checks, even though there is some discrepancy with the fitted model and the data in the posterior predictive checks.

Inference (10pts)

Based on the result so far please perform statistical inference to compare the comparison of interest.

The distribution of the coefficients

```
sims=as.matrix(fit)
mcmc_dens(sims)
```

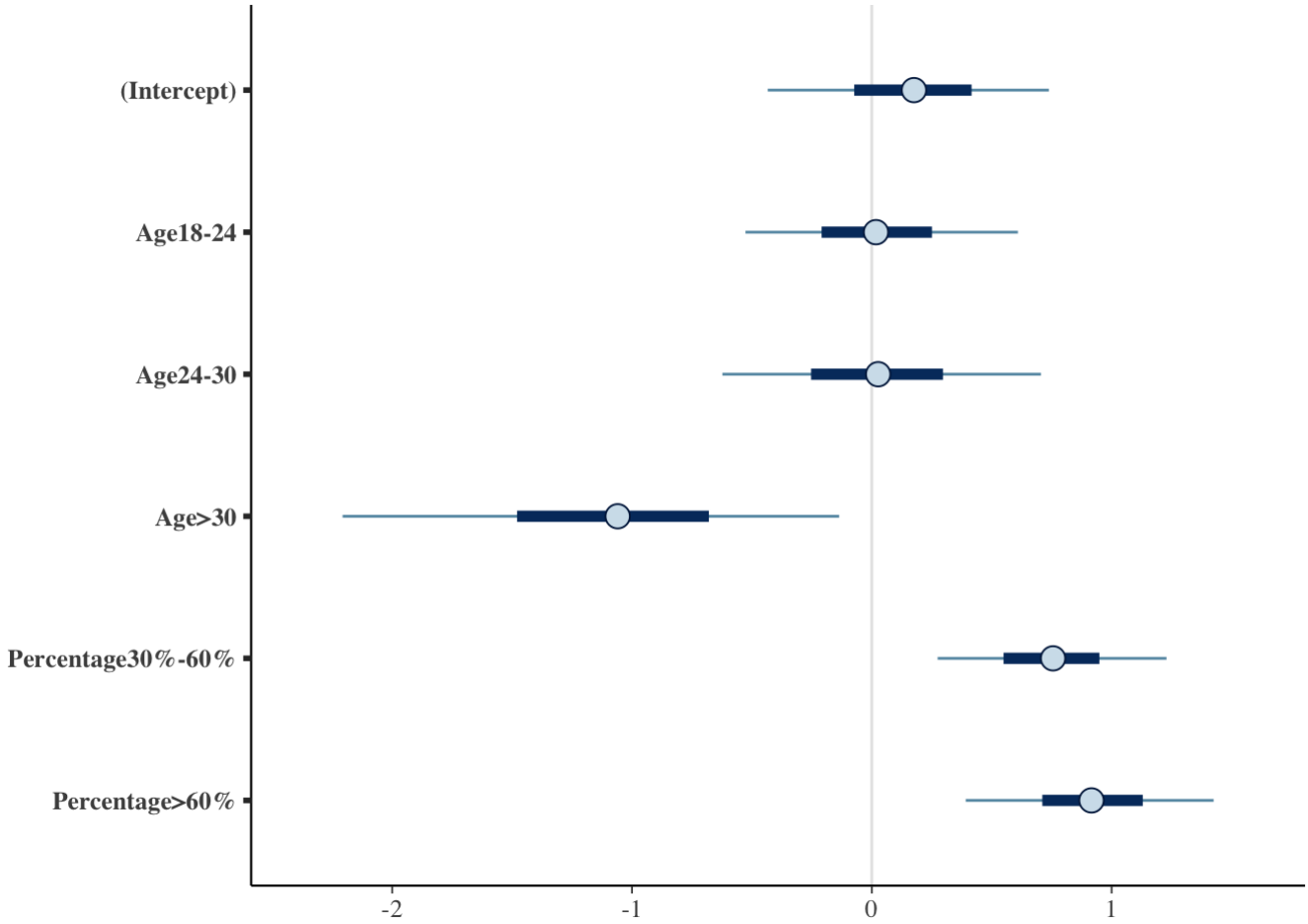


Posterior intervals

```
posterior_interval(fit)
```

```
##                           5%        95%
## (Intercept)       -0.4338464  0.7383276
## Age18-24          -0.5265921  0.6090817
## Age24-30          -0.6229044  0.7051973
## Age>30            -2.2066514 -0.1361478
## Percentage30%-60%  0.2745430  1.2286647
## Percentage>60%     0.3917722  1.4252833
```

```
mcmc_intervals(sims)
```



According to posterior intervals, the intercept and the coefficients of Age18-24 and Age24-30 are not significant because they are across 0. But we should not ignore them.

Discussion (10pts)

Please clearly state your conclusion and the implication of the result.

The fitted model

$$log(Frequency) = 0.18 + 0.02Age18 - 24 + 0.03Age24 - 30 - 1.07Age > 30 + 0.75Percentage30$$

Interpreting the coefficients

- The intercept of 0.18 represents frequency of 1.20 when people are less than 18 years old and have less than 30% total number of drinking bubble tea, cheese tea and fruit tea among drinks in one week.
- The coefficient of Age18-24 represents the conditional expectation of the frequency differs by multiplicative exp(0.02) on average when people are between 18 and 24 years old.
- The coefficient of Age24-30 represents the conditional expectation of the frequency differs by multiplicative exp(0.03) on average when people are between 24 and 30 years old.
- The coefficient of Age>30 represents the conditional expectation of the frequency differs by multiplicative exp(-1.07) on average when people are more than 30 years old.
- The coefficient of Percentage30%-60% represents the conditional expectation of the frequency differs by multiplicative exp(0.75) on average when people have 30%-60% total number of drinking bubble tea, cheese tea and fruit tea among drinks in one week.
- The coefficient of Percentage>60% represents the conditional expectation of the frequency differs by multiplicative exp(0.92) on average when people have more than 60% total number of drinking bubble tea, cheese tea and fruit tea among drinks in one week.

Conclusion

- When people are less than 30 years old, the age has a weak positive influence on the frequency. But when people are more than 30 years old, the age has a negative influence on the frequency.
- The percentage has more influence on the frequency than the age. The percentage has a positive influence on the frequency.

Limitations and future opportunity. (10pts)

Please list concerns about your analysis. Also, please state how you might go about fixing the problem in your future study.

- The sample size is not enough for the problem. Besides, I have little data which is male and has a job so that I can't analyze the relationship between sex, income and the frequency. I should collect more data with wider range in my future study.
- Posterior predictive checks suggest some discrepancy with the fitted model and the data. I should detect outliers and process data to fit the model better.
- I could try other type of model to fit the model better.

Comments or questions

If you have any comments or questions, please write them here.