Midterm Exam

ChenxunLi

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.

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.

1.What the data is

My data set is about drinking milk, and I set 6 variables as below.

'frequency': how many times a week of drinking milk

'gender': 1 represents male and 2 represents female

'skim': whether drinking skim milk, 1 represents yes and 2 represents no

'weight': person's weight in kg

'cost': the cost of drinking of milk at a time in yuan

'quantity': the number of milliliters of milk you drink at one time

2.What the comparison of interest is

I am interested in (1)the relationship between the frequency of drinking milk a week and whether drinking skim milk. (2)the relationship between the frequency of drinking milk a week and the quantity they drink one time.

```
original_data <- read_csv("original data.csv")
```

```
## Parsed with column specification:
## cols(
##   frequency = col_double(),
##   gender = col_double(),
##   skim = col_double(),
##   weight = col_double(),
##   cost = col_double(),
##   quantity = col_double()
## )
```

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.

1.Data processing

(1)First, I need clean the data because some samples are contradictory.

The sample which 'frequency' is 0 with 'cost' and 'quantity' filled with 0. The 'quantity' means the quantity of milk one drink one time if he/she drinks milk, so if 'frequency=0', 'quantity' may not be 0, so as the 'cost'.

```
milk <- filter(original_data,frequency!=0|cost!=0|quantity!=0)
```

(2)Then, I need to adjust the class of variables.

```
str(milk)
```

```
## tibble [46 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ frequency: num [1:46] 7 4 5 14 3 3 2 5 3 0 ...
##  $ gender   : num [1:46] 1 2 2 1 2 2 1 1 1 1 ...
##  $ skim     : num [1:46] 2 1 2 2 1 1 1 2 2 2 ...
##  $ weight   : num [1:46] 98 53 53 68 75 50 76 80 70 70 ...
##  $ cost     : num [1:46] 5 5 3.8 6 3.5 8 20 0 4 0 ...
##  $ quantity : num [1:46] 300 300 350 250 220 300 450 200 300 250 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   frequency = col_double(),
##   ..   gender = col_double(),
##   ..   skim = col_double(),
##   ..   weight = col_double(),
##   ..   cost = col_double(),
##   ..   quantity = col_double()
##   .. )
```

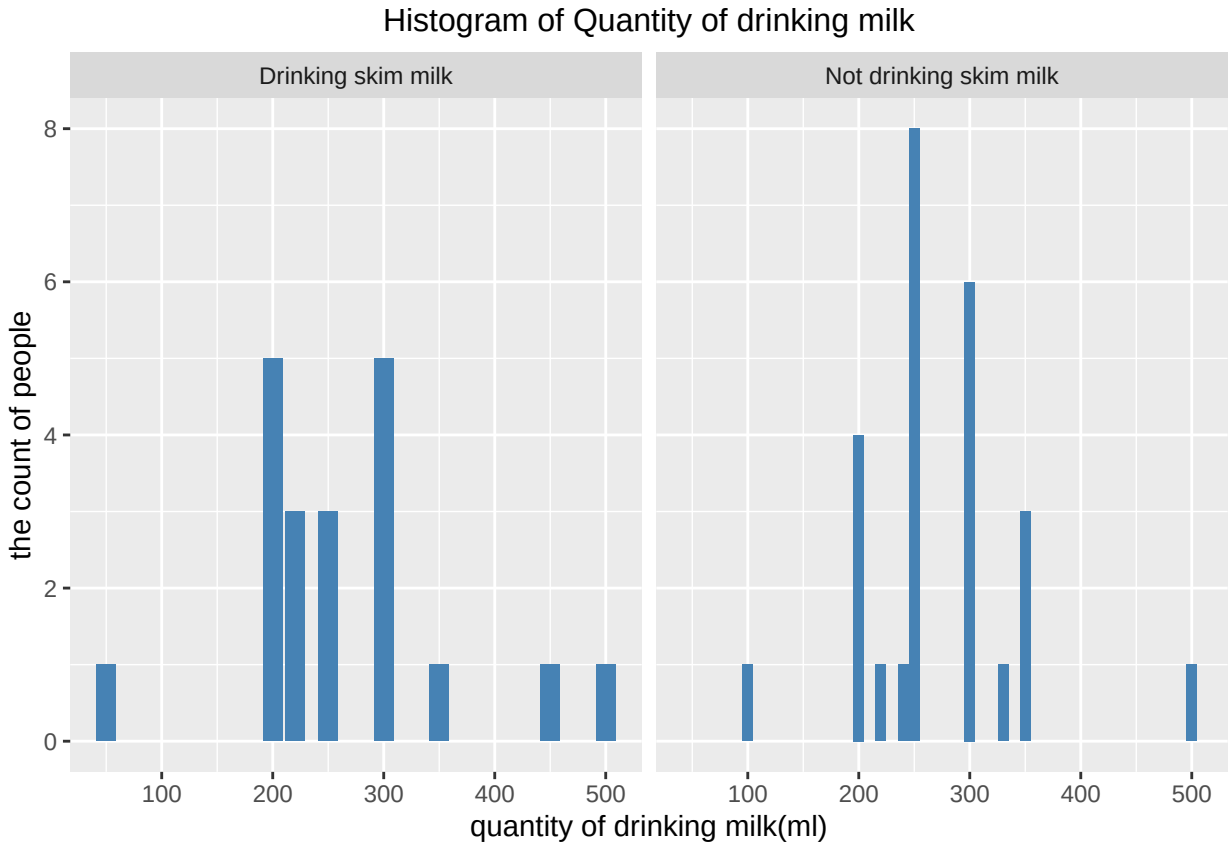
```
#make the gender and skim class factor, make the frequency integer
milk$frequency <- as.integer(milk$frequency)
milk$gender <- as.character(milk$gender)
milk$skim <- as.character(milk$skim)

str(milk)
```

```
## tibble [46 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ frequency: int [1:46] 7 4 5 14 3 3 2 5 3 0 ...
##  $ gender   : chr [1:46] "1" "2" "2" "1" ...
##  $ skim     : chr [1:46] "2" "1" "2" "2" ...
##  $ weight   : num [1:46] 98 53 53 68 75 50 76 80 70 70 ...
```

```
##  $ cost     : num [1:46] 5 5 3.8 6 3.5 8 20 0 4 0 ...
##  $ quantity : num [1:46] 300 300 350 250 220 300 450 200 300 250 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   frequency = col_double(),
##   ..   gender = col_double(),
##   ..   skim = col_double(),
##   ..   weight = col_double(),
##   ..   cost = col_double(),
##   ..   quantity = col_double()
##   .. )
```

2.The contrast of interest

```
milk$skim <- factor(milk$skim,labels = c("Drinking skim milk", "Not drinking skim milk"))
ggplot(data = milk)+
  geom_histogram(stat = "count", mapping = aes(x=milk$quantity),
      fill="steelblue",binwidth = 50, bins = 10)+
  facet_wrap(~skim)+
  xlab("quantity of drinking milk(ml)")+
  ylab("the count of people")+
  ggtitle("Histogram of Quantity of drinking milk")+
  theme(plot.title = element_text(hjust=0.5,size = 12))+
  scale_x_continuous(breaks=c(100,200,300,400,500))
```

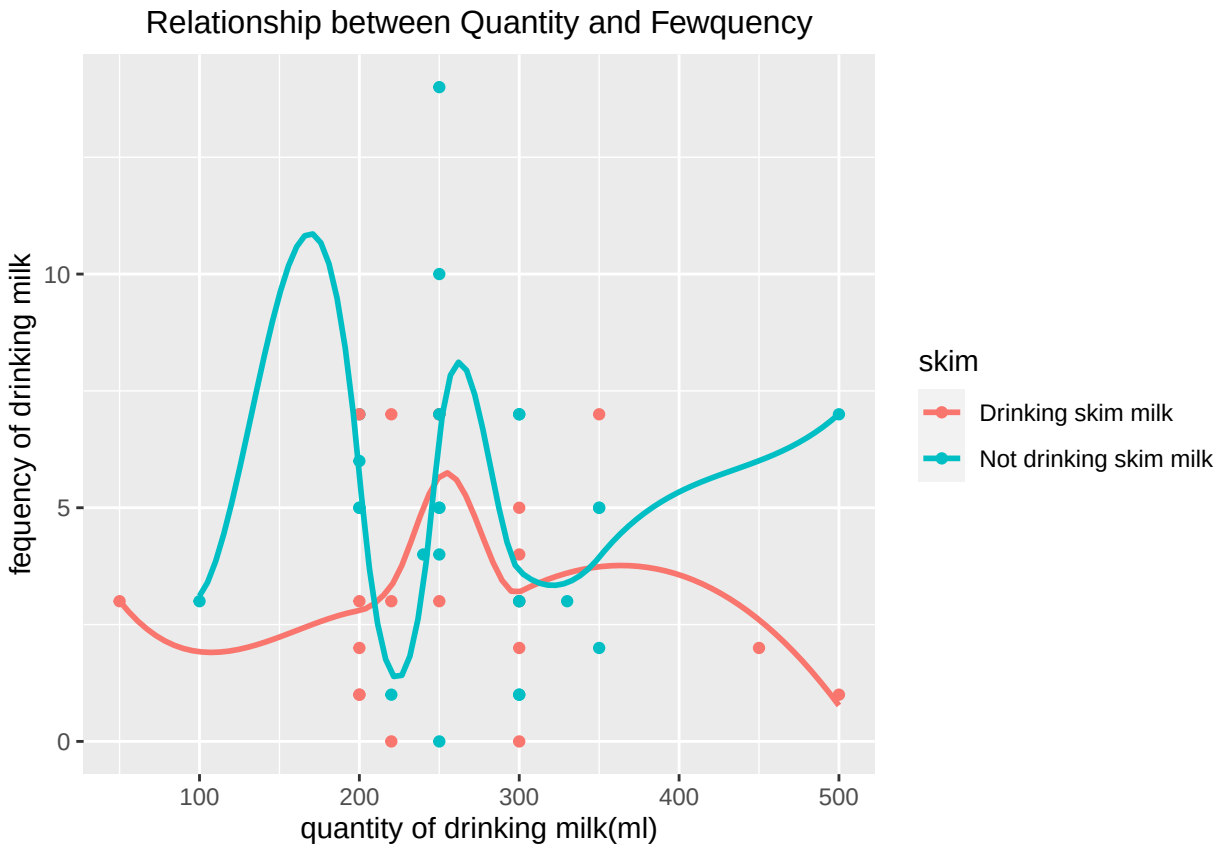


From the graph, we can clearly find most people drink milk between 200ml-300ml.

Besides, we can also find people who not drink skim milk usually drink more than people who drink skim milk.

```
ggplot(data=milk)+
  geom_point(aes(x=quantity, y= frequency, color=skim))+
  geom_smooth(aes(x=quantity, y= frequency, color=skim),se=F)+
  xlab("quantity of drinking milk(ml)")+
  ylab("fequency of drinking milk")+
  ggtitle("Relationship between Quantity and Fewquency")+
  theme(plot.title = element_text(hjust=0.5,size = 12))
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



From the graph, we can find that people who not drink skim milk usually drink more times a week than people who drink skim milk, and this situation is most obvious in quantity from 100ml-200ml.

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.

1.T test(infer the level of effect size)

```
milk_skim <- filter(milk, skim=="Drinking skim milk")
milk_noskim <- filter(milk, skim=="Not drinking skim milk")
n1 <- nrow(milk_skim)
n2 <- nrow(milk_noskim)
pwr.t2n.test(n1=n1,n2=n2, d=NULL, sig.level = 0.05,power = 0.8)
```

```
##
```

```
##      t test power calculation
##
##              n1 = 20
##              n2 = 26
##               d = 0.8520089
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

From the result, we can see that effect size is 0.85 which is a big effect size. That means another person who does the same experiment can have 85% probability to come to the same conclusion.

So, my sample size is enough for the problem at hand.

2.ANOVA

```
pwr.anova.test(k=2, n=min(n1,n2),f=NULL, sig.level = 0.05,power = 0.8)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 2
##               n = 20
##               f = 0.4545483
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

```
pwr.anova.test(k=2, n=max(n1,n2),f=NULL, sig.level = 0.05,power = 0.8)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 2
##               n = 26
##               f = 0.3961747
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

From the result, we can find no matter what n is, the effect size is around 0.4 which is big effect size. That can also provide evidence that my sample size is enough for the problem at hand.

3.Correlation

```
pwr.r.test(n=nrow(milk),r=NULL, sig.level = 0.05,power = 0.8)
```

```
##
##      approximate correlation power calculation (arctangh transformation)
##
##               n = 46
##               r = 0.3996617
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
```

From the result, we can find that linear correlation coefficient is 0.4, which is a medium effect size. That also

means my sample size is enough for the problem at hand.

4.chi-square test

```
pwr.chisq.test(w=NULL,N=nrow(milk),df=nrow(milk)-ncol(milk),sig.level = 0.05,power = 0.8)
```

```
##
##      Chi squared power calculation
##
##               w = 0.7739927
##               N = 46
##              df = 40
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: N is the number of observations
```

From the result, we can find that effect size is 0.77, which is a big effect. So, I can still draw the conclusion that my sample size is enough for the problem at hand.

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.

I choose linear regression because my variables are all numeric and I can use original regression to build the model.

```
fit_stanglm <- stan_glm(frequency~as.numeric(skim)+quantity, data=milk, refresh=0)
summary(fit_stanglm,digits = 4)
```

```
##
## Model Info:
##  function:     stan_glm
##  family:       gaussian [identity]
##  formula:      frequency ~ as.numeric(skim) + quantity
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 46
##  predictors:   3
##
## Estimates:
##                     mean    sd      10%     50%     90%
## (Intercept)       2.2287  1.9270 -0.2406  2.2516  4.6775
## as.numeric(skim)  1.5149  0.8622  0.3817  1.5111  2.6334
## quantity         -0.0013  0.0051 -0.0078 -0.0014  0.0051
## sigma             2.8792  0.3151  2.5014  2.8457  3.2962
##
## Fit Diagnostics:
##            mean   sd     10%    50%    90%
## mean_PPD 4.2357 0.6007 3.4770 4.2379 4.9992
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for de
##
## MCMC diagnostics
```

```
##                   mcse   Rhat   n_eff
## (Intercept)      0.0274 0.9999 4949
## as.numeric(skim) 0.0122 1.0000 5034
## quantity         0.0001 1.0002 4449
## sigma            0.0048 1.0002 4336
## mean_PPD         0.0095 0.9995 3977
## log-posterior    0.0366 1.0024 1637
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

```
fit_glm <- glm(frequency~as.numeric(skim)+quantity, data=milk)
summary(fit_glm)
```

```
##
## Call:
## glm(formula = frequency ~ as.numeric(skim) + quantity, data = milk)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.9146  -1.8424  -0.3827   2.0674   9.0854
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.283704   1.892350   1.207   0.2341
## as.numeric(skim)  1.495827   0.843142   1.774   0.0831 .
## quantity         -0.001443   0.005061  -0.285   0.7769
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 8.018652)
##
##     Null deviance: 370.37  on 45  degrees of freedom
## Residual deviance: 344.80  on 43  degrees of freedom
## AIC: 231.2
##
## Number of Fisher Scoring iterations: 2
```

Validation (10pts)

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

1.Cross Validation

Cross Validation is one method which can estimate the model's test error with its fitted data, and this can help us compare models easily.

(1)Cross Validation with Stan

```
loo <- loo(fit_stanglm)

kfold <- kfold(fit_stanglm)
```

```
## Fitting model 1 out of 10

## Fitting model 2 out of 10

## Fitting model 3 out of 10

## Fitting model 4 out of 10
```

```
## Fitting model 5 out of 10
## Fitting model 6 out of 10
## Fitting model 7 out of 10
## Fitting model 8 out of 10
## Fitting model 9 out of 10
## Fitting model 10 out of 10
```

'elpd_loo' is "Estimated Log Predictive Density" calculated through Loo and if it is closer to 0, the model will have a lower test error.

'looic' is double of 'elpd_loo', so if it is closer to 0, the model will have a lower test error.

From the result, 'elpd_loo' is not very big, so it is a appropriate model.

(2).Cross Validation without STAN

```
cv.glm(milk, fit_glm)$delta[1]
```

```
## [1] 8.41264
```

'cv.glm' compute the mean MSW across each fold and if this outcome is lower, the model will have a lower expected prediction.

fit1's 'cv.glm' is 8.41, which is small, so it fits a appropriate model.

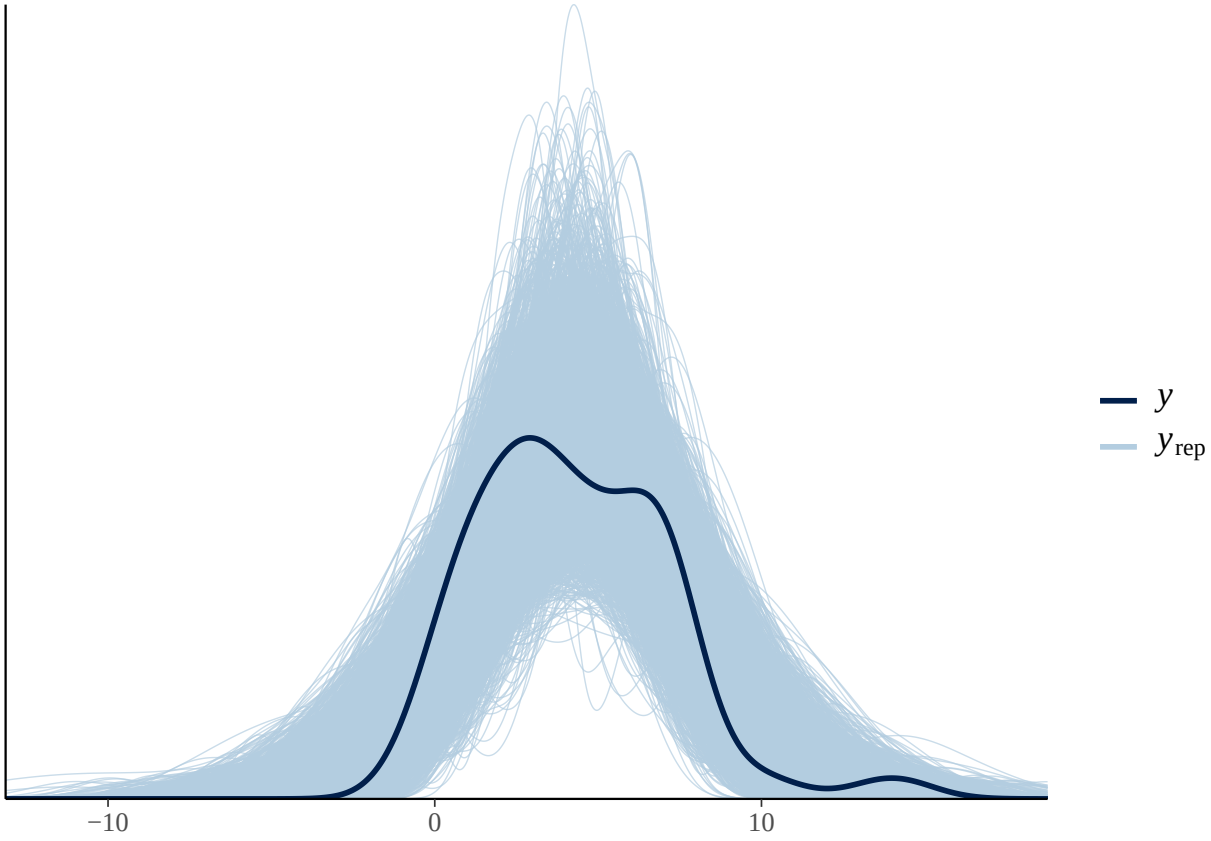
2.Posterior Predictive Checks

First, we use 'posterior_predict' to produce 4000 simulations of our data.

```
frequency_rep <- posterior_predict(fit_stanglm)
```

Then, we use 'ppc_dens_overlay' to visualize our data

```
ppc_dens_overlay(milk$frequency,frequency_rep)+scale_y_continuous(breaks = NULL)
```



Finally, we can see something from our data's distribution.

The dark line is our data's distribution, and the blue lines are our simulations. We can see that the dark line is wholly within the blue line, that means our model fits our data well.

Inference (10pts)

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

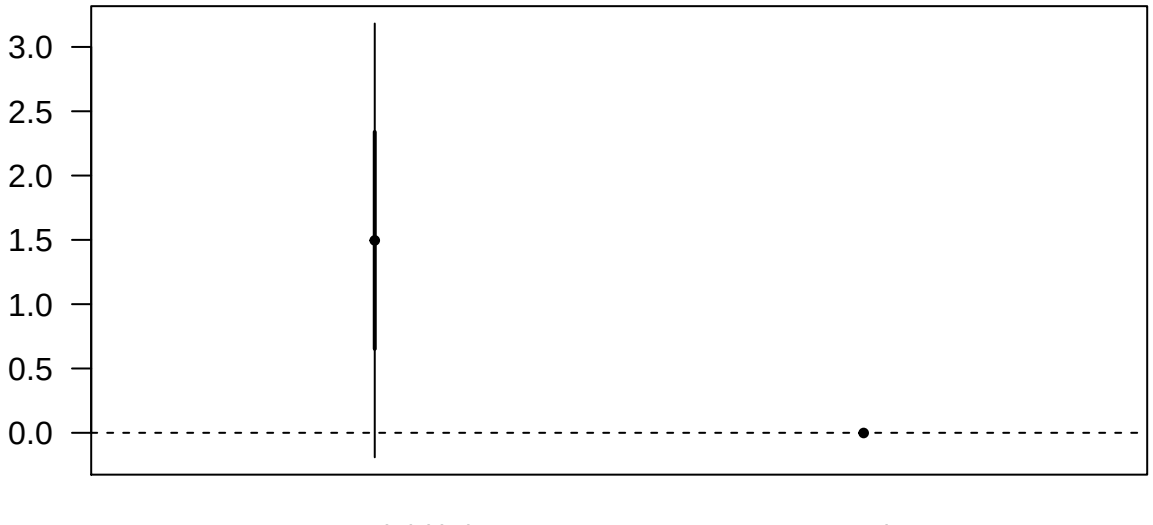
1. Classical Statistical Inference

```
confint(fit_glm)
```

```
## Waiting for profiling to be done...
```

```
##                        2.5 %      97.5 %
## (Intercept)      -1.42523297 5.992641661
## as.numeric(skim) -0.15670110 3.148355191
## quantity         -0.01136335 0.008477183
```

```
coefplot(fit_glm, vertical=FALSE, var.las=1, frame.plot=TRUE)
```



Though the 95% confidence interval for 'skim' does not crosses 0, its p-value is 0.2, which is bigger than 0.05, so it does not mean that there is a strong effect.

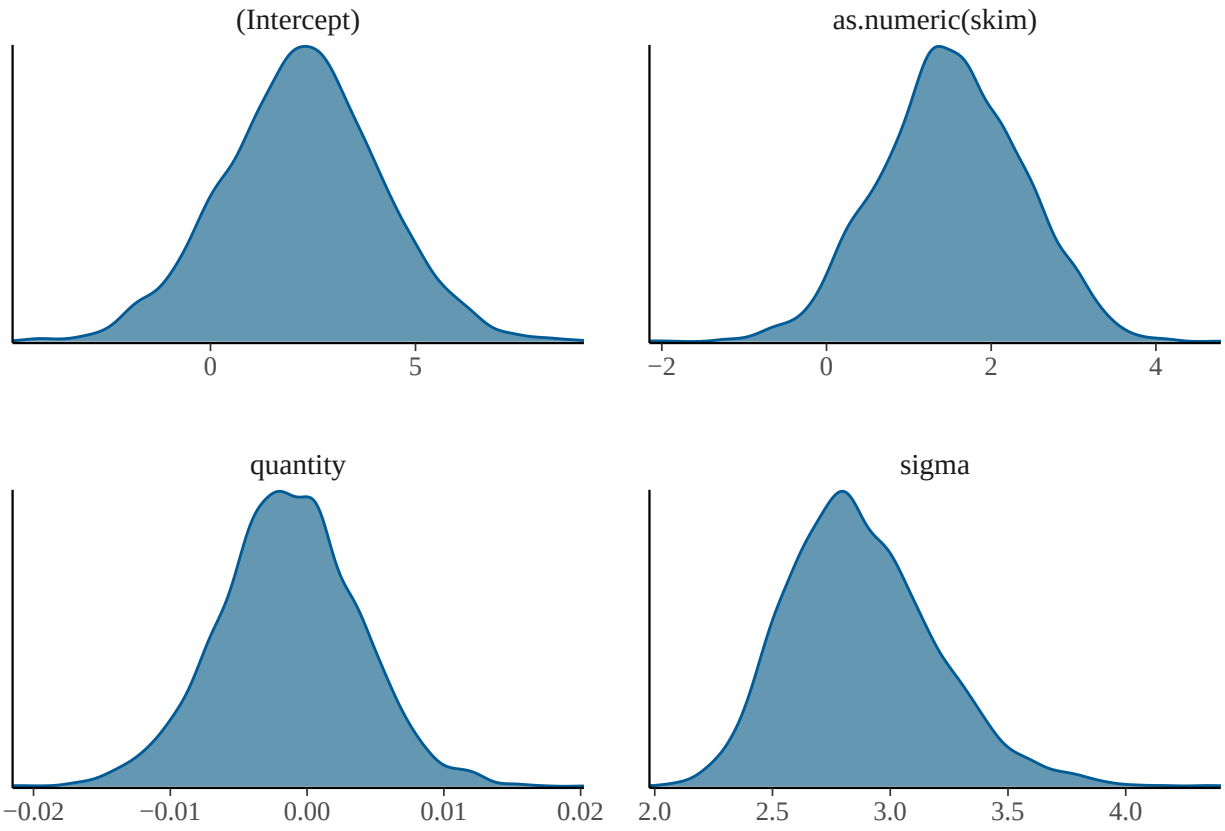
2. Bayesian Inference First, I use 'as.matrix' to build a matrix of posterior simulations

```
sims <- as.matrix(fit_stanglm)
head(sims)
```

```
##           parameters
## iterations (Intercept) as.numeric(skim)      quantity    sigma
##       [1,]   2.9963521        0.8847854  0.0008936288 3.229853
##       [2,]   2.5978420        0.3946832  0.0023195663 2.884649
##       [3,]   1.7991780        2.6164679 -0.0046122825 2.666049
##       [4,]  -0.2586194        1.8432330  0.0044884544 2.773909
##       [5,]   3.3604852        1.2682437 -0.0026583132 2.593885
##       [6,]   1.7014842        1.7881196 -0.0011360116 2.448887
```

Then, we use 'mcmc_dens' to see the distribution of the coefficients.

```
mcmc_dens(sims)
```



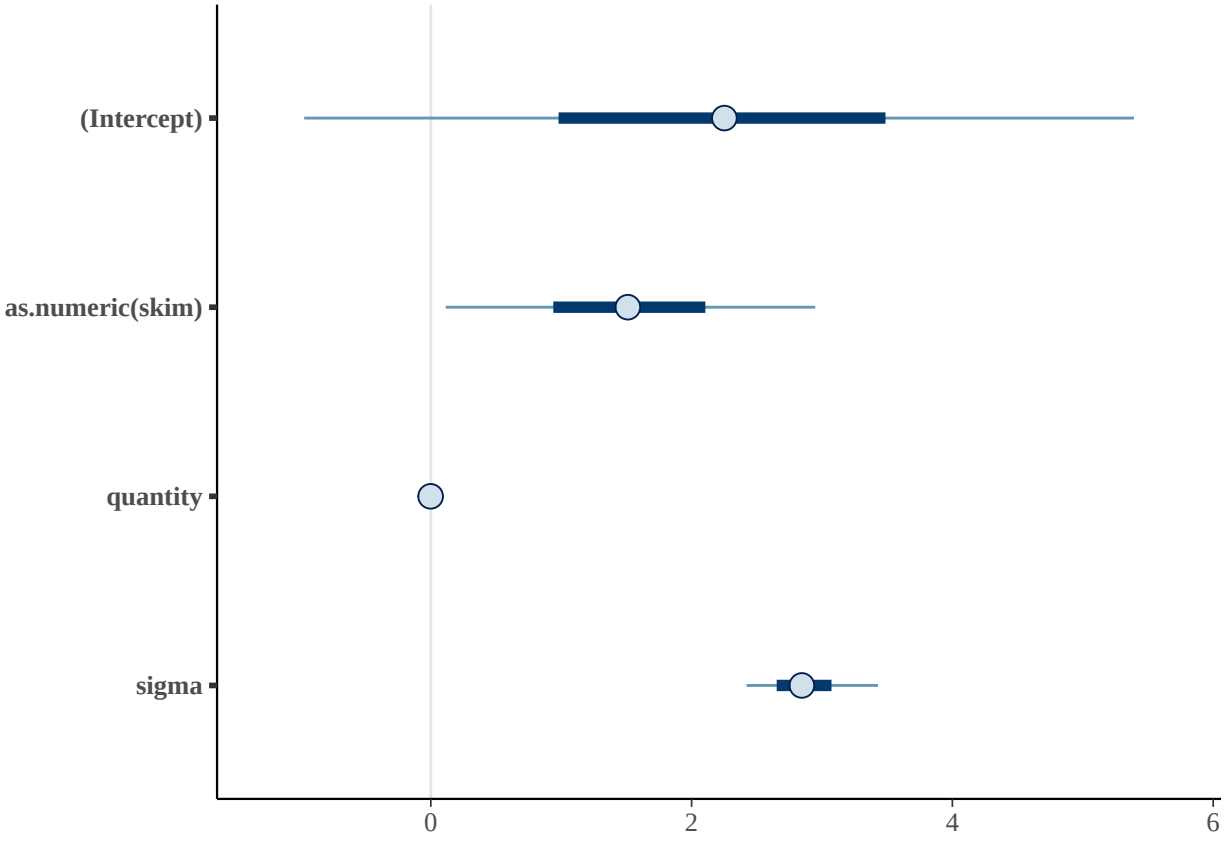
Next, we use 'posterior_interval' to show credible intervals.

```
posterior_interval(fit_stanglm)
```

```
##                             5%         95%
## (Intercept)      -0.971172131 5.392500075
## as.numeric(skim)  0.114470258 2.948242416
## quantity         -0.009760806 0.006947889
## sigma             2.421486981 3.429500016
```

Besides, we can use 'mcmc_intervals' to see the posterior intervals.

```
mcmc_intervals(sims)
```



Discussion (10pts)

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

1.My conclusion:

The intercept of model 'fit_stanglm' is 2.29, the coefficient of 'skim' is 1.47, and the coefficient of 'quantity' is -0.0013.

2.Implication of the result:

Intercept: Under the same situation, if one drinks skim milk and drinks average quantity milk one time, he/she will drink milk 2.29 times a week.

skim: Under the same situation, if one does not drink skim milk, he/she will drink more 1.47 times a week than who drinks skim milk.

quantity: Under the same situation, if one drinks 1(ml) more a time, he/she will drink 0.0013 less time a week. 0.0013 may seem not significant, however, usually the quantity of milk one drinks more than milk the other drinks is not just 1(ml), and at least 10(ml) I think.

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.

1.Limitations

(1)P-value The P-value of this model is 0.2 which is bigger than 0.05, so the outcome is not much significant.

(2)Cross Validation with Stan In cross Validation with Stan, 'elpd_loo' is not much closer to 0.

(3)Logical analysis In this model, I still have much more control variables. Though I get the conclusion that whether drinking skim milk really effect the frequency of drinking milk a week, I cannot be sure that other

factors does not effect,for example, maybe gender effect and male originally have more milk than female and the male usually does not like skim milk. I have not control these variables, so the conclusion still has some limitations.

(4)Samples I get the data set through the way of electronic questionnaire, and I send the questionnaire to my friends who have the same age period and works in similar industry. It is not wholly random at first, so the data set need to be more random.

2.Fix the problem

(1)First, I need to make the data set more random, for example, I should give them to different age period's people or I add a new variable called 'age' and so on.

(2)Next, I need to control variables. The more variables I come up with, the more accurate the model will be.

(3)Besides, the sample size still need to be bigger to make the model significant.

(4)Finally, notice the unit when designing the variables for example the quantity's unit, it may be better to use 10(ml) as one unit, which is more practical.

Comments or questions

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

Reference

[1]R in action- Data Analysis and Graphics with R Second Edition-Robert I. Kabacoff

[2]MA678 discussion - TA's Rmd file in Blackboard

[3]https://zhuanlan.zhihu.com/p/129375307