Assignment6\_bayes

mkhulekeli Nkosi 2017159092

2025-05-24

# 1.

The residual error is caused by the difference in groups that will do presentation and by the different assessors. Also not having enough variables to explain the the final mark.

# 2.

There is always a factor that influences an outcome but according to the above assumptions I would say in this case they are enough for the average assessor mark to be correct on average.

# 3.

data<- readxl::read\_xlsx("BayesAssignment6of2025.xlsx")  
summary(data)

## Group LecturerA LecturerB LecturerC   
## Length:15 Min. :60.00 Min. :49.00 Min. :60.00   
## Class :character 1st Qu.:72.00 1st Qu.:62.00 1st Qu.:63.50   
## Mode :character Median :74.00 Median :68.00 Median :67.50   
## Mean :74.14 Mean :65.33 Mean :69.75   
## 3rd Qu.:76.75 3rd Qu.:70.00 3rd Qu.:77.50   
## Max. :88.00 Max. :82.00 Max. :85.00   
## NA's :1 NA's :6 NA's :3   
## LecturerD LecturerE LecturerF LecturerG Proposal   
## Min. :60.00 Min. :52.00 Min. :53.00 Min. :60.0 Min. :57.00   
## 1st Qu.:68.00 1st Qu.:61.50 1st Qu.:71.75 1st Qu.:64.5 1st Qu.:63.50   
## Median :70.00 Median :68.00 Median :78.00 Median :69.5 Median :74.00   
## Mean :70.50 Mean :67.71 Mean :72.25 Mean :68.0 Mean :71.13   
## 3rd Qu.:76.25 3rd Qu.:76.00 3rd Qu.:78.50 3rd Qu.:73.0 3rd Qu.:78.00   
## Max. :78.00 Max. :79.00 Max. :80.00 Max. :73.0 Max. :84.00   
## NA's :5 NA's :8 NA's :11 NA's :11   
## Literature Quiz Interview   
## Min. :55.0 Min. :48.00 Min. :49.00   
## 1st Qu.:65.5 1st Qu.:66.50 1st Qu.:64.00   
## Median :69.0 Median :75.00 Median :71.00   
## Mean :69.4 Mean :72.47 Mean :68.13   
## 3rd Qu.:74.5 3rd Qu.:80.00 3rd Qu.:72.00   
## Max. :91.0 Max. :85.00 Max. :77.00   
##

sapply(data, class)

## Group LecturerA LecturerB LecturerC LecturerD LecturerE   
## "character" "numeric" "numeric" "numeric" "numeric" "numeric"   
## LecturerF LecturerG Proposal Literature Quiz Interview   
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"

colSums(is.na(data))

## Group LecturerA LecturerB LecturerC LecturerD LecturerE LecturerF   
## 0 1 6 3 5 8 11   
## LecturerG Proposal Literature Quiz Interview   
## 11 0 0 0 0

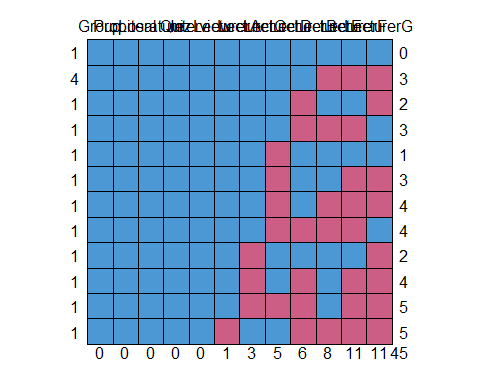
library(mice)

##   
## Attaching package: 'mice'

## The following object is masked from 'package:stats':  
##   
## filter

## The following objects are masked from 'package:base':  
##   
## cbind, rbind

md.pattern(data)



## Group Proposal Literature Quiz Interview LecturerA LecturerC LecturerD  
## 1 1 1 1 1 1 1 1 1  
## 4 1 1 1 1 1 1 1 1  
## 1 1 1 1 1 1 1 1 1  
## 1 1 1 1 1 1 1 1 1  
## 1 1 1 1 1 1 1 1 0  
## 1 1 1 1 1 1 1 1 0  
## 1 1 1 1 1 1 1 1 0  
## 1 1 1 1 1 1 1 1 0  
## 1 1 1 1 1 1 1 0 1  
## 1 1 1 1 1 1 1 0 1  
## 1 1 1 1 1 1 1 0 0  
## 1 1 1 1 1 1 0 1 1  
## 0 0 0 0 0 1 3 5  
## LecturerB LecturerE LecturerF LecturerG   
## 1 1 1 1 1 0  
## 4 1 0 0 0 3  
## 1 0 1 1 0 2  
## 1 0 0 0 1 3  
## 1 1 1 1 1 1  
## 1 1 1 0 0 3  
## 1 1 0 0 0 4  
## 1 0 0 0 1 4  
## 1 1 1 1 0 2  
## 1 0 1 0 0 4  
## 1 0 1 0 0 5  
## 1 0 0 0 0 5  
## 6 8 11 11 45

All the columns are numeric excerpt for the Group column. There are 45 missing values are observed in total. The missingness patterns given the visualizations, one can conclude that the missingness is MAR because the missing values in lecturer E are missing when lecturer F And G are missing missingness is dependent on the two variables also lecturer D only one is not dependent on G and F. This could be because the the groups have already been assessed

# 4.

library(tidyr)  
  
long\_data <- pivot\_longer(data, cols= c(LecturerA,LecturerB,LecturerC,LecturerD,LecturerE,LecturerF,LecturerG) ,names\_to = c("Lecturer"), values\_to = "Score")  
  
new\_data<- na.omit(long\_data)  
unique(new\_data$Lecturer)

## [1] "LecturerA" "LecturerC" "LecturerD" "LecturerG" "LecturerB" "LecturerE"  
## [7] "LecturerF"

# 5.

In our case the group of students is our Fixed effect because we not interested in how the next possible group will affect the final mark, with the lecturer as the random effect each group will experience the lecture effect and one would like to know how a different lecturer not included in this fit will grade each group.(read slides)

# 6.

The prior for the group intercepts, and intercept is a normal prior and also the sigma as a cauchy.

library(brms)

## Loading required package: Rcpp

## Loading 'brms' package (version 2.22.0). Useful instructions  
## can be found by typing help('brms'). A more detailed introduction  
## to the package is available through vignette('brms\_overview').

##   
## Attaching package: 'brms'

## The following object is masked from 'package:stats':  
##   
## ar

model <- brm(  
 formula = Score ~ Group + (1 | Lecturer),  
 data = new\_data,  
 prior = c(  
 set\_prior("normal(0, 10)", class = "b"),   
 set\_prior("normal(0, 5)", class = "Intercept"),  
 set\_prior("cauchy(0, 5)", class = "sd")   
 ),  
 iter = 5000  
)

## Compiling Stan program...

## Start sampling

summary(model)

## Family: gaussian   
## Links: mu = identity; sigma = identity   
## Formula: Score ~ Group + (1 | Lecturer)   
## Data: new\_data (Number of observations: 60)   
## Draws: 4 chains, each with iter = 5000; warmup = 2500; thin = 1;  
## total post-warmup draws = 10000  
##   
## Multilevel Hyperparameters:  
## ~Lecturer (Number of levels: 7)   
## Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS Tail\_ESS  
## sd(Intercept) 69.06 20.59 40.84 119.87 1.00 1382 2954  
##   
## Regression Coefficients:  
## Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS Tail\_ESS  
## Intercept 0.48 5.49 -10.20 11.24 1.00 7665 6305  
## GroupGroup10 7.40 3.12 1.10 13.52 1.00 5076 6638  
## GroupGroup11 6.27 3.15 0.01 12.41 1.00 4916 6058  
## GroupGroup12 -2.53 3.99 -10.42 5.34 1.00 7646 6798  
## GroupGroup13 6.75 3.14 0.52 12.81 1.00 5302 6132  
## GroupGroup14 2.51 3.02 -3.50 8.35 1.00 4867 5914  
## GroupGroup15 -11.87 3.99 -19.64 -4.14 1.00 6896 6774  
## GroupGroup2 -0.37 3.40 -7.30 6.34 1.00 5284 6523  
## GroupGroup3 14.70 3.17 8.43 20.98 1.00 5060 6988  
## GroupGroup4 -3.66 2.85 -9.33 1.84 1.00 4532 6004  
## GroupGroup5 8.54 2.69 3.26 13.80 1.00 4245 6002  
## GroupGroup6 4.64 3.13 -1.56 10.82 1.00 5272 6289  
## GroupGroup7 4.21 3.44 -2.60 10.93 1.00 6003 6154  
## GroupGroup8 -10.56 2.97 -16.38 -4.69 1.00 4246 5388  
## GroupGroup9 4.25 3.44 -2.54 10.95 1.00 5895 6776  
##   
## Further Distributional Parameters:  
## Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS Tail\_ESS  
## sigma 5.22 0.60 4.20 6.57 1.00 6369 6671  
##   
## Draws were sampled using sampling(NUTS). For each parameter, Bulk\_ESS  
## and Tail\_ESS are effective sample size measures, and Rhat is the potential  
## scale reduction factor on split chains (at convergence, Rhat = 1).

# 7.

fixed\_effects <- fixef(model, summary = TRUE)   
fit<-fitted(model)  
pred\_vals <- predict(model, summary = TRUE)  
data\_est<- cbind(new\_data$Group,fit,pred\_vals)  
data\_est<- data\_est[,c(-3,-6,-7)]  
colnames(data\_est) <- c("Groups","estimates","CI2.5", "CI97.5", "PI2.5","PI97.5")  
data\_est<- data.frame(data\_est)

# 8.

lecturer B is least biased

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

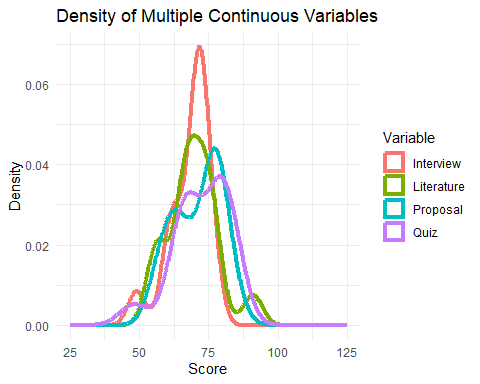
assessor\_biases <- ranef(model)$Lecturer %>%  
 as.data.frame() %>%  
 arrange(Estimate.Intercept)  
  
least\_biased <- assessor\_biases[which.min(abs(assessor\_biases$Estimate.Intercept)), ]  
least\_biased

## Estimate.Intercept Est.Error.Intercept Q2.5.Intercept Q97.5.Intercept  
## LecturerB 61.05436 5.52734 50.22019 71.47939

# 9.

library(tidyr)  
library(dplyr)  
  
long\_data\_mark<- data[,c(1,9,10,11,12)] %>%  
 pivot\_longer(cols = c(Proposal, Quiz,Literature,Interview),  
 names\_to = "mark\_type",  
 values\_to = "mark")  
library(ggplot2)  
  
ggplot(long\_data\_mark, aes(x = mark, color = mark\_type)) +  
 geom\_density(size = 1.5) +  
 labs(title = "Density of Multiple Continuous Variables",  
 x = "Score",  
 y = "Density",  
 color = "Variable") +  
 xlim(25, 125) +  
 theme\_minimal()

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



library(fitdistrplus)

## Loading required package: MASS

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

## Loading required package: survival

##   
## Attaching package: 'survival'

## The following object is masked from 'package:brms':  
##   
## kidney

historical\_data <- new\_data %>%  
 dplyr::select(Group, Proposal, Literature, Quiz, Interview) %>%  
 pivot\_longer(-Group, names\_to = "Component", values\_to = "Score") %>%  
 mutate(Score\_prop = Score / 100) %>%   
 group\_by(Group) %>%  
 summarise(  
 mean\_prop = mean(Score\_prop, na.rm = TRUE),  
 var\_prop = var(Score\_prop, na.rm = TRUE)  
 ) %>%  
 rowwise() %>%  
 mutate(  
   
 a = (mean\_prop^2 - mean\_prop^3 - mean\_prop \* var\_prop) / var\_prop,  
 b = (mean\_prop \* (1 - mean\_prop)^2 - (1 - mean\_prop) \* var\_prop) / var\_prop,  
 # Ensure valid parameters (avoid negatives)  
 a = pmax(a, 0.1),  
 b = pmax(b, 0.1),  
 # Convert back to score scale (0-100)  
 mean\_score = 100 \* (a / (a + b)),  
 sd\_score = 100 \* sqrt((a \* b) / ((a + b)^2 \* (a + b + 1)))  
 )  
  
priors <- historical\_data %>%  
 mutate(  
 coef = paste0("Group", Group),   
 prior = sprintf("normal(%s, %s)", round(mean\_score, 1), round(sd\_score, 1))  
 ) %>%  
 dplyr::select(coef, prior)  
  
  
priors\_list <- set\_prior(  
 priors$prior,   
 class = "b",   
 coef = priors$coef   
)

## Model

model\_beta\_prior <- brm(  
 bf(Score ~ 0+ Group + (1 | Lecturer)),  
 data = long\_data,  
 prior = c(  
 priors\_list,   
 set\_prior("cauchy(0, 5)", class = "sd"),   
 set\_prior("cauchy(0, 5)", class = "sigma")   
 ),  
   
)

## Compiling Stan program...

## Start sampling

## output and model comparison

summary(model\_beta\_prior)

## Family: gaussian   
## Links: mu = identity; sigma = identity   
## Formula: Score ~ 0 + Group + (1 | Lecturer)   
## Data: long\_data (Number of observations: 60)   
## Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;  
## total post-warmup draws = 4000  
##   
## Multilevel Hyperparameters:  
## ~Lecturer (Number of levels: 7)   
## Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS Tail\_ESS  
## sd(Intercept) 3.50 1.40 1.37 6.69 1.00 1696 2005  
##   
## Regression Coefficients:  
## Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS Tail\_ESS  
## GroupGroup1 63.12 2.16 58.85 67.48 1.00 5222 2986  
## GroupGroup10 75.22 2.34 70.60 79.72 1.00 4976 3307  
## GroupGroup11 73.94 2.12 69.80 78.23 1.00 4723 2913  
## GroupGroup12 64.14 3.40 57.42 70.90 1.00 4782 2977  
## GroupGroup13 73.60 2.40 68.84 78.14 1.00 4994 3294  
## GroupGroup14 70.95 2.41 66.34 75.57 1.00 5140 3069  
## GroupGroup15 55.47 3.56 48.46 62.65 1.00 4957 2933  
## GroupGroup2 66.81 2.94 61.03 72.50 1.00 5378 2933  
## GroupGroup3 79.91 2.17 75.57 84.16 1.00 4430 3072  
## GroupGroup4 63.91 2.21 59.51 68.28 1.00 4120 3128  
## GroupGroup5 76.47 2.13 72.22 80.66 1.00 3742 2398  
## GroupGroup6 71.16 2.37 66.49 75.79 1.00 5489 3121  
## GroupGroup7 70.59 3.02 64.36 76.50 1.00 4786 2834  
## GroupGroup8 57.40 2.46 52.58 62.10 1.00 4201 3031  
## GroupGroup9 72.97 2.96 67.23 78.88 1.00 4549 2909  
##   
## Further Distributional Parameters:  
## Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS Tail\_ESS  
## sigma 5.19 0.60 4.18 6.57 1.00 2718 2729  
##   
## Draws were sampled using sampling(NUTS). For each parameter, Bulk\_ESS  
## and Tail\_ESS are effective sample size measures, and Rhat is the potential  
## scale reduction factor on split chains (at convergence, Rhat = 1).

loo\_compare(loo(model),loo( model\_beta\_prior))

## elpd\_diff se\_diff  
## model\_beta\_prior 0.0 0.0   
## model -2.6 2.8

The density plots an the fact that they data is bounded one can assume that the data are beta distributed for each group thus it makes sense to use beta distribution as our subjective priors for each group. i used AI to fit the model and AI changed the beta distribution information to a normal distribution for each group. (OpenAI. (2024). ChatGPT (May 2024 version) [Large language model]. <https://chat.openai.com>)

And given the loo the subjective prior model fits much better than that of the vague prior.

# 10.

One can make the Group effect to be random to account for the difference in students, also more variables can be introduced like peer evaluations and maybe include give different roles to each student to contribute to the research project this will make it easy to assess individual student.

(OpenAI. (2024). ChatGPT (May 2024 version) [Large language model]. <https://chat.openai.com>)

# 11.

<https://github.com/mkhulekelinkosi/Bayes6.git>