University of Edinburgh

School of Mathematics

Bayesian Data Analysis, 2022/2023, Semester 2

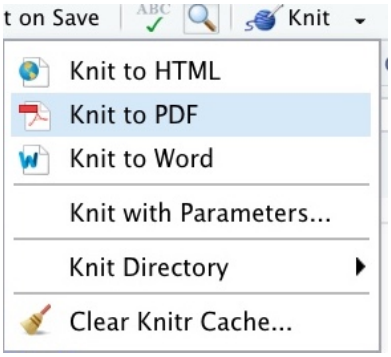
Assignment 2

IMPORTANT INFORMATION ABOUT THE ASSIGNMENT

In this paragraph, we summarize the essential information about this assignment. The format and rules for this assignment are different from your other courses, so please pay attention.

1) Deadline: The deadline for submitting your solutions to this assignment is the 17 March 12:00 noon Edinburgh time.

2) Format: You will need to submit your work as 2 components: a PDF report, and your R Markdown (.Rmd) notebook. There will be two separate submission systems on Learn: Gradescope for the report in PDF format, and a Learn assignment for the code in Rmd format. You need to write your solutions into this R Markdown notebook (code in R chunks and explanations in Markdown chunks), and then select Knit/Knit to PDF in RStudio to create a PDF report.



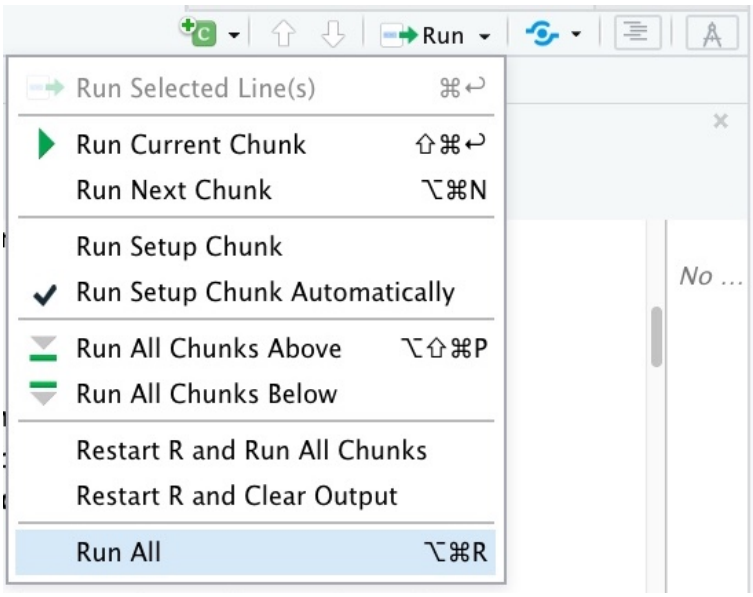
The compiled PDF needs to contain everything in this notebook, with your code sections clearly visible (not hidden), and the output of your code included. Reports without the code displayed in the PDF, or without the output of your code included in the PDF will be marked as 0, with the only feedback "Report did not meet submission requirements".

You need to upload this PDF in Gradescope submission system, and your Rmd file in the Learn assignment submission system. You will be required to tag every sub question on Gradescope.

Some key points that are different from other courses:

a) Your report needs to contain written explanation for each question that you solve, and some numbers or plots showing your results. Solutions without written explanation that clearly demonstrates that you understand what you are doing will be marked as 0 irrespectively whether the numerics are correct or not.

b) Your code has to be possible to run for all questions by the Run All in RStudio, and reproduce all of the numerics and plots in your report (up to some small randomness due to stochasticity of Monte Carlo simulations). The parts of the report that contain material that is not reproduced by the code will not be marked (i.e. the score will be 0), and the only feedback in this case will be that the results are not reproducible from the code.



c) Multiple Submissions are allowed BEFORE THE DEADLINE are allowed for both the report, and the code.
However, multiple submissions are NOT ALLOWED AFTER THE DEADLINE.
YOU WILL NOT BE ABLE TO MAKE ANY CHANGES TO YOUR SUBMISSION AFTER THE DEADLINE.
Nevertheless, if you did not submit anything before the deadline, then you can still submit your work after the deadline, but late penalties will apply. The timing of the late penalties will be determined by the time you have submitted BOTH the report, and the code (i.e. whichever was submitted later counts).

We illustrate these rules by some examples:

Alice has spent a lot of time and effort on her assignment for BDA. Unfortunately she has accidentally introduced a typo in her code in the first question, and it did not run using Run All in RStudio. - Alice will get 0 for the whole assignment, with the only feedback "Results are not reproducible from the code".

Bob has spent a lot of time and effort on his assignment for BDA. Unfortunately he forgot to submit his code. - Bob will get no personal reminder to submit his code. Bob will get 0 for the whole assignment, with the only feedback "Results are not reproducible from the code, as the code was not submitted."

Charles has spent a lot of time and effort on his assignment for BDA. He has submitted both his code and report in the correct formats. However, he did not include any explanations in the report. Charles will get 0 for the whole assignment, with the only feedback "Explanation is missing."

Denise has spent a lot of time and effort on her assignment for BDA. She has submitted her report in the correct format, but thought that she can include her code as a link in the report, and upload it online (such as Github, or Dropbox). - Denise will get 0 for the whole assignment, with the only feedback "Code was not uploaded on Learn."

3) Group work: This is an INDIVIDUAL ASSIGNMENT, like a 2 week exam for the course. Communication between students about the assignment questions is not permitted. Students who submit work that has not been done individually will be reported for Academic Misconduct, that can lead to serious consequences. Each problem will be marked by a single instructor, so we will be able to spot students who copy.

4) Piazza: During the periods of the assignments, the instructor will change Piazza to allow messaging the instructors only, i.e. students will not see each others messages and replies.

Only questions regarding clarification of the statement of the problems will be answered by the instructors. The instructors will not give you any information related to the solution of the problems, such questions will be simply answered as "This is not about the statement of the problem so we cannot answer your question."

THE INSTRUCTORS ARE NOT GOING TO DEBUG YOUR CODE, AND YOU ARE ASSESSED ON YOUR ABILITY TO RESOLVE ANY CODING OR TECHNICAL DIFFICULTIES THAT YOU ENCOUNTER ON YOUR OWN.

5) Office hours: There will be two office hours per week (Monday 14:00-15:00, and Wednesdays 15:00-16:00) during the 2 weeks for this assignment. The links are available on Learn / Course Information. I will be happy to discuss the course/workshop materials. However, I will only answer questions about the assignment that require clarifying the statement of the problems, and will not give you any information about the solutions. Students who ask for feedback on their assignment solutions during office hours will be removed from the meeting.

6) Late submissions and extensions: NO EXTENSIONS ARE ALLOWED FOR THIS ASSIGNMENT, AND THERE IS NO SUCH OPTION PROVIDED IN THE ESC SYSTEM. Students who have existing Learning Adjustments in Euclid will be allowed to have the same adjustments applied to this course as well, but they need to apply for this BEFORE THE DEADLINE on the website

https://www.ed.ac.uk/student-administration/extensions-special-circumstances

by clicking on "Access your learning adjustment". This will be approved automatically.

Students who submit their work late will have late submission penalties applied by the ESC team automatically (this means that even if you are 1 second late because of your internet connection was slow, the penalties will still apply). The penalties are 5% of the total mark deduced for every day of delay started (i.e. one minute of delay counts for 1 day). The course instructors do not have any role in setting these penalties, we will not be able to change them.

7) Please make sure to tag all pages in your submission on Gradescope, otherwise we may miss some of your work. Once your upload is complete, tagging does not counts towards your submission time (i.e. you won't get any late penalties for doing it).

```
rm(list = ls(all = TRUE))
#Do not delete this!
#It clears all variables to ensure reproducibility
```

Problem 1

In this problem, we study a dataset about car insurance. This data set is based on one-year vehicle insurance policies taken out in 2004 or 2005. In total, there are 67856 policies, of which 4624 have claims.

```
require(insuranceData)
```

```
## Loading required package: insuranceData
```

```
data(dataCar)

#You may need to set the working directory first before loading the dataset
#setwd("location of Assignment 1")
#The first 6 rows of the dataframe
print.data.frame(dataCar[1:6,])
```

```
##   veh_value  exposure clm numclaims claimcst0 veh_body veh_age gender area
## 1      1.06 0.3039014   0         0         0    HBACK       3      F    C
## 2      1.03 0.6488706   0         0         0    HBACK       2      F    A
## 3      3.26 0.5694730   0         0         0      UTE       2      F    E
## 4      4.14 0.3175907   0         0         0    STNWG       2      F    D
## 5      0.72 0.6488706   0         0         0    HBACK       4      F    C
## 6      2.01 0.8542094   0         0         0    HDTOP       3      M    C
##   agecat            X_OBSTAT_
## 1      2 01101    0    0    0
```

```
## 2      4 01101    0    0    0
## 3      2 01101    0    0    0
## 4      2 01101    0    0    0
## 5      2 01101    0    0    0
## 6      4 01101    0    0    0
```

Description of the columns.

veh_value: vehicle value in $10000s

exposure: maximum portion of the vehicle value the insurer may need to pay out in case of an incident

claimcst0: claim amount (0 if no claim)

clm: whether there was a claim during the 1 year duration

numclaims: number of claims during the 1 year duration

veh_body types: BUS = bus CONVT = convertible COUPE = coupe HBACK = hatchback HDTOP = hardtop MCARA = motorized caravan MIBUS = minibus PANVN = panel van RDSTR = roadster SEDAN = sedan STNWG = station wagon TRUCK = truck UTE = utility

gender: F- female, M - male

area: a factor with levels A,B,C,D,E, F

agecat: age category, 1 (youngest), 2, 3, 4, 5, 6

You can use either JAGS, Stan, or INLA for this question.

a)[10 marks] Fit a Bayesian logistic regression model on the dataset dataCar with

- **clm as response,**
- **a link function of your choice,**
- **using veh_value, exposure, veh_body, veh_age, gender, area, and agecat as covariates (you can use categorical covariates by converting integers to factors if appropriate).**

Center and scale the non-categorical covariates.

Choose your own prior distributions (do not use default priors), and explain the rationale your prior choices, and ensure that the posterior is not too sensitive to your prior choice [Hint: look at the induced prior on the linear predictor and on the response.]

Compute the posterior means of the model parameters, and discuss the results.

```r
# Center and scale the non-categorical covariates
dataCar$veh_value <- scale(dataCar$veh_value)[,1]
dataCar$exposure <- scale(dataCar$exposure)[,1]
# Convert integers to factors for categorical covariates
dataCar$veh_age <- as.factor(dataCar$veh_age)
dataCar$agecat <- as.factor(dataCar$agecat)

str(dataCar)
```

```
## 'data.frame':    67856 obs. of  11 variables:
##  $ veh_value: num  -0.595 -0.62 1.23 1.961 -0.877 ...
##  $ exposure : num  -0.568 0.621 0.348 -0.521 0.621 ...
##  $ clm      : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ numclaims: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ claimcst0: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ veh_body : Factor w/ 13 levels "BUS","CONVT",..: 4 4 13 11 4 5 8 4 4 4 ...
##  $ veh_age  : Factor w/ 4 levels "1","2","3","4": 3 2 2 2 4 3 3 2 4 4 ...
##  $ gender   : Factor w/ 2 levels "F","M": 1 1 1 1 1 2 2 2 1 1 ...
##  $ area     : Factor w/ 6 levels "A","B","C","D",..: 3 1 5 4 3 3 1 2 1 2 ...
##  $ agecat   : Factor w/ 6 levels "1","2","3","4",..: 2 4 2 2 2 4 4 6 3 4 ...
##  $ X_OBSTAT_: Factor w/ 1 level "01101    0    0    0": 1 1 1 1 1 1 1 1 1 1 ...
```

```
#prior
library(INLA)
```

```
## Loading required package: Matrix

## Loading required package: foreach

## Loading required package: parallel

## Loading required package: sp

## This is INLA_22.12.16 built 2022-12-23 13:36:23 UTC.
##  - See www.r-inla.org/contact-us for how to get help.
##  - To enable PARDISO sparse library; see inla.pardiso()
```

```
mm = model.matrix(clm~0+veh_value + exposure + veh_body +
  veh_age + gender + area + agecat ,data=dataCar)
var.beta=25/quantile(rowSums(mm^2),0.05)

formula_clm <- clm ~ veh_value + exposure + veh_body +
  veh_age + gender + area + agecat

prior.beta <- list(mean.intercept = 0, prec.intercept = 4e-2,
                   mean = 0, prec = var.beta)
prior.beta2 <- list(mean.intercept = 0, prec.intercept = 0.01,
                   mean = 0, prec = 5)

model <- inla(formula_clm, family = "binomial", data = dataCar,
              control.fixed = prior.beta,
              control.family = list(link = "logit"),
              control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
              )
model_test <- inla(formula_clm, family = "binomial", data = dataCar,
              control.fixed = prior.beta2,
              control.family = list(link = "logit"),
              control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
              )
summary(model)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.85, Running = 2.8, Post = 0.754, Total = 7.41
## Fixed effects:
##                 mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept)   -2.367 0.134     -2.629   -2.367     -2.106 -2.367   0
## veh_value      0.033 0.022     -0.009    0.033      0.076  0.033   0
## exposure       0.539 0.016      0.507    0.539      0.570  0.539   0
## veh_bodyCONVT -0.298 0.290     -0.867   -0.298      0.271 -0.298   0
## veh_bodyCOUPE  0.255 0.159     -0.056    0.255      0.567  0.255   0
## veh_bodyHBACK -0.128 0.120     -0.363   -0.128      0.106 -0.128   0
## veh_bodyHDTOP  0.037 0.140     -0.237    0.037      0.311  0.037   0
## veh_bodyMCARA  0.291 0.245     -0.189    0.291      0.772  0.291   0
## veh_bodyMIBUS -0.144 0.173     -0.483   -0.144      0.196 -0.144   0
## veh_bodyPANVN -0.052 0.161     -0.368   -0.052      0.264 -0.052   0
## veh_bodyRDSTR -0.026 0.324     -0.661   -0.026      0.608 -0.026   0
## veh_bodySEDAN -0.107 0.118     -0.338   -0.107      0.124 -0.107   0
## veh_bodySTNWG -0.070 0.117     -0.300   -0.070      0.161 -0.070   0
## veh_bodyTRUCK -0.140 0.141     -0.417   -0.140      0.137 -0.140   0
## veh_bodyUTE   -0.289 0.128     -0.539   -0.289     -0.038 -0.289   0
## veh_age2       0.068 0.047     -0.025    0.068      0.161  0.068   0
## veh_age3      -0.035 0.051     -0.134   -0.035      0.065 -0.035   0
## veh_age4      -0.109 0.059     -0.225   -0.109      0.008 -0.109   0
## genderM       -0.018 0.032     -0.081   -0.018      0.046 -0.018   0
## areaB          0.089 0.046      0.000    0.089      0.179  0.089   0
## areaC          0.037 0.042     -0.045    0.037      0.119  0.037   0
## areaD         -0.087 0.056     -0.197   -0.087      0.022 -0.087   0
## areaE         -0.013 0.061     -0.133   -0.013      0.108 -0.013   0
## areaF          0.077 0.071     -0.062    0.077      0.216  0.077   0
## agecat2       -0.180 0.057     -0.293   -0.180     -0.068 -0.180   0
## agecat3       -0.229 0.056     -0.338   -0.229     -0.119 -0.229   0
## agecat4       -0.261 0.056     -0.370   -0.261     -0.151 -0.261   0
## agecat5       -0.472 0.062     -0.593   -0.472     -0.350 -0.472   0
## agecat6       -0.465 0.071     -0.603   -0.465     -0.326 -0.465   0
##
## Deviance Information Criterion (DIC) ...............: 32457.17
```

```
## Deviance Information Criterion (DIC, saturated) ....: 32457.17
## Effective number of parameters .....................: 25.32
##
## Marginal log-Likelihood:  -16263.26
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

```
summary(model_test)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.58, Running = 2.45, Post = 0.431, Total = 6.46
## Fixed effects:
##                 mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept)   -2.333 0.154     -2.634   -2.333     -2.032 -2.333   0
## veh_value      0.033 0.022     -0.010    0.033      0.076  0.033   0
## exposure       0.539 0.016      0.508    0.539      0.570  0.539   0
## veh_bodyCONVT -0.397 0.335     -1.054   -0.397      0.259 -0.397   0
## veh_bodyCOUPE  0.242 0.177     -0.105    0.242      0.589  0.242   0
## veh_bodyHBACK -0.155 0.141     -0.432   -0.155      0.123 -0.155   0
## veh_bodyHDTOP  0.013 0.159     -0.299    0.013      0.325  0.013   0
## veh_bodyMCARA  0.323 0.269     -0.204    0.323      0.850  0.323   0
## veh_bodyMIBUS -0.175 0.192     -0.551   -0.175      0.202 -0.175   0
## veh_bodyPANVN -0.079 0.180     -0.431   -0.079      0.274 -0.079   0
## veh_bodyRDSTR -0.048 0.381     -0.796   -0.048      0.699 -0.048   0
## veh_bodySEDAN -0.133 0.140     -0.406   -0.133      0.141 -0.133   0
## veh_bodySTNWG -0.095 0.139     -0.368   -0.095      0.178 -0.095   0
## veh_bodyTRUCK -0.169 0.161     -0.484   -0.169      0.146 -0.169   0
## veh_bodyUTE   -0.318 0.148     -0.608   -0.318     -0.027 -0.318   0
## veh_age2       0.068 0.048     -0.026    0.068      0.161  0.068   0
## veh_age3      -0.035 0.051     -0.136   -0.035      0.065 -0.035   0
## veh_age4      -0.109 0.060     -0.227   -0.109      0.009 -0.109   0
## genderM       -0.018 0.032     -0.081   -0.018      0.046 -0.018   0
## areaB          0.090 0.046      0.000    0.090      0.180  0.090   0
## areaC          0.037 0.042     -0.045    0.037      0.119  0.037   0
```

```
## areaD         -0.088 0.056     -0.198   -0.088      0.023 -0.088   0
## areaE         -0.012 0.062     -0.133   -0.012      0.109 -0.012   0
## areaF          0.078 0.071     -0.062    0.078      0.218  0.078   0
## agecat2       -0.190 0.058     -0.304   -0.190     -0.077 -0.190   0
## agecat3       -0.238 0.056     -0.349   -0.238     -0.128 -0.238   0
## agecat4       -0.271 0.056     -0.381   -0.271     -0.160 -0.271   0
## agecat5       -0.483 0.063     -0.606   -0.483     -0.360 -0.483   0
## agecat6       -0.478 0.072     -0.618   -0.478     -0.337 -0.478   0
##
## Deviance Information Criterion (DIC) ...............: 32457.23
## Deviance Information Criterion (DIC, saturated) ....: 32457.23
## Effective number of parameters .....................: 25.90
##
## Marginal log-Likelihood:  -16267.27
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation): This study use these covariates, including veh_value, exposure, veh_body, veh_age, gender, area, and agecat. Also, we used clm as response, and then established Bayesian logistic regression model with logit link function. Firstly, variables are divided into discrete and continuous types for different preprocessing. By observing the data structure, it can be determined that veh_body, veh_age, gender, area, and agecat are discrete variables, so the as.factor function is used to convert the content of the variables into different levels. On the other hand, veh_value and exposure are continuous variables. To eliminate the influence of dimensions and improve interpretability, the scale function is used to center and scale these non-categorical covariates. When setting the prior, I use quantile function to decide the precision of beta .To ensure that the posterior is not too sensitive to your prior choice, I conducted a sensitivity test. By comparing models established using different priors, it was found that the posterior means of different variables were basically similar, and the indicators of fit such as DIC were also similar, indicating that the model is not sensitive to the prior setting. In the posterior mean analysis of the different variables, the top three variables in absolute value of the coefficients are selected here in order to avoid the interference caused by the large number of variables. Exposure with a posterior mean of 0.54, which has a positive association with the likelihood of a claim. This suggests that a higher exposure (i.e., more time spent driving or being at risk) increases the probability of filing a claim during the 1-year period. Also, the agecat5 and agecat6 variable has a posterior mean about -0.45, indicating a negative association with the likelihood of a claim. This suggests that individuals in age category 5 and 6(typically older drivers) are less likely to file a claim during the 1-year period compared to the reference age group. This could be due to their more extensive driving experience or more cautious driving habits.

b)[10 marks] Fit a Bayesian Poisson regression model on numclaims as response with

- **log link function,**
- **using veh_value, exposure, veh_body, veh_age, gender, area, and agecat as covariates.**

Center and scale the non-categorical covariates.

Choose your own prior distributions (do not use default priors), and explain the rationale your prior choices, and ensure that the posterior is not too sensitive to your prior choice [Hint: look at the induced prior on the linear predictor and the response.]

Compute the posterior means of the model parameters, and discuss the results.

```
formula_numclaims <- numclaims ~ veh_value + exposure +
  veh_body + veh_age + gender + area + agecat
prior.beta <- list(mean.intercept = 0, prec.intercept = 1/log(4)^2,
                   mean = 0, prec = 1/(log(5)/2)^2)
prior.beta2 <- list(mean.intercept = 0, prec.intercept = 1,
                   mean = 0, prec = 1)
model_poisson <- inla(formula_numclaims, family = "poisson", data = dataCar,
                      control.fixed = prior.beta,
                      control.family = list(link = "log"),
                      control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
                      )
model_poisson_test <- inla(formula_numclaims, family = "poisson", data = dataCar,
                      control.fixed = prior.beta2,
                      control.family = list(link = "log"),
                      control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
                      )
summary(model_poisson)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.49, Running = 1.53, Post = 0.506, Total = 5.52
## Fixed effects:
##                 mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept)   -2.125 0.217     -2.550   -2.125     -1.700 -2.125   0
## veh_value      0.029 0.020     -0.011    0.029      0.069  0.029   0
## exposure       0.522 0.015      0.493    0.522      0.551  0.522   0
## veh_bodyCONVT -0.850 0.449     -1.730   -0.850      0.029 -0.850   0
## veh_bodyCOUPE  0.045 0.233     -0.412    0.045      0.502  0.045   0
## veh_bodyHBACK -0.388 0.210     -0.800   -0.388      0.024 -0.388   0
## veh_bodyHDTOP -0.242 0.222     -0.676   -0.242      0.193 -0.242   0
## veh_bodyMCARA  0.164 0.311     -0.446    0.164      0.774  0.164   0
## veh_bodyMIBUS -0.413 0.248     -0.899   -0.413      0.073 -0.413   0
## veh_bodyPANVN -0.284 0.236     -0.746   -0.284      0.178 -0.284   0
## veh_bodyRDSTR -0.070 0.488     -1.027   -0.070      0.888 -0.070   0
## veh_bodySEDAN -0.335 0.209     -0.745   -0.335      0.075 -0.335   0
```

```
## veh_bodySTNWG -0.320 0.209     -0.729   -0.320      0.089 -0.320   0
## veh_bodyTRUCK -0.373 0.223     -0.809   -0.373      0.063 -0.373   0
## veh_bodyUTE   -0.528 0.215     -0.948   -0.528     -0.107 -0.528   0
## veh_age2       0.068 0.044     -0.020    0.068      0.155  0.068   0
## veh_age3      -0.042 0.048     -0.136   -0.042      0.052 -0.042   0
## veh_age4      -0.102 0.056     -0.212   -0.102      0.009 -0.102   0
## genderM       -0.023 0.030     -0.082   -0.023      0.036 -0.023   0
## areaB          0.053 0.043     -0.031    0.053      0.136  0.053   0
## areaC          0.004 0.039     -0.072    0.004      0.080  0.004   0
## areaD         -0.111 0.053     -0.215   -0.111     -0.008 -0.111   0
## areaE         -0.028 0.058     -0.141   -0.028      0.085 -0.028   0
## areaF          0.071 0.066     -0.058    0.071      0.200  0.071   0
## agecat2       -0.174 0.054     -0.280   -0.174     -0.068 -0.174   0
## agecat3       -0.223 0.053     -0.326   -0.223     -0.120 -0.223   0
## agecat4       -0.249 0.052     -0.351   -0.249     -0.146 -0.249   0
## agecat5       -0.465 0.059     -0.580   -0.465     -0.349 -0.465   0
## agecat6       -0.450 0.067     -0.582   -0.450     -0.319 -0.450   0
##
## Deviance Information Criterion (DIC) ...............: 34789.30
## Deviance Information Criterion (DIC, saturated) ....: 25354.60
## Effective number of parameters .....................: 27.27
##
## Marginal log-Likelihood:  -17446.78
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

```
summary(model_poisson_test)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.37, Running = 1.51, Post = 0.294, Total = 5.17
## Fixed effects:
##                mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept)  -1.968 0.238     -2.435   -1.968     -1.500 -1.968   0
## veh_value     0.029 0.020     -0.011    0.029      0.069  0.029   0
```

```
## exposure       0.522 0.015      0.493    0.522      0.551  0.522   0
## veh_bodyCONVT -1.073 0.490     -2.034   -1.073     -0.113 -1.073   0
## veh_bodyCOUPE -0.106 0.255     -0.605   -0.106      0.393 -0.106   0
## veh_bodyHBACK -0.541 0.233     -0.997   -0.541     -0.084 -0.541   0
## veh_bodyHDTOP -0.394 0.244     -0.872   -0.394      0.084 -0.394   0
## veh_bodyMCARA  0.027 0.331     -0.622    0.027      0.675  0.027   0
## veh_bodyMIBUS -0.567 0.269     -1.094   -0.567     -0.039 -0.567   0
## veh_bodyPANVN -0.437 0.257     -0.940   -0.437      0.067 -0.437   0
## veh_bodyRDSTR -0.203 0.523     -1.228   -0.203      0.822 -0.203   0
## veh_bodySEDAN -0.488 0.232     -0.942   -0.488     -0.033 -0.488   0
## veh_bodySTNWG -0.473 0.232     -0.927   -0.473     -0.018 -0.473   0
## veh_bodyTRUCK -0.526 0.245     -1.005   -0.526     -0.047 -0.526   0
## veh_bodyUTE   -0.681 0.237     -1.146   -0.681     -0.216 -0.681   0
## veh_age2       0.067 0.045     -0.021    0.067      0.154  0.067   0
## veh_age3      -0.043 0.048     -0.137   -0.043      0.051 -0.043   0
## veh_age4      -0.103 0.056     -0.214   -0.103      0.007 -0.103   0
## genderM       -0.024 0.030     -0.083   -0.024      0.035 -0.024   0
## areaB          0.052 0.043     -0.032    0.052      0.136  0.052   0
## areaC          0.003 0.039     -0.073    0.003      0.079  0.003   0
## areaD         -0.112 0.053     -0.216   -0.112     -0.009 -0.112   0
## areaE         -0.029 0.058     -0.142   -0.029      0.084 -0.029   0
## areaF          0.070 0.066     -0.059    0.070      0.199  0.070   0
## agecat2       -0.177 0.054     -0.283   -0.177     -0.071 -0.177   0
## agecat3       -0.226 0.053     -0.330   -0.226     -0.123 -0.226   0
## agecat4       -0.252 0.052     -0.355   -0.252     -0.149 -0.252   0
## agecat5       -0.468 0.059     -0.583   -0.468     -0.353 -0.468   0
## agecat6       -0.454 0.067     -0.586   -0.454     -0.322 -0.454   0
##
## Deviance Information Criterion (DIC) ...............: 34788.54
## Deviance Information Criterion (DIC, saturated) ....: 25353.84
## Effective number of parameters .....................: 27.59
##
## Marginal log-Likelihood:  -17452.35
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation): This part use same covariates as above. Also, we used numclaims as response, and then established Bayesian Poisson model with log link function.
For the prior selection, we can see that the usual values for the number of sadness are between 0 and 4, then we can beta0 on the interval [- ln(4), ln(4)].For the covariates, the most widely spaced variable is exposure, mainly distributed between -2 and 2, so max |xi - mean(x)| = 2. then beta would be in [-ln(5)/2, ln(5)/2]. To ensure that the posterior is not too sensitive to your prior choice, I conducted a sensitivity test. By comparing models established using different priors, it was found that the posterior means of different variables were basically similar, and the indicators of fit such as DIC were also similar, indicating that the model is not sensitive to the prior setting. In the posterior mean analysis of the different variables, the top three variables in absolute value of the coefficients are selected here in order to avoid the interference caused by the large number of variables. 'veh_bodyCONVT' with a posterior mean of -0.85, which indicates that convertibles are associated with fewer claims compared to the reference category, This could be due to various factors, such as convertible owners being more cautious drivers or driving less often, leading to fewer accidents and claims. The posterior mean of exposure is 0.52, which suggests a positive association between the exposure variable and the number of claims, and implies that as the exposure increases, the number of claims is also likely to increase, which is a reasonable expectation. 'veh_bodyUTE' with a posterior mean

of -0.53, which indicates that utility vehicles are associated with fewer claims compared to the reference category. This might be attributed to utility vehicle owners using their vehicles for specific purposes or driving less frequently, leading to a lower probability of accidents and claims.

c)[10 marks] Fit a zero-inflated Bayesian Poisson regression model (https://en.wikipedia.org/wiki/Zero-inflated_model) on

- **numclaims as response,**
- **with log link function,**
- **using veh_value, exposure, veh_body, veh_age, gender, area, and agecat as covariates.**

Center and scale the non-categorical covariates.

Choose your own prior distributions (do not use default priors), and explain the rationale your prior choices, and ensure that the posterior is not too sensitive to your prior choice [Hint: look at the induced prior on the linear predictor and the response.]

Compute the posterior means of the model parameters, and discuss the results.

```
formula_numclaims <- numclaims ~ veh_value + exposure +
  veh_body + veh_age + gender + area + agecat
prior.beta <- list(mean.intercept = 0, prec.intercept = 1/log(4)^2,
                   mean = 0, prec = 1/(log(5)/2)^2)
prior.beta2 <- list(mean.intercept = 0, prec.intercept = 1,
                   mean = 0, prec = 1)
model_zero <- inla(formula_numclaims,
                        family = "zeroinflatedpoisson1",
                        data = dataCar,
                        control.fixed = prior.beta,
                        control.family = list(link = "log"),
                        control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
)
model_zero_test <- inla(formula_numclaims,
                        family = "zeroinflatedpoisson1",
                        data = dataCar,
                        control.fixed = prior.beta2,
                        control.family = list(link = "log"),
                        control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
)
summary(model_zero)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
```

```
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.54, Running = 21.3, Post = 0.414, Total = 25.3
## Fixed effects:
##                 mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept)   -1.795 0.223     -2.234   -1.795     -1.358 -1.795   0
## veh_value      0.030 0.021     -0.011    0.030      0.072  0.030   0
## exposure       0.524 0.015      0.494    0.524      0.553  0.524   0
## veh_bodyCONVT -0.823 0.455     -1.715   -0.823      0.069 -0.823   0
## veh_bodyCOUPE  0.066 0.235     -0.395    0.066      0.526  0.066   0
## veh_bodyHBACK -0.365 0.211     -0.778   -0.365      0.048 -0.365   0
## veh_bodyHDTOP -0.224 0.223     -0.661   -0.224      0.213 -0.224   0
## veh_bodyMCARA  0.179 0.317     -0.442    0.179      0.800  0.179   0
## veh_bodyMIBUS -0.396 0.250     -0.886   -0.396      0.094 -0.396   0
## veh_bodyPANVN -0.263 0.237     -0.729   -0.263      0.203 -0.263   0
## veh_bodyRDSTR -0.045 0.498     -1.021   -0.045      0.931 -0.045   0
## veh_bodySEDAN -0.313 0.210     -0.724   -0.313      0.098 -0.313   0
## veh_bodySTNWG -0.301 0.209     -0.711   -0.301      0.110 -0.301   0
## veh_bodyTRUCK -0.354 0.224     -0.793   -0.354      0.084 -0.354   0
## veh_bodyUTE   -0.508 0.215     -0.930   -0.508     -0.086 -0.508   0
## veh_age2       0.071 0.045     -0.018    0.071      0.160  0.071   0
## veh_age3      -0.037 0.049     -0.133   -0.037      0.059 -0.037   0
## veh_age4      -0.095 0.058     -0.207   -0.095      0.018 -0.095   0
## genderM       -0.023 0.031     -0.084   -0.023      0.037 -0.023   0
## areaB          0.053 0.044     -0.033    0.053      0.138  0.053   0
## areaC          0.004 0.040     -0.073    0.004      0.082  0.004   0
## areaD         -0.110 0.054     -0.215   -0.110     -0.004 -0.110   0
## areaE         -0.025 0.059     -0.140   -0.025      0.091 -0.025   0
## areaF          0.073 0.067     -0.059    0.073      0.205  0.073   0
## agecat2       -0.176 0.055     -0.284   -0.176     -0.068 -0.176   0
## agecat3       -0.224 0.054     -0.329   -0.224     -0.118 -0.224   0
## agecat4       -0.249 0.054     -0.354   -0.249     -0.144 -0.249   0
## agecat5       -0.466 0.060     -0.584   -0.466     -0.348 -0.466   0
## agecat6       -0.452 0.069     -0.587   -0.452     -0.318 -0.452   0
##
## Model hyperparameters:
##                                                          mean    sd 0.025quant
## zero-probability parameter for zero-inflated poisson_1 0.294 0.038      0.215
##                                                        0.5quant 0.975quant mode
## zero-probability parameter for zero-inflated poisson_1    0.296      0.365 0.30
##
## Deviance Information Criterion (DIC) ...............: 34753.99
## Deviance Information Criterion (DIC, saturated) ....: NA
## Effective number of parameters .....................: 28.28
##
## Marginal log-Likelihood:  -17429.68
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
```

```
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

```
summary(model_zero_test)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.21, Running = 22, Post = 0.42, Total = 25.6
## Fixed effects:
##                 mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept)   -1.651 0.245     -2.131   -1.651     -1.171 -1.651   0
## veh_value      0.030 0.021     -0.011    0.030      0.072  0.030   0
## exposure       0.524 0.015      0.494    0.524      0.553  0.524   0
## veh_bodyCONVT -1.034 0.497     -2.009   -1.034     -0.059 -1.034   0
## veh_bodyCOUPE -0.071 0.257     -0.575   -0.071      0.434 -0.071   0
## veh_bodyHBACK -0.504 0.235     -0.964   -0.504     -0.044 -0.504   0
## veh_bodyHDTOP -0.362 0.246     -0.845   -0.362      0.120 -0.362   0
## veh_bodyMCARA  0.056 0.337     -0.605    0.056      0.718  0.056   0
## veh_bodyMIBUS -0.536 0.272     -1.069   -0.536     -0.003 -0.536   0
## veh_bodyPANVN -0.402 0.260     -0.911   -0.402      0.108 -0.402   0
## veh_bodyRDSTR -0.162 0.535     -1.211   -0.162      0.887 -0.162   0
## veh_bodySEDAN -0.451 0.234     -0.909   -0.451      0.007 -0.451   0
## veh_bodySTNWG -0.439 0.233     -0.897   -0.439      0.019 -0.439   0
## veh_bodyTRUCK -0.493 0.247     -0.977   -0.493     -0.009 -0.493   0
## veh_bodyUTE   -0.647 0.239     -1.116   -0.647     -0.179 -0.647   0
## veh_age2       0.070 0.045     -0.019    0.070      0.159  0.070   0
## veh_age3      -0.038 0.049     -0.134   -0.038      0.058 -0.038   0
## veh_age4      -0.096 0.058     -0.209   -0.096      0.017 -0.096   0
## genderM       -0.024 0.031     -0.084   -0.024      0.036 -0.024   0
## areaB          0.052 0.044     -0.033    0.052      0.138  0.052   0
## areaC          0.004 0.040     -0.074    0.004      0.082  0.004   0
## areaD         -0.111 0.054     -0.216   -0.111     -0.005 -0.111   0
## areaE         -0.026 0.059     -0.141   -0.026      0.090 -0.026   0
## areaF          0.071 0.067     -0.061    0.071      0.204  0.071   0
## agecat2       -0.179 0.055     -0.287   -0.179     -0.071 -0.179   0
## agecat3       -0.227 0.054     -0.333   -0.227     -0.121 -0.227   0
## agecat4       -0.252 0.054     -0.357   -0.252     -0.147 -0.252   0
```

```
## agecat5       -0.469 0.060     -0.587   -0.469     -0.352 -0.469   0
## agecat6       -0.456 0.069     -0.591   -0.456     -0.321 -0.456   0
##
## Model hyperparameters:
##                                                          mean    sd 0.025quant
## zero-probability parameter for zero-inflated poisson_1 0.295 0.038      0.217
##                                                        0.5quant 0.975quant
## zero-probability parameter for zero-inflated poisson_1    0.297      0.366
##                                                          mode
## zero-probability parameter for zero-inflated poisson_1 0.301
##
## Deviance Information Criterion (DIC) ...............: 34753.40
## Deviance Information Criterion (DIC, saturated) ....: NA
## Effective number of parameters .....................: 28.60
##
## Marginal log-Likelihood:  -17435.01
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation): This part use same covariates as above. Also, we used numclaims as response, and then established Bayesian zero-inflated model with log link function. Since the target variable and covariates are the same as b, the same prior as above is used. To ensure that the posterior is not too sensitive to your prior choice, I conducted a sensitivity test. By comparing models established using different priors, it was found that the posterior means of different variables were basically similar, and the indicators of fit such as DIC were also similar, indicating that the model is not sensitive to the prior setting. In the posterior mean analysis of the different variables, the top three variables in absolute value of the coefficients are selected here in order to avoid the interference caused by the large number of variables. 'veh_bodyCONVT' with a posterior mean of -0.82, which indicates that convertibles are associated with fewer claims compared to the reference category, This could be due to various factors, such as convertible owners being more cautious drivers or driving less often, leading to fewer accidents and claims. The posterior mean of exposure is 0.52, which suggests a positive association between the exposure variable and the number of claims, and implies that as the exposure increases, the number of claims is also likely to increase, which is a reasonable expectation. 'veh_bodyUTE' with a posterior mean of -0.51, which indicates that utility vehicles are associated with fewer claims compared to the reference category. This might be attributed to utility vehicle owners using their vehicles for specific purposes or driving less frequently, leading to a lower probability of accidents and claims.

d)[10 marks] Fit a new model on numclaims in terms of the same covariates to improve on the models in part b) or part c) by considering interactions between covariates, as well as random effects. Describe your new model and justify your choices.

Choose your own prior distributions (do not use default priors), and explain the rationale your prior choices, and ensure that the posterior is not too sensitive to your prior choice [Hint: look at the induced prior on the linear predictor and the response.]

Compute the posterior means of the model parameters, and discuss the results.

```
# sigma_obs:
sigma.unif.prior = "expression:
b = 20;
log_dens= (theta>=(-2*log(b)))*(-log(b)-theta/2-log(2))+
(theta<(-2*log(b)))*(-Inf); return(log_dens);"
#sigma_alpha:
```

```
sigma.unif.prior.random.eff = "expression:
b = 20;
log_dens = (theta>=(-2*log(b)))*(-log(b)-theta/2-log(2)) +
(theta<(-2*log(b)))*(-Inf); return(log_dens);"
b=20;
prec.prior <- list(prec=list(prior = sigma.unif.prior,
                             initial = -2*log(b)+1,fixed = FALSE))
prec.prior.random.eff <- list(prec=list(prior =
                                          sigma.unif.prior.random.eff,
                                        initial = -2*log(b)+1, fixed = FALSE))
formula_random <- numclaims ~
  veh_value + exposure + veh_age + gender + area + agecat +
  veh_value:exposure + veh_age:gender + area:gender +
  f(veh_body, model = "iid",hyper= prec.prior.random.eff)

prior.beta <- list(mean.intercept = 0, prec.intercept = 1/log(4)^2,
                   mean = 0, prec = 1/(log(5)/2)^2)
prior.beta2 <- list(mean.intercept = 0, prec.intercept = 1,
                   mean = 0, prec = 1)

model_random <- inla(formula_random, family = "poisson", data = dataCar,
                     control.fixed=prior.beta,
                     control.inla = list(strategy = "laplace", npoints = 40),
                     control.compute = list(config = TRUE,dic = TRUE, cpo=TRUE))
model_random_test <- inla(formula_random, family = "poisson", data = dataCar,
                     control.fixed=prior.beta2,
                     control.inla = list(strategy = "laplace", npoints = 40),
                     control.compute = list(config = TRUE,dic = TRUE, cpo=TRUE))
summary(model_random)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 4, Running = 8.11, Post = 0.424, Total = 12.5
## Fixed effects:
##                       mean    sd 0.025quant 0.5quant 0.975quant   mode kld
```

```
## (Intercept)        -2.429 0.101     -2.617   -2.433     -2.218 -2.440   0
## veh_value           0.007 0.022     -0.036    0.007      0.049  0.007   0
## exposure            0.519 0.015      0.490    0.519      0.548  0.519   0
## veh_age2            0.063 0.056     -0.048    0.063      0.173  0.063   0
## veh_age3           -0.054 0.059     -0.168   -0.054      0.061 -0.054   0
## veh_age4           -0.114 0.066     -0.242   -0.114      0.015 -0.113   0
## genderM             0.002 0.084     -0.163    0.002      0.166  0.002   0
## areaB               0.034 0.057     -0.076    0.034      0.145  0.034   0
## areaC               0.028 0.051     -0.072    0.028      0.128  0.028   0
## areaD              -0.072 0.068     -0.205   -0.072      0.061 -0.072   0
## areaE              -0.008 0.076     -0.157   -0.008      0.141 -0.008   0
## areaF               0.178 0.085      0.012    0.178      0.345  0.178   0
## agecat2            -0.173 0.054     -0.279   -0.173     -0.068 -0.173   0
## agecat3            -0.221 0.053     -0.324   -0.221     -0.118 -0.221   0
## agecat4            -0.245 0.052     -0.348   -0.245     -0.142 -0.245   0
## agecat5            -0.461 0.059     -0.577   -0.461     -0.346 -0.461   0
## agecat6            -0.448 0.067     -0.580   -0.448     -0.317 -0.448   0
## veh_value:exposure  0.038 0.014      0.010    0.038      0.066  0.038   0
## veh_age2:genderM    0.010 0.087     -0.161    0.010      0.181  0.010   0
## veh_age3:genderM    0.019 0.086     -0.149    0.019      0.188  0.019   0
## veh_age4:genderM    0.023 0.088     -0.150    0.023      0.195  0.023   0
## genderM:areaB       0.040 0.085     -0.127    0.040      0.207  0.040   0
## genderM:areaC      -0.058 0.078     -0.211   -0.058      0.094 -0.058   0
## genderM:areaD      -0.101 0.105     -0.308   -0.101      0.106 -0.101   0
## genderM:areaE      -0.057 0.113     -0.280   -0.057      0.165 -0.057   0
## genderM:areaF      -0.253 0.128     -0.503   -0.253     -0.002 -0.253   0
##
## Random effects:
##   Name     Model
##     veh_body IID model
##
## Model hyperparameters:
##                          mean     sd 0.025quant 0.5quant 0.975quant  mode
## Precision for veh_body 94.69 269.81       6.75    41.47     452.32 16.77
##
## Deviance Information Criterion (DIC) ...............: 34795.10
## Deviance Information Criterion (DIC, saturated) ....: 25360.40
## Effective number of parameters .....................: 33.42
##
## Marginal log-Likelihood:  -17458.02
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

```
summary(model_random_test)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
```

```
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.55, Running = 17.9, Post = 0.456, Total = 21.9
## Fixed effects:
##                       mean    sd 0.025quant 0.5quant 0.975quant   mode kld
## (Intercept)        -2.416 0.103     -2.605   -2.420     -2.200 -2.428   0
## veh_value           0.006 0.022     -0.036    0.006      0.049  0.006   0
## exposure            0.519 0.015      0.490    0.519      0.548  0.519   0
## veh_age2            0.060 0.056     -0.051    0.060      0.171  0.060   0
## veh_age3           -0.056 0.059     -0.172   -0.056      0.059 -0.056   0
## veh_age4           -0.117 0.066     -0.246   -0.117      0.012 -0.117   0
## genderM            -0.001 0.084     -0.166   -0.001      0.164 -0.001   0
## areaB               0.033 0.057     -0.078    0.033      0.144  0.033   0
## areaC               0.027 0.051     -0.073    0.027      0.127  0.027   0
## areaD              -0.073 0.068     -0.207   -0.073      0.060 -0.073   0
## areaE              -0.009 0.076     -0.158   -0.009      0.141 -0.009   0
## areaF               0.178 0.085      0.011    0.178      0.346  0.178   0
## agecat2            -0.177 0.054     -0.283   -0.177     -0.071 -0.177   0
## agecat3            -0.224 0.053     -0.328   -0.224     -0.121 -0.224   0
## agecat4            -0.249 0.052     -0.352   -0.249     -0.146 -0.249   0
## agecat5            -0.465 0.059     -0.581   -0.465     -0.350 -0.465   0
## agecat6            -0.453 0.067     -0.585   -0.453     -0.321 -0.453   0
## veh_value:exposure  0.038 0.014      0.010    0.038      0.066  0.038   0
## veh_age2:genderM    0.012 0.088     -0.160    0.012      0.184  0.012   0
## veh_age3:genderM    0.022 0.086     -0.148    0.022      0.191  0.022   0
## veh_age4:genderM    0.025 0.088     -0.148    0.025      0.198  0.025   0
## genderM:areaB       0.041 0.085     -0.127    0.041      0.208  0.041   0
## genderM:areaC      -0.058 0.078     -0.211   -0.058      0.095 -0.058   0
## genderM:areaD      -0.101 0.106     -0.309   -0.101      0.107 -0.101   0
## genderM:areaE      -0.057 0.114     -0.281   -0.057      0.167 -0.057   0
## genderM:areaF      -0.254 0.129     -0.507   -0.254     -0.002 -0.254   0
##
## Random effects:
##   Name     Model
##     veh_body IID model
##
## Model hyperparameters:
##                          mean     sd 0.025quant 0.5quant 0.975quant  mode
## Precision for veh_body 90.36 257.80       6.54    40.12     427.10 16.26
##
## Deviance Information Criterion (DIC) ...............: 34795.10
## Deviance Information Criterion (DIC, saturated) ....: 25360.40
## Effective number of parameters .....................: 33.55
##
```

```
## Marginal log-Likelihood:  -17464.21
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation): This part used same covariates as above. Moreover, veh_body as a random effect and added veh_value:exposure , veh_age:gender , area:gender intersections to the original covariates Also, we used numclaims as response, and then established Bayesian Poisson model with log link function as part b . Since the target variable and covariates are the same as b, the same prior as above is used. To ensure that the posterior is not too sensitive to your prior choice, I conducted a sensitivity test. By comparing models established using different priors, it was found that the posterior means of different variables were basically similar, and the indicators of fit such as DIC were also similar, indicating that the model is not sensitive to the prior setting. In the posterior mean analysis of the different variables, the top three variables in absolute value of the coefficients are selected here in order to avoid the interference caused by the large number of variables. The posterior mean of exposure is 0.52, which suggests a positive association between the exposure variable and the number of claims, and implies that as the exposure increases, the number of claims is also likely to increase, which is a reasonable expectation.'agecat5' and 'agecat6' with the posterior mean of -0.45 ,which represents the fifth and sixth age category. The negative association suggests that policyholders in this age category have fewer claims than those in the reference age category. This could be due to older drivers being more experienced and cautious, leading to fewer accidents and claims.

e)[10 marks] Perform posterior predictive model checks for your models b, c, d (i.e. using replicates).

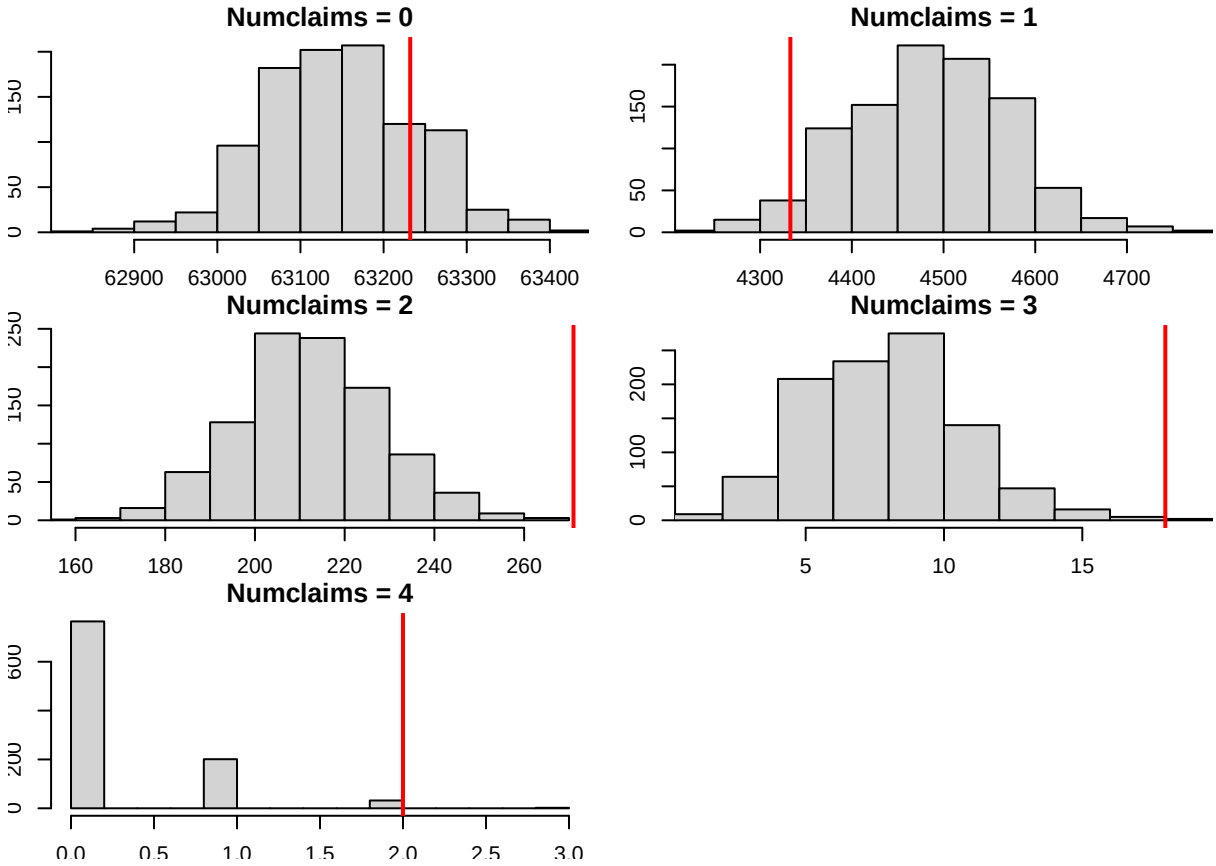
As test functions, use the number of rows in the dataset with numclaims equal 0, 1, 2, 3, and 4 (5 test functions).

Compute the RMSE values for predicting numclaims based on all 3 models.

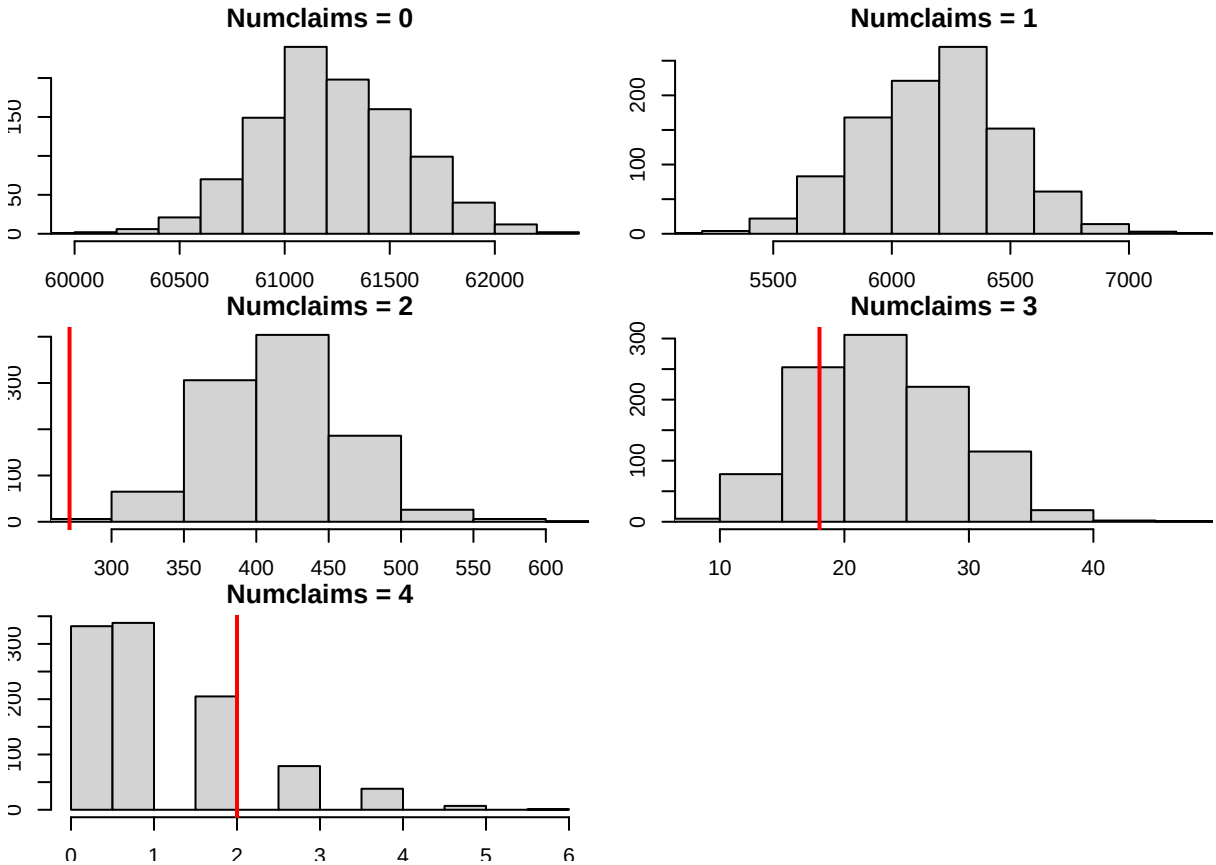
Discuss the results.

```
nbsamp=1000
n=nrow(dataCar)
yrep1 = matrix(0,nrow=n,ncol=nbsamp)
yrep2 = matrix(0,nrow=n,ncol=nbsamp)
yrep3 = matrix(0,nrow=n,ncol=nbsamp)

poisson.samples=inla.posterior.sample(nbsamp, result=model_poisson)
zero.samples=inla.posterior.sample(nbsamp, result=model_zero)
random.samples=inla.posterior.sample(nbsamp, result=model_random)
predictor.samples.poisson=inla.posterior.sample.eval(function(...) {Predictor},
                                                     poisson.samples)
predictor.samples.zero=inla.posterior.sample.eval(function(...) {Predictor},
                                                  zero.samples)
predictor.samples.random=inla.posterior.sample.eval(function(...) {Predictor},
                                                    random.samples)

for (row.num in 1:n){
  yrep1[row.num,]<- rpois(nbsamp,
                          lambda=exp(predictor.samples.poisson[row.num,]))
  yrep2[row.num,]<- rpois(nbsamp,
                          lambda=exp(predictor.samples.zero[row.num,]))
  yrep3[row.num,]<- rpois(nbsamp,
                          lambda=exp(predictor.samples.random[row.num,]))
}
```

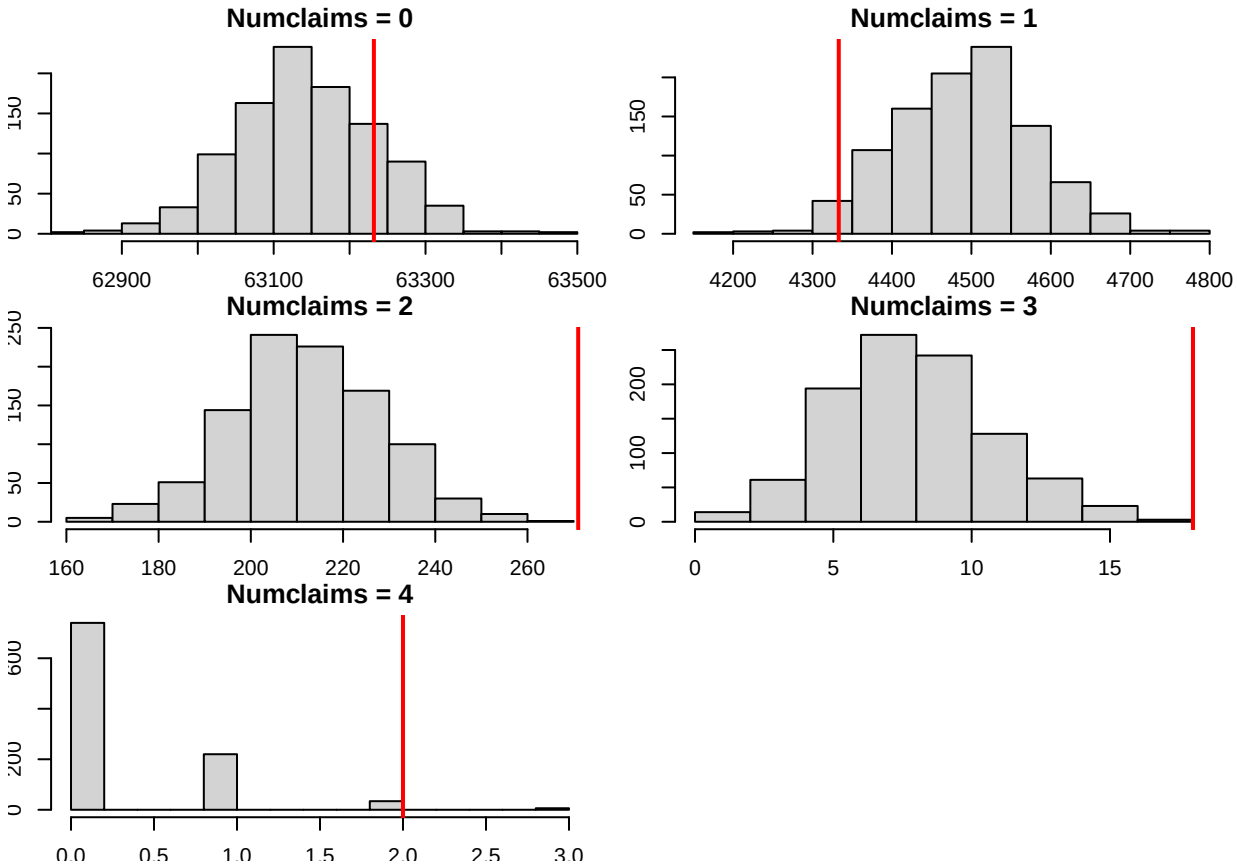
```
plot.post.pred.test <- function(yrep) {
  par(mfrow = c(3, 2))
  par(mar = c(1.7, 1.7, 1.7, 1.7))

  for (i in 0:4) {
    numclaims <- apply(yrep, 2, function(x) sum(x == i))
    hist(numclaims, xlim = c(min(numclaims), max(numclaims)),
         main = paste("Numclaims =", i), xlab = "Value", ylab = "Frequency")
    abline(v = sum(dataCar$numclaims == i), col = 'red', lwd = 2)
  }

  par(mfrow = c(1, 1))
}

plot.post.pred.test(yrep1)
```




```
plot.post.pred.test(yrep2)
```




```
plot.post.pred.test(yrep3)
```




```
y_hat_poisson = model_poisson$summary.fitted.values[,1]
y_hat_zero = model_zero$summary.fitted.values[,1]
y_hat_random = model_random$summary.fitted.values[,1]
rmse_poisson <- sqrt(mean((y_hat_poisson - dataCar$numclaims)^2))
rmse_zero <- sqrt(mean((y_hat_zero - dataCar$numclaims)^2))
rmse_random <- sqrt(mean((y_hat_random - dataCar$numclaims)^2))
rmse_poisson;rmse_zero;rmse_random
```

```
## [1] 0.2754645
```

```
## [1] 0.2779885
```

```
## [1] 0.2754854
```

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation): For model b, it performs well when numclaim is equal to 0, 1, and 4, as indicated by the red line on the histogram. For model c, it performs well when numclaim is equal to 2, 3, and 4, as indicated by the red line on the histogram. For model d, it performs well when numclaim is between 0 and 4, as indicated by the red line on the histogram. However, the RMSE errors for all three models are relatively similar, possibly due to the same priors being used.

Problem 2 - Barcelona study

In this problem, we will use a dataset from the CitieS-Health project that provides insight into the impact of air pollution on humans. It is comprised of data collected in Barcelona, Spain, and examines various environmental variables, such as air pollution levels, and their effects on mental health and wellbeing. In addition to environmental factors, this dataset also captures self-reported survey data on mental health, physical activity, diet habits, and more. From performance in a Stroop test (a type of psychological test evaluating attention capacity and processing speed) to information on total noise exposure at 55 dB - this dataset contains interesting information to understand the link between air pollution and human health.

We start by loading the dataset.

```
study<-read.csv("Barcelona.csv")
head(study)
```

```
##   Person_ID date_all year month day dayoftheweek hour sadness wellbeing energy
## 1       115    22222 2020    11   3            1   18      14         3      2
## 2       212    22247 2020    11  28            5   18       4         9      9
## 3       104    22208 2020    10  20            1   20       1         6      6
## 4       216    22247 2020    11  28            5   18       2         8      8
## 5        94    22213 2020    10  25            6   19      12         8      4
## 6       215    22258 2020    12   9            2   20       4         7      7
##   stress sleep hours_out physical_activity computer_use on_a_diet alcohol drugs
## 1      5     2         5                No          Yes       Yes      No    No
## 2      1     9         5               Yes           No        No     Yes    No
## 3      7     9        11                No          Yes       Yes      No    No
## 4      1     3         2               Yes           No       Yes     Yes    No
```

```
## 5      2     8         1                No          Yes        No      No   Yes
## 6      7     9         5               Yes           No       Yes     Yes    No
##   sick other_factors stroop_test_performance no2bcn_24h no2bcn_12h no2gps_24h
## 1   No           Yes                58.17712   33.81250  33.666667   24.32836
## 2   No            No                40.35988   15.80159  18.333333   15.48938
## 3   No           Yes                36.79430   47.52778  34.888889   48.59409
## 4  Yes            No                36.32432   15.80159  18.333333   15.64394
## 5   No            No                42.78266   12.35065   9.595238   17.03566
## 6   No           Yes                42.36540   16.91071  23.011905   22.38318
##   no2gps_12h no2bcn_12h_x30 no2bcn_24h_x30 no2gps_12h_x30 no2gps_24h_x30
## 1   22.66778      1.1222222      1.1270833      0.8109452      0.8109452
## 2   18.20557      0.6111111      0.5267196      0.5163127      0.5163127
## 3   28.62250      1.1629629      1.5842593      1.6198030      1.6198030
## 4   18.28909      0.6111111      0.5267196      0.5214648      0.5214648
## 5   15.02632      0.3198413      0.4116883      0.5678554      0.5678554
## 6   29.95232      0.7670635      0.5636905      0.7461060      0.7461060
##     pm25bcn  BCmicrog sec_noise55_day sec_noise65_day sec_greenblue_day
## 1 16.533333 1.1670614               0               0                 0
## 2  8.916667 0.2854848               0               0                 0
## 3 11.516667 1.0294803               0               0                 0
## 4  8.916667 0.2854848           37430            1426              6343
## 5 11.150000 0.4683368           12185               0                 0
## 6 10.460000 0.2532321           20596           14601                 0
##   tmean_24h tmean_12h humi_24h humi_12h pressure_24h pressure_12h precip_24h
## 1  18.05417  18.25833 82.97917 78.20833     1020.179     1020.983          0
## 2  13.89167  14.36667 86.47917 81.79167     1002.600     1001.575         37
## 3  18.98958  20.58750 76.12500 74.50000     1013.992     1012.621          0
## 4  13.89167  14.36667 86.47917 81.79167     1002.600     1001.575         37
## 5  18.57609  19.87083 51.00000 49.16667     1009.852     1007.842          0
## 6  10.19375  11.70833 47.77083 45.62500     1005.508     1006.933          0
##   maxwindspeed_24h access_greenbluespaces_300mbuff      incidence_cat age_yrs
## 1                0                             Yes       No incidence      29
## 2                4                             Yes Mobility incidence      28
## 3                0                             Yes Physical incidence      50
## 4                4                              No Mobility incidence      25
## 5                0                             Yes Physical incidence      35
## 6                0                              No       No incidence      48
##   yearbirth smoke gender     district  education microgram3
## 1      1991    No  Woman   Sant Martí University      15.72
## 2      1992    No  Woman Ciutat Vella University      37.50
## 3      1970   Yes    Man     Eixample University      41.97
## 4      1995   Yes    Man       Gràcia University      33.49
## 5      1985   Yes    Man   Sant Martí University      33.47
## 6      1972    No  Woman Ciutat Vella University      25.91
```

Descriptions of some of the covariates:

Column name	Description
Person_ID	ID of person filling out the survey (integer). Multiple rows for most persons, at different dates.
date_all	Date of the survey. (Date)
year	Year of the survey. (Integer)
month	Month of the survey. (Integer)

Column name	Description
day	Day of the survey. (Integer)
dayoftheweek	Day of the week of the survey. (Integer)
hour	Hour of the survey. (Integer)
sadness	Sadness score. (Integer)
wellbeing	Self-reported survey responses regarding wellbeing. (Integer)
energy	Self-reported survey responses regarding energy levels. (Integer)
stress	Self-reported survey responses regarding stress levels. (Integer)
sleep	Self-reported survey responses regarding sleep quality. (Integer)
hours_out	Self-reported survey responses regarding time spent outdoors. (Integer)
computer_use	Self-reported survey responses regarding computer use. (Yes/No)
on_a_diet	Self-reported survey responses regarding diet. (Yes/No)
alcohol	Self-reported survey responses regarding alcohol consumption. (Yes/No)
drugs	Self-reported survey responses regarding drug use. (Yes/No)
sick	Self-reported survey responses regarding illness. (Yes/No)
other_factors	Self-reported survey responses regarding other factors. (Yes/No)
stroop_test_performance	Performance in the Stroop test. (Float)
no2bcn_24h	Nitrogen dioxide (NO2) levels in Barcelona over 24 hours. (Float)
no2bcn_12h	Nitrogen dioxide (NO2) levels in Barcelona over 12 hours. (Float)
no2gps_24h	Nitrogen dioxide (NO2) levels in GPS locations over 24 hours. (Float)
no2gps_12h	Nitrogen dioxide (NO2) levels in GPS locations over 12 hours. (Float)
no2bcn_12h_x30	Nitrogen dioxide (NO2) levels in Barcelona over 12 hours multiplied by 30. (Float)
no2bcn_24h_x30	Nitrogen dioxide (NO2) levels in Barcelona over 24 hours multiplied by 30. (Float)
no2gps_12h_x30	Nitrogen dioxide (NO2) levels in GPS locations over 12 hours multiplied by 30. (Float)
no2gps_24h_x30	Nitrogen dioxide (NO2) levels in GPS locations over 24 hours multiplied by 30. (Float)
min_gps	Minimum GPS location. (Float)
district	District of Barcelona where the survey was conducted. (String)
education	Educational level of the participant. (String)
maxwindspeed_12h	Maximum wind speed over 12 hours. (Float)
access_greenbluespaces_300mbuff	Access to green and blue spaces within a 300m buffer. (Yes/No)
microgram3	Micrograms per cubic meter of pollutants. (Float)
age_yrs	Age of the participant in years. (Integer)
yearbirth	Year of birth of the participant. (Integer)
smoke	Self-reported survey responses regarding smoking status. (Yes/No)
gender	Gender of the participant. (Woman/Man)
hour_gps	Hour of the GPS location. (Integer)
pm25bcn	Particulate matter (PM2.5) levels in Barcelona. (Float)
BCmicrog	Black carbon (BC) levels in micrograms. (Float)
sec_noise55_day	Seconds of noise over 55 minutes in a day. (Integer)
sec_noise65_day	Seconds of noise over 65 minutes in a day. (Integer)
tmean_24h	Mean temperature over 24 hours. (Float)
tmean_12h	Mean temperature over 12 hours. (Float)
humi_24h	Humidity over 24 hours. (Float)
humi_12h	Humidity over 12 hours. (Float)
pressure_24h	Pressure over 24 hours. (Float)

Column name	Description
pressure_12h	Pressure over 12 hours. (Float)
precip_24h	Precipitation over 24 hours. (Float)
precip_12h	Precipitation over 12 hours. (Float)
precip_12h_binary	Binary value for precipitation over 12 hours. (Integer)
precip_24h_binary	Binary value for precipitation over 24 hours. (Integer)
maxwindspeed_24h	Maximum wind speed over 24 hours. (Float)

You can use either JAGS, Stan, or INLA for this question.

a)[10 marks] Fit a Bayesian linear regression model

- **on the logarithm of stroop_test_performance as response,**
- **using the following covariates: gender, on_a_diet, alcohol, drugs, sick, other_factors, educational, smoke, no2gps_24h, maxwindspeed_24h, precip_24h, sec_noise55_day, access_greenbluespaces_300mbuff, age_yrs, tmean_24h (you can use categorical covariates by converting integers to factors if appropriate).**

Center and scale the non-categorical covariates.

Choose your own prior distributions (do not use default priors), and explain the rationale your prior choices, and ensure that the posterior is not too sensitive to your prior choice [Hint: look at the induced prior on the response.]

Compute the posterior means of the model parameters, and interpret their meaning.

```
#remove the missing value
study <- study[apply(study != "", 1, all), ]
#calculate the mean and sd
vars_to_scale <- c('no2gps_24h', 'maxwindspeed_24h', 'precip_24h', 'sec_noise55_day', 'age_yrs', 'tmean_
mean_sd_list <- lapply(vars_to_scale, function(var) {
  list(mean = mean(study[[var]]), sd = sd(study[[var]]))
})
# Center and scale the non-categorical covariates
study$no2gps_24h <- scale(study$no2gps_24h)[,1]
study$maxwindspeed_24h <- scale(study$maxwindspeed_24h)[,1]
study$precip_24h <- scale(study$precip_24h)[,1]
study$sec_noise55_day <- scale(study$sec_noise55_day)[,1]
study$age_yrs <- scale(study$age_yrs)[,1]
study$tmean_24h <- scale(study$tmean_24h)[,1]

# Convert integers to factors for categorical covariates
cols_to_factor <- c("gender", "on_a_diet", "alcohol", "drugs", "sick", "other_factors",
                    "education", "smoke", "access_greenbluespaces_300mbuff")
study[cols_to_factor] <- lapply(study[cols_to_factor], as.factor)

str(study)
```

```
## 'data.frame':    1765 obs. of  51 variables:
##  $ Person_ID                     : int  115 212 104 216 94 215 151 172 94 203 ...
##  $ date_all                      : int  22222 22247 22208 22247 22213 22258 22240 22222 22211 22237
##  $ year                          : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
```

```
##  $ month                           : int  11 11 10 11 10 12 11 11 10 11 ...
##  $ day                             : int  3 28 20 28 25 9 21 3 23 18 ...
##  $ dayoftheweek                    : int  1 5 1 5 6 2 5 1 4 2 ...
##  $ hour                            : int  18 18 20 18 19 20 21 18 18 22 ...
##  $ sadness                         : int  14 4 1 2 12 4 12 1 10 6 ...
##  $ wellbeing                       : int  3 9 6 8 8 7 7 8 8 3 ...
##  $ energy                          : int  2 9 6 8 4 7 7 8 7 6 ...
##  $ stress                          : int  5 1 7 1 2 7 2 5 3 10 ...
##  $ sleep                           : int  2 9 9 3 8 9 6 6 8 6 ...
##  $ hours_out                       : num  5 5 11 2 1 5 2 7 1 9 ...
##  $ physical_activity               : chr  "No" "Yes" "No" "Yes" ...
##  $ computer_use                    : chr  "Yes" "No" "Yes" "No" ...
##  $ on_a_diet                       : Factor w/ 2 levels "No","Yes": 2 1 2 2 1 2 2 1 1 2 ...
##  $ alcohol                         : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 2 1 1 1 ...
##  $ drugs                           : Factor w/ 2 levels "No","Yes": 1 1 1 1 2 1 1 1 2 1 ...
##  $ sick                            : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
##  $ other_factors                   : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 2 1 2 1 2 ...
##  $ stroop_test_performance         : num  58.2 40.4 36.8 36.3 42.8 ...
##  $ no2bcn_24h                      : num  33.8 15.8 47.5 15.8 12.4 ...
##  $ no2bcn_12h                      : num  33.7 18.3 34.9 18.3 9.6 ...
##  $ no2gps_24h                      : num  -0.521 -1.233 1.434 -1.22 -1.108 ...
##  $ no2gps_12h                      : num  22.7 18.2 28.6 18.3 15 ...
##  $ no2bcn_12h_x30                  : num  1.122 0.611 1.163 0.611 0.32 ...
##  $ no2bcn_24h_x30                  : num  1.127 0.527 1.584 0.527 0.412 ...
##  $ no2gps_12h_x30                  : num  0.811 0.516 1.62 0.521 0.568 ...
##  $ no2gps_24h_x30                  : num  0.811 0.516 1.62 0.521 0.568 ...
##  $ pm25bcn                         : num  16.53 8.92 11.52 8.92 11.15 ...
##  $ BCmicrog                        : num  1.167 0.285 1.029 0.285 0.468 ...
##  $ sec_noise55_day                 : num  -0.834 -0.834 -0.834 1.281 -0.146 ...
##  $ sec_noise65_day                 : int  0 0 0 1426 0 14601 5016 1128 2368 18109 ...
##  $ sec_greenblue_day               : int  0 0 0 6343 0 0 134 0 1167 51702 ...
##  $ tmean_24h                       : num  0.746 -0.693 1.069 -0.693 0.926 ...
##  $ tmean_12h                       : num  18.3 14.4 20.6 14.4 19.9 ...
##  $ humi_24h                        : num  83 86.5 76.1 86.5 51 ...
##  $ humi_12h                        : num  78.2 81.8 74.5 81.8 49.2 ...
##  $ pressure_24h                    : num  1020 1003 1014 1003 1010 ...
##  $ pressure_12h                    : num  1021 1002 1013 1002 1008 ...
##  $ precip_24h                      : num  -0.3 4.34 -0.3 4.34 -0.3 ...
##  $ maxwindspeed_24h                : num  -0.298 1.171 -0.298 1.171 -0.298 ...
##  $ access_greenbluespaces_300mbuff: Factor w/ 2 levels "No","Yes": 2 2 2 1 2 1 2 1 2 2 ...
##  $ incidence_cat                   : chr  "No incidence" "Mobility incidence" "Physical incidence" "M
##  $ age_yrs                         : num  -0.728 -0.809 0.984 -1.054 -0.239 ...
##  $ yearbirth                       : int  1991 1992 1970 1995 1985 1972 1995 1984 1985 1980 ...
##  $ smoke                           : Factor w/ 2 levels "No","Yes": 1 1 2 2 2 1 1 1 2 1 ...
##  $ gender                          : Factor w/ 3 levels "Man","Otra","Woman": 3 3 1 1 1 3 1 3 1 3 ...
##  $ district                        : chr  "Sant Martí" "Ciutat Vella" "Eixample" "Gràcia" ...
##  $ education                       : Factor w/ 3 levels "Baccalaureate",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ microgram3                      : num  15.7 37.5 42 33.5 33.5 ...
```

```
library(INLA)
formula<- log(stroop_test_performance) ~ gender+on_a_diet+
  alcohol+drugs+sick+other_factors+education+
  smoke+no2gps_24h+maxwindspeed_24h+precip_24h+sec_noise55_day+
  access_greenbluespaces_300mbuff+age_yrs+tmean_24h
```

```r
prior.beta <- list(mean.intercept = 0, prec.intercept = 0.001,
                   mean = 0, prec = 0.001)
prior.beta2 <- list(mean.intercept = 0, prec.intercept = 1,
                   mean = 0, prec = 1)

model_linear <- inla(formula, family = "gaussian", data =study,
                     control.fixed = prior.beta,
                     control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
)
model_linear_test <- inla(formula, family = "gaussian", data =study,
                     control.fixed = prior.beta2,
                     control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
)
summary(model_linear)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 4.69, Running = 0.808, Post = 0.266, Total = 5.76
## Fixed effects:
##                                    mean    sd 0.025quant 0.5quant 0.975quant
## (Intercept)                       3.706 0.023      3.662    3.706      3.750
## genderOtra                       -0.033 0.099     -0.227   -0.033      0.161
## genderWoman                      -0.011 0.012     -0.034   -0.011      0.012
## on_a_dietYes                      0.031 0.013      0.006    0.031      0.055
## alcoholYes                        0.009 0.013     -0.015    0.009      0.034
## drugsYes                         -0.067 0.041     -0.147   -0.067      0.013
## sickYes                          -0.011 0.015     -0.040   -0.011      0.017
## other_factorsYes                 -0.043 0.012     -0.065   -0.043     -0.020
## educationPrimary or less          0.039 0.038     -0.034    0.039      0.113
## educationUniversity               0.139 0.020      0.099    0.139      0.179
## smokeYes                          0.026 0.015     -0.003    0.026      0.054
## no2gps_24h                       -0.001 0.005     -0.012   -0.001      0.010
## maxwindspeed_24h                  0.023 0.009      0.005    0.023      0.042
## precip_24h                       -0.030 0.009     -0.048   -0.030     -0.012
## sec_noise55_day                  -0.011 0.005     -0.021   -0.011      0.000
```

```
## access_greenbluespaces_300mbuffYes  0.005 0.011     -0.016    0.005      0.026
## age_yrs                             -0.124 0.006     -0.136   -0.124     -0.111
## tmean_24h                           -0.013 0.006     -0.024   -0.013     -0.002
##                                       mode kld
## (Intercept)                          3.706   0
## genderOtra                          -0.033   0
## genderWoman                         -0.011   0
## on_a_dietYes                         0.031   0
## alcoholYes                           0.009   0
## drugsYes                            -0.067   0
## sickYes                             -0.011   0
## other_factorsYes                    -0.043   0
## educationPrimary or less             0.039   0
## educationUniversity                  0.139   0
## smokeYes                             0.026   0
## no2gps_24h                          -0.001   0
## maxwindspeed_24h                     0.023   0
## precip_24h                          -0.030   0
## sec_noise55_day                     -0.011   0
## access_greenbluespaces_300mbuffYes  0.005   0
## age_yrs                             -0.124   0
## tmean_24h                           -0.013   0
##
## Model hyperparameters:
##                                         mean    sd 0.025quant 0.5quant
## Precision for the Gaussian observations 21.43 0.725      20.03    21.42
##                                         0.975quant  mode
## Precision for the Gaussian observations      22.87 21.40
##
## Deviance Information Criterion (DIC) ...............: -378.20
## Deviance Information Criterion (DIC, saturated) ....: 1786.36
## Effective number of parameters .....................: 19.00
##
## Marginal log-Likelihood:  56.72
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

```
summary(model_linear_test)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
```

```
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.86, Running = 0.393, Post = 0.0856, Total = 4.34
## Fixed effects:
##                                    mean    sd 0.025quant 0.5quant 0.975quant
## (Intercept)                       3.704 0.023      3.660    3.704      3.749
## genderOtra                       -0.031 0.098     -0.224   -0.031      0.162
## genderWoman                      -0.011 0.012     -0.034   -0.011      0.013
## on_a_dietYes                      0.031 0.013      0.007    0.031      0.056
## alcoholYes                        0.009 0.013     -0.015    0.009      0.034
## drugsYes                         -0.067 0.041     -0.147   -0.067      0.013
## sickYes                          -0.011 0.015     -0.040   -0.011      0.017
## other_factorsYes                 -0.042 0.012     -0.065   -0.042     -0.020
## educationPrimary or less          0.040 0.038     -0.033    0.040      0.114
## educationUniversity               0.140 0.020      0.101    0.140      0.180
## smokeYes                          0.026 0.015     -0.003    0.026      0.055
## no2gps_24h                       -0.001 0.005     -0.012   -0.001      0.010
## maxwindspeed_24h                  0.023 0.009      0.005    0.023      0.042
## precip_24h                       -0.030 0.009     -0.048   -0.030     -0.012
## sec_noise55_day                  -0.011 0.005     -0.021   -0.011     -0.001
## access_greenbluespaces_300mbuffYes  0.005 0.011     -0.016    0.005      0.027
## age_yrs                          -0.124 0.006     -0.136   -0.124     -0.111
## tmean_24h                        -0.013 0.006     -0.024   -0.013     -0.002
##                                    mode kld
## (Intercept)                       3.704   0
## genderOtra                       -0.031   0
## genderWoman                      -0.011   0
## on_a_dietYes                      0.031   0
## alcoholYes                        0.009   0
## drugsYes                         -0.067   0
## sickYes                          -0.011   0
## other_factorsYes                 -0.042   0
## educationPrimary or less          0.040   0
## educationUniversity               0.140   0
## smokeYes                          0.026   0
## no2gps_24h                       -0.001   0
## maxwindspeed_24h                  0.023   0
## precip_24h                       -0.030   0
## sec_noise55_day                  -0.011   0
## access_greenbluespaces_300mbuffYes  0.005   0
## age_yrs                          -0.124   0
## tmean_24h                        -0.013   0
##
## Model hyperparameters:
##                                       mean    sd 0.025quant 0.5quant
## Precision for the Gaussian observations 21.43 0.725      20.03    21.42
##                                       0.975quant  mode
## Precision for the Gaussian observations      22.87 21.40
##
## Deviance Information Criterion (DIC) ...............: -378.22
```

```
## Deviance Information Criterion (DIC, saturated) ....: 1786.34
## Effective number of parameters .....................: 18.98
##
## Marginal log-Likelihood:  112.00
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

```
cat("The marginal log-likelihood value is:", model_linear$mlik[1], "\n")
```

```
## The marginal log-likelihood value is: 57.09852
```

```
cat("The NSLCPO value is:", -sum(log(model_linear$cpo$cpo)), "\n")
```

```
## The NSLCPO value is: -186.6895
```

```
cat("The DIC value is:", model_linear$dic$dic, "\n")
```

```
## The DIC value is: -378.1975
```

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation): In order to understand the link between air pollution and human health, a study has been conducted using data collected in Barcelona, Spain. This study incorporates partial environmental variables and self-reported survey data as covariates, including gender, on_a_diet, alcohol, drugs, sick, other_factors, educational, smoke, no2gps_24h, maxwindspeed_24h, precip_24h, sec_noise55_day, access_greenbluespaces_300mbuff, age_yrs, and tmean_24h. The participants' performance in a Stroop test, which is a psychological test that evaluates attention capacity and processing speed, was assessed and scored. A Bayesian linear regression model was then established to predict the logarithm of stroop_test_performance based on these variables. Firstly, variables are divided into discrete and continuous types for different preprocessing. By observing the data structure, it can be determined that gender, on_a_diet, alcohol, drugs, sick, other_factors, education, smoke, and access_greenbluespaces_300mbuff are discrete variables, so the as.factor function is used to convert the content of the variables into different levels. On the other hand, no2gps_24h, maxwindspeed_24h, precip_24h, sec_noise55_day, age_yrs, and tmean_24h are continuous variables. To eliminate the influence of dimensions and improve interpretability, the scale function is used to center and scale these non-categorical covariates. When setting the prior, make the intercept and beta use the classical setting method with a mean of 0 and an precision of 0.001.To ensure that the posterior is not too sensitive to your prior choice, I conducted a sensitivity test. By comparing models established using different priors, it was found that the posterior means of different variables were basically similar, and the indicators of fit such as DIC were also similar, indicating that the model is not sensitive to the prior setting. In the posterior mean analysis of the different variables, the top three variables in absolute value of the coefficients are selected here in order to avoid the interference caused by the large number of variables. With a posterior mean of 0.14, a university-level education has a positive association with the logarithm of Stroop test performance. This suggests that people with higher education levels tend to perform better on the Stroop test, potentially due to enhanced cognitive abilities. 'age_yrs' has a posterior mean about -0.12,which is a strong negative association with the logarithm of Stroop test performance. This implies that as individuals grow older, their attention capacity and processing speed tend to decline. 'drugsYes' has a posterior mean of -0.07, which indicates a negative association between drug usage and the logarithm of Stroop test performance. This implies that individuals who use drugs are likely to experience reduced attention capacity and processing speed.

b)[10 marks] Fit a Bayesian Poisson GLM

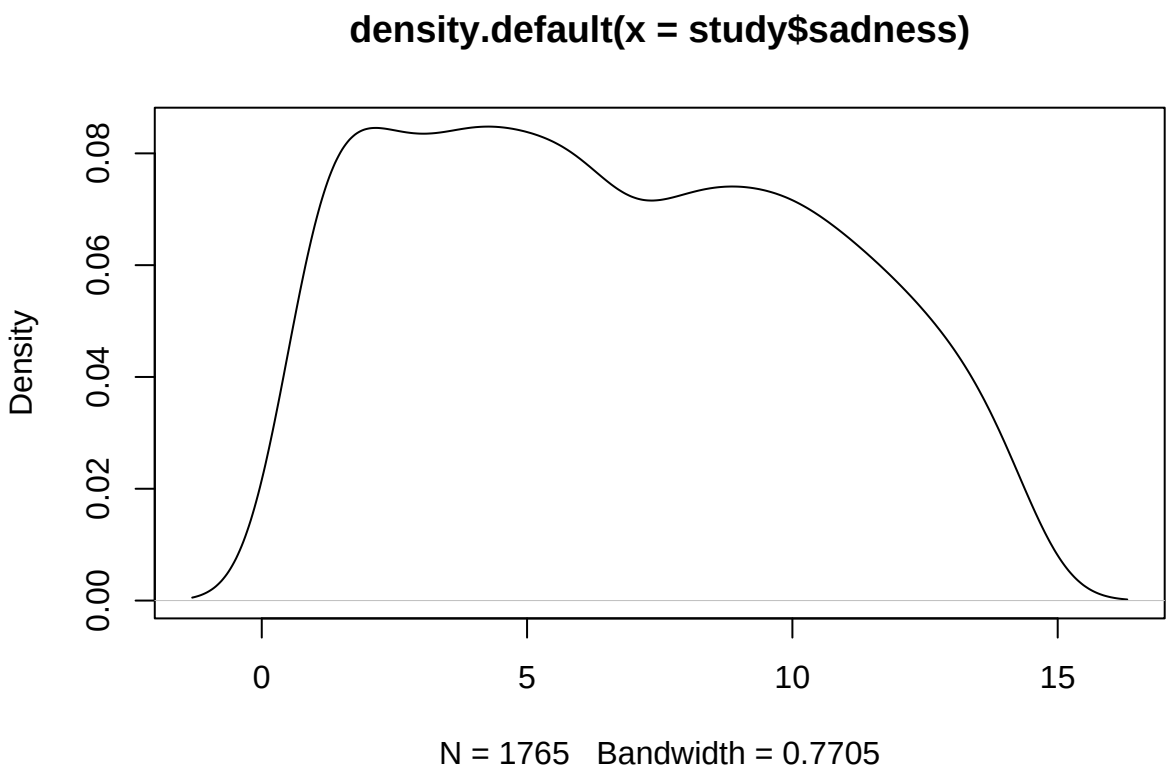
- **for sadness as response,**
- **log link function,**
- **using the following covariates: gender, on_a_diet, alcohol, drugs, sick, other_factors, educational, smoke, no2gps_24h, maxwindspeed_24h, precip_24h, sec_noise55_day, access_greenbluespaces_300mbuff, age_yrs, tmean_24h (you can use categorical covariates by converting integers to factors if appropriate).**

Center and scale the non-categorical covariates.

Choose your own prior distributions (do not use default priors), and explain the rationale your prior choices, and ensure that the posterior is not too sensitive to your prior choice [Hint: look at the induced prior on the response.]

Compute the posterior means of the model parameters, and interpret their meaning.

```
plot(density(study$sadness))
```




```
formula_poisson<- sadness ~ gender+on_a_diet+alcohol+drugs+sick+other_factors+education+
  smoke+no2gps_24h+maxwindspeed_24h+precip_24h+sec_noise55_day+
  access_greenbluespaces_300mbuff+age_yrs+tmean_24h

prior.beta <- list(mean.intercept = 0, prec.intercept =  1/(log(15)^2),
                   mean = 0, prec = 1/(log(5)/2)^2)
prior.beta2 <- list(mean.intercept = 0, prec.intercept = 1,
                   mean = 0, prec = 1)

model_poisson <- inla(formula_poisson, family = "poisson", data =study,
                      control.fixed = prior.beta,
```

```
                      control.family = list(link = "log"),
                      control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
)
model_poisson_test <- inla(formula_poisson, family = "poisson", data =study,
                      control.fixed = prior.beta2,
                      control.family = list(link = "log"),
                      control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
)
summary(model_poisson)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.18, Running = 0.326, Post = 0.0738, Total = 3.58
## Fixed effects:
##                                    mean    sd 0.025quant 0.5quant 0.975quant
## (Intercept)                       1.761 0.041      1.680    1.761      1.842
## genderOtra                       -0.520 0.243     -0.996   -0.520     -0.043
## genderWoman                       0.037 0.021     -0.004    0.037      0.079
## on_a_dietYes                      0.012 0.023     -0.032    0.012      0.056
## alcoholYes                       -0.054 0.023     -0.099   -0.054     -0.010
## drugsYes                          0.173 0.071      0.035    0.173      0.312
## sickYes                          -0.013 0.027     -0.065   -0.013      0.039
## other_factorsYes                 -0.139 0.021     -0.181   -0.139     -0.098
## educationPrimary or less         -0.130 0.071     -0.268   -0.130      0.009
## educationUniversity               0.167 0.037      0.095    0.167      0.240
## smokeYes                         -0.027 0.026     -0.078   -0.027      0.025
## no2gps_24h                        0.093 0.010      0.074    0.093      0.111
## maxwindspeed_24h                  0.028 0.019     -0.009    0.028      0.064
## precip_24h                       -0.058 0.019     -0.096   -0.058     -0.021
## sec_noise55_day                   0.020 0.009      0.002    0.020      0.039
## access_greenbluespaces_300mbuffYes 0.021 0.019     -0.017    0.021      0.058
## age_yrs                           0.019 0.011     -0.003    0.019      0.042
## tmean_24h                        -0.130 0.010     -0.149   -0.130     -0.110
##                                    mode kld
## (Intercept)                       1.761   0
```

```
## genderOtra                          -0.520   0
## genderWoman                          0.037   0
## on_a_dietYes                         0.012   0
## alcoholYes                          -0.054   0
## drugsYes                             0.173   0
## sickYes                             -0.013   0
## other_factorsYes                    -0.139   0
## educationPrimary or less            -0.130   0
## educationUniversity                  0.167   0
## smokeYes                            -0.027   0
## no2gps_24h                           0.093   0
## maxwindspeed_24h                     0.028   0
## precip_24h                          -0.058   0
## sec_noise55_day                      0.020   0
## access_greenbluespaces_300mbuffYes   0.021   0
## age_yrs                              0.019   0
## tmean_24h                           -0.130   0
##
## Deviance Information Criterion (DIC) ...............: 10048.41
## Deviance Information Criterion (DIC, saturated) ....: 3776.35
## Effective number of parameters .....................: 17.90
##
## Marginal log-Likelihood:  -5072.82
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

```
summary(model_poisson_test)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 4.89, Running = 0.378, Post = 0.402, Total = 5.67
## Fixed effects:
##                                    mean    sd 0.025quant 0.5quant 0.975quant
## (Intercept)                       1.758 0.041      1.677    1.758      1.840
## genderOtra                       -0.537 0.249     -1.025   -0.537     -0.049
```

```
## genderWoman                          0.038 0.021     -0.004    0.038      0.079
## on_a_dietYes                         0.013 0.023     -0.031    0.013      0.057
## alcoholYes                          -0.054 0.023     -0.099   -0.054     -0.010
## drugsYes                             0.174 0.071      0.035    0.174      0.312
## sickYes                             -0.013 0.027     -0.065   -0.013      0.039
## other_factorsYes                    -0.139 0.021     -0.181   -0.139     -0.098
## educationPrimary or less            -0.128 0.071     -0.267   -0.128      0.010
## educationUniversity                  0.169 0.037      0.097    0.169      0.242
## smokeYes                            -0.027 0.026     -0.078   -0.027      0.025
## no2gps_24h                           0.093 0.010      0.074    0.093      0.111
## maxwindspeed_24h                     0.028 0.019     -0.009    0.028      0.064
## precip_24h                          -0.058 0.019     -0.096   -0.058     -0.021
## sec_noise55_day                      0.020 0.009      0.002    0.020      0.039
## access_greenbluespaces_300mbuffYes   0.021 0.019     -0.017    0.021      0.059
## age_yrs                              0.019 0.011     -0.003    0.019      0.042
## tmean_24h                           -0.130 0.010     -0.149   -0.130     -0.110
##                                      mode kld
## (Intercept)                         1.758   0
## genderOtra                         -0.537   0
## genderWoman                         0.038   0
## on_a_dietYes                        0.013   0
## alcoholYes                         -0.054   0
## drugsYes                            0.174   0
## sickYes                            -0.013   0
## other_factorsYes                   -0.139   0
## educationPrimary or less           -0.128   0
## educationUniversity                 0.169   0
## smokeYes                           -0.027   0
## no2gps_24h                          0.093   0
## maxwindspeed_24h                    0.028   0
## precip_24h                         -0.058   0
## sec_noise55_day                     0.020   0
## access_greenbluespaces_300mbuffYes  0.021   0
## age_yrs                             0.019   0
## tmean_24h                          -0.130   0
##
## Deviance Information Criterion (DIC) ...............: 10048.48
## Deviance Information Criterion (DIC, saturated) ....: 3776.43
## Effective number of parameters .....................: 17.93
##
## Marginal log-Likelihood:  -5076.72
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

```
cat("The marginal log-likelihood value is:", model_poisson$mlik[1], "\n")
```

```
## The marginal log-likelihood value is: -5072.82
```

```
cat("The NSLCPO value is:", -sum(log(model_poisson$cpo$cpo)), "\n")
```

```
## The NSLCPO value is: 5033.498
```

```
cat("The DIC value is:", model_poisson$dic$dic, "\n")
```

```
## The DIC value is: 10048.41
```

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation): This part use same covariates as above. Also, we used sadness score as response, and then established Bayesian Poisson model with log link function.
For the prior selection, from the density plot, we can see that the usual values for the number of sadness are between 0 and 15, then we can beta0 on the interval [-ln(15), ln(15)].For the covariates, the most widely spaced variable is no2gps_24h, mainly distributed between -2 and 2, so max |xi - mean(x)| = 2. then beta would be in [-ln(5)/2, ln(5)/2]. To ensure that the posterior is not too sensitive to your prior choice, I conducted a sensitivity test. By comparing models established using different priors, it was found that the posterior means of different variables were basically similar, and the indicators of fit such as DIC were also similar, indicating that the model is not sensitive to the prior setting. In the posterior mean analysis of the different variables, the top three variables in absolute value of the coefficients are selected here in order to avoid the interference caused by the large number of variables. 'genderOtra' With a posterior mean of -0.52, which means individuals with non-binary genders have a strong negative association with the sadness score. This suggests that non-binary individuals might report lower levels of sadness compared to male and female. Also,'drugsYes' has a posterior mean of 0.17, which indicates a positive association between drug usage and the sadness score. This implies that individuals who use drugs tend to report higher levels of sadness. Drug usage might have an impact on mental health, potentially causing increased feelings of sadness or worsening pre-existing mental health conditions. Thirdly, 'EducationUniversity' with a posterior mean of 0.17, which has a positive association with the sadness score. This suggests that people with higher education levels might report higher levels of sadness. This may be due to various factors, such as increased stress levels or higher expectations.

c)[10 marks] Incorporate Person_ID as a random effects into the models a.) and b.).

Choose your own prior distributions for this random effect (do not use default priors).

Compare the posterior means of the parameter values with a) and b).

Discuss the changes that happened due to using random effects.

```
sigma.unif.prior = "expression:
b = 20;
log_dens= (theta>=(-2*log(b)))*(-log(b)-theta/2-log(2))+
(theta<(-2*log(b)))*(-Inf); return(log_dens);"
#sigma_alpha:
sigma.unif.prior.random.eff = "expression:
b = 20;
log_dens = (theta>=(-2*log(b)))*(-log(b)-theta/2-log(2)) +
(theta<(-2*log(b)))*(-Inf); return(log_dens);"
b=20;
prec.prior <- list(prec=list(prior = sigma.unif.prior,
                             initial = -2*log(b)+1,fixed = FALSE))
prec.prior.random.eff <- list(prec=list(prior =
                                          sigma.unif.prior.random.eff,
                                        initial = -2*log(b)+1, fixed = FALSE))
```

```
study$Person_ID <- as.factor(study$Person_ID)
formula.linear.random<- log(stroop_test_performance) ~
  gender+on_a_diet+alcohol+drugs+sick+other_factors+education+
  smoke+no2gps_24h+maxwindspeed_24h+precip_24h+sec_noise55_day+
```

```
  access_greenbluespaces_300mbuff+age_yrs+tmean_24h+
  f(Person_ID, model = "iid",hyper= prec.prior.random.eff)

prior.beta <- list(mean.intercept = 0, prec.intercept = 0.001,
                   mean = 0, prec = 0.001)

model.linear.random <- inla(formula.linear.random, family = "gaussian", data =study,
                     control.fixed = prior.beta,
                     control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
)
summary(model.linear.random)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.89, Running = 0.645, Post = 0.237, Total = 4.77
## Fixed effects:
##                                    mean    sd 0.025quant 0.5quant 0.975quant
## (Intercept)                       3.684 0.046      3.594    3.684      3.774
## genderOtra                       -0.017 0.182     -0.374   -0.017      0.340
## genderWoman                       0.004 0.027     -0.049    0.004      0.058
## on_a_dietYes                      0.004 0.011     -0.018    0.004      0.026
## alcoholYes                       -0.001 0.011     -0.023   -0.001      0.021
## drugsYes                         -0.041 0.045     -0.130   -0.041      0.048
## sickYes                          -0.020 0.012     -0.043   -0.020      0.004
## other_factorsYes                 -0.032 0.010     -0.053   -0.032     -0.012
## educationPrimary or less          0.039 0.083     -0.124    0.039      0.201
## educationUniversity               0.163 0.043      0.079    0.163      0.249
## smokeYes                          0.014 0.032     -0.049    0.014      0.076
## no2gps_24h                        0.002 0.004     -0.006    0.002      0.011
## maxwindspeed_24h                  0.016 0.007      0.002    0.016      0.030
## precip_24h                       -0.019 0.007     -0.033   -0.019     -0.005
## sec_noise55_day                  -0.008 0.007     -0.021   -0.008      0.005
## access_greenbluespaces_300mbuffYes  0.005 0.025     -0.044    0.005      0.054
## age_yrs                          -0.124 0.014     -0.152   -0.124     -0.097
## tmean_24h                        -0.032 0.006     -0.043   -0.032     -0.020
```

```
##                                     mode kld
## (Intercept)                        3.684   0
## genderOtra                        -0.017   0
## genderWoman                        0.004   0
## on_a_dietYes                       0.004   0
## alcoholYes                        -0.001   0
## drugsYes                          -0.041   0
## sickYes                           -0.020   0
## other_factorsYes                  -0.032   0
## educationPrimary or less           0.039   0
## educationUniversity                0.163   0
## smokeYes                           0.014   0
## no2gps_24h                         0.002   0
## maxwindspeed_24h                   0.016   0
## precip_24h                        -0.019   0
## sec_noise55_day                   -0.008   0
## access_greenbluespaces_300mbuffYes  0.005   0
## age_yrs                           -0.124   0
## tmean_24h                         -0.032   0
##
## Random effects:
##   Name     Model
##     Person_ID IID model
##
## Model hyperparameters:
##                                       mean   sd 0.025quant 0.5quant
## Precision for the Gaussian observations 39.90 1.44      37.12    39.88
## Precision for Person_ID                38.58 4.57      30.30    38.34
##                                       0.975quant  mode
## Precision for the Gaussian observations      42.80 39.84
## Precision for Person_ID                      48.26 37.90
##
## Deviance Information Criterion (DIC) ...............: -1482.13
## Deviance Information Criterion (DIC, saturated) ....: 1778.20
## Effective number of parameters .....................: 104.14
##
## Marginal log-Likelihood:  375.18
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

```
cat("The marginal log-likelihood value is:", model.linear.random$mlik[1], "\n")
```

```
## The marginal log-likelihood value is: 374.7367
```

```
cat("The NSLCPO value is:", -sum(log(model.linear.random$cpo$cpo)), "\n")
```

```
## The NSLCPO value is: -633.8518
```

```
cat("The DIC value is:", model.linear.random$dic$dic, "\n")
```

```
## The DIC value is: -1482.128
```

```
formula.poisson.random<- sadness ~ gender+on_a_diet+alcohol+
  drugs+sick+other_factors+education+
  smoke+no2gps_24h+maxwindspeed_24h+precip_24h+sec_noise55_day+
  access_greenbluespaces_300mbuff+age_yrs+tmean_24h+
  f(Person_ID, model = "iid",hyper= prec.prior.random.eff)

prior.beta <- list(mean.intercept = 0, prec.intercept =  1/(log(15)^2),
                   mean = 0, prec = 1/(log(5)/2)^2)

model.poisson.random <- inla(formula.poisson.random, family = "poisson", data =study,
                      control.fixed = prior.beta,
                      control.family = list(link = "log"),
                      control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
)
summary(model.poisson.random)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.3, Running = 0.487, Post = 0.146, Total = 3.93
## Fixed effects:
##                                    mean    sd 0.025quant 0.5quant 0.975quant
## (Intercept)                       1.529 0.091      1.348    1.529      1.705
## genderOtra                       -0.302 0.356     -0.999   -0.303      0.397
## genderWoman                       0.080 0.052     -0.022    0.080      0.183
## on_a_dietYes                      0.010 0.027     -0.043    0.010      0.062
## alcoholYes                       -0.041 0.027     -0.094   -0.041      0.012
## drugsYes                          0.074 0.106     -0.135    0.074      0.282
## sickYes                          -0.057 0.030     -0.115   -0.057      0.002
## other_factorsYes                 -0.164 0.026     -0.214   -0.164     -0.114
## educationPrimary or less         -0.181 0.161     -0.498   -0.181      0.135
## educationUniversity               0.338 0.084      0.174    0.337      0.505
## smokeYes                         -0.024 0.062     -0.145   -0.024      0.097
## no2gps_24h                        0.115 0.010      0.094    0.115      0.135
## maxwindspeed_24h                 -0.009 0.019     -0.047   -0.009      0.030
## precip_24h                       -0.030 0.020     -0.068   -0.030      0.009
```

```
## sec_noise55_day                       0.005 0.015     -0.025    0.005      0.034
## access_greenbluespaces_300mbuffYes    0.033 0.048     -0.061    0.033      0.127
## age_yrs                               0.012 0.027     -0.041    0.012      0.065
## tmean_24h                            -0.242 0.015     -0.271   -0.242     -0.214
##                                       mode kld
## (Intercept)                          1.531   0
## genderOtra                           -0.304   0
## genderWoman                           0.080   0
## on_a_dietYes                          0.010   0
## alcoholYes                           -0.041   0
## drugsYes                              0.075   0
## sickYes                              -0.057   0
## other_factorsYes                     -0.164   0
## educationPrimary or less             -0.180   0
## educationUniversity                   0.336   0
## smokeYes                             -0.024   0
## no2gps_24h                            0.115   0
## maxwindspeed_24h                     -0.009   0
## precip_24h                           -0.030   0
## sec_noise55_day                       0.005   0
## access_greenbluespaces_300mbuffYes    0.032   0
## age_yrs                               0.012   0
## tmean_24h                            -0.242   0
##
## Random effects:
##   Name     Model
##     Person_ID IID model
##
## Model hyperparameters:
##                           mean   sd 0.025quant 0.5quant 0.975quant  mode
## Precision for Person_ID 12.01 1.88       8.77    11.86      16.12 11.55
##
## Deviance Information Criterion (DIC) ...............: 9638.25
## Deviance Information Criterion (DIC, saturated) ....: 3366.20
## Effective number of parameters .....................: 174.81
##
## Marginal log-Likelihood:  -4956.78
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

```
cat("The marginal log-likelihood value is:", model.poisson.random$mlik[1], "\n")
```

```
## The marginal log-likelihood value is: -4956.395
```

```
cat("The NSLCPO value is:", -sum(log(model.poisson.random$cpo$cpo)), "\n")
```

```
## The NSLCPO value is: 4877.971
```

```
cat("The DIC value is:", model.poisson.random$dic$dic, "\n")
```

```
## The DIC value is: 9638.253
```

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation): Based on the first two parts, this section incorporates Person_ID as a random effect into the covariates, while keeping the response and model family unchanged. In the posterior mean analysis of the different variables, the top three variables in absolute value of the coefficients are selected here in order to avoid the interference caused by the large number of variables. The inclusion of the person ID as a random effect accounts for the variability between individuals, which can affect the relationships between the covariates and the outcome variable. By doing so, the model can better capture the within-person variations and isolate the effects of the other covariates. For the Bayesian linear model, after adding person_id as a random effect to the model, we can see that the top three posterior means of importance do not change, namely educationUniversity, age_yrs and drugsYes. Compare with above model, the posterior mean for age remains similar, suggesting that the relationship between age and the outcome variable is consistent, regardless of the inclusion of the person ID as a random effect. On the other hand, Education (university level) now has a higher posterior mean, indicating a stronger positive association with the outcome variable. In contrast, the drugs covariate has a reduced posterior mean, suggesting a weaker negative association. The change in the posterior means for education and drug use indicates that accounting for individual variability has an impact on the associations between these covariates and the outcome variable. This highlights the importance of considering random effects in models when there is a potential for unobserved individual differences that may influence the relationships between the covariates and the outcome. For the Bayesian poisson model with random effect, with the addition of person_id as a random effect in the model, we can see that the first three posterior means of importance change, from genderOtra, drugsYes, educationUniversity to educationUniversity, genderOtra, tmean_24h. The posterior mean for non-binary individuals(genderOtra) changed closer to 0, indicating that the negative association with the sadness score is still present but less strong. This could be due to individual differences among non-binary individuals that were not captured in the original model. For 'educationUniversity', the posterior mean increased, suggesting that the positive association between higher education levels and sadness scores becomes stronger when accounting for individual differences. Additionally, with a posterior mean of -0.24, there is a negative association between mean temperature and sadness scores. Finally, We can see that the value of dic has become smaller in both model and that the addition of the random effect has made the model more effective.

d)[10 marks] Do posterior predictive checks (i.e. using replicates) for the sadness score for your models with or without random effects. Explain the choice of test functions that you used.

Compute the posterior means of the response variable using the original covariates, and use this to compute the RMSE values for both models (i.e. with, or without random effects).

Discuss the results.

```
summary(model_poisson)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
```

```
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.18, Running = 0.326, Post = 0.0738, Total = 3.58
## Fixed effects:
##                                    mean    sd 0.025quant 0.5quant 0.975quant
## (Intercept)                       1.761 0.041      1.680    1.761      1.842
## genderOtra                       -0.520 0.243     -0.996   -0.520     -0.043
## genderWoman                       0.037 0.021     -0.004    0.037      0.079
## on_a_dietYes                      0.012 0.023     -0.032    0.012      0.056
## alcoholYes                       -0.054 0.023     -0.099   -0.054     -0.010
## drugsYes                          0.173 0.071      0.035    0.173      0.312
## sickYes                          -0.013 0.027     -0.065   -0.013      0.039
## other_factorsYes                 -0.139 0.021     -0.181   -0.139     -0.098
## educationPrimary or less         -0.130 0.071     -0.268   -0.130      0.009
## educationUniversity               0.167 0.037      0.095    0.167      0.240
## smokeYes                         -0.027 0.026     -0.078   -0.027      0.025
## no2gps_24h                        0.093 0.010      0.074    0.093      0.111
## maxwindspeed_24h                  0.028 0.019     -0.009    0.028      0.064
## precip_24h                       -0.058 0.019     -0.096   -0.058     -0.021
## sec_noise55_day                   0.020 0.009      0.002    0.020      0.039
## access_greenbluespaces_300mbuffYes 0.021 0.019     -0.017    0.021      0.058
## age_yrs                           0.019 0.011     -0.003    0.019      0.042
## tmean_24h                        -0.130 0.010     -0.149   -0.130     -0.110
##                                    mode kld
## (Intercept)                       1.761   0
## genderOtra                       -0.520   0
## genderWoman                       0.037   0
## on_a_dietYes                      0.012   0
## alcoholYes                       -0.054   0
## drugsYes                          0.173   0
## sickYes                          -0.013   0
## other_factorsYes                 -0.139   0
## educationPrimary or less         -0.130   0
## educationUniversity               0.167   0
## smokeYes                         -0.027   0
## no2gps_24h                        0.093   0
## maxwindspeed_24h                  0.028   0
## precip_24h                       -0.058   0
## sec_noise55_day                   0.020   0
## access_greenbluespaces_300mbuffYes 0.021   0
## age_yrs                           0.019   0
## tmean_24h                        -0.130   0
##
## Deviance Information Criterion (DIC) ...............: 10048.41
## Deviance Information Criterion (DIC, saturated) ....: 3776.35
## Effective number of parameters .....................: 17.90
##
## Marginal log-Likelihood:  -5072.82
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

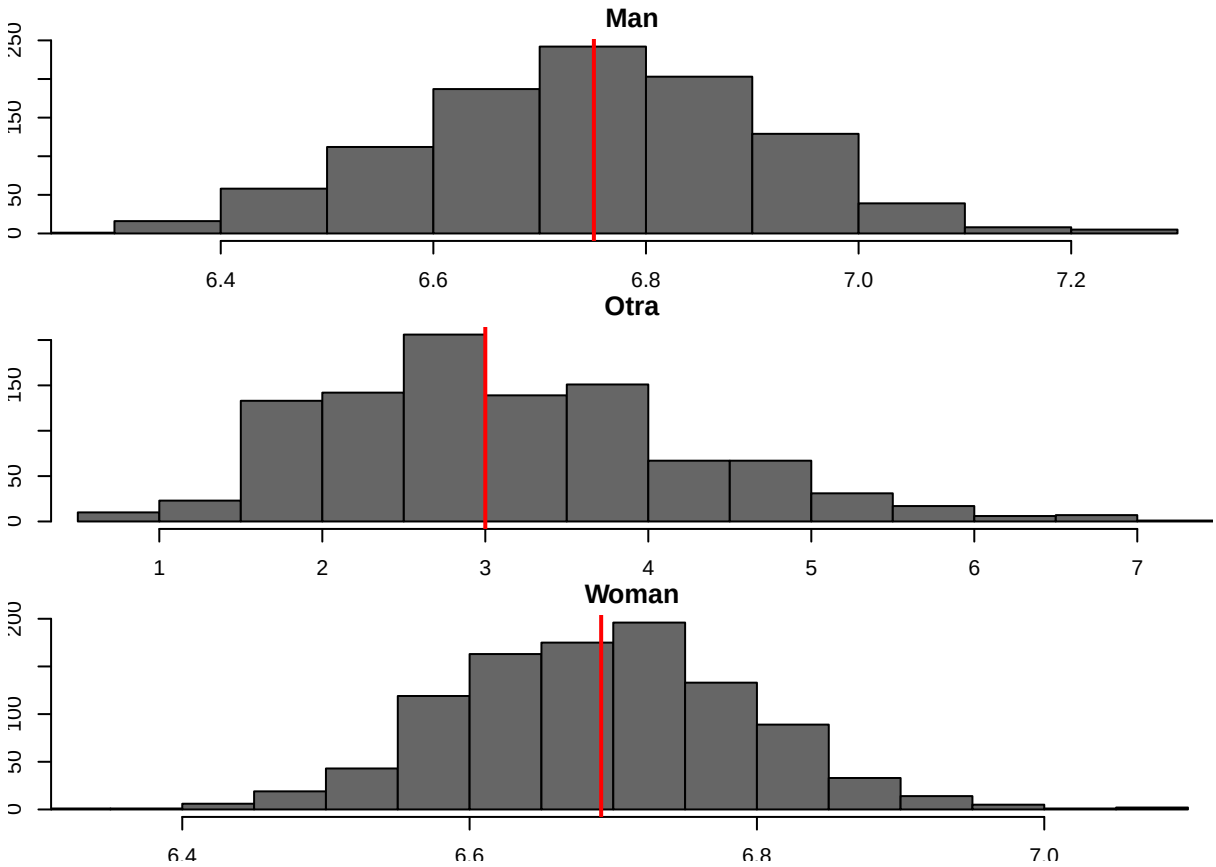
```
summary(model.poisson.random)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.3, Running = 0.487, Post = 0.146, Total = 3.93
## Fixed effects:
##                                    mean    sd 0.025quant 0.5quant 0.975quant
## (Intercept)                       1.529 0.091      1.348    1.529      1.705
## genderOtra                       -0.302 0.356     -0.999   -0.303      0.397
## genderWoman                       0.080 0.052     -0.022    0.080      0.183
## on_a_dietYes                      0.010 0.027     -0.043    0.010      0.062
## alcoholYes                       -0.041 0.027     -0.094   -0.041      0.012
## drugsYes                          0.074 0.106     -0.135    0.074      0.282
## sickYes                          -0.057 0.030     -0.115   -0.057      0.002
## other_factorsYes                 -0.164 0.026     -0.214   -0.164     -0.114
## educationPrimary or less         -0.181 0.161     -0.498   -0.181      0.135
## educationUniversity               0.338 0.084      0.174    0.337      0.505
## smokeYes                         -0.024 0.062     -0.145   -0.024      0.097
## no2gps_24h                        0.115 0.010      0.094    0.115      0.135
## maxwindspeed_24h                 -0.009 0.019     -0.047   -0.009      0.030
## precip_24h                       -0.030 0.020     -0.068   -0.030      0.009
## sec_noise55_day                   0.005 0.015     -0.025    0.005      0.034
## access_greenbluespaces_300mbuffYes  0.033 0.048     -0.061    0.033      0.127
## age_yrs                           0.012 0.027     -0.041    0.012      0.065
## tmean_24h                        -0.242 0.015     -0.271   -0.242     -0.214
##                                    mode kld
## (Intercept)                       1.531   0
## genderOtra                       -0.304   0
## genderWoman                       0.080   0
## on_a_dietYes                      0.010   0
## alcoholYes                       -0.041   0
## drugsYes                          0.075   0
## sickYes                          -0.057   0
## other_factorsYes                 -0.164   0
## educationPrimary or less         -0.180   0
```

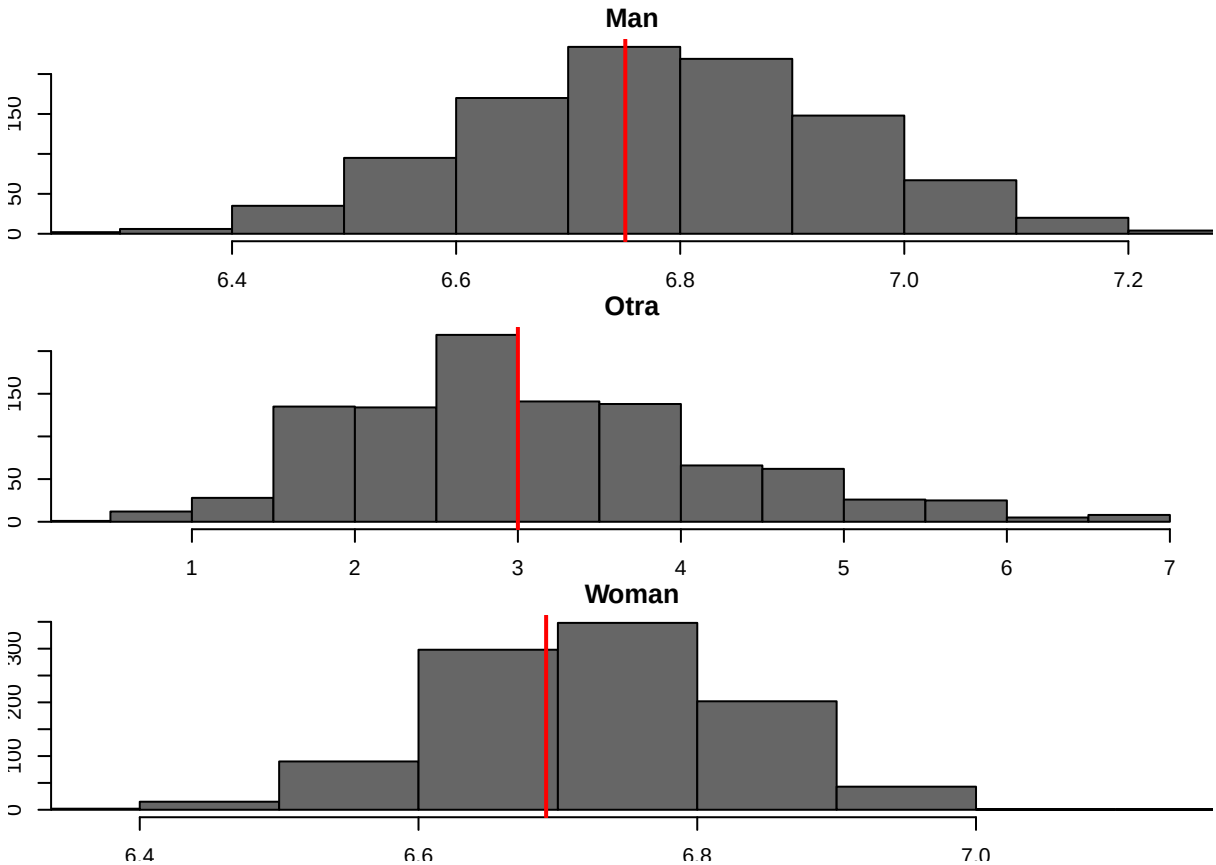
```
## educationUniversity                 0.336   0
## smokeYes                           -0.024   0
## no2gps_24h                          0.115   0
## maxwindspeed_24h                   -0.009   0
## precip_24h                         -0.030   0
## sec_noise55_day                     0.005   0
## access_greenbluespaces_300mbuffYes  0.032   0
## age_yrs                             0.012   0
## tmean_24h                          -0.242   0
##
## Random effects:
##   Name     Model
##     Person_ID IID model
##
## Model hyperparameters:
##                           mean   sd 0.025quant 0.5quant 0.975quant  mode
## Precision for Person_ID 12.01 1.88       8.77    11.86      16.12 11.55
##
## Deviance Information Criterion (DIC) ...............: 9638.25
## Deviance Information Criterion (DIC, saturated) ....: 3366.20
## Effective number of parameters .....................: 174.81
##
## Marginal log-Likelihood:  -4956.78
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

```
nbsamp=1000
n=nrow(study)
yrep1 = matrix(0,nrow=n,ncol=nbsamp)
yrep2 = matrix(0,nrow=n,ncol=nbsamp)

poisson.samples=inla.posterior.sample(n=nbsamp, result=model_poisson)
random.samples=inla.posterior.sample(n=nbsamp, result=model.poisson.random)

predictor.samples.poisson=inla.posterior.sample.eval(function(...) {Predictor},
                                                     poisson.samples)
predictor.samples.random=inla.posterior.sample.eval(function(...) {Predictor},
                                                    random.samples)

for (row.num in 1:n){
  yrep1[row.num,]<- rpois(n=nbsamp,
                          lambda=exp(predictor.samples.poisson[row.num,]))
  yrep2[row.num,]<- rpois(n=nbsamp,
                          lambda=exp(predictor.samples.random[row.num,]))
}

plot.post.pred.test<-function(yrep){
  sadness.per.gender.samples=aggregate(yrep,list(study$gender), mean)
  sadness.per.gender.in.data=aggregate(study$sadness,list(study$gender), mean)
  par(mfrow=c(3,1))
  par(mar=c(1.7,1.7,1.7,1.7))
  for(it in 1:3) {
    x=as.numeric(sadness.per.gender.samples[it,2:(nbsamp+1)])
```

```
    sadness.on.data=sadness.per.gender.in.data[it,2]
    xmin=min(min(x),sadness.on.data)
    xmax=max(max(x),sadness.on.data)
    hist(as.numeric(sadness.per.gender.samples[it,2:(nbsamp+1)]),
       col="gray40",main=sadness.per.gender.samples[it,1],xlim=c(xmin,xmax))
    abline(v=sadness.per.gender.in.data[it,2],col="red",lwd=2)
  }
  par(mfrow=c(1,1))
}
plot.post.pred.test(yrep1)
```




```
plot.post.pred.test(yrep2)
```




```
y_hat_poisson = model_poisson$summary.fitted.values[,1]
y_hat_random = model.poisson.random$summary.fitted.values[,1]
rmse_poisson <- sqrt(mean((y_hat_poisson - study$sadness)^2))
rmse_random <- sqrt(mean((y_hat_random - study$sadness)^2))
cat("RMSE without random effect:", rmse_poisson, "\n")
```

```
## RMSE without random effect: 3.588648
```

```
cat("RMSE with random effect:", rmse_random, "\n")
```

```
## RMSE with random effect: 3.202793
```

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation): In this part, we need to do posterior predictive checks for the sadness score for models with or without random effects. Observing the posterior means, it can be seen that the variable with the largest absolute value is "genderOtra". Therefore, I chose this variable as the test selection. I have used histograms to represent the distribution of sadness predictions by gender, while the lines represent the mean values for the three genders. From the figure, it can be observed that for both models, the red lines for different genders fall in the middle of the histograms, indicating a good fit of the models. For the three different genders, the sadness index for men and women are similar, around 6.8, while the sadness index for "other" is significantly lower, around 3. In terms of RMSE, the value without considering random effects is around 3.58, while after incorporating Person_ID as a random effect, the RMSE decreases to around 3.2. This indicates that the random effect has a positive impact on the model.

e)[10 marks]

Plot the posterior predictive distributions for stroop_test_performance and sadness for the random effect models in part c) for the following new person in the dataset:

Person_ID=286, gender="Woman", on_a_diet="Yes", alcohol="No", drugs="No", sick="No", other_factors="No", education="University", smoke="Yes", no2gps_24h=80, maxwindspeed_24h=10, precip_24h=50, sec_noise55_day=10000, access_greenbluespaces_300mbuff="Ye age_yrs=40, tmean_24h=25

In the case of stroop_test_performance, plot the estimated density, while for sadness, plot a histogram.

Compute the posterior predictive mean, and standard deviation.

Discuss the results.

```
new_person <- data.frame(
  Person_ID = factor("286"),
  gender = factor("Woman"),
  on_a_diet = factor("Yes"),
  alcohol = factor("No"),
  drugs = factor("No"),
  sick = factor("No"),
  other_factors = factor("No"),
  education = factor("University"),
  smoke = factor("Yes"),
  no2gps_24h = 80,
  maxwindspeed_24h = 10,
  precip_24h = 50,
  sec_noise55_day = 10000,
  access_greenbluespaces_300mbuff = factor("Yes"),
  age_yrs = 40,
  tmean_24h = 25,
  stroop_test_performance = NA,
  sadness = NA
)

# select the used variable
select.var = c('gender', 'on_a_diet', 'alcohol', 'drugs', 'sick', 'other_factors',
'education', 'smoke', 'no2gps_24h', 'maxwindspeed_24h', 'precip_24h', 'sec_noise55_day',
'access_greenbluespaces_300mbuff', 'age_yrs', 'tmean_24h','Person_ID','stroop_test_performance','sadnes
study.sub = study[, select.var]

# scale the new data
mean_array <- c()
sd_array <- c()

for (i in 1:length(mean_sd_list)) {
  mean_array[i] <- mean_sd_list[[i]]$mean
  sd_array[i] <- mean_sd_list[[i]]$sd
}
mean_sd_df <- data.frame(variable = vars_to_scale, Mean = mean_array, SD =sd_array )
for (i in 1:nrow(mean_sd_df)) {
  var <- mean_sd_df$variable[i]
  new_person[[var]] <- (new_person[[var]] - mean_sd_df$Mean[i]) / mean_sd_df$SD[i]
}
#combine the data
```

```r
study.sub = rbind(study.sub, new_person)

# use the model with new data
model.linear.random.new <- inla(formula.linear.random, family = "gaussian", data =study.sub,
                            control.fixed = list(mean = 0, prec = 1/100),
                            control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
)

model.poisson.random.new <- inla(formula.poisson.random, family = "poisson", data =study.sub,
                             control.fixed = list(mean = 0, prec = 1/100),
                             control.family = list(link = "log"),
                             control.compute = list(config = TRUE,cpo=TRUE, dic = TRUE)
)
summary(model.linear.random.new)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 4.11, Running = 0.937, Post = 0.259, Total = 5.31
## Fixed effects:
##                                    mean    sd 0.025quant 0.5quant 0.975quant
## (Intercept)                       3.684 0.046      3.594    3.684      3.774
## genderOtra                       -0.017 0.182     -0.374   -0.017      0.339
## genderWoman                       0.004 0.027     -0.049    0.004      0.058
## on_a_dietYes                      0.004 0.011     -0.018    0.004      0.026
## alcoholYes                       -0.001 0.011     -0.023   -0.001      0.021
## drugsYes                         -0.041 0.045     -0.130   -0.041      0.048
## sickYes                          -0.020 0.012     -0.043   -0.020      0.004
## other_factorsYes                 -0.032 0.010     -0.053   -0.032     -0.012
## educationPrimary or less          0.039 0.083     -0.124    0.039      0.201
## educationUniversity               0.163 0.043      0.079    0.163      0.249
## smokeYes                          0.014 0.032     -0.049    0.014      0.076
## no2gps_24h                        0.002 0.004     -0.006    0.002      0.011
## maxwindspeed_24h                  0.016 0.007      0.002    0.016      0.030
## precip_24h                       -0.019 0.007     -0.033   -0.019     -0.005
## sec_noise55_day                  -0.008 0.007     -0.021   -0.008      0.005
```

```
## access_greenbluespaces_300mbuffYes  0.005 0.025     -0.044    0.005      0.054
## age_yrs                             -0.124 0.014     -0.152   -0.124     -0.097
## tmean_24h                           -0.032 0.006     -0.043   -0.032     -0.020
##                                       mode kld
## (Intercept)                          3.684   0
## genderOtra                          -0.017   0
## genderWoman                          0.004   0
## on_a_dietYes                         0.004   0
## alcoholYes                          -0.001   0
## drugsYes                            -0.041   0
## sickYes                             -0.020   0
## other_factorsYes                    -0.032   0
## educationPrimary or less             0.039   0
## educationUniversity                  0.163   0
## smokeYes                             0.014   0
## no2gps_24h                           0.002   0
## maxwindspeed_24h                     0.016   0
## precip_24h                          -0.019   0
## sec_noise55_day                     -0.008   0
## access_greenbluespaces_300mbuffYes  0.005   0
## age_yrs                             -0.124   0
## tmean_24h                           -0.032   0
##
## Random effects:
##   Name     Model
##     Person_ID IID model
##
## Model hyperparameters:
##                                          mean   sd 0.025quant 0.5quant
## Precision for the Gaussian observations 39.90 1.44      37.12    39.88
## Precision for Person_ID                 38.58 4.57      30.30    38.34
##                                         0.975quant  mode
## Precision for the Gaussian observations      42.80 39.84
## Precision for Person_ID                      48.26 37.90
##
## Deviance Information Criterion (DIC) ...............: -1482.13
## Deviance Information Criterion (DIC, saturated) ....: 1778.20
## Effective number of parameters .....................: 104.14
##
## Marginal log-Likelihood:  399.13
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

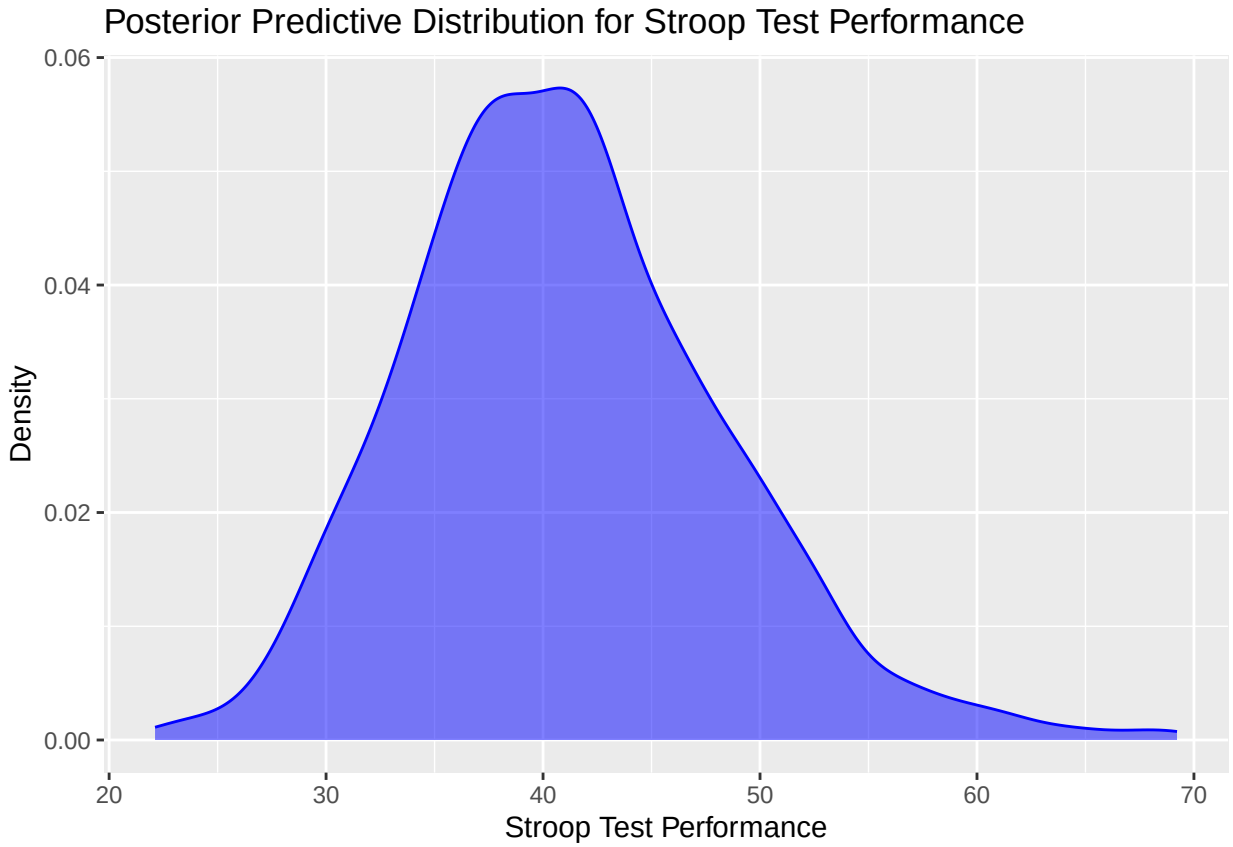
```
summary(model.poisson.random.new)
```

```
##
## Call:
##    c("inla.core(formula = formula, family = family, contrasts = contrasts,
##    ", " data = data, quantiles = quantiles, E = E, offset = offset, ", "
##    scale = scale, weights = weights, Ntrials = Ntrials, strata = strata,
##    ", " lp.scale = lp.scale, link.covariates = link.covariates, verbose =
##    verbose, ", " lincomb = lincomb, selection = selection, control.compute
##    = control.compute, ", " control.predictor = control.predictor,
```

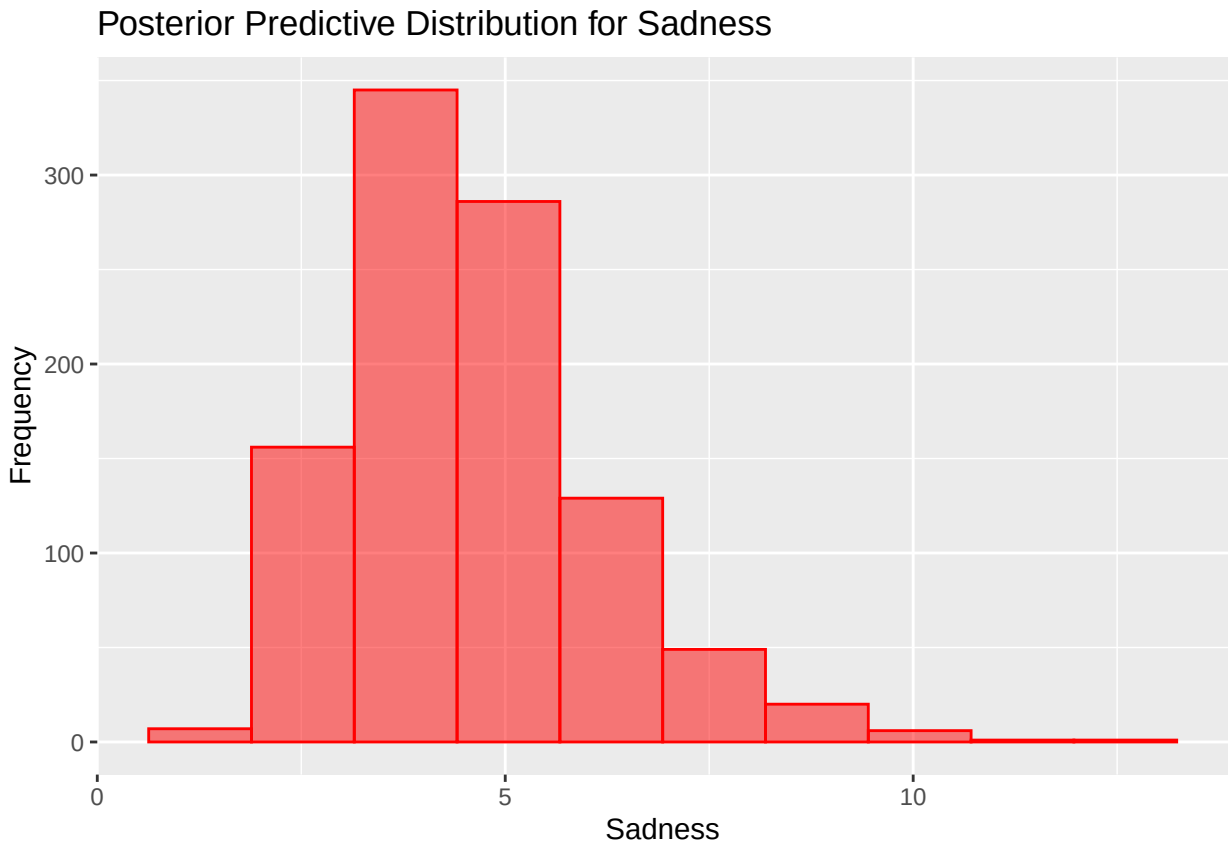
```
##    control.family = control.family, ", " control.inla = control.inla,
##    control.fixed = control.fixed, ", " control.mode = control.mode,
##    control.expert = control.expert, ", " control.hazard = control.hazard,
##    control.lincomb = control.lincomb, ", " control.update =
##    control.update, control.lp.scale = control.lp.scale, ", "
##    control.pardiso = control.pardiso, only.hyperparam = only.hyperparam,
##    ", " inla.call = inla.call, inla.arg = inla.arg, num.threads =
##    num.threads, ", " blas.num.threads = blas.num.threads, keep = keep,
##    working.directory = working.directory, ", " silent = silent, inla.mode
##    = inla.mode, safe = FALSE, debug = debug, ", " .parent.frame =
##    .parent.frame)")
## Time used:
##     Pre = 3.4, Running = 0.603, Post = 0.218, Total = 4.22
## Fixed effects:
##                                    mean    sd 0.025quant 0.5quant 0.975quant
## (Intercept)                       1.531 0.092      1.349    1.532      1.709
## genderOtra                       -0.378 0.399     -1.159   -0.379      0.406
## genderWoman                       0.079 0.052     -0.023    0.079      0.183
## on_a_dietYes                      0.009 0.027     -0.043    0.009      0.062
## alcoholYes                       -0.041 0.027     -0.094   -0.041      0.012
## drugsYes                          0.075 0.107     -0.136    0.075      0.285
## sickYes                          -0.057 0.030     -0.115   -0.057      0.002
## other_factorsYes                 -0.164 0.026     -0.215   -0.164     -0.114
## educationPrimary or less         -0.188 0.165     -0.512   -0.188      0.135
## educationUniversity               0.336 0.085      0.171    0.336      0.505
## smokeYes                         -0.025 0.062     -0.146   -0.025      0.097
## no2gps_24h                        0.115 0.010      0.094    0.115      0.135
## maxwindspeed_24h                 -0.009 0.019     -0.047   -0.009      0.030
## precip_24h                       -0.030 0.020     -0.068   -0.030      0.009
## sec_noise55_day                   0.005 0.015     -0.025    0.005      0.034
## access_greenbluespaces_300mbuffYes 0.032 0.048     -0.062    0.032      0.126
## age_yrs                           0.012 0.027     -0.041    0.012      0.065
## tmean_24h                        -0.242 0.015     -0.271   -0.242     -0.214
##                                    mode kld
## (Intercept)                       1.534   0
## genderOtra                       -0.380   0
## genderWoman                       0.079   0
## on_a_dietYes                      0.009   0
## alcoholYes                       -0.041   0
## drugsYes                          0.076   0
## sickYes                          -0.057   0
## other_factorsYes                 -0.164   0
## educationPrimary or less         -0.187   0
## educationUniversity               0.335   0
## smokeYes                         -0.025   0
## no2gps_24h                        0.115   0
## maxwindspeed_24h                 -0.009   0
## precip_24h                       -0.030   0
## sec_noise55_day                   0.005   0
## access_greenbluespaces_300mbuffYes 0.032   0
## age_yrs                           0.012   0
## tmean_24h                        -0.242   0
##
## Random effects:
```

```
##   Name     Model
##     Person_ID IID model
##
## Model hyperparameters:
##                           mean   sd 0.025quant 0.5quant 0.975quant  mode
## Precision for Person_ID 11.99 1.88       8.75    11.84      16.10 11.54
##
## Deviance Information Criterion (DIC) ...............: 9638.43
## Deviance Information Criterion (DIC, saturated) ....: 3366.38
## Effective number of parameters .....................: 174.99
##
## Marginal log-Likelihood:  -4997.09
## CPO, PIT is computed
## Posterior summaries for the linear predictor and the fitted values are computed
## (Posterior marginals needs also 'control.compute=list(return.marginals.predictor=TRUE)')
```

```
study.linear.samp=inla.posterior.sample(n=nbsamp, result=model.linear.random.new ,selection= list(Predi
study.poisson.samp=inla.posterior.sample(n=nbsamp, result=model.poisson.random.new,selection= list(Pred

predictor.linear.samples=exp(unlist(lapply(study.linear.samp, function(x)(x$latent[1]))))
predictor.poisson.samples=exp(unlist(lapply(study.poisson.samp, function(x)(x$latent[1]))))
library(ggplot2)

# Plot estimated density for stroop_test_performance
ggplot() +
  geom_density(aes(x = predictor.linear.samples), color = "blue", fill = "blue", alpha = 0.5) +
  labs(title = "Posterior Predictive Distribution for Stroop Test Performance",
       x = "Stroop Test Performance",
       y = "Density")
```




```
# Plot histogram for sadness
ggplot() +
  geom_histogram(aes(x = predictor.poisson.samples), color = "red", fill = "red", alpha = 0.5, bins = 1
  labs(title = "Posterior Predictive Distribution for Sadness",
       x = "Sadness",
       y = "Frequency")
```




```
mean(predictor.linear.samples);sd(predictor.linear.samples)
```

```
## [1] 41.02193
```

```
## [1] 7.214152
```

```
mean(predictor.poisson.samples);sd(predictor.poisson.samples)
```

```
## [1] 4.574077
```

```
## [1] 1.527744
```

Explanation (min 300 characters in your own words, otherwise -5 marks for insufficient explanation): We can see that for the posterior predictive density plot of Stroop Test Performance, there is a high probability that this person's score will be around 40. For the histogram, it is highly likely that this person's sadness value is around 4.5. For both models, I calculated the mean and standard deviation. For the Bayesian linear regression, the mean of stroop is 41 and the standard deviation is 7. For the Bayesian Poisson model, the mean of sadness is 4.6 and the standard deviation is 1.4.