Assignment

General Information

- **Points**: Assignment 4 comprises of 6 tasks, 2 points each (12 in total). 2 points are obtained for complete and correct answers. 1 point is obtained for a proper approach or if only part of the task is solved.
- **Submission**: Hand in the assignment as a `Markdown` report (RMarkdown or Quarto) rendered as PDF. The PDF report should show the result(s), the code that produced the result(s), and possibly additional text or comment. Also indicate your name. The report should be uploaded on Moodle until Friday, July 5, 6 pm.
- **Working in teams**: Everyone needs to hand in a report on Moodle. However, the report can be handed in as a team work (max. 2 people). When working in teams, state at the beginning of the document, who you worked with. It Ideally, teams use GitHub and add a link to the GitHub repository to which both contributed.

Additional remarks

ulam and Quarto

Running MCMC with `ulam()` takes some time and produces many messages. This can be annoying when you repeatedly render the `Quarto` document and the goal is to have a clean report. Here are some tips to avoid long rendering times and an ugly document:

- Write the ulam() model in a separate code chunk and give the chunk a name `{r name}`.In addition, specify the following settings at the top of the code chunk to avoid printing the MCMC progress messages of `ulam()` in the PDF document,

```
#| echo: true
#| eval: true
#| output: false

# data list and model
```

- Set the caching option `cache: true` in the YAML header at the start of the document. The first time the code chunk is rendered, its results are cached (stored) in the background. If you leave the code chunk untouched, the results are directly retrieved the next time you render the document. This avoids rerunning a model with every new rendering of the document. The chunk will only be newly evaluated if you actually change its code. To allow `Quarto` to recognize a code chunk, it needs a name.
- If you could not not install the full `rethinking` package, solve the tasks with the `quap()` function instead of the `ulam()` function. However, if `ulam()` is available, make use of it.

Data list and index variables

- Provide `ulam()` a list of variables and values that you need for estimating the model rather than the entire data frame. `ulam()` works more reliable with lists. Moreover, while creating the list, you can also recode variables (e.g., from dummy values (0,1) or group names (male, female) to index values (1,2))
- Using indices for group specific parameters such as in `y = a[G] + b[G]*X`, only works for integer values larger than 0, that is: `G = {1,2,3,...}`. When the variable values are names or include 0, you have to recode it first. Use the function `as.integer(variable)` if the variable is of type `factor` or `as.integer(as.factor(variable))` if the variable is of type `character`. Use `variable + 1` when the variable is dummy coded with 0 and 1 to get values 1 and 2.

```r
#load packages here
library(dplyr)
```

```

Attaching package: 'dplyr'


The following objects are masked from 'package:stats':

    filter, lag


The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v forcats   1.0.0     v readr     2.1.5
v ggplot2   3.5.1     v stringr   1.5.1
v lubridate 1.9.3     v tibble    3.2.1
v purrr     1.0.2     v tidyr     1.3.1

-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
library(ggplot2)
library(tinytex)
library(rethinking)
```

```
Zorunlu paket yükleniyor: cmdstanr
This is cmdstanr version 0.8.0
- CmdStanR documentation and vignettes: mc-stan.org/cmdstanr
- CmdStan path: C:/Users/kerem/.cmdstan/cmdstan-2.35.0
- CmdStan version: 2.35.0
Zorunlu paket yükleniyor: posterior
This is posterior version 1.5.0
```

```

Attaching package: 'posterior'

The following objects are masked from 'package:stats':

    mad, sd, var

The following objects are masked from 'package:base':

    %in%, match

Zorunlu paket yükleniyor: parallel
rethinking (Version 2.40)

Attaching package: 'rethinking'

The following object is masked from 'package:purrr':

    map

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

    rstudent
```

```r
library(rstan)
```

```
Zorunlu paket yükleniyor: StanHeaders

rstan version 2.32.6 (Stan version 2.32.2)

For execution on a local, multicore CPU with excess RAM we recommend calling
options(mc.cores = parallel::detectCores()).
To avoid recompilation of unchanged Stan programs, we recommend calling
rstan_options(auto_write = TRUE)
For within-chain threading using `reduce_sum()` or `map_rect()` Stan functions,
change `threads_per_chain` option:
rstan_options(threads_per_chain = 1)

Do not specify '-march=native' in 'LOCAL_CPPFLAGS' or a Makevars file

Attaching package: 'rstan'

The following objects are masked from 'package:rethinking':
    stan, traceplot

The following objects are masked from 'package:posterior':

    ess_bulk, ess_tail

The following object is masked from 'package:tidyr':

    extract
```

```
# load the data set 'heart.csv' here
heart <- read.csv("heart.csv")
```

Task Set 1

Task 1.1

Run a Bayesian logistic regression model to estimate the risk of men and women to develop a coronary heart disease (TenYearCHD). Provide a summary of the posterior distributions. What is the average probability of men and women to develop the disease?

```
# write data list and model here
heart.chd.gender <- na.omit(heart[, c("male", "TenYearCHD")])

model <- ulam(
  alist(
    TenYearCHD ~ dbinom(1, p),
    logit(p) <- a + bm * male,
    a ~ dnorm(0, 1.5), # Check standard deviations of a and bm
    bm ~ dnorm(0, 0.5)
  ), data = heart.chd.gender, chains = 4, cores = 4
)
```

```
#write code here
#traceplot_ulam(model)
precis(model, depth = 2)
```

```
         mean         sd      5.5%     94.5%     rhat ess_bulk
a  -1.9405531 0.05979012 -2.036445 -1.844219 1.003101 680.7122
bm  0.4714967 0.08293539  0.338348  0.601174 1.005648 647.9497
```

```
samples <- extract.samples(model)
cat("Avg. probability of CHD for women:", inv_logit(mean(samples$a)), "\n")
```

```
Avg. probability of CHD for women: 0.1255871
```

```
cat("Avg. probability of CHD for men:", inv_logit(mean(samples$a + samples$bm)), "\n")
```

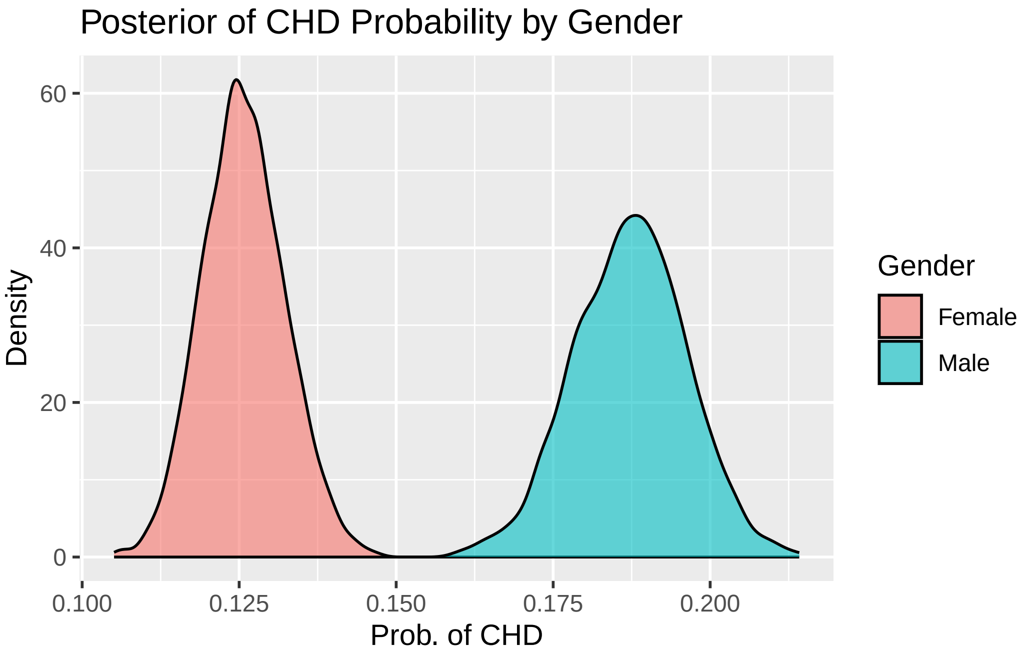
```
Avg. probability of CHD for men: 0.1870861
```

Task 1.2

For the model of `Task 1.1`, visualize the posterior distribution of gender-differences to assess the credibility of the gender difference.

```
samples.df <- data.frame(
  Female = inv_logit(samples$a),
  Male = inv_logit(samples$a + samples$bm)) %>%
  pivot_longer(cols = c(Female, Male), names_to = "Gender", values_to = "Probability")

# Plot the posterior distributions
ggplot(samples.df, aes(x = Probability, fill = Gender)) +
  geom_density(alpha = 0.6) +
  labs(title = "Posterior of CHD Probability by Gender",
       x = "Prob. of CHD",
       y = "Density")
```


Posterior of CHD Probability by Gender
Density
60
40
20
0
0.100
0.125
0.150
0.175
0.200
Prob. of CHD
Gender
Female
Male

Task Set 2

Task 2.1

Run a Bayesian logistic regression model to estimate the risk of men and women with and without diabetes to develop a coronary heart disease (TenYearCHD). Provide a summary of the posterior distributions. Does the effect of diabetes differ between men and women?

```
# write data list and model here
heart.chd.gender.diabetes <- na.omit(heart[, c("male", "diabetes", "TenYearCHD")])

model <- ulam(
  alist(
    TenYearCHD ~ dbinom(1, p),
    logit(p) <- a + bm * male + bd * diabetes + bmd * male * diabetes,
    a ~ dnorm(0, 1.5),
    bm ~ dnorm(0, 0.5),
    bd ~ dnorm(0, 0.5),
    bmd ~ dnorm(0, 0.5)
  ), data = list(
    TenYearCHD = heart.chd.gender.diabetes$TenYearCHD,
    male = heart.chd.gender.diabetes$male,
    diabetes = heart.chd.gender.diabetes$diabetes
  ), chains = 4, cores = 4
)
```

```
# write code here
# Summarize the posterior distributions
precis(model, depth = 2)
```

```
          mean         sd       5.5%      94.5%     rhat ess_bulk
a   -1.9744964 0.06035556 -2.0681585 -1.8777796 1.003144 1233.363
bm   0.4643483 0.08720059  0.3253579  0.6020065 1.001359 1279.046
bd   0.9327300 0.24428969  0.5373650  1.3068867 1.003370 1238.321
bmd  0.2037555 0.31312142 -0.3105971  0.7118741 1.001906 1415.511
```

```
# Extract samples
samples <- extract.samples(model)

# Define the inverse logit function
inv.logit <- function(x) { exp(x) / (1 + exp(x)) }
```

```
# Directly calculate and print probabilities and differences
cat("Avg. probability of CHD for women without diabetes:", mean(inv.logit(samples$a)), "\n",
    "Avg. probability of CHD for men without diabetes:", mean(inv.logit(samples$a + samples$b
    "Avg. probability of CHD for women with diabetes:", mean(inv.logit(samples$a + samples$bd
    "Avg. probability of CHD for men with diabetes:", mean(inv.logit(samples$a + samples$bm +
    "Increase for women:", (increase_women <- mean(inv.logit(samples$a + samples$bd)) - mean
    "Increase for men:", (increase_men <- mean(inv.logit(samples$a + samples$bm + samples$bd
    "Difference:", (difference <- increase_men - increase_women), "\n",
    "So the effect of diabetes on the probability of developing CHD is greater for men than f
```

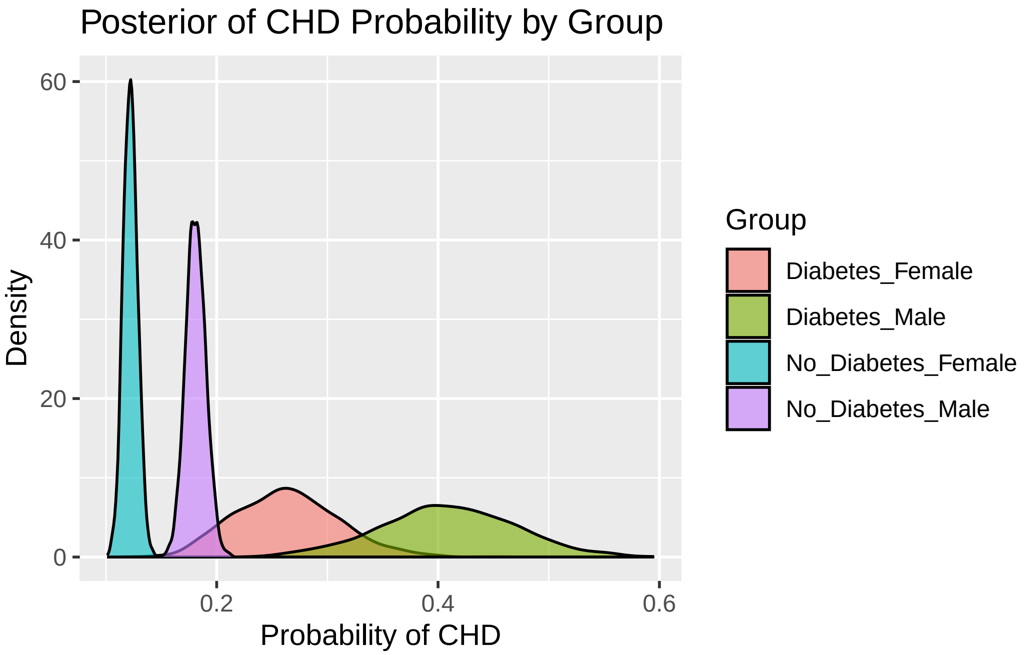
```
Avg. probability of CHD for women without diabetes: 0.122054
 Avg. probability of CHD for men without diabetes: 0.1810941
 Avg. probability of CHD for women with diabetes: 0.2634851
 Avg. probability of CHD for men with diabetes: 0.4091169
 Increase for women: 0.1414311
 Increase for men: 0.2280228
 Difference: 0.08659176
 So the effect of diabetes on the probability of developing CHD is greater for men than for w
```

Task 2.2

For the model of `Task 2.1`, visualize the posterior distributions of each group in one plot to better assess the credibility of the group differences.

```
# write code here
# Create a data frame for visualization
samples.df <- data.frame(
  No_Diabetes_Female = inv.logit(samples$a),
  No_Diabetes_Male = inv.logit(samples$a + samples$bm),
  Diabetes_Female = inv.logit(samples$a + samples$bd),
  Diabetes_Male = inv.logit(samples$a + samples$bm + samples$bd + samples$bmd)
) %>%
  pivot_longer(cols = everything(), names_to = "Group", values_to = "Probability")

# Plot the posterior distributions
ggplot(samples.df, aes(x = Probability, fill = Group)) +
  geom_density(alpha = 0.6) +
  labs(title = "Posterior of CHD Probability by Group",
       x = "Probability of CHD",
       y = "Density")
```


Posterior of CHD Probability by Group
Density
60
40
20
0
0.2
0.4
0.6
Probability of CHD
Group
Diabetes_Female
Diabetes_Male
No_Diabetes_Female
No_Diabetes_Male

Task Set 3

Task 3.1

Run a Bayesian logistic regression model to estimate the effect of age on the risk of developing a coronary heart disease (TenYearCHD), separately for women and men. Ensure that the regression intercept represents the risk of women and men with average age. Provide a summary of the posterior distributions.

```r
# write data list and model here
heart <- na.omit(heart[, c("male", "age", "TenYearCHD")])

# Center the age variable around the mean to interpret the intercept as the risk at average a
heart <- heart %>%
  mutate(age_centered = age - mean(age))

# Split data by gender
heart.male <- heart %>% filter(male == 1)
heart.female <- heart %>% filter(male == 0)

# Fit the Bayesian logistic regression model for males using ulam
model.male <- ulam(
  alist(
    TenYearCHD ~ dbinom(1, p),
    logit(p) <- a + b_age * age_centered,
    a ~ dnorm(0, 1.5),
    b_age ~ dnorm(0, 0.5)
  ), data = list(
    TenYearCHD = heart.male$TenYearCHD,
    age_centered = heart.male$age_centered
  ), chains = 4, cores = 4
)

# Fit the Bayesian logistic regression model for females using ulam
model.female <- ulam(
  alist(
    TenYearCHD ~ dbinom(1, p),
    logit(p) <- a + b_age * age_centered,
    a ~ dnorm(0, 1.5),
    b_age ~ dnorm(0, 0.5)
  ), data = list(
    TenYearCHD = heart.female$TenYearCHD,
    age_centered = heart.female$age_centered
  ), chains = 4, cores = 4
)
```

```r
# write code here
# Summarize the posterior distributions
precis(model.male, depth = 2)
```

```
             mean          sd        5.5%       94.5%     rhat  ess_bulk
a     -1.54447866 0.064537344 -1.64869255 -1.44008270 1.005217  992.6127
b_age  0.06838808 0.007233704  0.05686918  0.07965403 1.002391 1296.4652
```

```r
precis(model.female, depth = 2)
```

```
            mean          sd        5.5%       94.5%     rhat  ess_bulk
a     -2.1653803 0.075442131 -2.28892715 -2.04666900 1.003372  812.8664
b_age  0.0848521 0.007864394  0.07218019  0.09768639 1.001775 1127.3047
```

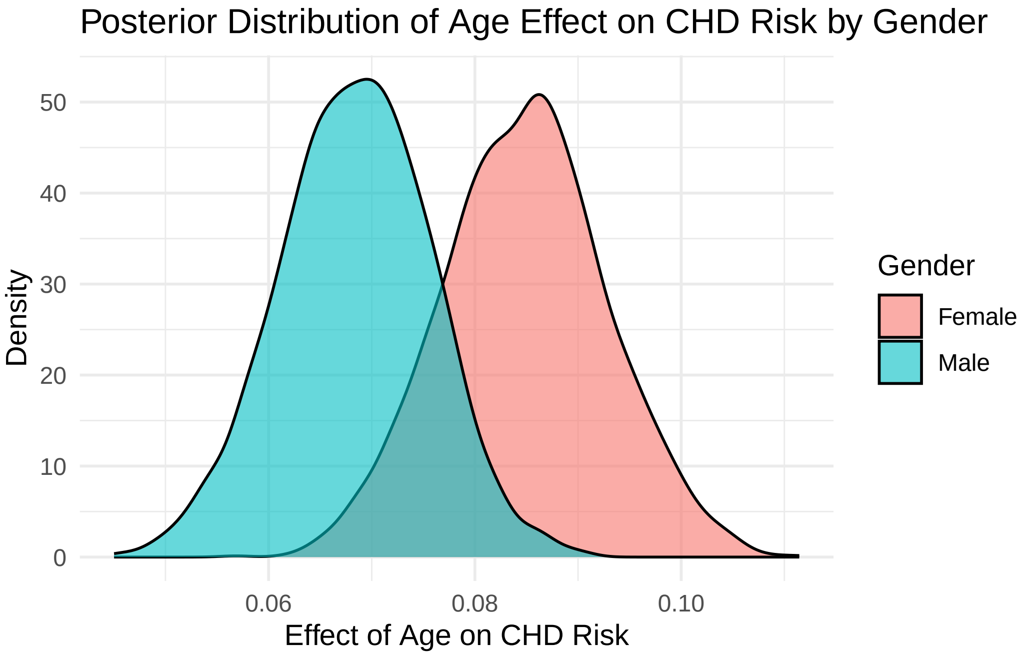
```r
# Extract samples
samples.male <- extract.samples(model.male)
samples.female <- extract.samples(model.female)
```

Task 3.2

For the model of `Task 3.1`, visualize the posterior distribution of differences in the age effect between women and men. Does age increase the risk of developing the disease and does this effect differ between women and men?

```r
# write code here
samples.df <- data.frame(
  Male = samples.male$b_age,
  Female = samples.female$b_age
) %>%
  pivot_longer(cols = c(Male, Female), names_to = "Gender", values_to = "Age.Effect")

# Plot the posterior distributions
ggplot(samples.df, aes(x = Age.Effect, fill = Gender)) +
  geom_density(alpha = 0.6) +
  labs(title = "Posterior Distribution of Age Effect on CHD Risk by Gender",
       x = "Effect of Age on CHD Risk",
       y = "Density") +
  theme_minimal()
```



```
# Calculate and print differences in the age effect
age.effect.diff <- samples.male$b_age - samples.female$b_age

cat("Average age effect for men:", mean(samples.male$b_age), "\n")
```

```
Average age effect for men: 0.06838808
```

```
cat("Average age effect for women:", mean(samples.female$b_age), "\n")
```

```
Average age effect for women: 0.0848521
```

```
cat("Difference in age effect between men and women:", mean(age.effect.diff), "\n")
```

```
Difference in age effect between men and women: -0.01646402
```

```
cat("So the effect of age on the probability of developing CHD is",
    ifelse(mean(age.effect.diff) > 0, "greater for men than for women", "greater for women th
```

```
So the effect of age on the probability of developing CHD is greater for women than for men
```