Group project

For the project, we use the following packages:

```
## ...
library(tidyr)
library(dplyr)
library(glue)
library(brms)
library(ggplot2)
library(priorsense)
library(mlr3measures)
library(testthat)
library(rstudioapi)

options(mc.cores = parallel::detectCores()) # paralellize if possible
options(brms.file_refit = "on_change") # save the files if the model has changed
ggplot2::theme_set(ggplot2::theme_light()) # nicer theme
```

1. Dataset Selection (0.5pt)

Select a dataset with clusters such as schools, regions, or people with multiple observations per individual. (From for example, https://www.kaggle.com/) It would be a good idea to choose a smallish dataset (not too many rows, e.g., less than 1000) or subset it so that fitting the models doesn't take too long.

a. Describe the dataset with a couple of short sentences. What was its intended use? Are there papers that reference it? Provide information on how to obtain the dataset, including its source and any necessary preprocessing steps/feature engineering.

The following "Sleep Health and Lifestyle Dataset" (Tharmalingam, 2023) is designed to provide insights into various factors influencing sleep patterns and overall health.

The data is synthetic, and the team received permission from the course professor to use it for this project. To the best of the team's knowledge, it has not been referenced in any prior literature.

However, it is accompanied by 162 related notebooks on Kaggle. The dataset can be accessed on https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset.

Data Processing:

We began by one-hot encoding the categorical variable "Gender," assigning binary values: 0 for male and 1 for female. Similarly, we addressed the categories in the "BMI" variable, combining "overweight" and "obese" into one category, and merging "normal" and "normal weight" into another.

Afterward, we one-hot encoded the "BMI" variable into binary values: 0 for normal weight and 1 for overweight. Additionally, we one-hot encoded the "Sleep Disorder" variable, using 0 to represent no sleep disorder and 1 for having a sleep disorder (including insomnia or sleep apnea).

Finally, we split the "Blood Pressure" variable into its systolic and diastolic components. Systolic blood pressure represents the highest arterial pressure during systole, while diastolic blood pressure represents the lowest pressure during diastole.

Data Preprocessing

```
# Load the data and check distinct values for variables that need processing
sleep <- read.csv("Sleep_health_and_lifestyle_dataset.csv")
distinct(sleep, Gender)
```

```
  Gender
1   Male
2 Female
```

```
distinct(sleep, BMI.Category)
```

```
   BMI.Category
1    Overweight
2        Normal
3         Obese
4 Normal Weight
```

```
distinct(sleep, Sleep.Disorder)
```

```
  Sleep.Disorder
1           None
2    Sleep Apnea
3       Insomnia
```

```
# Group variables in BMI.Category and Sleep.Disorder, then one-hot encode them
sleep <- sleep %>%
  mutate(
    Gender = ifelse(Gender == "Male", 0, 1),  # One-Hot Encode Gender
    BMI.Category = ifelse(BMI.Category %in% c("Overweight", "Obese"), 1, 0),  # Group and encode BMI ca
    Sleep.Disorder = ifelse(Sleep.Disorder == "None", 0, 1)  # Encode Sleep Disorder
  )

# Split Blood.Pressure into Systolic and Diastolic Blood Pressure components
sleep <- sleep %>%
  separate(Blood.Pressure, into = c("systolic.BP", "diastolic.BP"), sep = "/")

# Convert systolic and diastolic values to numeric
sleep$systolic.BP <- as.numeric(sleep$systolic.BP)
sleep$diastolic.BP <- as.numeric(sleep$diastolic.BP)
```

b. Report the number of observations, columns (with their meaning) and their data types. Indicate clearly what you will use as dependent variable/label.

Dataset Overview

The dataset consists of **374 observations** and **13 columns**, containing the following variables:

- **Person ID**: A unique identifier for each individual.

- **Gender (Integer)**: The gender of the person (Male = 0, Female = 1).
- **Age (Integer)**: The age of the individual in years.
- **Occupation (String)**: The person's occupation or profession.
- **Sleep Duration (Float, hours)**: The number of hours the person sleeps per day.
- **Quality of Sleep (Integer, scale: 1-10)**: A subjective rating of sleep quality, ranging from 1 (poor) to 10 (excellent).
- **Physical Activity Level (Integer, minutes/day)**: The number of minutes the individual engages in physical activity each day.
- **Stress Level (Integer, scale: 1-10)**: A subjective rating of the person's stress level, ranging from 1 (low) to 10 (high).
- **BMI Category (Integer)**: The BMI category of the person (e.g., Underweight, Normal, Overweight).
- **Blood Pressure (Integer, systolic/diastolic)**: Blood pressure measurements, indicated as systolic over diastolic pressure.
- **Heart Rate (Integer, bpm)**: The person's resting heart rate in beats per minute.
- **Daily Steps (Integer)**: The number of steps the individual takes per day.
- **Sleep Disorder (Integer)**: Indicates whether the person has a sleep disorder (None, Insomnia, Sleep Apnea).

We will use **Sleep Duration** as the dependent variable, with all other variables serving as independent variables.

2. Split the data and tranform columns as necessary. (0.5pt)

Split the data into training (80%) and test set (80%). Transform the columns if necessary.

```
# Set the random seed for reproducibility
set.seed(235711)

# Create a vector to randomly assign observations to training (80%) and testing (20%) sets
ind <- sample(c(rep("train", nrow(sleep) * 0.80), rep("test", nrow(sleep) * 0.20)))

# Split the data into training and testing sets
tmp <- split(sleep, ind)
```

```
Warning in split.default(x = seq_len(nrow(x)), f = f, drop = drop, ...): Datenlänge ist
kein Vielfaches der Split-Variablen
```

```
# Create data.frames for the training and testing sets
sleep_train <- tmp$train
sleep_test <- tmp$test

# Save the training and testing sets as .rds files
saveRDS(sleep_train, "sleep_train.rds")
saveRDS(sleep_test, "sleep_test.rds")

# Define the working directory
thiswork_directory <- "C:/Users/Xuxu/Downloads"
```

3. Model Exploration (3pt)

a. Fit multiple appropriate models to the dataset (as many models as there are members in the group, with a minimum of two models). Models might vary in the multilevel structure, informativeness of

their priors (but not just trivial changes), model of the data/likelihood, etc. (I recommend not to use no pooling models since they tend to take a long time and it's very hard to assign good priors).

```
# # Define Person_3 Custom Priors
my_prior <- c(
  prior(normal(7.5, 1.3), class = Intercept),  # Prior for intercept
  prior(normal(120, 71), class = b, coef = systolic.BP),  # Prior for systolic blood pressure
  prior(normal(80, 45), class = b, coef = diastolic.BP),  # Prior for diastolic blood pressure
  prior(exponential(1.2), class = sigma)  # Prior for residual standard deviation
)

# Set the file path for saving the model
fileyt <- file.path(thiswork_directory, "Person_1_model")

# Build the Person_1 model
Person_1_model <- brm(
  formula = Sleep.Duration ~ BMI.Category + systolic.BP + diastolic.BP + Stress.Level + (Stress.Level |
  data = sleep_train,
  prior = my_prior,
  control = list(adapt_delta = 0.9),
  file = fileyt
)
```

```
summary(Person_1_model)
```

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: Sleep.Duration ~ BMI.Category + systolic.BP + diastolic.BP + Stress.Level + (Stress.Level || O
   Data: sleep_train (Number of observations: 300)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Multilevel Hyperparameters:
~Occupation (Number of levels: 10)
                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)       1.43      0.41     0.86     2.44 1.00     1829     2545
sd(Stress.Level)    0.23      0.07     0.14     0.41 1.00     1108     1265

Regression Coefficients:
             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept        8.52      0.79     6.99    10.09 1.00     3137     3104
BMI.Category    -0.55      0.09    -0.73    -0.37 1.00     5218     2991
systolic.BP     -0.07      0.01    -0.09    -0.05 1.00     4155     3117
diastolic.BP     0.11      0.01     0.08     0.14 1.00     4064     3014
Stress.Level    -0.27      0.08    -0.44    -0.11 1.00     1696     1923

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.22      0.01     0.21     0.24 1.00     5638     3121

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

```r
# Define Person_2 Gelman Priors
gelman_priors <- c(
  prior(normal(7, 2), class = Intercept),  # Prior for intercept
  prior(normal(0, 2), class = b),  # Apply the same prior to all slopes
  prior(exponential(1.2458), class = sd)  # Prior for standard deviation
)

# Set the file path for saving the model
filexj <- file.path(thiswork_directory, "Person_2_model")

# Build the Person_2 model
Person_2_model <- brm(
  formula = Sleep.Duration ~ Quality.of.Sleep + Stress.Level + BMI.Category + Heart.Rate + Sleep.Disorde
            Sleep.Disorder:Stress.Level + (1 + Stress.Level || Occupation),
  data = sleep_train,
  prior = gelman_priors,
  family = lognormal(),
  control = list(adapt_delta = 0.99, max_treedepth = 12),
  seed = 123,
  file = filexj
)
```

```r
summary(Person_2_model)
```

```
 Family: lognormal
  Links: mu = identity; sigma = identity
Formula: Sleep.Duration ~ Quality.of.Sleep + Stress.Level + BMI.Category + Heart.Rate + Sleep.Disorder +
   Data: sleep_train (Number of observations: 300)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Multilevel Hyperparameters:
~Occupation (Number of levels: 10)
                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)       0.16      0.05     0.09     0.29 1.00     1676     1949
sd(Stress.Level)    0.03      0.01     0.01     0.05 1.00     1802     2436

Regression Coefficients:
                           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                      1.80      0.12     1.56     2.05 1.00     3192     2920
Quality.of.Sleep               0.05      0.01     0.03     0.06 1.00     5723     3161
Stress.Level                  -0.02      0.01    -0.04     0.00 1.00     1609     2210
BMI.Category                  -0.04      0.01    -0.06    -0.02 1.00     6128     3447
Heart.Rate                    -0.00      0.00    -0.00     0.00 1.00     4243     3585
Sleep.Disorder                -0.03      0.02    -0.08     0.01 1.00     4528     2996
Age                           -0.00      0.00    -0.00     0.00 1.00     5158     3634
Stress.Level:Sleep.Disorder    0.01      0.00    -0.00     0.01 1.00     4357     3169

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.03      0.00     0.03     0.04 1.00     5907     2794

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
```

and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).

```
# Define Person_3 Gelman Priors
gelmanPrior <- c(
  prior(normal(7.158667, 2.006821), class = Intercept),  # Prior for intercept
  prior(normal(0, 1.635375), class = b, coef = Quality.of.Sleep),  # Prior for Quality of Sleep
  prior(normal(0, 4.063786), class = b, coef = Sleep.Disorder),  # Prior for Sleep Disorder
  prior(exponential(1.245751), class = sigma),  # Prior for residual standard deviation
  prior(exponential(1.245751), class = sd)  # Prior for standard deviation of random effects
)

# Set the file path for saving the model
filejn <- file.path(thiswork_directory, "Person_3_model")

# Build the Person_3 model
Person_3_model <- brm(
  formula = Sleep.Duration ~ 1 + Quality.of.Sleep + Sleep.Disorder +
            (1 + Quality.of.Sleep || Occupation),
  data = sleep_train,
  family = lognormal(),
  prior = gelmanPrior,
  iter = 3000,
  control = list(adapt_delta = 0.99, max_treedepth = 12),
  seed = 123,
  file = filejn
)
```

```
summary(Person_3_model)
```

```
 Family: lognormal
  Links: mu = identity; sigma = identity
Formula: Sleep.Duration ~ 1 + Quality.of.Sleep + Sleep.Disorder + (1 + Quality.of.Sleep || Occupation)
   Data: sleep_train (Number of observations: 300)
  Draws: 4 chains, each with iter = 3000; warmup = 1500; thin = 1;
         total post-warmup draws = 6000

Multilevel Hyperparameters:
~Occupation (Number of levels: 10)
                    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)           0.18      0.06     0.08     0.33 1.00     1660     2634
sd(Quality.of.Sleep)    0.03      0.01     0.01     0.06 1.00     1712     2536

Regression Coefficients:
                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept           1.39      0.07     1.27     1.55 1.00     3243     3691
Quality.of.Sleep    0.08      0.01     0.06     0.10 1.00     2169     2652
Sleep.Disorder     -0.01      0.01    -0.03     0.00 1.00     6977     4137

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.04      0.00     0.04     0.04 1.00     6592     4525
```

```
Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

```r
# Scale the numerical features for the training set
k_sleep_train <- sleep_train %>%
  mutate(across(c(Age, Quality.of.Sleep, Physical.Activity.Level, Stress.Level, BMI.Category, systolic.

# Scale the numerical features for the test set
k_sleep_test <- sleep_test %>%
  mutate(across(c(Age, Quality.of.Sleep, Physical.Activity.Level, Stress.Level, BMI.Category, systolic.

# Define Person_4's custom priors
priors_person_4 <- c(
  set_prior("student_t(3, 7.34, 1.2)", class = "Intercept"),
  set_prior("student_t(3, 0, 1.5)", class = "sd"),
  set_prior("student_t(3, 0, 2)", class = "sd", group = "Occupation"),
  set_prior("student_t(3, 0, 1.5)", class = "sigma")
)

# Set the file path for saving the model
file0 <- file.path(thiswork_directory, "Person_4_model")

# Build the Person_4 model
Person_4_model <- brm(
  formula = Sleep.Duration ~ Stress.Level + Physical.Activity.Level +
            (1 + Stress.Level + Physical.Activity.Level || Occupation),
  data = k_sleep_train,
  prior = priors_person_4,
  control = list(adapt_delta = 0.99, max_treedepth = 13),
  file = file0
)
```

```r
summary(Person_4_model)
```

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: Sleep.Duration ~ Stress.Level + Physical.Activity.Level + (1 + Stress.Level + Physical.Activity
   Data: k_sleep_train (Number of observations: 300)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Multilevel Hyperparameters:
~Occupation (Number of levels: 10)
                           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)                  0.20      0.10     0.08     0.45 1.00     1359     2094
sd(Stress.Level)               0.44      0.16     0.24     0.85 1.00     1491     1898
sd(Physical.Activity.Level)    0.41      0.15     0.21     0.77 1.00     1717     2461

Regression Coefficients:
                      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept                 7.20      0.09     7.01     7.36 1.00     1393     2302
Stress.Level             -0.34      0.18    -0.71    -0.02 1.01     1374     1673
```

```
Physical.Activity.Level     0.51      0.17     0.21     0.86 1.00     1434     2048

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.18      0.01     0.17     0.20 1.00     5639     3086

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

```r
# Dominik's Gelman Priors
Dominik_priors <- c(
  prior(normal(7.13, 1.98), class = Intercept),  # Prior for intercept
  prior(normal(0, 1.12), class = b, coef = "Stress.Level"),  # Prior for Stress Level
  prior(normal(0, 1.635375), class = b, coef = "Quality.of.Sleep"),  # Prior for Quality of Sleep
  prior(normal(0, 0.48), class = b, coef = "Heart.Rate"),  # Prior for Heart Rate
  prior(exponential(1.245751), class = sigma),  # Prior for residual standard deviation
  prior(exponential(1.245751), class = sd)  # Prior for standard deviation of random effects (Tau)
)

# Set the file path for saving the model
filedmk <- file.path(thiswork_directory, "Dominik_model")

# Dominik's Model
Dominik_Model <- brm(
  formula = Sleep.Duration ~ Stress.Level + Heart.Rate + Quality.of.Sleep + (1 | Occupation),
  data = sleep_train,
  family = gaussian(),
  prior = Dominik_priors,
  seed = 123,
  control = list(adapt_delta = 0.999, max_treedepth = 12),
  file = filedmk
)
```

```r
summary(Dominik_Model)
```

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: Sleep.Duration ~ Stress.Level + Heart.Rate + Quality.of.Sleep + (1 | Occupation)
   Data: sleep_train (Number of observations: 300)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Multilevel Hyperparameters:
~Occupation (Number of levels: 10)
              Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)     0.28      0.08     0.17     0.48 1.00     1087     1818

Regression Coefficients:
                Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept           3.23      0.72     1.87     4.70 1.00     2408     1861
Stress.Level       -0.15      0.03    -0.21    -0.09 1.00     2198     2127
Heart.Rate          0.02      0.01     0.01     0.03 1.00     4128     2843
```

```
Quality.of.Sleep     0.45      0.05     0.35     0.54 1.00     2060     2487

Further Distributional Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.28      0.01     0.25     0.30 1.00     2762     2485

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

b. Explain each model and describe its structure (what they assume about potential population- level or group-level effects), and the type of priors used.

Person_1 model:

Person_1's model posits that sleep duration varies by stress levels across different occupations, with a focus on key variables such as BMI category, systolic blood pressure (BP), diastolic BP, and stress level. Custom priors were applied, where the average adult sleep duration was set to 7.5 hours with a standard deviation of 1.3, based on established research. For systolic and diastolic BP, the means were set to 120 and 80, with standard deviations of 71 and 45, respectively. The residual standard deviation was given an exponential prior distribution, with a rate of 1/sd(y).

The prior for average sleep duration was informed by the study conducted by Liu et al. (2016), while the systolic and diastolic blood pressure priors were derived from the recommendations provided by the World Health Organization (WHO, 2023).

Person_2 model:

The Person_2 model explores the relationship between Sleep.Duration and several independent variables, including Quality.of.Sleep, Stress.Level, BMI.Category, Heart.Rate, Sleep.Disorder, and Age. These variables were selected due to their high correlation (greater than 0.3) with the target variable, Sleep.Duration. The model also includes an interaction between Sleep.Disorder and Stress.Level, as well as a grouping effect for Occupation. This allows the model to account for both the general impact of stress on sleep duration and the variability of this effect across different occupation types.

To ensure that the posterior distribution contains only positive values, the model employs the lognormal family. The priors used in the model follow recommendations from Gelman et al. (2020), ensuring well-informed prior assumptions and improving the model's robustness.

Person_3 model:

Person_3's model examines the relationship between sleep duration (the dependent variable) and the independent variables: sleep quality and the presence of a sleep disorder. It employs a multi-level structure, grouping by occupation type, with the assumption that occupation influences both the intercept and the slope of the sleep quality variable. The model's priors are based on those suggested by Gelman et al. (2020). The lognormal family was chosen to ensure that the posterior distribution remains positive, as sleep duration cannot be negative.

Initially, the model encountered a warning related to the maximum tree depth. However, given that the model's Rhat and Effective Sample Size (ESS) values were appropriate, the issue was attributed not to model misspecification but rather to computational constraints. Following the recommendation in the error message, the maximum tree depth was increased to resolve the issue (Stan Development Team, 2022; R: Effective Sample Size (ESS), n.d.

Person_4 model:

Person_4's model is based on the assumption that Stress.Level and Physical.Activity.Level are two of the most significant predictors of Sleep.Duration. These variables may vary across different occupations, although it is assumed that their effect on Sleep.Duration remains similar across different occupation types.

For the priors, an article was used that reported the average sleep duration in 2011 to be 7.34 hours, with a standard deviation of less than 0.98. However, this standard deviation did not work well in the current model, so it was slightly increased to improve model fit.

Additionally, based on the assumption that the differences in sleep patterns between various occupations should not be substantial, as well as the expectation of less variability in Stress.Level and Physical.Activity.Level, the standard deviation for all priors was reduced to reflect this lower variability.

The reference for the average sleep duration in 2011 is Hublin et al. (2020).

Dominik_Model:

Dominik's model predicts Sleep.Duration based on Stress Level, Heart Rate, and Quality of Sleep, with varying intercepts for each Occupation. The priors for the model are informed by Gelman et al. (2020), and it utilizes the Gaussian family.

For the fixed effects (population-level effects), it is assumed that Stress Level, Heart Rate, and Quality of Sleep have a linear relationship with Sleep.Duration across all observations. By incorporating a random effect (group-level effect) for Occupation, the model allows the intercept of sleep duration to vary across different occupation types.

An estimate of 0.24 for sd(Intercept) suggests that sleep duration does vary between occupations, although the extent of this variation is relatively small. Similar to Person_3's model, there was an initial issue with running the model due to maximum tree depth constraints. This issue was resolved by increasing the maximum tree depth, as recommended.

4. Model checking (3pt)

a. Perform a prior sensitivity analysis for each model and modify the model if appropriate. Justify.

```
powerscale_sensitivity(Person_1_model)
```

```
Sensitivity based on cjs_dist:
# A tibble: 29 x 4
   variable                              prior likelihood diagnosis
   <chr>                                 <dbl>      <dbl> <chr>
 1 b_Intercept                         0.00956     0.0641 -
 2 b_BMI.Category                      0.00107     0.0934 -
 3 b_systolic.BP                       0.00221     0.0771 -
 4 b_diastolic.BP                      0.00214     0.0774 -
 5 b_Stress.Level                      0.00979     0.0144 -
 6 sd_Occupation__Intercept            0.0424      0.0560 -
 7 sd_Occupation__Stress.Level         0.00988     0.0530 -
 8 sigma                               0.00114     0.211  -
 9 Intercept                           0.0312      0.0119 -
10 r_Occupation[Accountant,Intercept] 0.0171      0.0266 -
# i 19 more rows
```

```
powerscale_sensitivity(Person_2_model)
```

```
Sensitivity based on cjs_dist:
# A tibble: 32 x 4
   variable                           prior likelihood diagnosis
   <chr>                              <dbl>      <dbl> <chr>
 1 b_Intercept                     0.00474      0.146  -
```

```
 2 b_Quality.of.Sleep            0.00106      0.170  -
 3 b_Stress.Level                0.00545      0.0257 -
 4 b_BMI.Category                0.000946     0.107  -
 5 b_Heart.Rate                  0.000852     0.167  -
 6 b_Sleep.Disorder              0.000481     0.135  -
 7 b_Age                         0.000820     0.0948 -
 8 b_Stress.Level:Sleep.Disorder 0.000621     0.131  -
 9 sd_Occupation__Intercept      0.0169       0.0933 -
10 sd_Occupation__Stress.Level   0.00424      0.0752 -
# i 22 more rows
```

```
powerscale_sensitivity(Person_3_model)
```

```
Sensitivity based on cjs_dist:
# A tibble: 27 x 4
   variable                            prior likelihood diagnosis
   <chr>                               <dbl>      <dbl> <chr>
 1 b_Intercept                       0.00748      0.117  -
 2 b_Quality.of.Sleep                0.0109       0.0826 -
 3 b_Sleep.Disorder                  0.00321      0.227  -
 4 sd_Occupation__Intercept          0.0190       0.255  -
 5 sd_Occupation__Quality.of.Sleep   0.00815      0.244  -
 6 sigma                             0.000944     0.294  -
 7 Intercept                         0.0133       0.0485 -
 8 r_Occupation[Accountant,Intercept] 0.00579     0.0761 -
 9 r_Occupation[Doctor,Intercept]    0.00303      0.249  -
10 r_Occupation[Engineer,Intercept]  0.00792      0.0258 -
# i 17 more rows
```

```
powerscale_sensitivity(Person_4_model)
```

```
Sensitivity based on cjs_dist:
# A tibble: 38 x 4
   variable                                prior likelihood diagnosis
   <chr>                                   <dbl>      <dbl> <chr>
 1 b_Intercept                           0.00330      0.0539 -
 2 b_Stress.Level                        0.00336      0.0292 -
 3 b_Physical.Activity.Level             0.00142      0.0315 -
 4 sd_Occupation__Intercept              0.00544      0.207  -
 5 sd_Occupation__Stress.Level           0.0106       0.0621 -
 6 sd_Occupation__Physical.Activity.Level 0.00736     0.0966 -
 7 sigma                                 0.000289     0.254  -
 8 Intercept                             0.00330      0.0539 -
 9 r_Occupation[Accountant,Intercept]    0.00333      0.0919 -
10 r_Occupation[Doctor,Intercept]        0.00317      0.0457 -
# i 28 more rows
```

```
powerscale_sensitivity(Dominik_Model)
```

```
Sensitivity based on cjs_dist:
# A tibble: 17 x 4
   variable                                     prior likelihood diagnosis
```

	<chr>	<dbl>	<dbl>	<chr>
1	b_Intercept	0.00125	0.104	-
2	b_Stress.Level	0.00151	0.0916	-
3	b_Heart.Rate	0.00102	0.107	-
4	b_Quality.of.Sleep	0.00113	0.0896	-
5	sd_Occupation__Intercept	0.0291	0.0325	-
6	sigma	0.00231	0.166	-
7	Intercept	0.00463	0.0139	-
8	r_Occupation[Accountant,Intercept]	0.00486	0.0335	-
9	r_Occupation[Doctor,Intercept]	0.00398	0.0280	-
10	r_Occupation[Engineer,Intercept]	0.00386	0.0202	-
11	r_Occupation[Lawyer,Intercept]	0.00387	0.0237	-
12	r_Occupation[Nurse,Intercept]	0.00445	0.0173	-
13	r_Occupation[Sales.Representative,Intercept]	0.00489	0.101	-
14	r_Occupation[Salesperson,Intercept]	0.00375	0.0246	-
15	r_Occupation[Scientist,Intercept]	0.00275	0.0407	-
16	r_Occupation[Software.Engineer,Intercept]	0.00363	0.0444	-
17	r_Occupation[Teacher,Intercept]	0.00465	0.0520	-

Person_1 model:

The prior sensitivity analysis for Person_1's model revealed no conflicts between the specified priors and the likelihood. This indicates that the priors were well-aligned with the data, and no adjustments were necessary.

Person_2 model:

The diagnostic analysis for Person_2's model showed no conflicts, suggesting that the model and priors are well-specified and appropriate for the data. No changes were made based on these results.

Person_3 model:

Person_3's model passed all checks using the powerscale_sensitivity function, with no issues flagged during diagnosis. Since the fitting process ran smoothly without errors, the model is considered properly specified, and no changes were required.

Person_4 model:

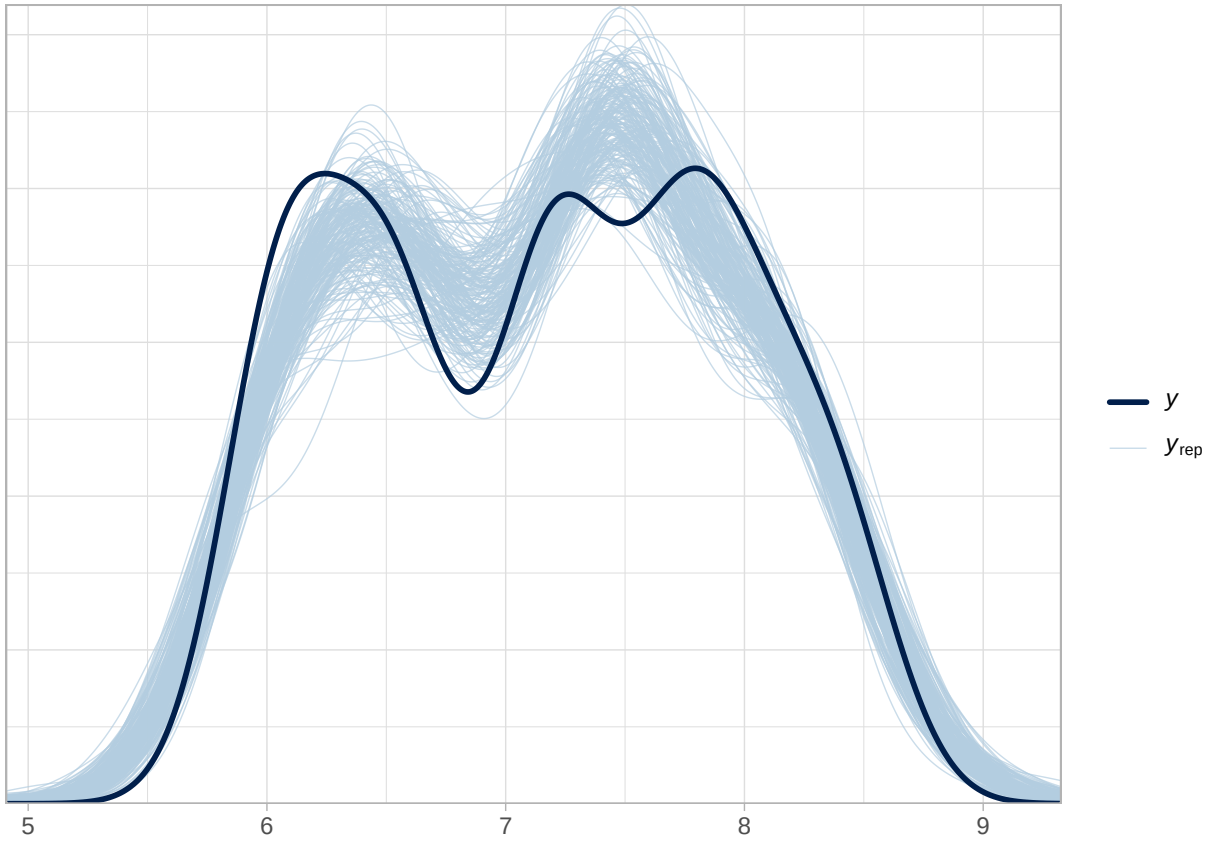
Person_4's model also showed no conflicts in the prior sensitivity analysis. This confirms that the priors were appropriately chosen, and the model was not altered as no adjustments were needed.

Dominik_Model:

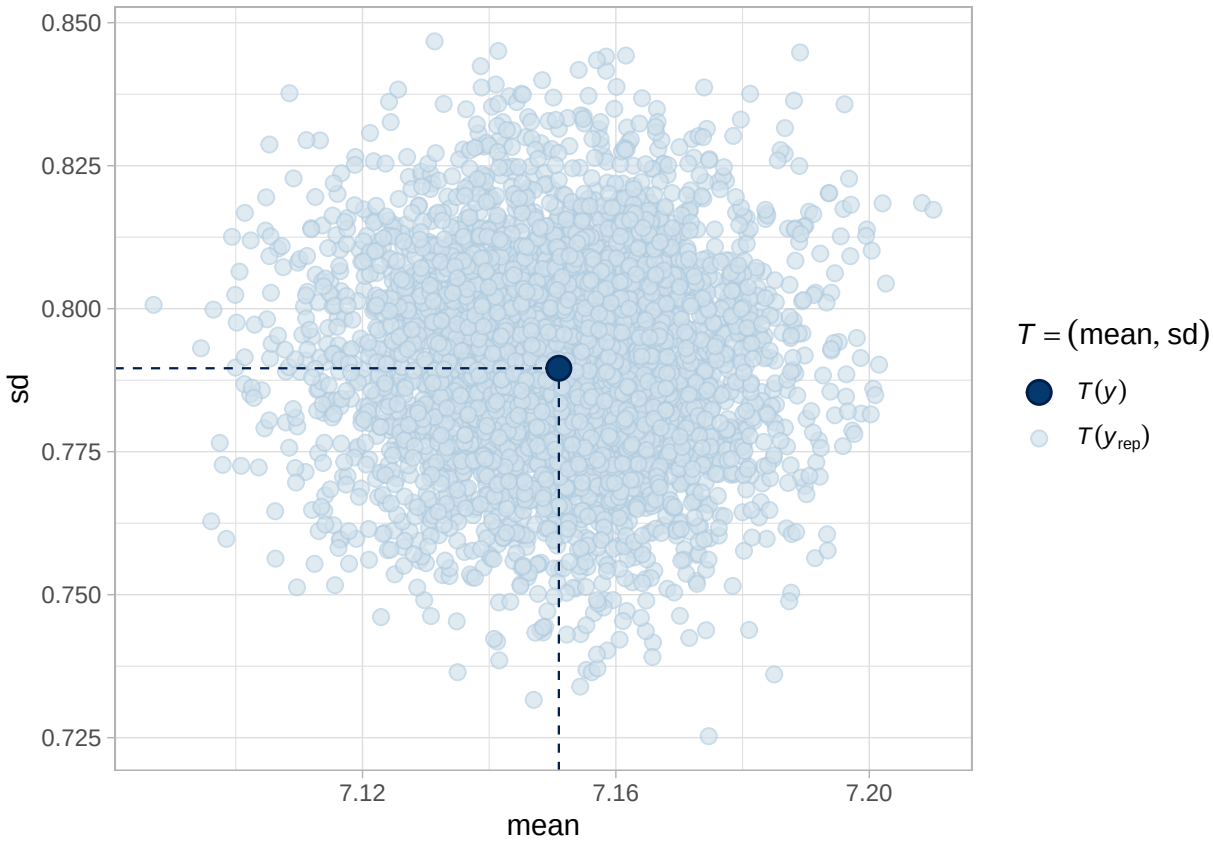
The sensitivity analysis for Dominik's model indicated no conflicts between the priors and the data. This included checks for both prior-data conflicts and the balance between strong and weak priors. As a result, the chosen priors were deemed suitable, and no bias was detected in the posterior inference, meaning no modifications were necessary.

b. Conduct posterior predictive checks for each model to assess how well they fit the data. Explain what you conclude.

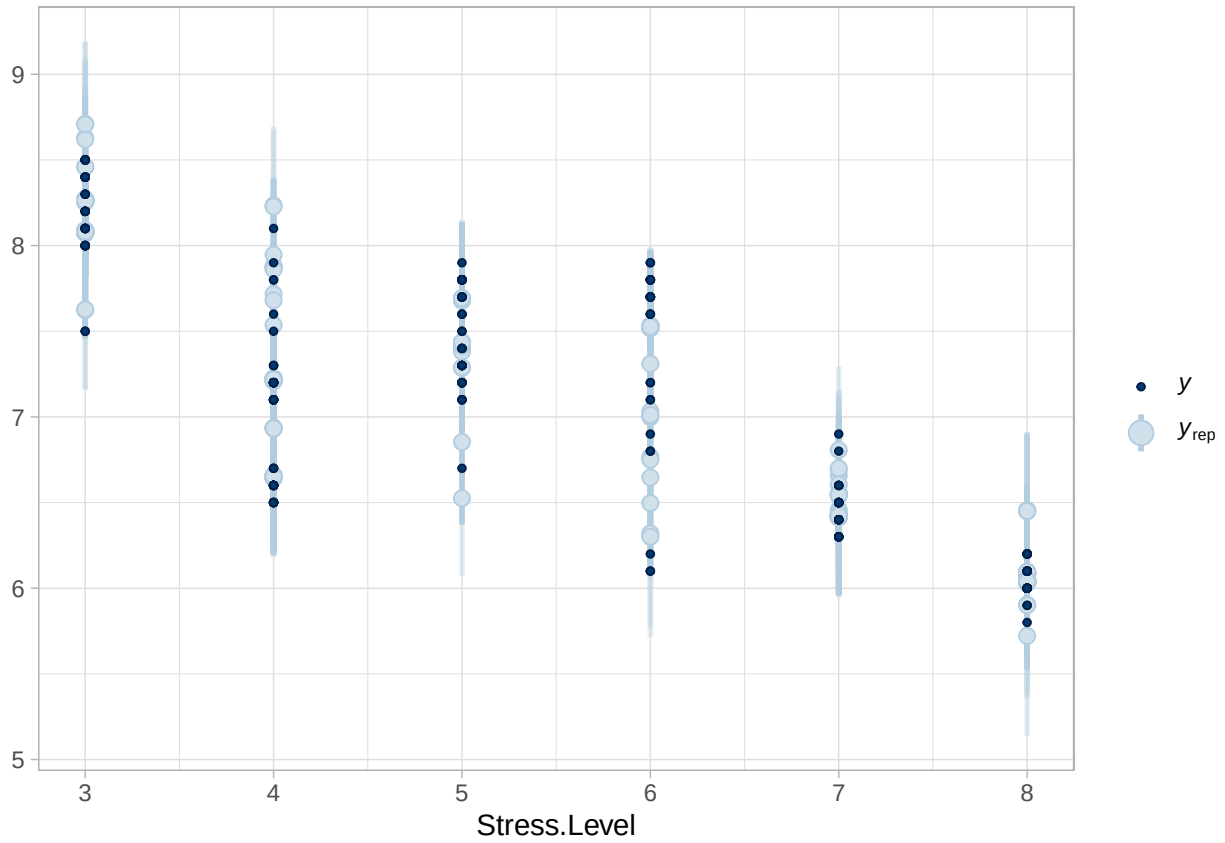
```
# Posterior Predictive Checks for Person_1_model
pp_check(Person_1_model, ndraws = 200)
```



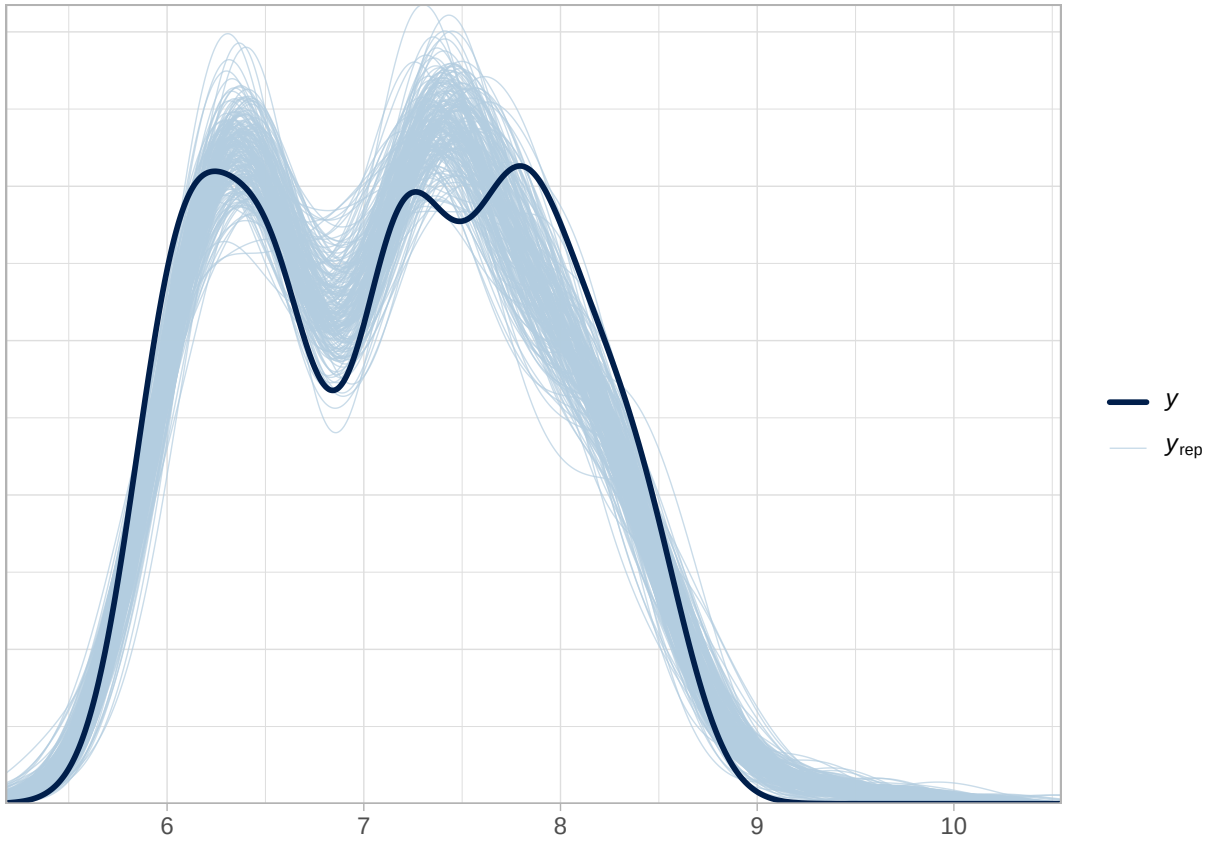

```
pp_check(Person_1_model, type = "stat_2d")
```



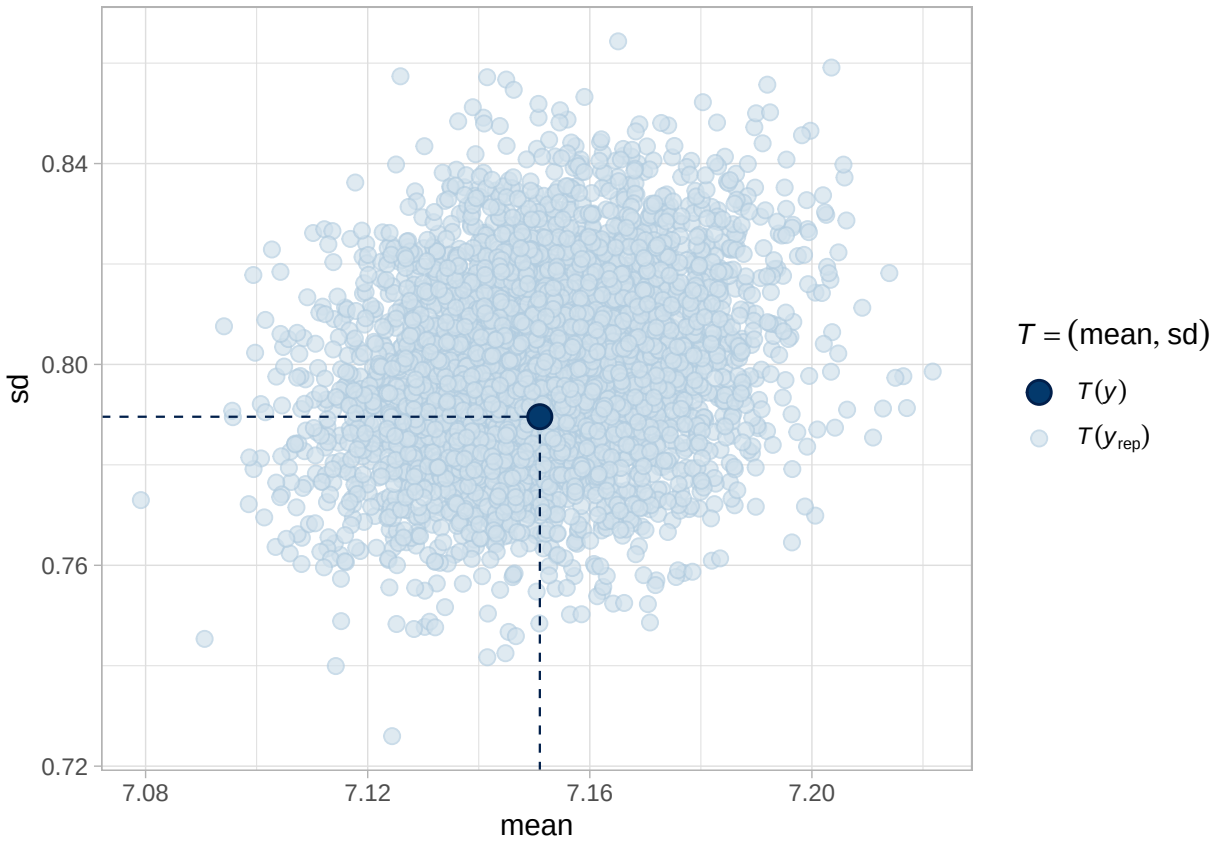

```
pp_check(Person_1_model, type = "intervals", x = "Stress.Level", prob_outer = 0.95) +
  xlab("Stress.Level")
```



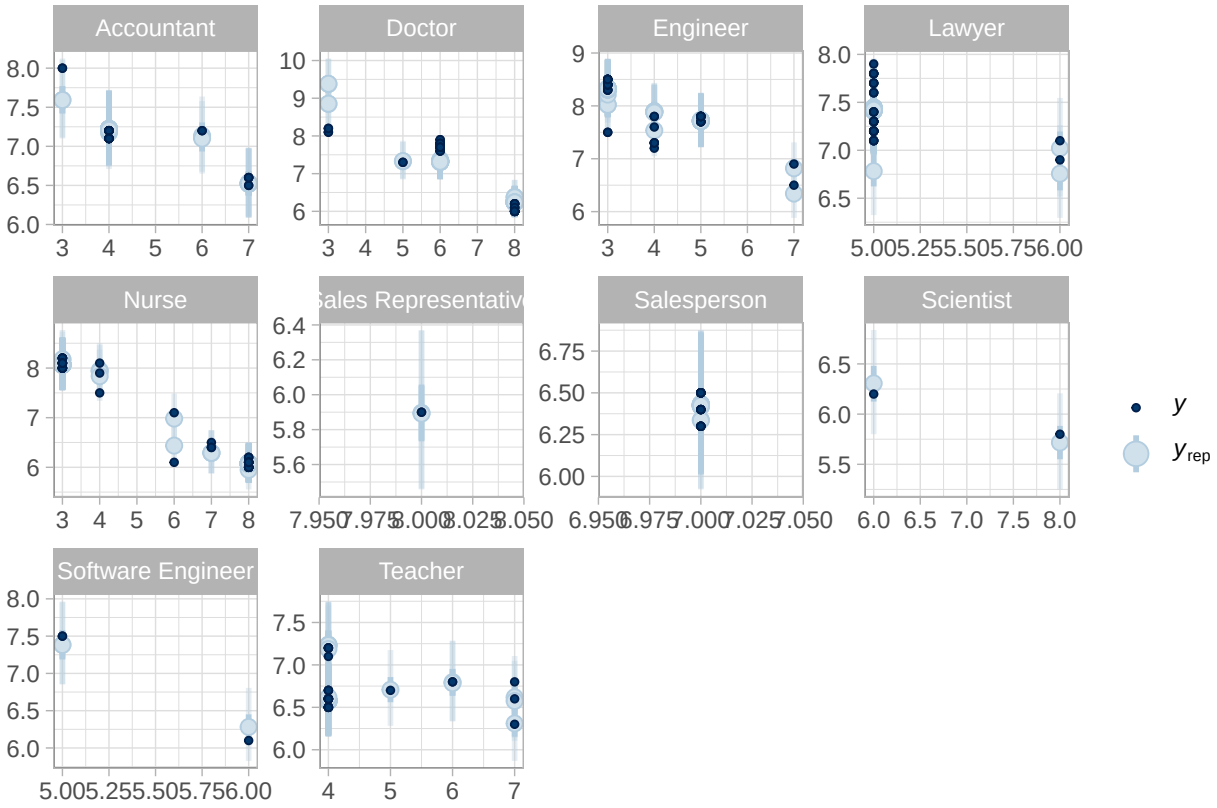

```
# Posterior Predictive Checks for Person_2_model
pp_check(Person_2_model, ndraws = 200)
```



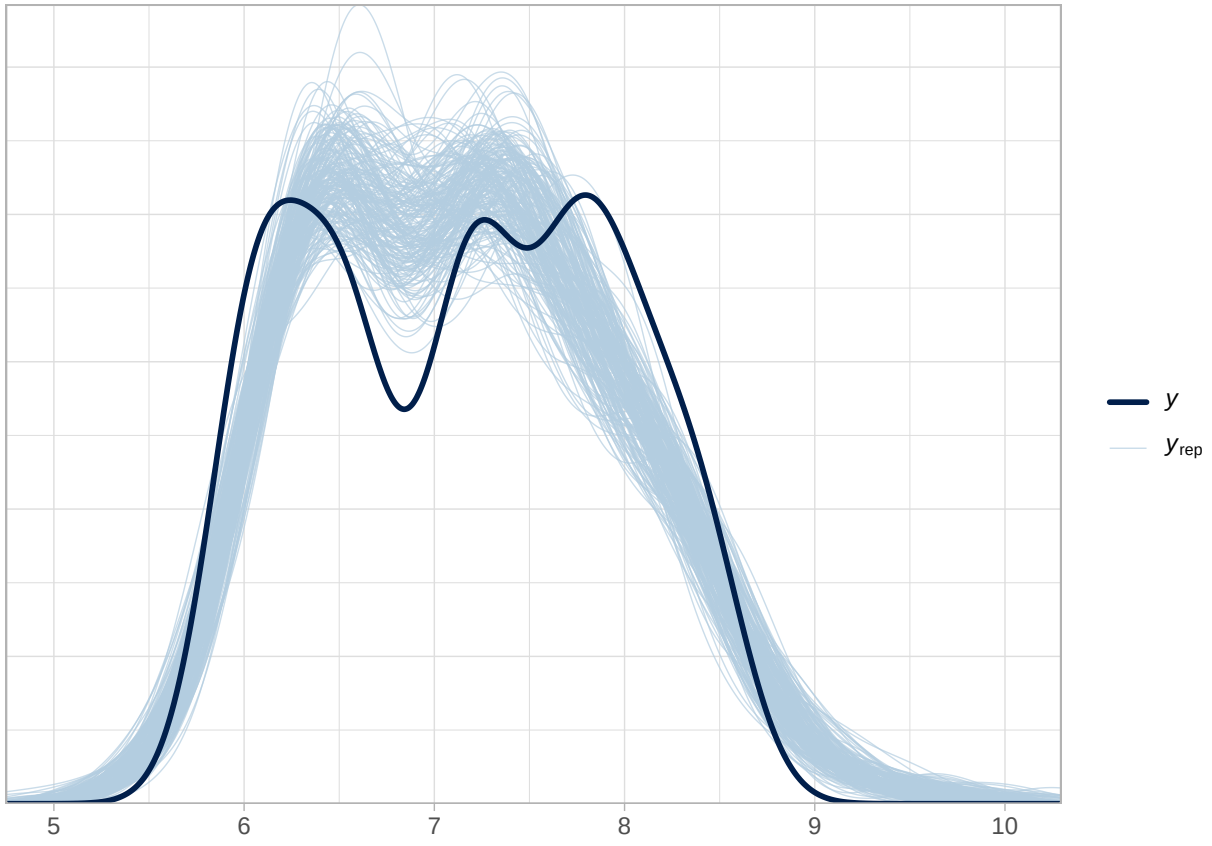

```
pp_check(Person_2_model, type = "stat_2d")
```



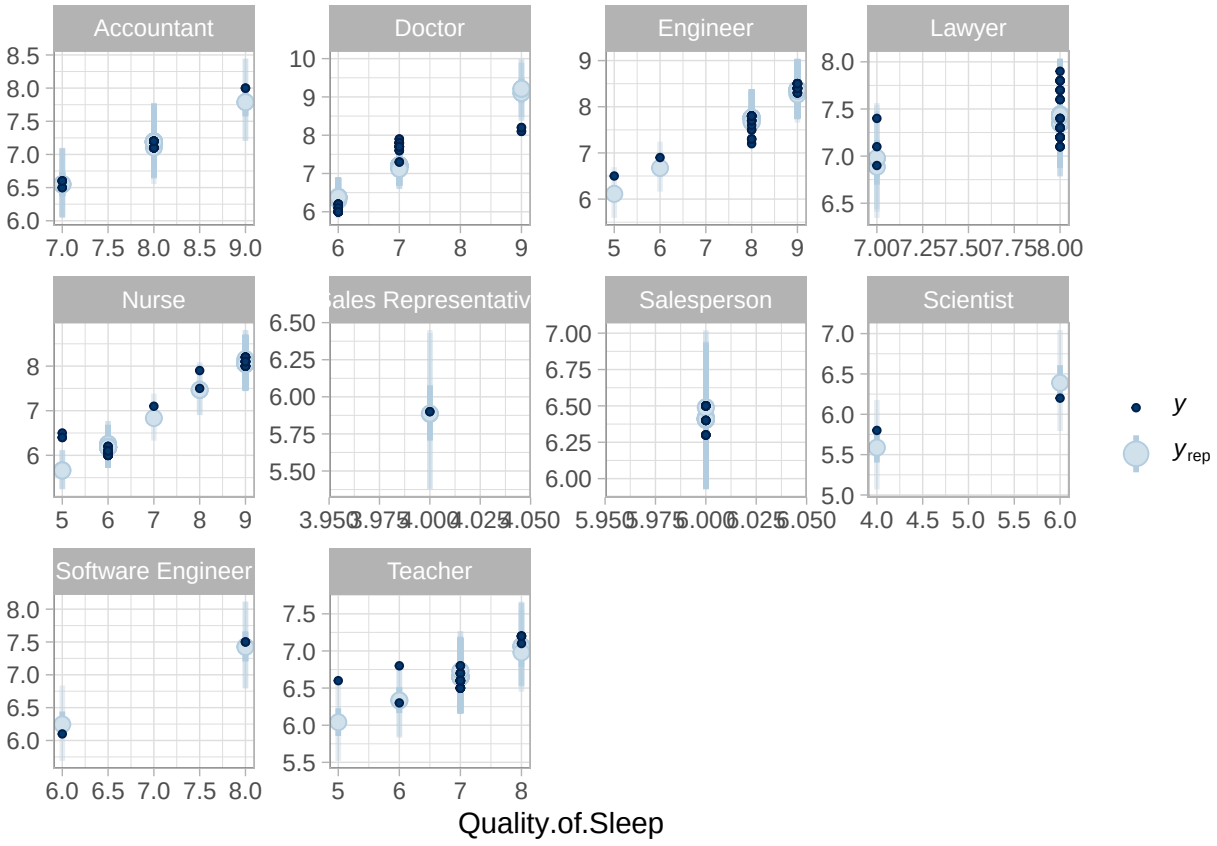

```
pp_check(Person_2_model, type = "intervals_grouped", x = "Stress.Level", group = "Occupation", prob_outer
  xlab("Stress.Level")
```



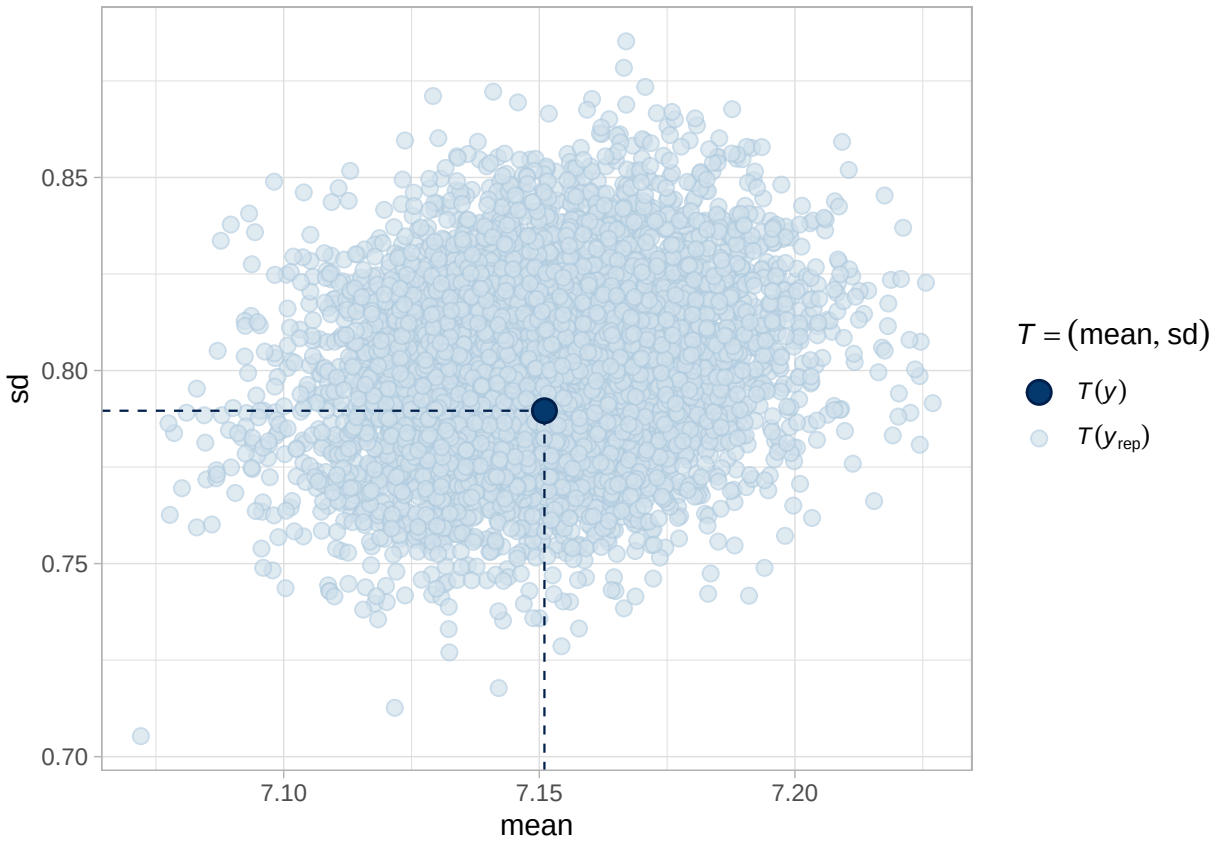

```
# Posterior Predictive Checks for Person_3_model
pp_check(Person_3_model, ndraws = 200)
```



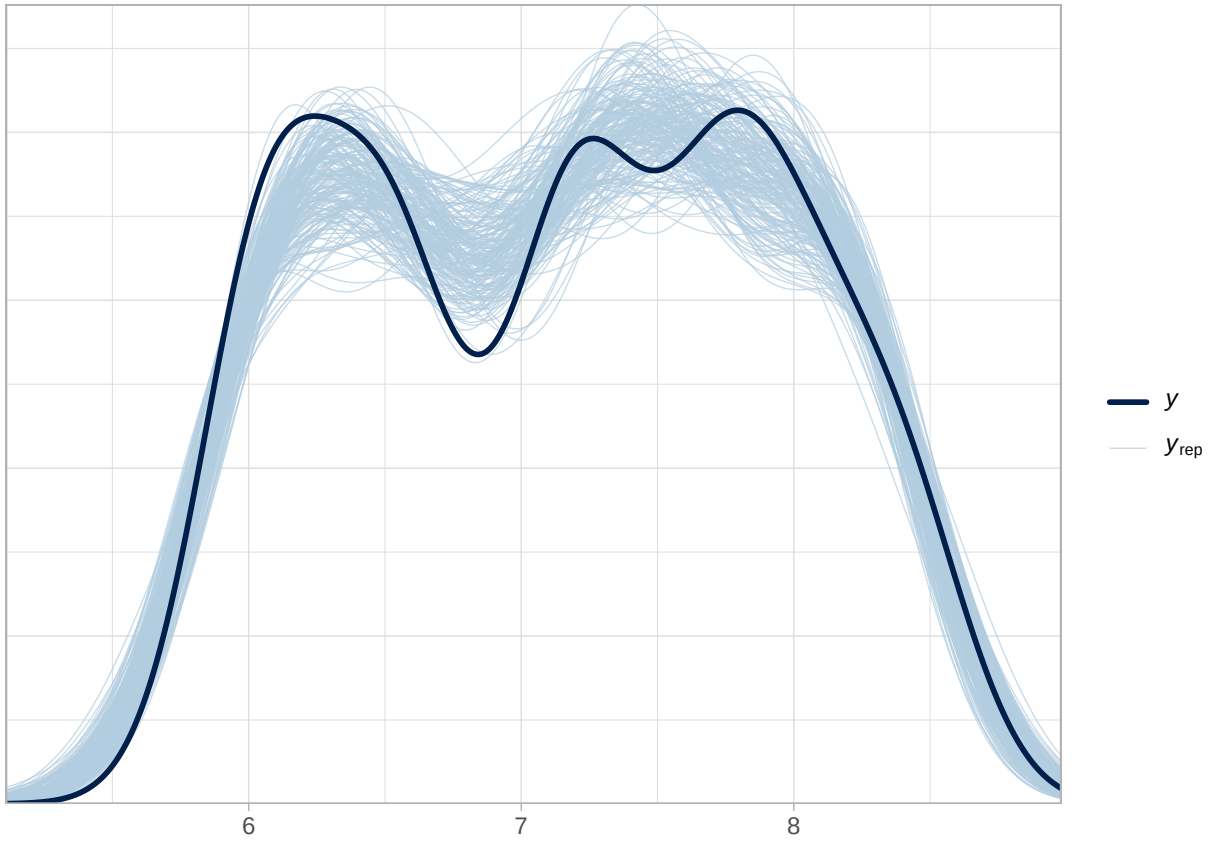

```
pp_check(Person_3_model, type = "intervals_grouped", x = "Quality.of.Sleep", group = "Occupation", prob
  xlab("Quality.of.Sleep")
```



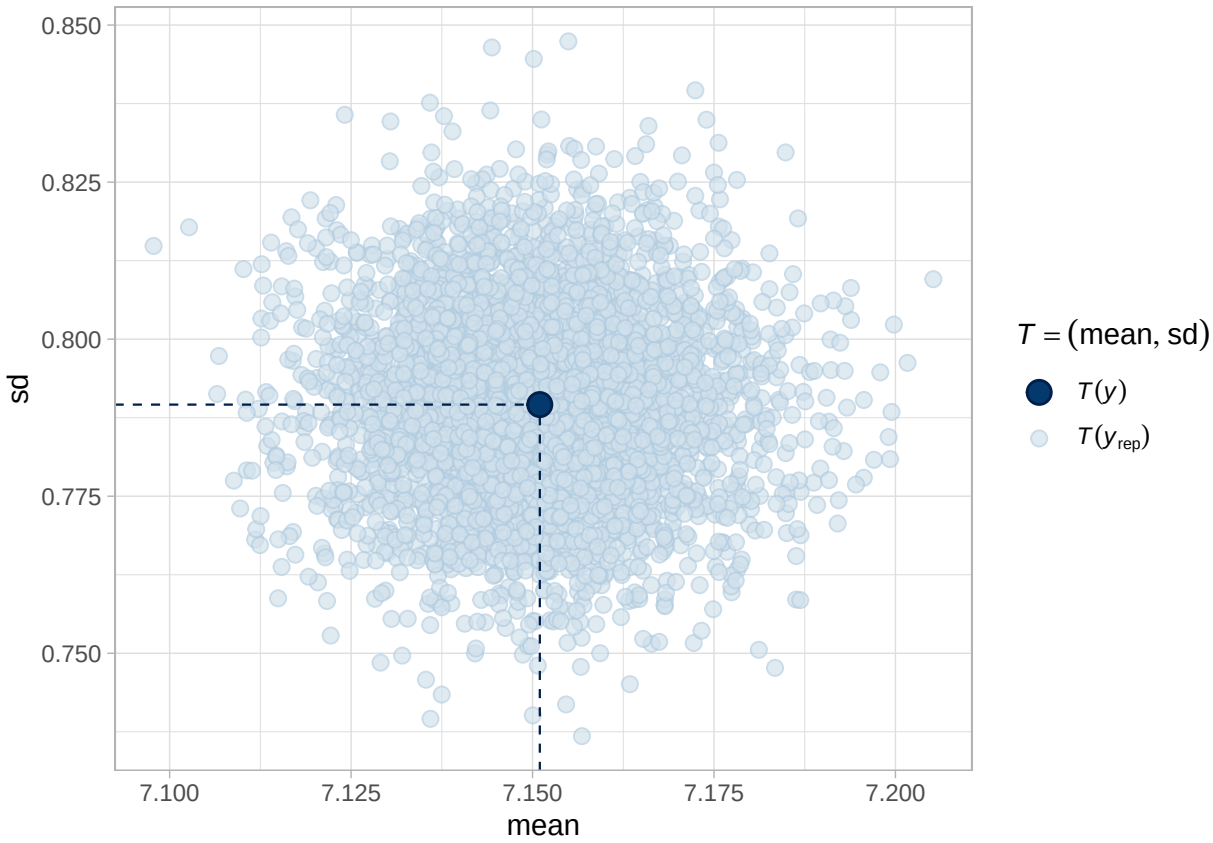

```
pp_check(Person_3_model, type = "stat_2d")
```



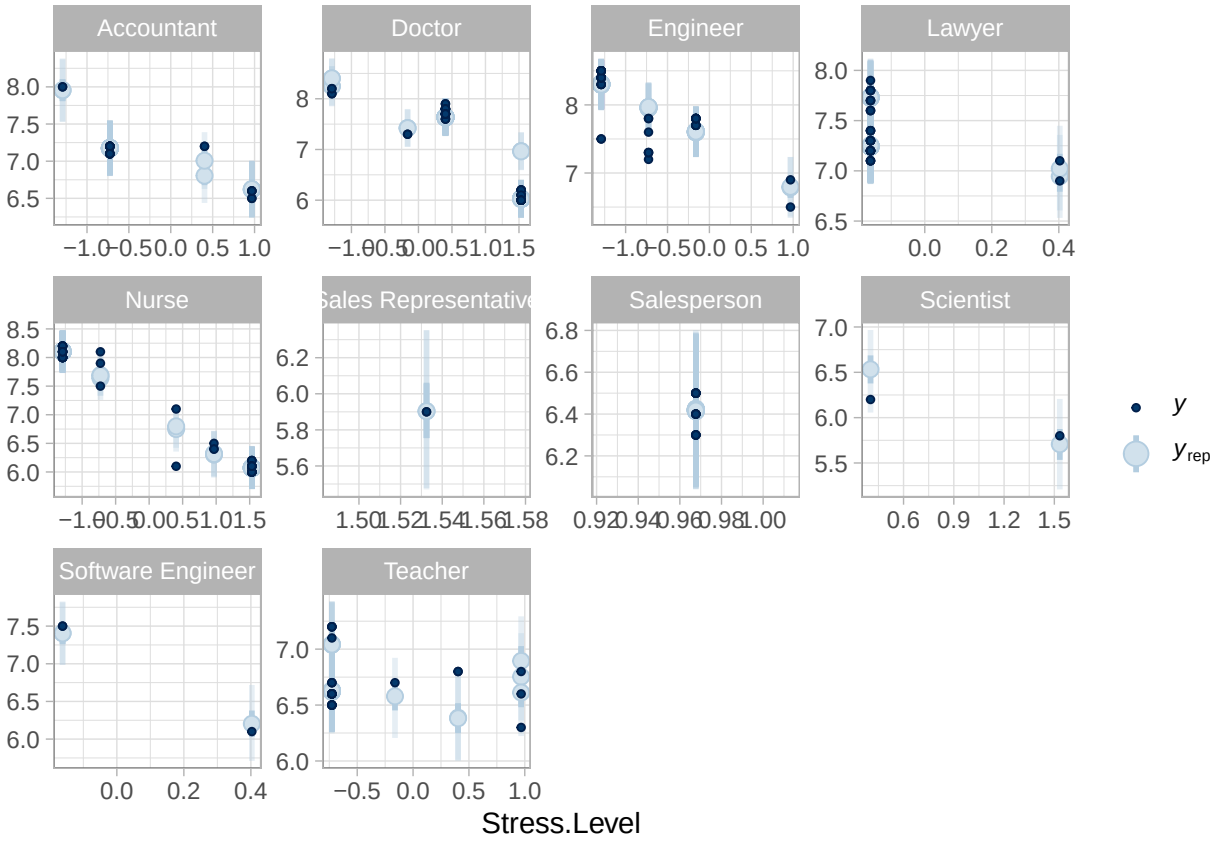

```
# Posterior Predictive Checks for Person_4_model
pp_check(Person_4_model, ndraws = 200)
```



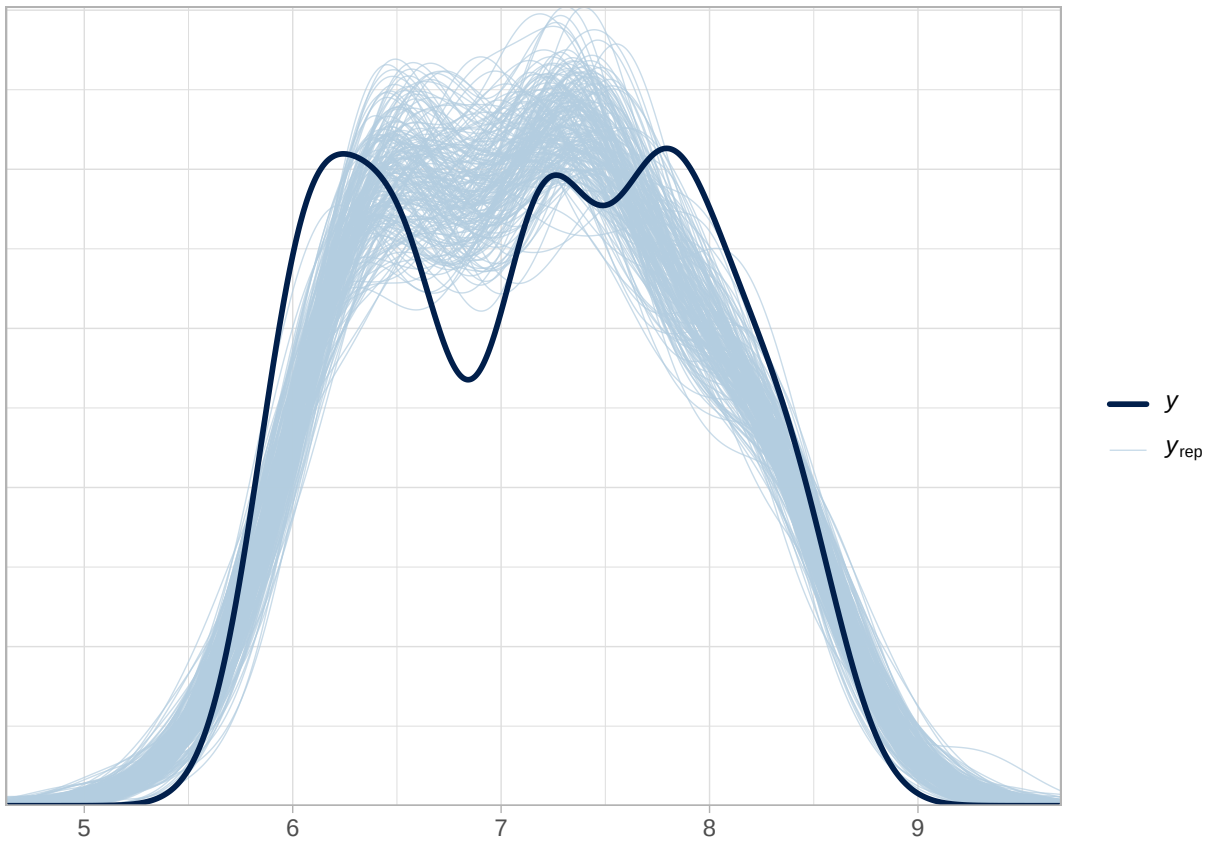

```
pp_check(Person_4_model, type = "stat_2d")
```



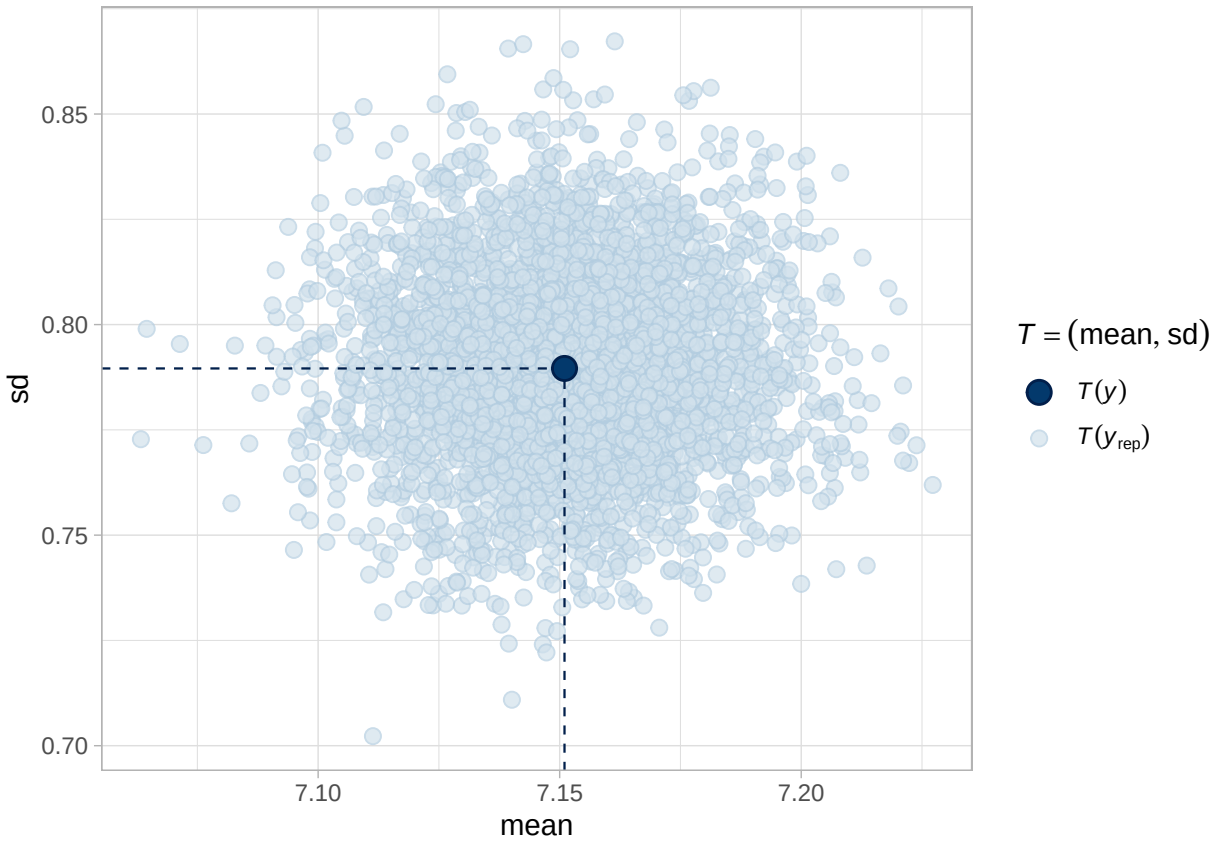

```
pp_check(Person_4_model, type = "intervals_grouped", x = "Stress.Level", group = "Occupation", prob_outer
  xlab("Stress.Level")
```



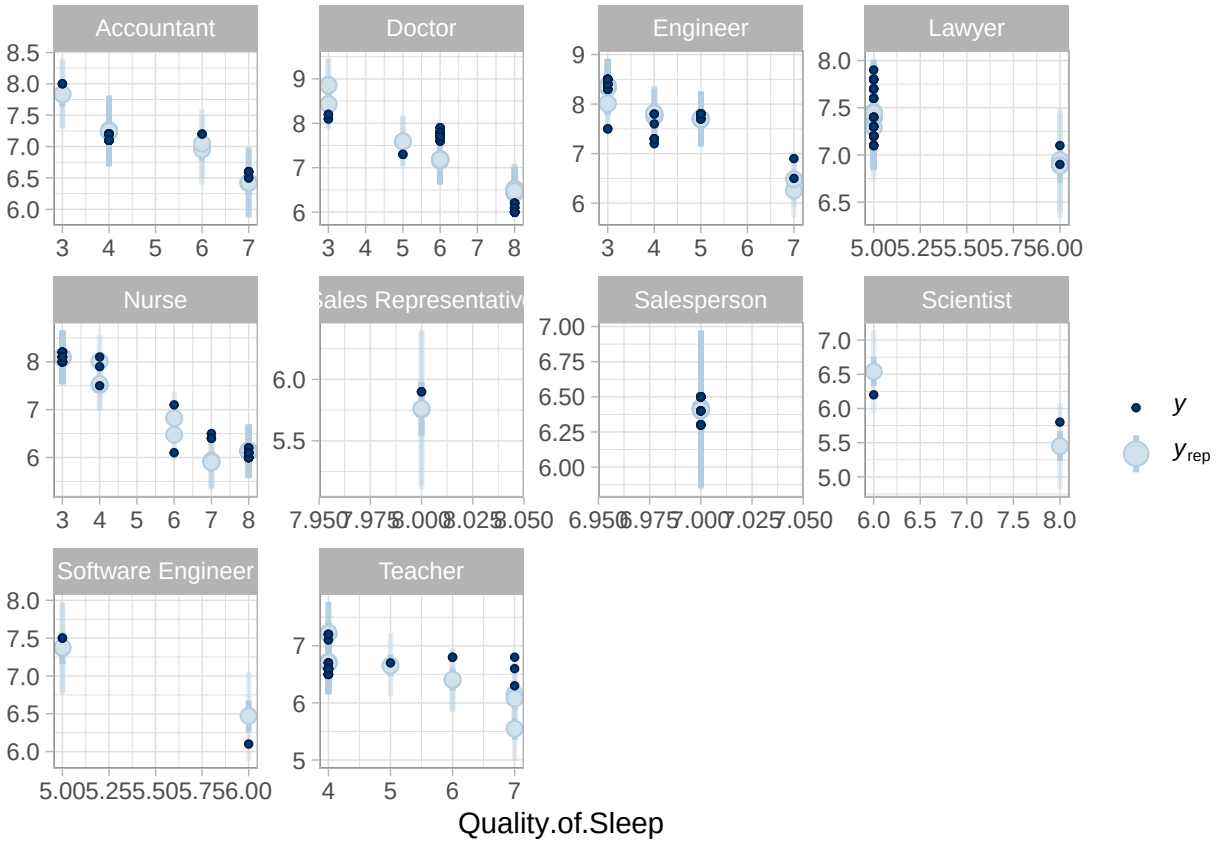

```
# Posterior Predictive Checks for Dominik_Model
pp_check(Dominik_Model, ndraws = 200)
```




```
pp_check(Dominik_Model, type = "stat_2d")
```




```
pp_check(Dominik_Model, type = "intervals_grouped", x = "Stress.Level", group = "Occupation", prob_outer
  xlab("Quality.of.Sleep")
```



Person_1 model:

The posterior predictive checks indicate that the model's fit is generally good. In the curve plot, the overall shape of the predicted values (ypred) aligns well with the observed values (y). The scatter plot further confirms this, as the observed mean and standard deviation are centered within the cloud of posterior predictive points. Although the ypred points in the interval plot do not fully overlap with the observed data, they remain within the 95% credible interval, suggesting a reasonable fit.

Person_2 model:

The posterior predictive checks suggest that Person_2's model fits the observed data quite well. There are no major discrepancies between the predicted and observed values, indicating that the model is a reasonable fit for the data.

Person_3 model:

Person_3's model shows an imperfect fit, but there are no signs of systematic bias. The posterior distribution, while slightly heavier-tailed and less peaked than the observed data, follows the overall shape of the data. The grouped intervals plot suggests no consistent errors. Although the scatter plot of mean and standard deviation shows that the observed data falls within the posterior distribution, it is not centered. This may be due to the relatively weak priors used in the model.

Person_4 model:

Compared to Dominik's model, Person_4's model demonstrates a better fit. The observed data closely overlaps with the 200 replicated lines in the posterior predictive checks. The grouped intervals plot shows that across all stress levels, the observed values are mostly surrounded by the 95% interval, indicating that the model effectively captures the relationships between the chosen features and the target variable.

Dominik_Model:

The posterior predictive checks for Dominik's model reveal that the observed data does not perfectly follow a normal distribution, as evidenced by two noticeable valleys. However, the replicated datasets show similar minor valleys and ultimately approximate a normal distribution. Although the fit is not perfect, adding more predictors, as done in Person_2's model, could potentially improve the model. The 2D summary statistic plot shows that the observed data is well-encapsulated within the dense cloud of replicated data, suggesting that the model captures the overall structure of the data fairly well.

5. Model Comparison (1.5pt)

a. Use k-fold cross-validation to compare the models.

```
set.seed(123)
k <- loo::kfold_split_random(K = 5, N = nrow(sleep_train))
```

```
kf_Person_1_model <- kfold(Person_1_model, folds = k, chains = 1, save_fits = TRUE)
kf_Person_1_model
```

```
Based on 5-fold cross-validation.

           Estimate   SE
elpd_kfold     10.9 21.8
p_kfold        24.2  4.5
kfoldic       -21.8 43.6
```

```
kf_Person_2_model <- kfold(Person_2_model, folds = k, chains = 1, save_fits = TRUE)
kf_Person_2_model
```

```
Based on 5-fold cross-validation.

           Estimate   SE
elpd_kfold    -11.0 20.2
p_kfold        41.2  6.9
kfoldic        22.0 40.3
```

```
kf_Person_3_model <- kfold(Person_3_model, folds = k, chains = 1, save_fits = TRUE)
kf_Person_3_model
```

```
Based on 5-fold cross-validation.

           Estimate   SE
elpd_kfold    -54.8 17.6
p_kfold        31.2  7.4
kfoldic       109.6 35.2
```

```
kf_Person_4_model <- kfold(Person_4_model, folds = k, chains = 1, save_fits = TRUE)
kf_Person_4_model
```

```
Based on 5-fold cross-validation.

           Estimate   SE
elpd_kfold     49.5 33.8
p_kfold        49.8 20.6
kfoldic       -99.0 67.6
```

```
kf_dominik <- kfold(Dominik_Model, folds = k, chains = 1, save_fits = TRUE)
kf_dominik
```

```
Based on 5-fold cross-validation.

           Estimate   SE
elpd_kfold    -57.3 17.9
p_kfold        26.4  5.4
kfoldic       114.5 35.7
```

```
loo_compare(kf_Person_3_model, kf_Person_4_model, kf_Person_1_model, kf_Person_2_model,kf_dominik)
```

```
               elpd_diff se_diff
Person_4_model    0.0       0.0
Person_1_model  -38.6      28.8
Person_2_model  -60.5      35.2
Person_3_model -104.3      35.5
Dominik_Model  -106.8      35.4
```

b. Determine the best model based on predictive accuracy and justify your decision.

We conducted a 5-fold cross-validation, dividing the data into folds of approximately 75 observations each. According to the results from the loo_compare function, kf_Person_4_model emerged as the best-performing model. Although the standard errors might seem large, they do not exceed the expected log pointwise predictive density (ELPD) differences. With ELPD differences greater than 4 and a sample size of 374, these results suggest that kf_Person_4_model provides noticeably better performance compared to the other models.

6. Interpretation of Important Parameters (1.5pt)

Choose one of the best models and interpret its most important parameters.

```
kf_Person_4_model
```

```
Based on 5-fold cross-validation.

           Estimate   SE
elpd_kfold     49.5 33.8
p_kfold        49.8 20.6
kfoldic       -99.0 67.6
```

We selected Occupation as a multilevel variable, with a standard deviation of 0.26 hours, indicating that sleep duration varies by approximately 0.26 hours across different occupations, which can be considered meaningful. The standard deviations for Stress.Level and Physical.Activity.Level as random effects are 0.39 and 0.42, respectively, suggesting that sleep duration is impacted differently by these factors across occupations.

When individuals have average Stress.Level and Physical.Activity.Level, the intercept suggests a mean sleep duration of 7.1 hours, with a 95% confidence interval ranging from 6.89 to 7.28 hours. However, the direction and strength of the relationships between Physical.Activity.Level and Stress.Level with sleep duration provide more valuable insights than the point estimates, especially given the standardization of the independent variables.

Contributions of each member

-**Person_1:** Data Processing, Data Splitting, Person_1's model and analysis

-**Person_2:** Answered question 4 , Person_2's model and analysis

-**Person_3:** found the data set on Kaggle, Person_3's model and analysis, Answered question 6

-**Person_4:** Combining all data models, batch sensitivity check and pp_check. Answered question 6 and 7

-**Dominik:** Answered question 1, Dominik's model and analysis

Statement of technology

The citation generator at Scribbr.com was used in creation of in text citations and the reference list (Citation Generator, 2024).

References

Citation generator. (2024, January 10). Scribbr. Retrieved May 31, 2024, from https://www.scribbr.com/citation/generator/

Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and other stories. https://doi.org/10.1017/9781139161879

Hublin, C., Haasio, L., & Kaprio, J. (2020). Changes in self-reported sleep duration with age - a 36-year longitudinal study of Finnish adults. BMC Public Health, 20(1). https://doi.org/10.1186/s12889-020-09376-z

Liu, Y., Wheaton, A. G., Chapman, D. P., Cunningham, T. J., Lu, H., & Croft, J. B. (2016). Prevalence of Healthy Sleep Duration among Adults — United States, 2014. Morbidity and Mortality Weekly Report, 65(6), 137–141. https://doi.org/10.15585/mmwr.mm6506a1

R: Effective Sample Size (ESS). (n.d.). https://search.r-project.org/CRAN/refmans/bayestestR/html/effective_sample.html

Stan Development Team. (2022, March 10). Runtime warnings and convergence problems. https://mc-stan.org/misc/warnings.html#maximum-treedepth-exceeded

Tharmalingam, L. (2023, September 18). Sleep Health and Lifestyle Dataset. Kaggle. https://www.kaggle.com/datasets/uom190346a/sleep-health-and-lifestyle-dataset

World Health Organization: WHO. (2023, March 16). Hypertension. https://www.who.int/news-room/fact-sheets/detail/hypertension#:~:text=Overview,get%20your%20blood%20pressure%20checked