# 2292HDSBS

2024-04-17

```
# Load dataset.
sgaj <- read.csv("sgaj.csv")
head(sgaj, 7)
```

```
##   patient female age time measured
## 1       1      1  19    0    199.5
## 2       1      1  19    6    239.4
## 3       1      1  19   12    163.5
## 4       1      1  19   18    268.1
## 5       1      1  19   24    228.9
## 6       2      0  19    0    172.4
## 7       2      0  19    6    159.3
```

## Part 1

```
# Get summary of the age variable.
summary(sgaj$age)
```

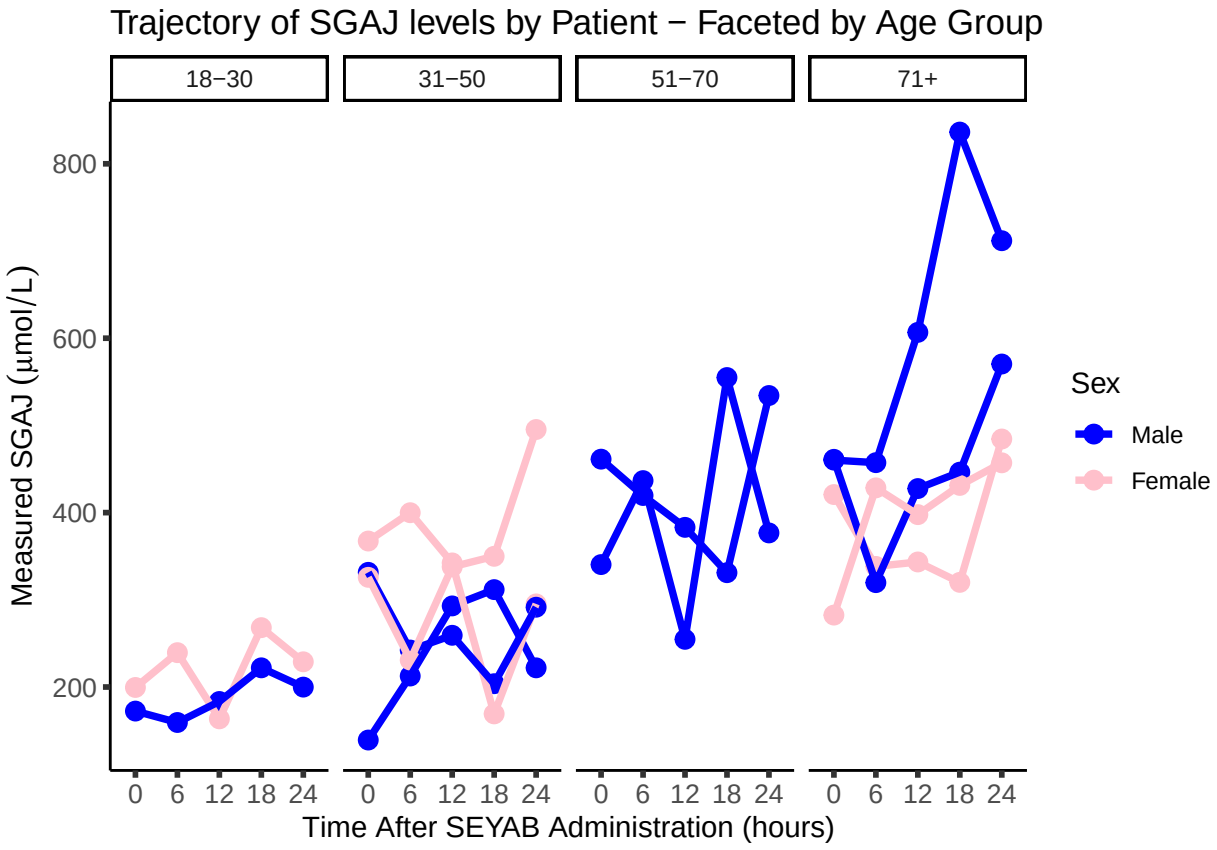
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   19.00   40.75   56.00   55.67   77.25   88.00
```

After reviewing the age distribution in the dataset, I decided to categorise the age variable into four groups: 18-30 (young adults), 31-50 (middle-aged adults), 51-70 (older adults), and 71+ (seniors). This is done solely for visualisation purposes (not used in the model) as it allows for the ordering (and separation in facets) of patients based on their age, making the exploration of how age and gender affect SGAJ easier to interpret.

```
# Create age groups
sgaj_with_age_groups <- sgaj %>%
  mutate(age_group = case_when(
    age >= 18 & age <= 30 ~ "18-30",
    age >= 31 & age <= 50 ~ "31-50",
    age >= 51 & age <= 70 ~ "51-70",
    age > 70 ~ "71+"
  ))
```

```
ggplot(
  sgaj_with_age_groups,
  aes(
    x = as.factor(time),
    y = measured,
```

```
    group = patient,
    color = as.factor(female)
)) +
geom_point(size = 3) +
geom_line(linewidth = 1.3) +
facet_wrap(~age_group, ncol = 4) +
theme_classic() +
theme(
  axis.text.x = element_text(size = 10),
  axis.text.y = element_text(size = 10),
  axis.ticks = element_line(size = 1)
) +
labs(
  title = "Trajectory of SGAJ levels by Patient - Faceted by Age Group",
  x = "Time After SEYAB Administration (hours)",
  y = expression(Measured~SGAJ~(mu*mol/L)),
  color = "Sex"
) +
scale_color_manual(
  values = c("blue", "pink"),
  labels = c("Male", "Female")  # Set labels for Male and Female
)
```




The plot shows the trajectory of SGAJ level (across the timepoints) for each patient, facetted by the different age groups (to aid visualisation). The following insights can be drawn from the plot:

- There appears to be a positive association between age and SGAJ levels, suggesting that older patients tend to have higher biomarker concentrations.
- In younger age groups, female patients display higher SGAJ levels than male patients (in most time points). In contrast, in the senior age group (71+ years), this pattern is reversed, with male patients presenting higher SGAJ levels.
- Each patient has a different starting level of SGAJ (time=0) which suggests the need for a random intercept. Furthermore, the rate at which SGAJ levels change over time is not constant across patients; In general, older patients have a faster rate SGAJ growth than younger patients but there is still a lot of variability between patients, suggesting the need for a random slope on time.
- I acknowledge that unmeasured factors (such as medication, lifestyle) might influence SGAJ levels and could confound the observed associations with age and sex.

# Part 2

## 2.1 Assumptions

- **Normality:** The SGAJ level for patient i at time t is assumed to follow a Normal distribution.
- **Random Intercept and Slope ($\alpha_i$ and $\gamma_i$):** Each patient has an inherent baseline SGAJ level and a specific rate of change in SGAJ levels over time, represented by $\alpha_i$ (random intercept) and $\gamma_i$ (random slope). These assumptions allow the model to capture individual differences in initial SGAJ levels and in the dynamics of response to the drug over the observation period.
- **Linearity:** The model assumes a linear relationship between the dependent variable (SGAJ) and each of the covariates, including time. The effect of time is modeled as a linear rate of change ($\gamma_i$), which is patient-specific but constant between time intervals for each patient.
- **Additive:** Simplifying assumption that there is no interaction between the covariates (model is additive).
- **Exchangeability:** Before considering the data, we treat all patients as essentially the same. The prior distributions for the group-specific parameters $\alpha_i$ and $\gamma_i$ reflect this assumption. After seeing the data, the model allows these parameters to differ, but the starting point is a belief in their similarity.

## 2.2 Model formulation

The SGAJ level for each patient $i$ at time $t$ is modeled as a normal distribution with mean $\mu_{it}$, the expected measurement at reference values, and a fixed standard deviation of 75, as per the brief's suggestion. This formulation utilizes a re-parameterization (reference conditions - age=50, female=1, time=12 hours) that aligns with the prior study's findings as detailed in the brief, ensuring the model leverages existing knowledge about SGAJ levels.

$$y_{it} \sim \mathcal{N}(\mu_{it}, 75^2)$$

$$\mu_{it} = \alpha_i + \beta_{\text{age}} \left(\text{age}_i - 50\right)/10 + \beta_{\text{female}}(1 - \text{female}_i) + \gamma_i \left(\text{time}_t - 12\right)/6$$

### 2.2.1 Priors on alpha

The variability between patients is represented by a random effect $\alpha_i$ (patient-specific random intercept), which is assumed to be drawn from a Normal distribution with mean $\mu_a$ and variance $\sigma_a^2$.

$$\alpha_i \sim \mathcal{N}(\mu_a, \sigma_a^2)$$

$$\mu_a \sim \mathcal{N}(225, 20.41^2)$$

$$\sigma_a \sim \text{Uniform}(80, 100)$$

- I am assuming that the random intercept $a_i$ representing the deviation in baseline SGAJ level for each patient from the population average under reference conditions (age=50, female=1, time=12 hours), follows a Normal distribution.
- $\mu_a$ represents the average value (intercept) across the population for the reference group when other predictors (e.g., age, female, time) are held at their reference levels. Being influenced by the information from the previous study, I am assuming that $\mu_a$ has an average value of 225. The standard deviation of $\mu_a$ is calculated as (265-185)/(1.96*2)=20.41, with a 95% upper and lower credible limit of 265 and 185 respectively.
- For the random effects standard deviation $\sigma_a$, I am setting it as a uniform distribution between 80 and 100 to account for some uncertainty around the value 90 suggested from the previous study.

#### 2.2.2 Priors on beta_age and beta_female

The terms involving $\beta_{age}$, and $\beta_{female}$ correspond to the fixed effects of age (centered around the age of 50 and scaled such that a unit increase is in 10 years), and female (has the value 1 if female and zero otherwise), respectively.

$$\beta_{\text{age}} \sim \mathcal{N}(0, 10^2)$$
$$\beta_{\text{female}} \sim \mathcal{N}(0.1, 10^2)$$

- As stated in the brief, the effect of $\beta_{age}$ and $\beta_{female}$ is not well understood, therefore I am assuming vague priors and relying on data to inform these relationships. For the prior of $\beta_{female}$, I am subtly incorporating the expectation that males may have slightly higher SGAJ levels than females (as explained by the principal investigator).

#### 2.2.3 Priors on gamma

$\gamma_i$ is modelled as a random effect and represents the patient-specific rate of change of SGAJ levels - capturing individual variability in response to the drug over time. The term $\gamma_i(time_t - 12)/6$ normalizes this rate of change to a 6-hour period; without this random slope, we would observe only different initial SGAJ levels but uniform rates of change across individuals. I am assuming that $\gamma_i$ follows a normal distribution with mean $\mu_\gamma$ and variance $\sigma^2_\gamma$.

$$\gamma_i \sim \mathcal{N}(\mu_\gamma, \sigma^2_\gamma)$$
$$\mu_\gamma \sim \mathcal{N}(0, 10^2)$$
$$\sigma_\gamma \sim \text{Uniform}(0, 100)$$

- For $mu_{gamma}$ and $sigma_{gamma}$ I chose vague priors due to the lack of specific prior information, ensuring that the inferences are driven predominantly by the observed data rather than by subjective assumptions. I chose wide uniform and large variance normal distributions for these priors to offer the flexibility needed for the model to adapt and learn from the data.

### 2.3 RJAGS

```
# Initialisation of data
data <- list(n_patients = length(unique(sgaj$patient)),
             patient = sgaj$patient,
             age = sgaj$age,
             time = sgaj$time,
```

```
            female = sgaj$female,
            n = nrow(sgaj),
            y = sgaj$measured
)
```

Constructing a hierarchical linear regression model based on the assumptions and model formulation described above. Checking convergence of the model from two different chains - starting from slightly different initial values.

```
# Model definition
model <- "model{
  # Linear regression model
  for (i in 1:n) {
    mu[i] <- alpha[patient[i]] + beta_age*(age[i] - 50)/10 + beta_female*(1-female[i]) + gamma[patient[
    y[i] ~ dnorm(mu[i], prec_psi)
    yrep[i] ~ dnorm(mu[i], prec_psi) # posterior-predictive distribution
  }

  # Random effects
  for (p in 1:n_patients) {
    alpha[p] ~ dnorm(mu_alpha, prec_alpha)
    gamma[p] ~ dnorm(mu_gamma, prec_gamma)
    rate_of_change[p] <- gamma[p] * 4
  }

  # Priors for fixed effects
  beta_age ~ dnorm(0, 1/(10^2))
  beta_female ~ dnorm(0.1, 1/(10^2))

  # Prior for the population average of the random intercepts
  mu_alpha ~ dnorm(225, 1/(20.41^2))

  # Prior for the standard deviation of the random intercepts
  sd_alpha ~  dunif(80,100)
  prec_alpha <- 1/(sd_alpha^2)

  # Prior for the population average of the random slope
  mu_gamma ~ dnorm(0, 1/(10^2))

  # Prior for the standard deviation of the random slope
  sd_gamma ~ dunif(0, 100)
  prec_gamma <- 1/(sd_gamma^2)

  prec_psi <- 1/(75^2)
}"


# Two different initial values - to check that the model is converging to the same conclusion
initial_values <-
  list(
    list(
      alpha = rep(200, data$n_patients),
      sd_alpha = 80,
```

```
      mu_gamma = 0,
      sd_gamma = 20,
      beta_age = 0,
      beta_female = 0,
      .RNG.name = c("base::Mersenne-Twister"),
      .RNG.seed = 20
    ),
    list(
      alpha = rep(100, data$n_patients),
      sd_alpha = 95,
      mu_gamma = 1,
      sd_gamma = 50,
      beta_age = 1,
      beta_female = 1,
      .RNG.name = c("base::Mersenne-Twister"),
      .RNG.seed = 42
    )
  )

# Model setup
jags_model <- jags.model(textConnection(model),
                         data = data,
                         inits = initial_values,
                         n.chains = 2)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 60
##    Unobserved stochastic nodes: 90
##    Total graph size: 650
##
## Initializing model
```

```
# Burn-in
update(jags_model, 1000)

pars <- c(
    "alpha",
    "mu_alpha",
    "sd_alpha",
    "beta_age",
    "beta_female",
    "gamma",
    "mu_gamma",
    "sd_gamma",
    "rate_of_change",
    "yrep"
  )

# Sampling - 200,000 iterations, keeping every 10th sample.
set.seed(2292)
```

```
samples <- coda.samples(jags_model,
                        variable.names = pars,
                        n.iter = 200000,
                        thin=10)


# Get draws
sgaj_draws <- as_draws(samples)
```

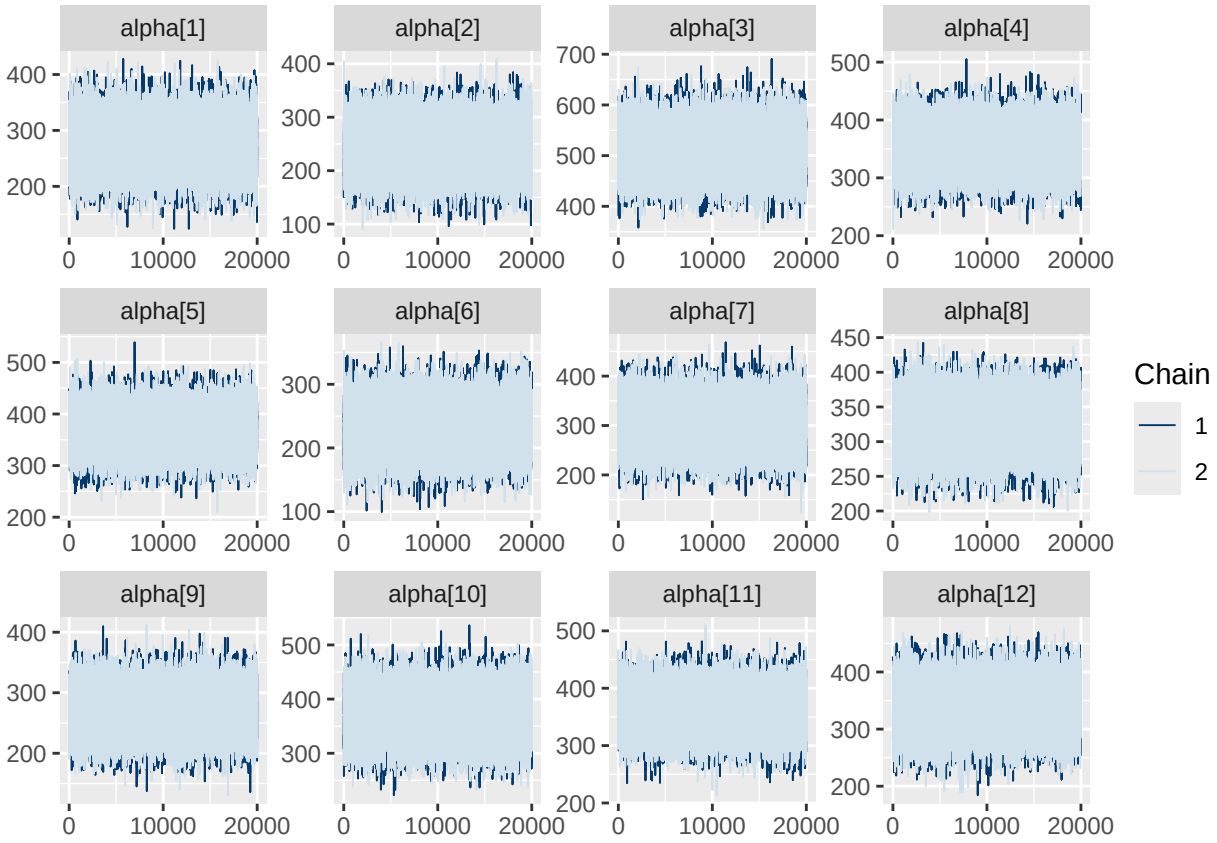
## 2.4 Summarising posterior distributions of key quantities of interest

```
# Function that returns posterior summary for only specified parameters.
# The summmary includes the 2.5%, 50%, 97.5% quantiles,
# convergence measures (rhat, ESS), and monte carlo standard error
get_posterior_summary <- function(draws, pars) {
  subsetted_draws <- subset_draws(draws, variable = pars)
  summary_subsetted_draws <-
    summary(
      subsetted_draws,
      ~ quantile(.x, probs = c(0.025, 0.5, 0.975)),
      default_convergence_measures(),
      c("mcse_mean", "mcse_median")
    )

  return(summary_subsetted_draws)
}
```

### 2.4.1 Random intercept

```
mcmc_trace(samples, pars=c(paste0("alpha[",1:12,"]"))) +
  scale_x_continuous(breaks = function(x) c(0, 10000, 20000))
```

```
## Scale for x is already present.
## Adding another scale for x, which will replace the existing scale.
```




```
print(get_posterior_summary(sgaj_draws, c(paste0("alpha[",1:12,"]"))))
```

```
## # A tibble: 12 x 9
##    variable  `2.5%` `50%` `97.5%`  rhat ess_bulk ess_tail mcse_mean mcse_median
##    <chr>      <dbl> <dbl>   <dbl> <dbl>    <dbl>    <dbl>      <dbl>       <dbl>
##  1 alpha[1]    204.  278.    352.  1.00   38116.   38740.     0.194       0.225
##  2 alpha[2]    167.  244.    320.  1.00   37900.   38360.     0.201       0.224
##  3 alpha[3]    433.  511.    591.  1.00   35746.   38252.     0.212       0.247
##  4 alpha[4]    286.  351.    418.  1.00   38928.   37033.     0.171       0.233
##  5 alpha[5]    301.  369.    438.  1.00   39216.   38603.     0.177       0.222
##  6 alpha[6]    169.  234.    298.  1.00   39627.   39686.     0.164       0.184
##  7 alpha[7]    223.  303.    383.  1.00   36839.   39089.     0.213       0.247
##  8 alpha[8]    260.  322.    385.  1.00   39046.   38667.     0.161       0.199
##  9 alpha[9]    205.  271.    335.  1.00   39411.   39466.     0.165       0.247
## 10 alpha[10]   296.  370.    445.  1.00   38252.   39222.     0.194       0.261
## 11 alpha[11]   295.  360.    424.  1.00   40269.   39863.     0.163       0.218
## 12 alpha[12]   265.  337.    409.  1.00   38511.   38781.     0.187       0.251
```

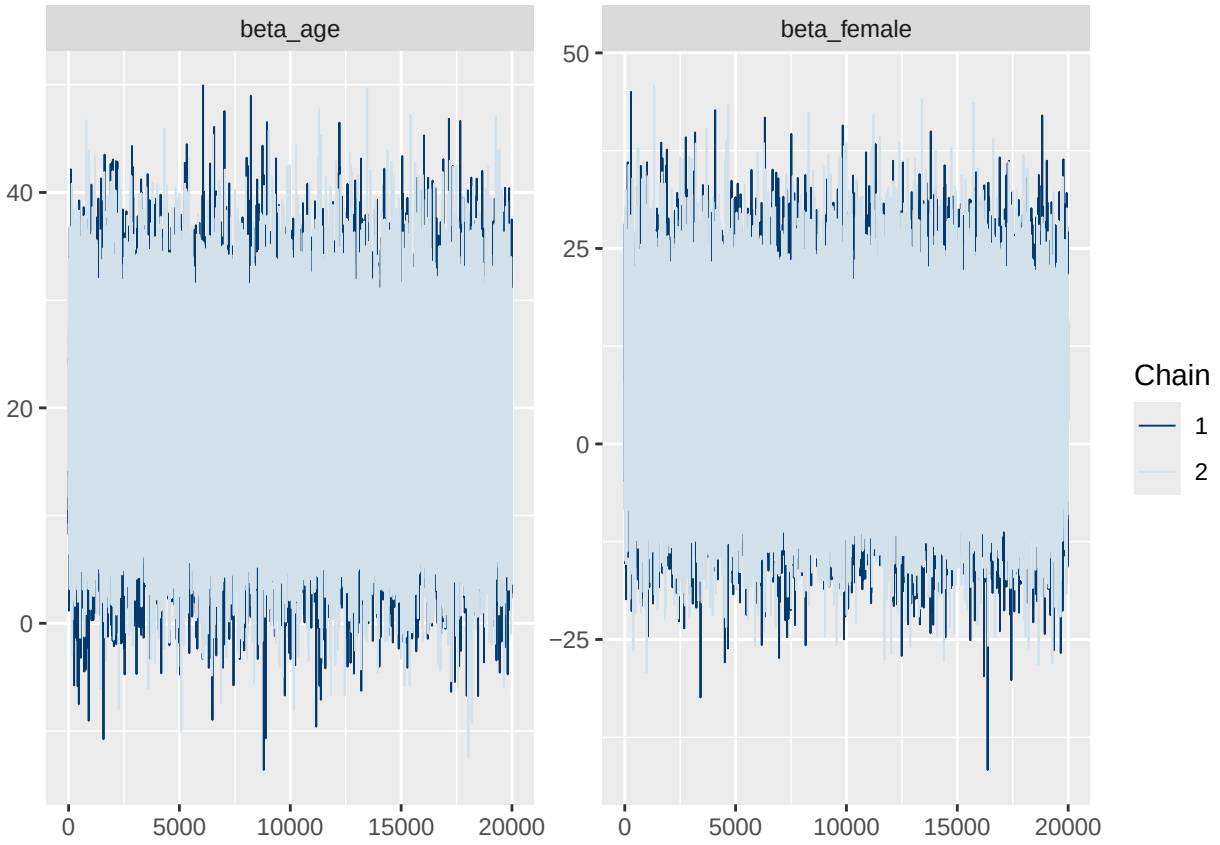
**Convergence:** The traceplots suggest a well-mixed distribution ("fat hairy caterpillar" appearance), indicating stable estimates with random dispersion around the mean. The overlap of both chains and an $\hat{R}$ value of 1.00 (3 s.f.) for all parameters show no signs of convergence issues. This implies that within-chain and between-chain variances are comparable. The ESS values range from approximately 36,000 to 40,000 for both bulk and tail. High ESS values suggest that the sample provides a reliable approximation of the posterior distribution, reducing the error in our estimates and increasing the confidence in statistical inferences. The monte carlo standard error (MCSE) values for the mean range from 0.16 to 0.21 and for the median from

0.18 to 0.26. These values indicate the variability of the estimate if the entire Monte Carlo simulation were repeated under the same conditions. Given the magnitude of the estimated parameters, the observed MCSE values are low, which substantiates the precision and reliability of our estimates.

**Interpretation:** The median values (50% quantile), representing the central tendency of the random intercept estimates, demonstrate considerable variation across patients. Notably, patient 6 exhibits both the lowest median at 234 and the narrowest 2.5% to 97.5% quantile interval, ranging from 169 to 298. Conversely, patient 3 shows the highest median at 510, accompanied by the broadest interval from 433 to 591.

### 2.4.2 Fixed effects

```
mcmc_trace(samples, pars=c("beta_age", "beta_female"))
```




```
print(get_posterior_summary(sgaj_draws, c("beta_age", "beta_female")))
```

```
## # A tibble: 2 x 9
##   variable    `2.5%` `50%` `97.5%`  rhat ess_bulk ess_tail mcse_mean mcse_median
##   <chr>        <dbl> <dbl>   <dbl> <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
## 1 beta_age      3.99 19.3     34.3  1.00   35109.   38743.    0.0414      0.0476
## 2 beta_female -13.0   6.14    25.3  1.00   40251.   39476.    0.0486      0.0667
```

**Convergence:** The traceplots show stable mean with random dispersion around the mean and both chains overlap. The $\hat{R}$ statistic is estimated to be 1.00, giving no suggestion of any lack of convergence. The mean MCSE values for both are smaller than 0.05.

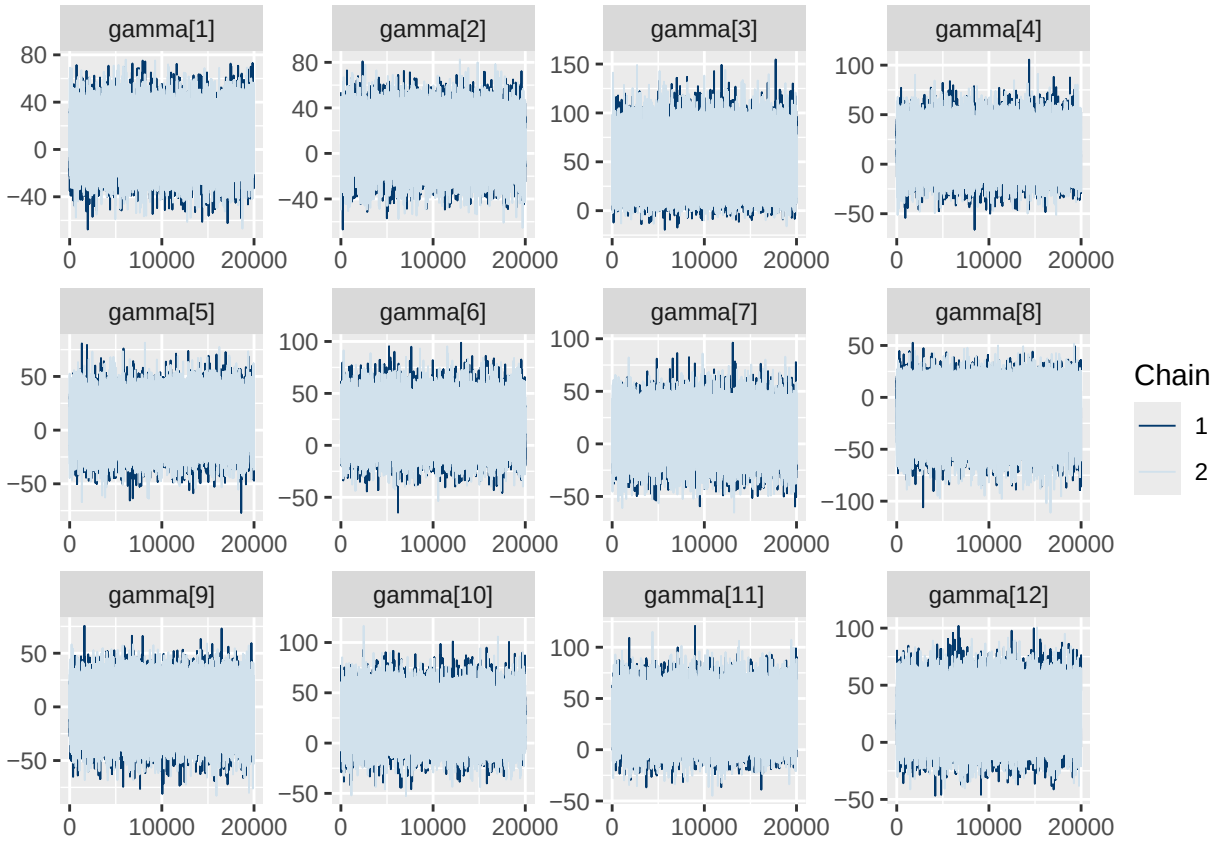
**Interpretation:** - The median value of the $\beta_{age}$ coefficient at 19.3 indicates that SGAJ levels typically rise by about 19.3 (3.99-34.3) units for each decade past 50 years, and decrease by the same amount per decade before 50 (all other values held constant at their reference levels).

- The $\beta_{female}$ coefficient's median at 6.14 suggests males have SGAJ levels that are typically 6.14 units higher than females (all other things constant), but with a broad 95% credible interval from -13.0 to 25.3, there is considerable uncertainty around this effect.

### 2.4.3 Random slope

```
mcmc_trace(samples, pars=c(paste0("gamma[",1:12,"]"))) +
  scale_x_continuous(breaks = function(x) c(0, 10000, 20000))
```

```
## Scale for x is already present.
## Adding another scale for x, which will replace the existing scale.
```




```
print(get_posterior_summary(sgaj_draws, c(paste0("gamma[",1:12,"]"))))
```

```
## # A tibble: 12 x 9
##    variable  `2.5%` `50%` `97.5%`  rhat ess_bulk ess_tail mcse_mean mcse_median
##    <chr>      <dbl> <dbl>   <dbl> <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
##  1 gamma[1]  -24.0  10.6     42.2  1.00   38234.   38389.    0.0827      0.0939
```

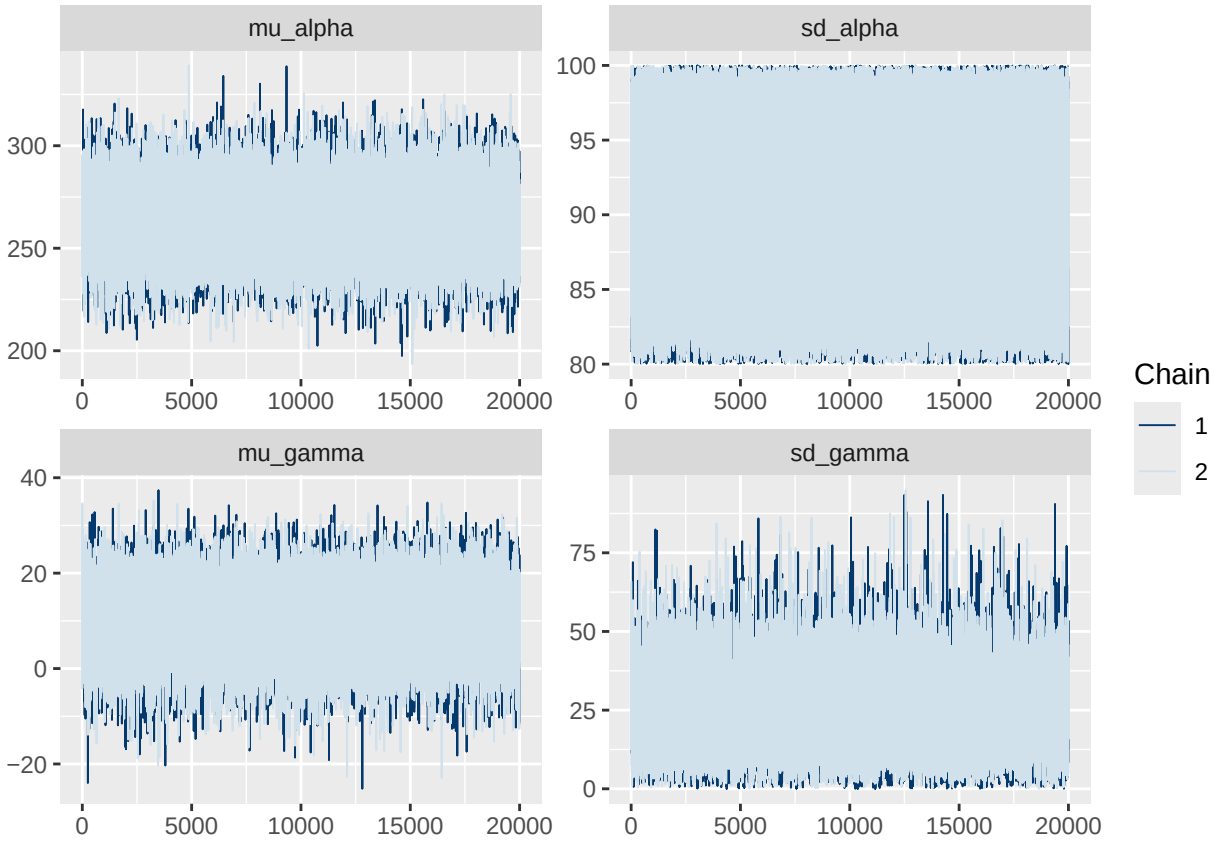
```
##  2 gamma[2]  -22.1  11.8     43.3  1.00   37968.   37280.    0.0828      0.112
##  3 gamma[3]    7.54 43.5     93.7  1.00   15788.   15971.    0.176       0.205
##  4 gamma[4]  -17.6  14.3     48.3  1.00   36857.   37262.    0.0847      0.0954
##  5 gamma[5]  -25.8   9.35    40.7  1.00   39321.   38165.    0.0826      0.0989
##  6 gamma[6]  -13.5  17.2     52.7  1.00   36764.   37723.    0.0854      0.109
##  7 gamma[7]  -22.3  11.5     43.4  1.00   37270.   37807.    0.0839      0.0998
##  8 gamma[8]  -53.7  -8.40    22.4  1.00   21141.   23795.    0.135       0.185
##  9 gamma[9]  -36.4   2.56    31.8  1.00   30418.   36593.    0.0998      0.129
## 10 gamma[10]  -9.50 20.1     58.1  1.00   33573.   38183.    0.0923      0.107
## 11 gamma[11]  -4.24 24.6     64.9  1.00   28559.   34198.    0.104       0.149
## 12 gamma[12]  -9.09 20.4     58.3  1.00   32642.   38553.    0.0933      0.113
```

*Convergence:* The traceplots show stable mean with random dispersion around the mean and both chains overlap. The $\hat{R}$ statistic is estimated to be 1.00 across all the gammas (no suggestion of any lack of convergence). The mean MCSE values are between 0.08 and 0.18.

*Interpretation:* The summary statistics for the patient-specific rate of change of SGAJ levels (gamma[1]-gamma[12]) across 12 patients reveal considerable diversity in response dynamics, both in magnitude and direction. Median estimates range from a decrease of -8.40 in patient 8 every 6 hours to a substantial increase of 43.5 in patient 3, illustrating the complex and individualized nature of biological responses to treatment.

### 2.4.4 Hyper-parameters

```r
# Check convergence for random intercept (alpha).
mcmc_trace(samples, pars=c("mu_alpha", "sd_alpha", "mu_gamma", "sd_gamma"))
```



```r
get_posterior_summary(sgaj_draws, c("mu_alpha", "sd_alpha", "mu_gamma", "sd_gamma"))
```

```
## # A tibble: 4 x 9
##   variable `2.5%` `50%` `97.5%`  rhat ess_bulk ess_tail mcse_mean mcse_median
##   <chr>     <dbl> <dbl>   <dbl> <dbl>    <dbl>    <dbl>     <dbl>       <dbl>
## 1 mu_alpha 231.   264.    297.   1.00   39818.   39049.    0.0846      0.113
## 2 sd_alpha  80.8   91.8    99.6  1.00   38800.   39690.    0.0286      0.0484
## 3 mu_gamma  -4.51  10.5    23.8  1.00   30845.   30542.    0.0409      0.0501
## 4 sd_gamma   2.52  22.2    50.1  1.00   10551.    6663.    0.105       0.108
```

*Convergence:* The traceplots show stable mean with random dispersion around the mean and both chains overlap. All hyper-parameters indicate convergence with $\hat{R}$ values at 1.00. The MCSE is under 0.1 for all the hyper-parameters.

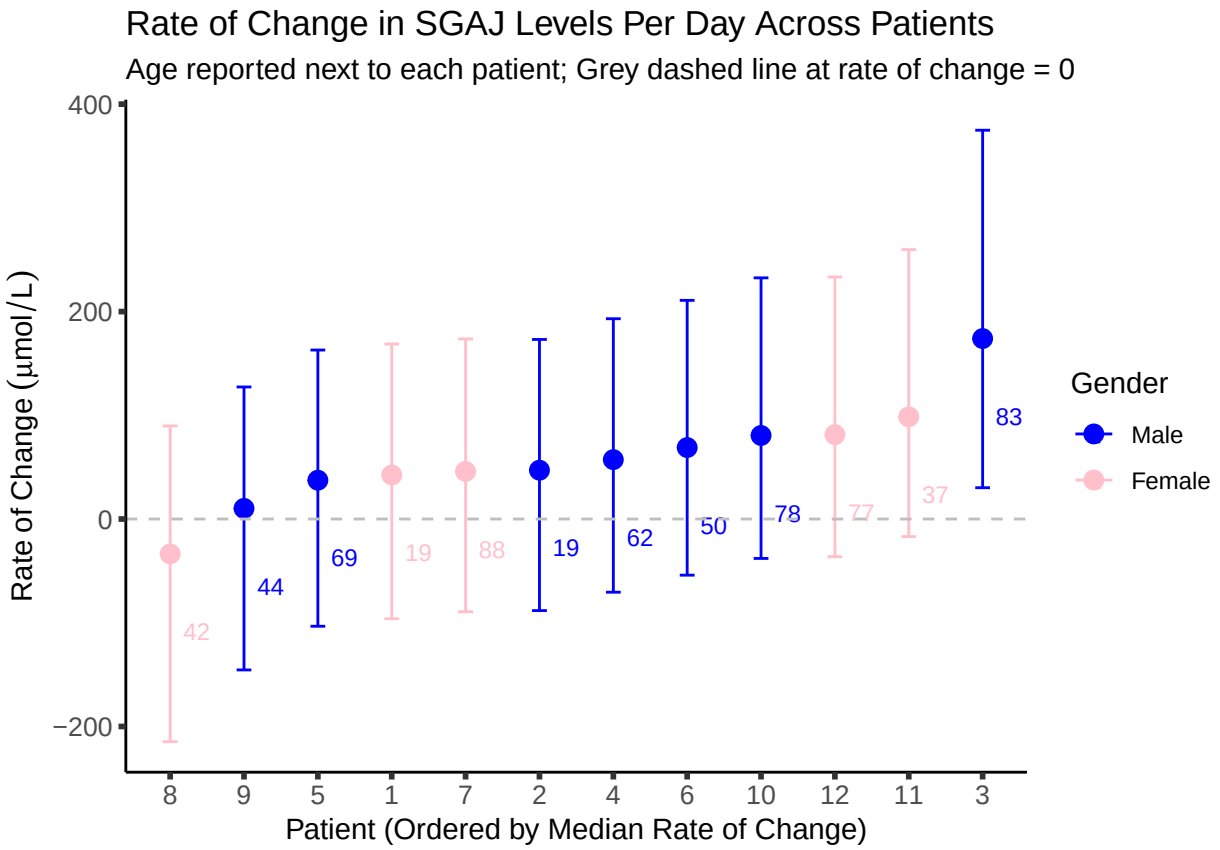
*Interpretation:* - The median posterior estimate for mu_alpha at 264 (231, 297) suggests that the average baseline SGAJ level across the population (reference conditions) is approximately 265 units, a higher value than the suggested average SGAJ (225) from the previous cohort. The posterior median for sigma_alpha (standard deviation across all individuals) at reference conditions is 91.8 (80.8, 99.6) which is also slightly higher than the value suggested from the previous study.

- The median value for mu_gamma at 10.5 implies an average rate of increase across the population (reference conditions) of about 10.5 units in SGAJ levels every 6 hours. However, the broad 95% credible interval from -4.51 to 23.8 indicates substantial uncertainty, suggesting that in some scenarios, SGAJ levels might even decrease. The median of sd_gamma at 22.2, with a range from 2.152 to 50.1, reflects high variability in the rate of change of SGAJ levels among patients. This suggests that individual responses to treatment vary notably.

## 2.5 Visualising rate of change per day per patient

```r
# Get dataframe for rate of change variable
rate_of_change_df <-
  get_posterior_summary(sgaj_draws, c("rate_of_change"))[, c("2.5%", "50%", "97.5%")]
colnames(rate_of_change_df) <- c("lower", "median", "upper")
rate_of_change_df$patient <- as.factor(1:12)

# Combine rate_of_change_df with the original dataset (to get patient details)
unique_sgaj <- sgaj %>%
  select(patient, female, age) %>%
  distinct() %>% mutate(patient = as.factor(patient))

rate_of_change_df <- rate_of_change_df %>%
  left_join(unique_sgaj, by = "patient") %>%
  mutate(patient = reorder(patient, median))

# Plotting the data with facets and ages as labels
ggplot(rate_of_change_df, aes(x = patient, y = median, color = as.factor(female))) +
  geom_point(size = 3) +
  geom_text(aes(label = age, vjust = 5, hjust=-0.5), size = 3) +  # Add age labels
  scale_color_manual(values = c("blue", "pink"), labels = c("Male", "Female")) +
  geom_errorbar(aes(ymin = lower, ymax = upper), width = 0.2) +
  geom_hline(yintercept = 0, linetype = "dashed", color = "grey") +
```

```r
  theme_classic() +
  theme(
    axis.text.x = element_text(size = 10),
    axis.text.y = element_text(size = 10),
    axis.ticks = element_line(size = 1)
  ) +
  labs(
    y = expression(Rate~of~Change~(mu*mol/L)),
    x = "Patient (Ordered by Median Rate of Change)",
    title = "Rate of Change in SGAJ Levels Per Day Across Patients",
    subtitle = "Age reported next to each patient; Grey dashed line at rate of change = 0",
    color = "Gender",
    size = "Age"
  )
```




The plot shows the rate of change of SGAJ levels per day for each patient. The age of each patient is placed close to their corresponding point. The following insights can be drawn:

- All patients except patient 8 show an increase in SGAJ levels per day as indicated by positive median values. However, the wide error bars, which represent the 95% confidence intervals, introduce a notable degree of uncertainty into these estimates. For almost all patients, except number 3, this uncertainty extends into the possibility of a decrease in SGAJ levels, as indicated by confidence intervals that straddle the zero line. Some patients (like patient 3) exhibit larger intervals, denoting less precise estimates, whereas others (e.g., patient 2) show narrower intervals.
- While the plot does not reveal a definitive linear relationship between age and the rate of change in SGAJ levels, there is a discernible cluster of older individuals—specifically patients 10, 12, and 3—who

exhibit a more substantial rate of change. This observation tentatively suggests that age may influence a heightened response within the initial 24-hour period. However, this potential trend is also conflicted by the presence of older patients 7 and 5, aged 88 and 69 respectively, who demonstrate a slower rate of change, indicating that additional factors may influence the rate of change beyond age alone.

- Differentiating patients by gender shows that both males (blue) and females (pink) are present across the spectrum of the rate of change. There isn't a clear gender-based trend;
- The inter-patient variability underscores the personalized nature of the biological response to treatment, suggesting that factors such as age, gender, and possibly others that were not measured (such as underlying health conditions or genetic factors) play a role in SGAJ level changes. Such variability highlights the potential need for personalized approaches to treatment dosing and monitoring to account for these differences.

### 2.6 Model limitations

- The model's assumption of linearity between covariates (age, gender, and time) and SGAJ levels may not accurately reflect complex biological interactions, potentially missing nonlinear relationships and interactions among these variables. Additionally, by assuming a uniform linear effect of time, the model simplifies the time dynamics. I acknowledge that this approach likely constitutes an oversimplification, as the drug's impact on SGAJ levels is expected to vary non-linearly across different time intervals and among individual patients, potentially affecting the precision of our interpretations.
- The choice of prior distributions, particularly the uniform distributions for the standard deviations and the normal priors with large variances for fixed effects and the random slope, introduces a level of subjectivity. These priors are intentionally broad, aimed to be vague and allow the data itself to play a major role in shaping the posterior distributions. This potentially resulted in wider credible intervals and less precise estimates, particularly given the small number of samples.
- I assumed (prompted by the laboratory technician) a constant measurement error (standard deviation of 75 $\mu mol/L$) for SGAJ across all times and patients. If the measurement error however varies by time point or between patients (e.g., due to different conditions under which measurements were taken), this could affect the accuracy of the estimates.
- Since we have data only for the first 24 hours post-administration, the model may not generalize well to later time points, such as the 30-hour mark required by Task 3. Biological behavior beyond the first 24 hours could differ significantly.
- While individual-specific intercepts and slopes tailor the model to each patient, more data points per patient and more patients overall would enhance the robustness of estimates. The limited number of patients we have (12) may skew interpretations of general trends and individual differences.

## Part 3

```
# Add new row for patient 12 with timepoint = 30 and measured = NA.
sgaj_miss<- rbind(sgaj, c(12, 1, 77, 30, NA))

data_miss <- list(n_patients = length(unique(sgaj_miss$patient)),
             patient = sgaj_miss$patient,
             age = sgaj_miss$age,
             time = sgaj_miss$time,
             female = sgaj_miss$female,
             n = nrow(sgaj_miss),
```

```
            y = sgaj_miss$measured
)
```

```
jags_model_miss <- jags.model(textConnection(model),
                          data = data_miss,
                          inits = initial_values,
                          n.chains = 2)
```

```
## Compiling model graph
##    Resolving undeclared variables
##    Allocating nodes
## Graph information:
##    Observed stochastic nodes: 60
##    Unobserved stochastic nodes: 92
##    Total graph size: 660
##
## Initializing model
```

```
# Burn-in.
update(jags_model_miss, 1000)

pars_miss <- c("mu", "yrep")

# Sampling - same as before 200,000 iterations with thin=10.
samples_miss <- coda.samples(jags_model_miss,
                         variable.names = pars_miss,
                         n.iter = 200000,
                         thin=10)
sgaj_miss_draws <- as_draws(samples_miss)

# Get last yrep - which is the one for patient 12 at 30 hours.
y.pred <- extract_variable(sgaj_miss_draws,  paste0("yrep[",data_miss$n,"]"))

# Find probability (based on posterior predictive distribution) that it is above 500.
mean(y.pred > 500)
```

```
## [1] 0.317275
```

The probability of SGAJ levels for patient id = 12 exceeding 500 $\mu mol/L$ at 30 hours after SEYAB administration is estimated at approximately 0.317 (31.7%). This prediction considers individual factors such as age, gender, and patient-specific drug response trajectories, which helps in tailoring the forecast to patient id = 12, potentially aiding in personalized clinical decision-making. Moreover, it integrates both the uncertainty of the model parameters and the observed variability among patients, offering a probabilistic assessment rather than a fixed estimate. However, it's important to note that this extrapolation to 30 hours carries additional uncertainty since the model was initially calibrated only with data from the first 24 hours. Furthermore, the reliability of this prediction is bounded by the model's assumptions and inherent limitations (as explained in section 2.6).

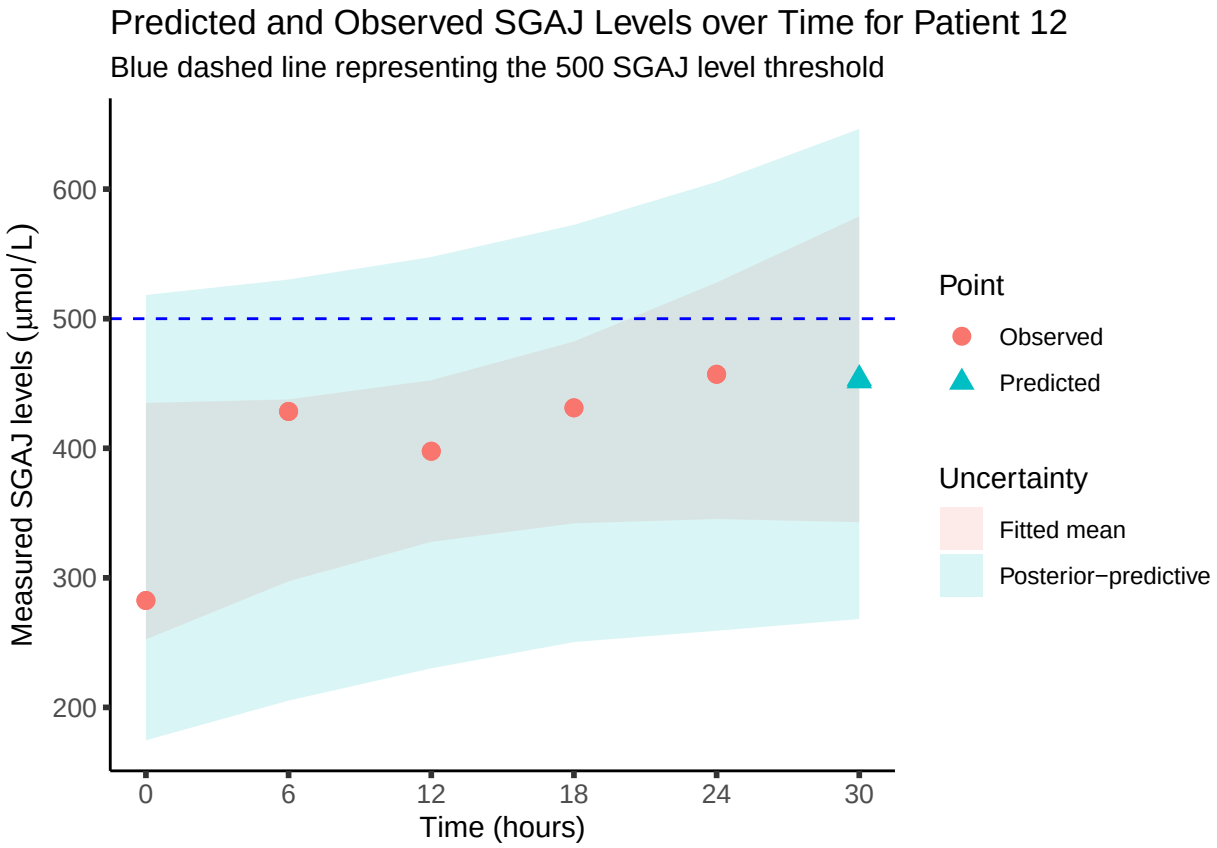
```
# get ids of patient 12 - to use to get yrep.
ids_patient_12 <- rownames(sgaj_miss[sgaj_miss$patient == 12, ])

# Get dataframe with observations only for patient 12.
```

```r
sgaj_miss_patient_12 <- sgaj_miss[sgaj_miss$patient == 12, ] %>%
  mutate(index = 1:6)

# Get summary for mu (fitted mean) and yrep (posterior predictive) for patient 12.
samples_miss_mu_pp <- summary(subset_draws(sgaj_miss_draws, c(
    paste0("mu[", ids_patient_12, "]"),
    paste0("yrep[", ids_patient_12, "]")
  )), ~ quantile(.x, probs = c(0.025, 0.5, 0.975))) %>% mutate(
    index = rep(1:nrow(sgaj_miss_patient_12), 2),
    Uncertainty = if_else(
      str_detect(variable, "mu"),
      "Fitted mean",
      "Posterior-predictive"
    )
)

# Combine dataframes.
combined_df_patient_12 <-
  samples_miss_mu_pp %>% left_join(sgaj_miss_patient_12, by = "index")  %>%
  mutate(
    Point = if_else(is.na(measured), "Predicted", "Observed"),
    measured = if_else(is.na(measured), `50%`, measured)
  )

# Get posterior summaries (mu and yrep) table for patient 12 at time 30.
posterior_summaries_patient_12_30_hours <-
  combined_df_patient_12[combined_df_patient_12$time == 30, c("variable",
                                                              "2.5%",
                                                              "50%",
                                                              "97.5%",
                                                              "patient",
                                                              "age",
                                                              "female",
                                                              "time")]
print(posterior_summaries_patient_12_30_hours)
```

```
## # A tibble: 2 x 8
##   variable `2.5%` `50%` `97.5%` patient   age female  time
##   <chr>     <dbl> <dbl>   <dbl>   <dbl> <dbl>  <dbl> <dbl>
## 1 mu[61]     343.  452.    579.      12    77      1    30
## 2 yrep[61]   268.  454.    646.      12    77      1    30
```

```r
# Create plot for predicted and observed values for patient 12.
ggplot(combined_df_patient_12, aes(x = time, y = measured)) +
  geom_ribbon(aes(ymin = `2.5%`, ymax = `97.5%`, fill = Uncertainty), alpha = 0.15) +
  geom_hline(yintercept = 500,
             linetype = "dashed",
             color = "blue") +
  geom_point(size = 3, aes(color = Point, shape = Point)) +
  scale_shape_manual(values = c(16, 17)) +
  scale_x_continuous(breaks = seq(0, max(combined_df_patient_12$time), by = 6)) +
  theme_classic() +
  theme(
```

```
    axis.text.x = element_text(size = 10),
    axis.text.y = element_text(size = 10),
    axis.ticks = element_line(size = 1)
  ) +
  labs(
    x = "Time (hours)",
    y = expression(Measured ~ SGAJ ~ levels ~ (mu * mol / L)),
    title = "Predicted and Observed SGAJ Levels over Time for Patient 12",
    subtitle = "Blue dashed line representing the 500 SGAJ level threshold",
  ) +
  guides(color = guide_legend(title = "Point"),
         shape = guide_legend(title = "Point"))
```




The plot shows patient id = 12's observed and predicted (time=30) SGAJ levels along with the 95% CI uncertainty bands for the fitted mean and posterior-predictive distributions. The following insights can be drawn from the plot and the table:

- For the fitted mean (mu), the 50% percentile is 452 (343, 579) $\mu mol/L$ which suggests the model's central estimate is below the 500 $\mu mol/L$ threshold, but the upper range of the credible interval does surpass it.
- The posterior predictive distribution (yrep), which accounts for both the uncertainty in the model parameters and the additional variability in future observations, has a median of 454 $\mu mol/L$ and a wider 95% credible interval from 268 to 646 $\mu mol/L$. The fact that the 97.5% percentile of the posterior-predicteve distribution is considerably higher than 500 confirms that there is indeed a possibility of substantially higher SGAJ levels, though not the most probable outcome.

- The slightly wider uncertainty band at the 30-hour prediction (compared to the other points) highlights the model's increasing uncertainty with time, emphasizing that predictions further from the last observed data point should be treated with more caution.

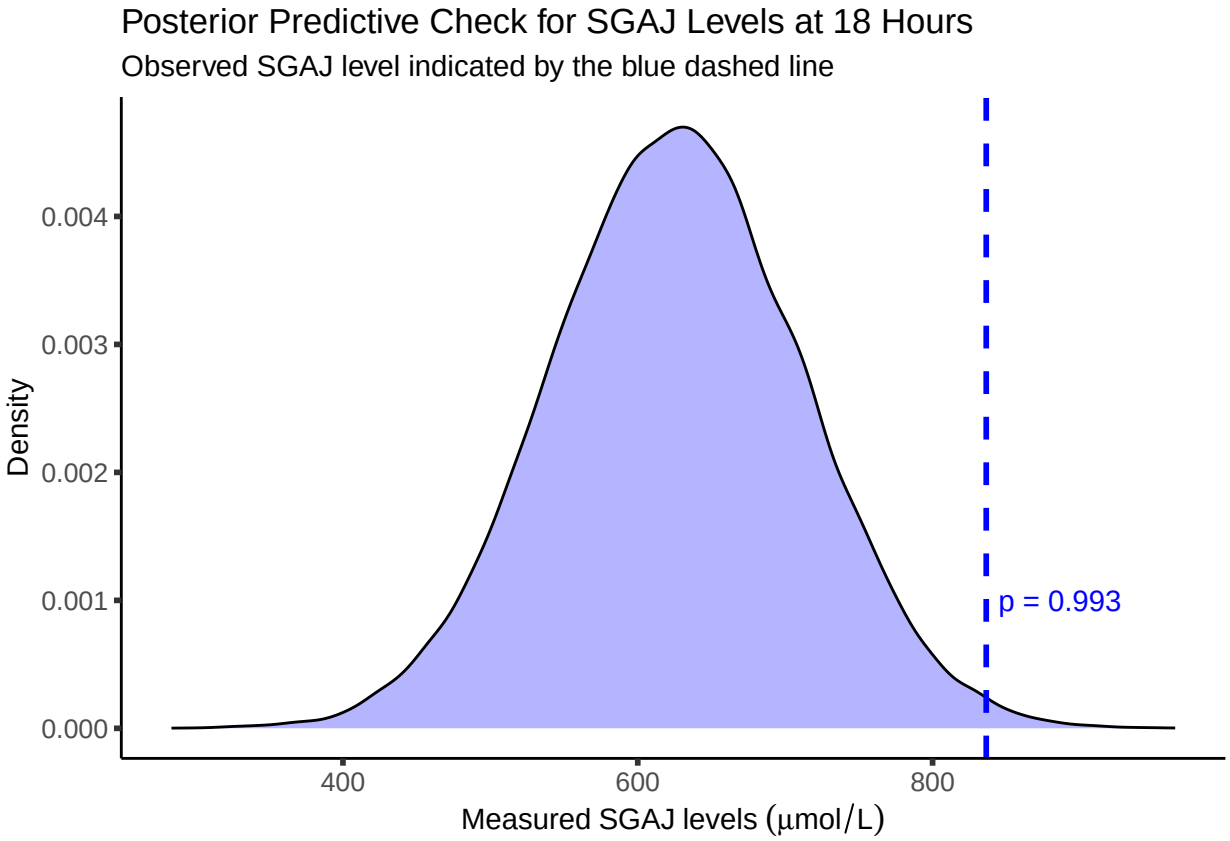
# Part 4

```
# get posterior predictive distribution for patient with id=3.
# From the original model (not the one with missing)
index_patient_3_time_18 <-
  rownames(sgaj[sgaj$patient == 3 & sgaj$time == 18, ])

# Get summary for posterior-predictive distribution of patient 3 at time=18.
yrep_patient_3_time_18 <-
  extract_variable(sgaj_draws, paste0("yrep[", index_patient_3_time_18, "]")) %>%
  as_tibble()
summary(yrep_patient_3_time_18)
```

```
##      value
##  Min.   :283.7
##  1st Qu.:569.0
##  Median :626.2
##  Mean   :626.2
##  3rd Qu.:683.2
##  Max.   :964.6
```

```
obs_patient_3_time_18 <-
  as.numeric(sgaj[index_patient_3_time_18, ]$measured)

# Calculate the p-value.
p_value <-
  mean(yrep_patient_3_time_18$value < obs_patient_3_time_18)

# Posterior-predictive check plot.
ggplot(yrep_patient_3_time_18, aes(x = value)) +
  geom_density(fill="blue", alpha=0.3) +
  geom_vline(
    aes(xintercept = obs_patient_3_time_18),
    color = "blue",
    linetype = "dashed",
    linewidth = 1
  ) +
  annotate(
    "text",
    x = obs_patient_3_time_18 + 50,
    y = 0.001,
    label = paste0("p = ", round(p_value, 3)),
    color = "blue"
  )+
  theme_classic() +
  labs(
    x = expression(Measured ~ SGAJ ~ levels ~ (mu * mol / L)),
```

```
    y = "Density",
    title = "Posterior Predictive Check for SGAJ Levels at 18 Hours",
    subtitle = "Observed SGAJ level indicated by the blue dashed line"
  ) +
  theme(
    axis.text.x = element_text(size = 10),
    axis.text.y = element_text(size = 10),
    axis.ticks = element_line(size = 1)
  )
```




The observed SGAJ level for patient id = 3 at 18 hours post-SEYAB administration, depicted by the dashed blue line on the plot, significantly deviates from the majority of the model's posterior predictive distribution (shown in the density plot), as reflected by a p-value of 0.993. There are two possible interpretations for this p-value: - This exceptionally high p-value suggests a potential lack of fit between the model and the observed data for this specific instance. This may necessitate a re-evaluation of the model's prior distributions, formulation, and the inclusion of additional data or covariates to improve the model's predictive accuracy. - Alternatively, this particular observation could be an outlier and may not be indicative of the typical patient's response. This could be due to unique individual factors influencing the patient's metabolism of SEYAB, an unusual interaction with other variables not accounted for in the model, or even measurement error or data recording anomalies.

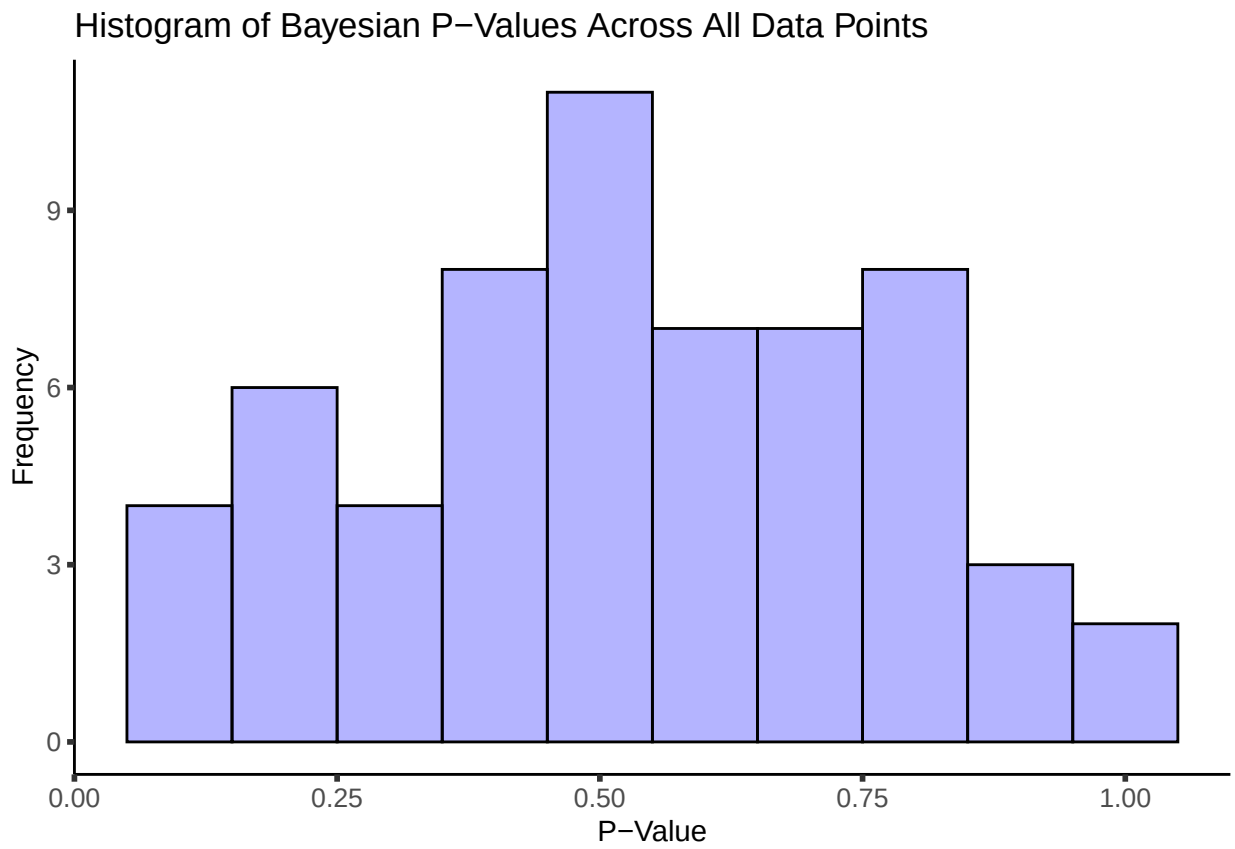
To discern whether the lack of fit is an isolated incident or indicative of a broader issue with the model, I calculated the Bayesian p-values for all the observed values for all patients.

```
# Calculate bayesian p-values for all observations.
bayesian_p_values <- c()
for (observation in 1:nrow(sgaj)) {
```

```
  yrep <- sgaj_draws %>%
    extract_variable(paste0("yrep[", observation,"]")) %>%
    as_tibble()

  # Calculate p-value for current observation and append it.
  value_measured <- sgaj$measured[observation]
  bayesian_p_values <- c(bayesian_p_values, mean(yrep < value_measured))
}

# Print summary distribution of p-values.
print(summary(bayesian_p_values))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05955 0.36673 0.51948 0.52772 0.70089 0.99333
```

```
# Plot p-values in a histogram
data <- data.frame(p_values = bayesian_p_values)

ggplot(data, aes(x = p_values)) +
  geom_histogram(binwidth = 0.1, fill = "blue", color="black", alpha=0.3) +
  theme_classic() +
  labs(title = "Histogram of Bayesian P-Values Across All Data Points",
       x = "P-Value",
       y = "Frequency") +
  theme(
    axis.text.x = element_text(size = 10),
    axis.text.y = element_text(size = 10),
    axis.ticks = element_line(size = 1)
  )
```




The inherent conservatism of p-values due to the dual use of observed data in model fitting and p-value calculation implies that exceptionally high or low p-values are indicative of notable model discrepancies which can signal areas where the model may not be accurately capturing the underlying data distribution. While it's obvious that there are a few instances with very high or low p-values, the histogram illustrates a majority cluster of Bayesian p-value around 0.5, suggesting a general concordance between the model's predictions and observed data across multiple cases.