# BDA Project

Arthur Aspelin, Jannica Savander, Christian Segercrantz

12/2021

## Contents



# 1. Introduction

The motivation for this project is to estimate the parameters for blood pressure data with the help of Bayesian methods. High blood pressure corresponds with different diseases, such as diabetes and heart diseases. This means that it is essential to predict distribution of blood pressure and its parameters in an accurate way.

Solving the problem, firstly, we want to estimate what type of distribution can describe blood pressure. Then with different Bayesian models estimate the parameters for the distribution. We will also investigate how the

parameters differ when dividing the data into different age groups. Higher blood pressure could correspond with higher age.

The main modeling idea is to test, with different Bayesian models, how to get accurate estimates of the parameters that could describe blood pressure. This answers the research question of whether older and younger people have different models describing their blood pressure, and therefore have different blood pressures.

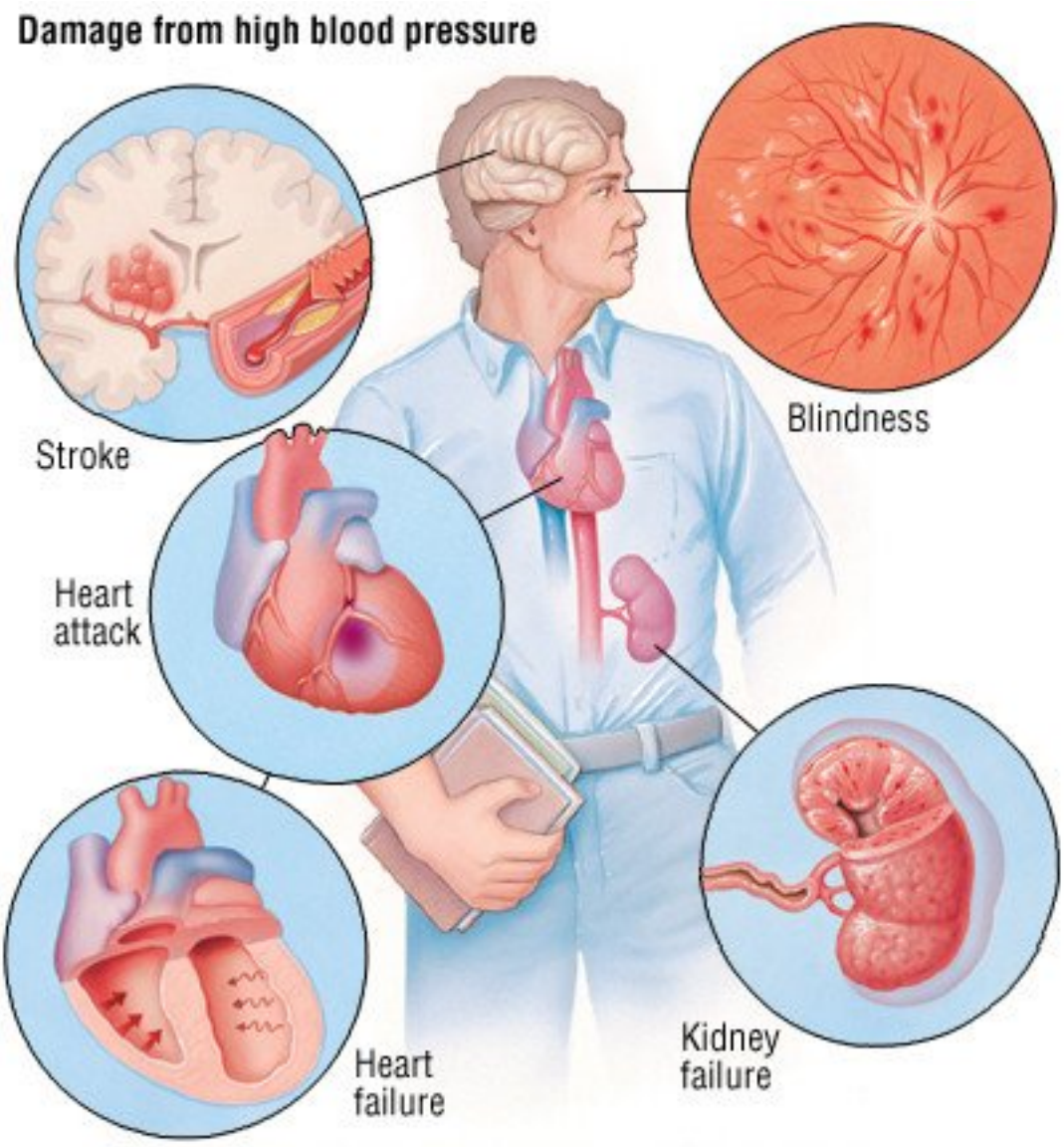


Figure 1: Damage that can be caused from high blood pressure.

## 2. Description of the data

We used blood pressure data combined with age data from the Diabetes Dataset from Kaggle. According to the data description, the dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases, and one data point corresponds to a female patient of Pima Indian heritage. All patients are at least 21 years old.

The dataset has more columns than we used, for example number of pregnancies, BMI and diabetes classification. We only used the columns BloodPressure, describing the diastolic blood pressure, and Age, describing the age of the patient in years. The diastolic blood pressure is the pressure the heart applies on the walls of the arteries between the beats (Mayo Clinic Staff (2021)). The unit for the diastolic blood pressure is mmHg, and a normal value is usually below 80. Higher values might indicate hypertension, which increases with age in Western countries (Gurven et al. (2012)).

We separated the data in two parts based on age group. Using the cutoff value 30 for age, we got an younger and an older group with 394 and 338 patients respectively. The groups will from here be referred to as "young group" and "old group." The code and the plot for the groups separately can be seen below.

```
data <- data %>%
  filter(BloodPressure > 0) %>%
  select(BloodPressure, Age) %>%
  mutate(AgeGroup = case_when(
  Age <= 30      ~ "Young",
  Age > 30       ~ "Old")
  )
knitr::kable(head(data),
             caption = "The first rows of the dataset, with the additinal 'AgeGroup' column.")
```

Table 1: The first rows of the dataset, with the additinal 'AgeGroup' column.

| BloodPressure | Age | AgeGroup |
|---|---|---|
| 72 | 50 | Old |
| 66 | 31 | Old |
| 64 | 32 | Old |
| 66 | 21 | Young |
| 40 | 33 | Old |
| 74 | 30 | Young |

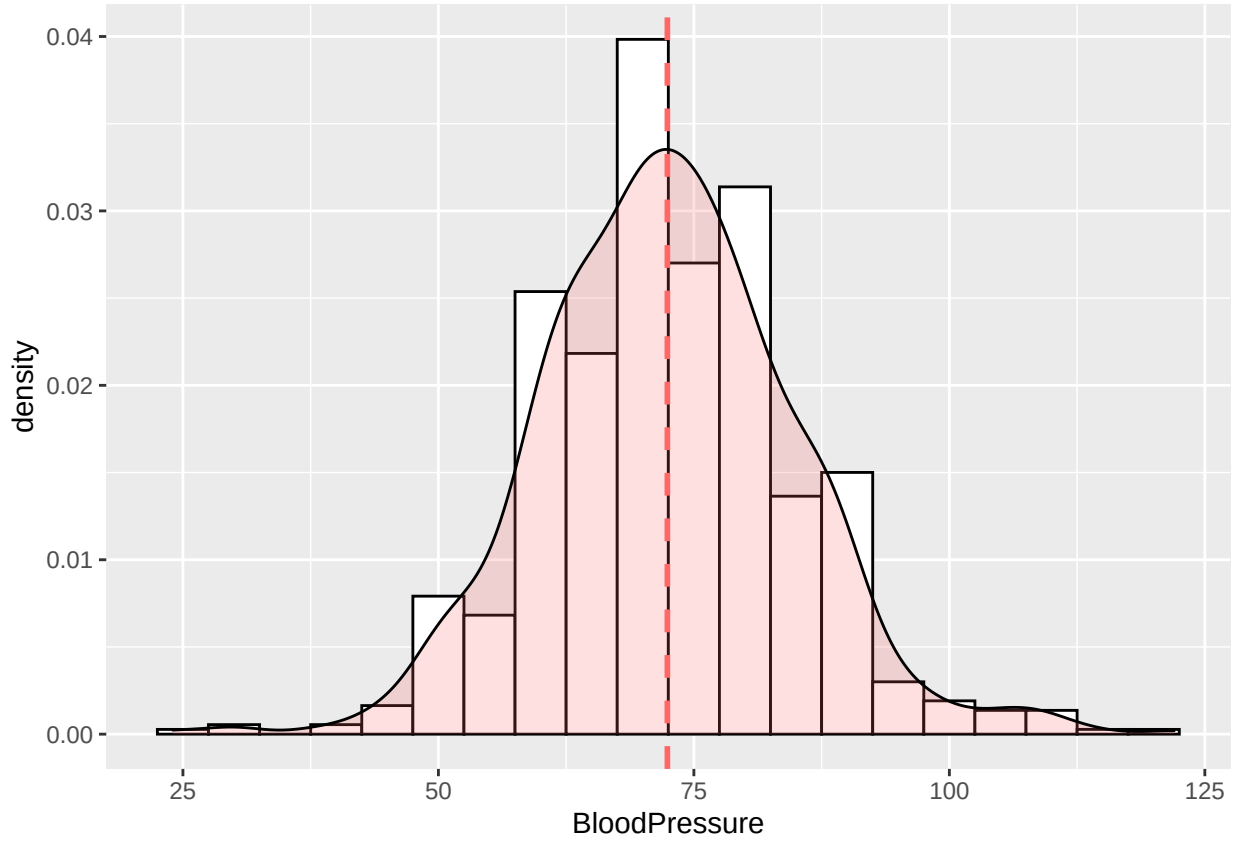


Figure 2: A histogram of the complete data set for the blood pressure. A density approximation plot and a mean line is included as well.

```
means <- data %>%
  group_by(AgeGroup) %>%
  summarise(mean = mean(BloodPressure), n = n())

ggplot(data, aes(x=BloodPressure, fill=AgeGroup)) +
  geom_histogram(aes(y=..density..),
                 binwidth = 5,
                 colour="black",
                 position = "identity",
                 alpha = 0.4) +
  geom_vline(data = means, aes(xintercept=mean, color = AgeGroup), linetype="dashed", size=1) +
  geom_density(alpha=.2)
```

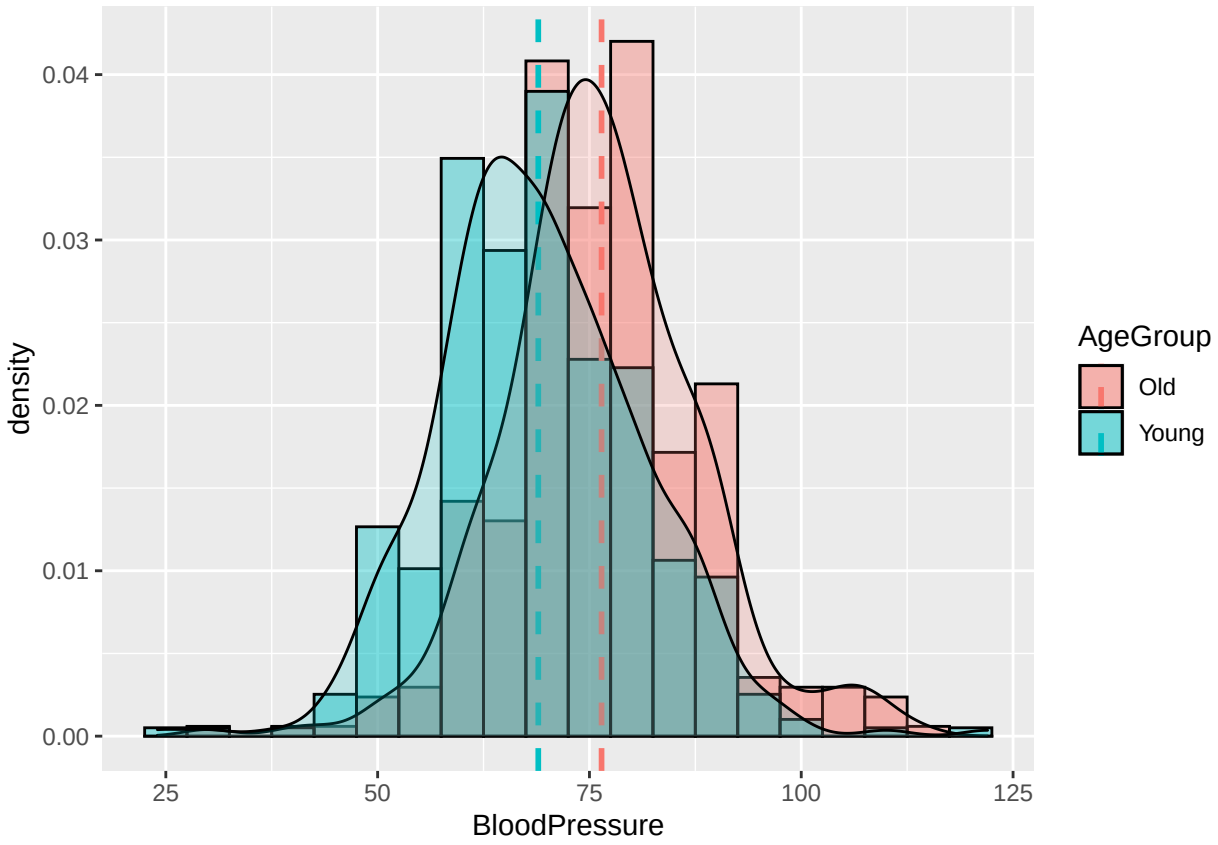


Figure 3: Histograms with density approximations and mean lines for both age groups.


Figure 4: The bloodpressure and age scatterplotted for each data point.

## 3. Description of the models

We used two models, one hierarchical and one nonhierarchical, and ran each model two times (once per age group). Therefore we got four fits: hierarchical for the old group, nonhierarchical for the old group, hierarchical for the young group, and nonhierarchical for the young group. The data looks normally distributed (which is very natural in its biological context), so we base our models on the normal distribution.

The nonhierarchical model can be mathematically summarized as

$$y_{ij} \sim \mathcal{N}(\mu_j, \sigma_j)$$

$$\mu_j \sim \mathcal{N}(\mu_{prior}, \sigma_{prior})$$

$$\sigma_j \sim \text{Inv-}\chi^2(Var_{prior})$$

.

Using the same notation we get the following hierarchical model:

$y_i \sim \mathcal{N}(\mu_i, \sigma)$

$\mu_i \sim \mathcal{N}(\mu, \tau)$

$\mu \sim \mathcal{N}(\mu_{prior}, \sigma_{prior})$

$\tau \sim$ Inv-$\chi^2(Var_{prior})$

$\sigma \sim$ Inv-$\chi^2(\tau)$.

# 4. Priors

We use weekly informative priors for parameters $\mu$, $\sigma$ and $Var$. They are chosen based on the blood pressure data that we get from the data set. From the plotted data we choose priors that possibly could describe a normal distribution describing the blood pressure. $\mu$, $\sigma$ and $Var$ for the distribution of our two different age groups:

- For $Age \leq 30 : \mu_{prior} = 65$, $\sigma_{prior} = 10$, $Var_{prior} = 20$. Our prior $\mu$ is slightly lower here than for the older age group. Other priors stay the same.
- For $Age > 30 : \mu_{prior} = 75$, $\sigma_{prior} = 10$, $Var_{prior} = 20$. Here the prior $\mu$ is higher, because it can be seen from the plot that the mean is higher, but other priors stay the same.

# 5. Stan code

Below the Stan code for the two models (nonhierarchical and hierarchical) can be seen.

### Nonhierarchical model

```
data {
  int<lower=0> N;                  //Amount of data points
  vector[N] y;                     //Data points
  real mean_mu_prior;              //Expected value of the mean prior
  real<lower=0> mean_sigma_prior;  //variance of the mean prior
  real<lower=0> var_prior;         //Variance of the variance prior
}

parameters {
  real mu;                         //The parameter mu describing the man
  real<lower=0> sigma;             //The parameter sigma describing the variance
}

model {
  //Priors:
  // The prior for the mean, normally distributed
  mu ~ normal(mean_mu_prior, mean_sigma_prior);
  // The prior for the sigma, inverse Chi-squared distributed
  sigma ~ inv_chi_square(var_prior);
 //likelihoods
  y ~ normal(mu, sigma);
}

generated quantities {
  real ypred;
  vector[N] log_lik;
  ypred = normal_rng(mu, sigma);
  for (n in 1:(N)){
    log_lik[n] = normal_lpdf(y[n] | mu, sigma);
  }
}
```

### Hierarchical model

```
data {
  int<lower=0> N;                  //Amount of data points
  vector[N] y;                     //Data points
  real mean_mu_prior;              //Expected value of the mean prior
  real<lower=0> mean_sigma_prior;  //Variance of the mean prior
  real<lower=0> var_prior;         //Variance of the variance prior
}

parameters {
  real mu;                         //The parameter mu describing the man
  real<lower=0> sigma;             //The parameter sigma describing the variance
  real mu_hypo;                    //The hyperparameter describing the mean of the hyperprior
  real<lower=0> tau;               //The hyperparameter describing the variance of the hyperpriors
}

model {
  //Hyperpriors
  //The hyperprior for the mean, normally distributed
  mu_hypo ~ normal(mean_mu_prior, mean_sigma_prior);
  //The hyperprior for the variance, inverse chi-square distributed
  tau ~ inv_chi_square(var_prior);
  //Priors:
  // The prior for the mean, normally distributed
  mu ~ normal(mu_hypo, tau);
  sigma ~ inv_chi_square(var_prior);
 //likelihoods
  y ~ normal(mu, sigma);
}

generated quantities {
  real ypred;
  vector[N] log_lik;
  ypred = normal_rng(mu, sigma);
  for (n in 1:(N)){
    log_lik[n] = normal_lpdf(y[n] | mu, sigma);
  }
}
```

# 6. Running the Stan model

We will need to run both of the models twice to get a model both for the young group and the old group. Therefore, we repeat almost the same process four times. The code is not included for all models for this reason. All models are ran with 4 chains and 2000 chains, of which one half (1000 iterations) are warmup.

### Nonhierarchical model

The nonhierarchical model uses $\mu_{old} = 75$, $\mu_{young} = 65$, $\sigma_{prior} = 10$, $\text{Var}_{prior} = 20$. The data used is the blood pressure data from the dataset for the age group respectively.

#### Old group

In this section we can see the data used in the separate model for the old age group.

```r
data_old <- data %>%
  filter(AgeGroup == "Old")

mean_mu_prior_old = 75
mean_sigma_prior_old = 10
var_prior_old = 20
data_nonhiera_old <- list(
  y = data_old$BloodPressure,
  N = length(data_old$BloodPressure),
  mean_mu_prior = mean_mu_prior_old,
  mean_sigma_prior = mean_sigma_prior_old,
  var_prior = var_prior_old
)
```

```r
fit_nonhiera_old = sampling(nonhieramodel,
  data = data_nonhiera_old,            # named list of data
  chains = 4,              # number of Markov chains
  warmup = 1000,           # number of warmup iterations per chain
  iter = 2000,             # total number of iterations per chain
  cores = 4,               # number of cores (could use one per chain)
  refresh = 0              # no progress shown
  )
```

### Young group

In this section we can see the data used in the separate model for the young age group.

```r
data_young <- data %>%
  filter(AgeGroup == "Young")

mean_mu_prior_young = 65
mean_sigma_prior_young = 10
var_prior_young = 20
data_nonhiera_young <- list(
  y = data_young$BloodPressure,
  N = length(data_young$BloodPressure),
  mean_mu_prior = mean_mu_prior_young,
  mean_sigma_prior = mean_sigma_prior_young,
  var_prior = var_prior_young
)
```

## Hierarchical model

The nonhierarchical model uses $\mu = 70$ (same for both groups). All other parameters are the same as in the nonhierarchical model.

### Old group

In this section we can see the data used in the hierarchical model for the old age group.

### Young group

In this section we can see the data used in the hierarchical model for the young age group.

```
mean_mu_prior = 70
mean_sigma_prior = 10
var_prior = 20
data_hiera_young <- list(
  y = data_young$BloodPressure,
  N = length(data_young$BloodPressure),
  mean_mu_prior = mean_mu_prior,
  mean_sigma_prior = mean_sigma_prior,
  var_prior = var_prior
)
```

# 7. Convergence diagnostics

We compute the convergence diagnostics for all models. All models have a $\widehat{R}$ value of <1.05, indicating that the chains have mixed well. The effective sample sizes (ESS) are all over 100, and can be considered good. When plotting the chains we can also see that they are converging well, and there is no need to edit the model or the priors.

Table 2: The diagnostics of the nonhierarcical model for the old age group.

| Variable | mean | se_mean | sd | n_eff | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| mu | 76.4 | 0.011 | 0.638 | 3204 | 1 | 3240 | 2465 |
| sigma | 11.4 | 0.007 | 0.429 | 3815 | 1 | 3839 | 2783 |
| ypred | 76.5 | 0.182 | 11.413 | 3912 | 1 | 3942 | 3680 |

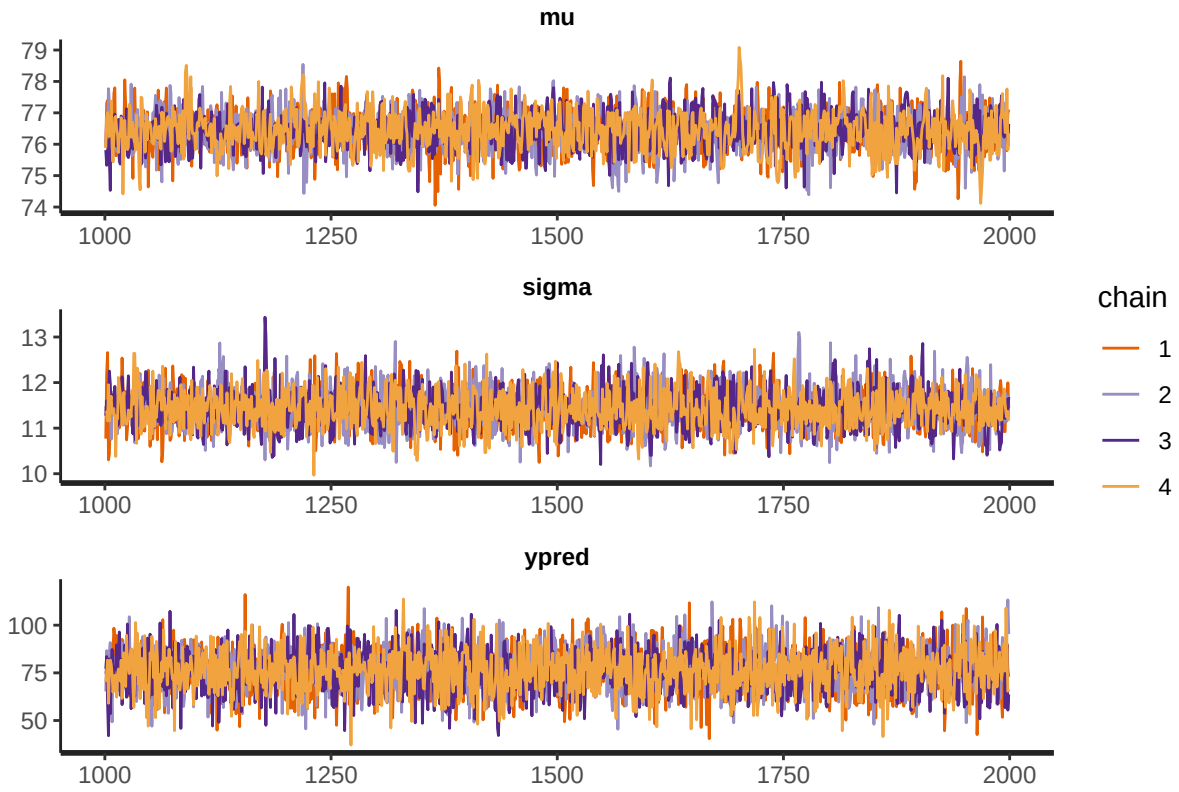


Figure 5: The trace plot of the nonhierarcical model for the old age group.

Table 3: The diagnostics of the nonhierarcical model for the young age group.

| Variable | mean | se_mean | sd | n_eff | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| mu | 69.0 | 0.010 | 0.605 | 3316 | 1 | 3324 | 2639 |
| sigma | 11.9 | 0.007 | 0.429 | 3895 | 1 | 3909 | 2557 |
| ypred | 68.8 | 0.186 | 11.816 | 4016 | 1 | 4025 | 3927 |

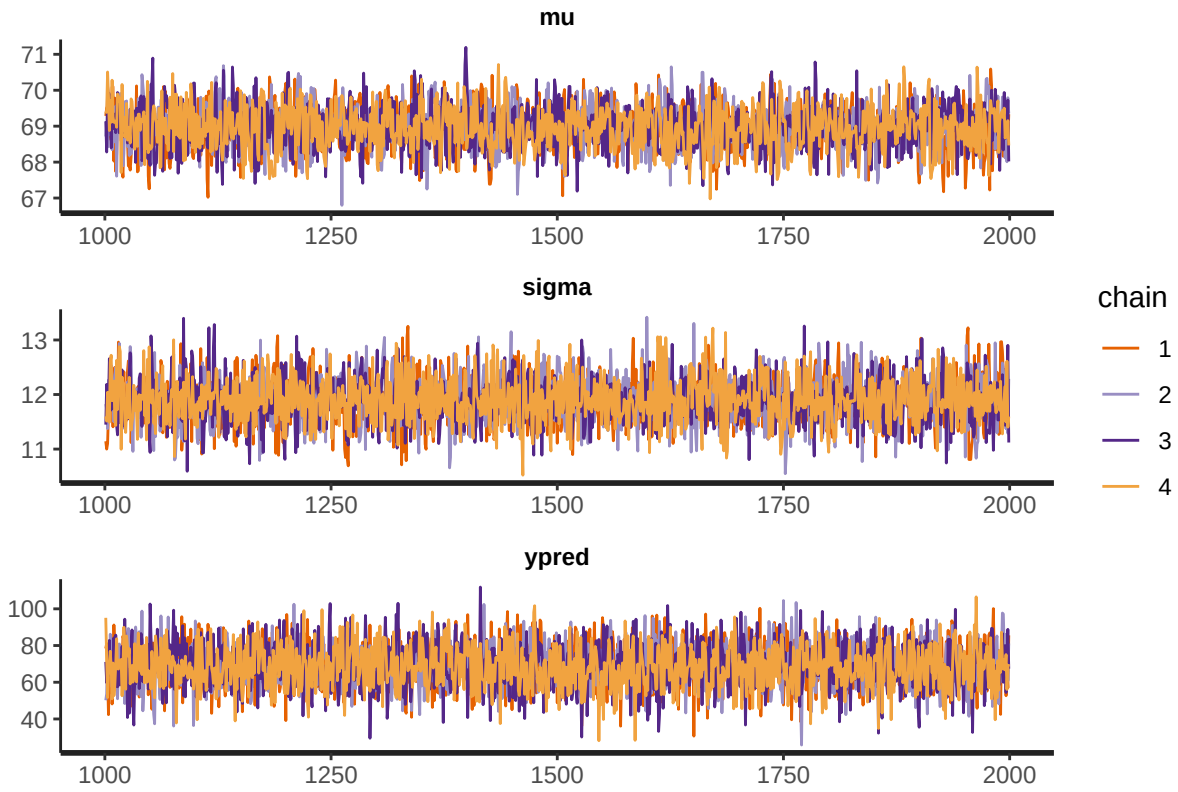


Figure 6: The trace plot of the nonhierarcical model for the young age group.

Table 4: The diagnostics of the hierarcical model for the old age group.

| Variable | mean | se_mean | sd | n_eff | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| mu | 76.402 | 0.016 | 0.618 | 1476 | 1 | 1506 | 1569 |
| sigma | 11.433 | 0.009 | 0.439 | 2311 | 1 | 2267 | 2071 |
| mu_hypo | 76.402 | 0.016 | 0.619 | 1466 | 1 | 1493 | 1569 |
| tau | 0.056 | 0.001 | 0.021 | 1284 | 1 | 1323 | 1398 |
| ypred | 76.202 | 0.185 | 11.414 | 3772 | 1 | 3799 | 3865 |

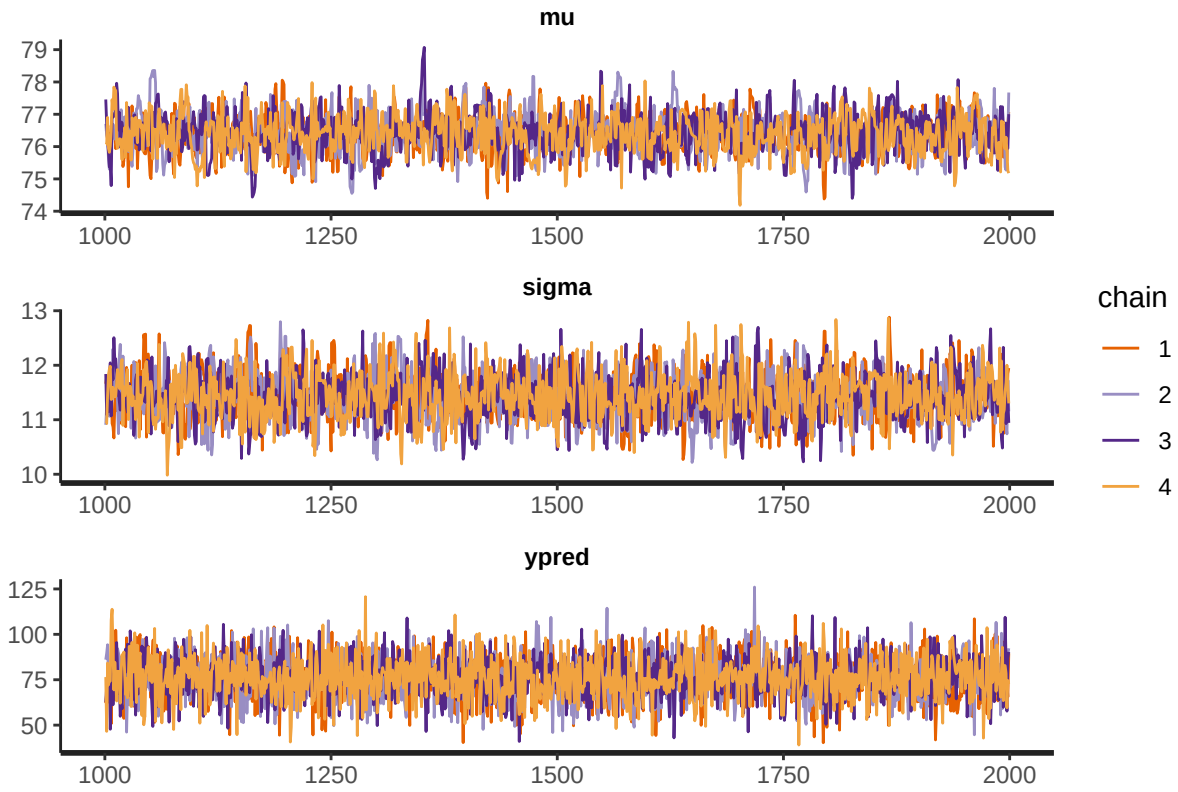


Figure 7: The trace plot of the hierarcical model for the old age group.

Table 5: The diagnostics of the hierarcical model for the young age group.

| Variable | mean | se_mean | sd | n_eff | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|
| mu | 68.974 | 0.016 | 0.601 | 1423 | 1 | 1429 | 1695 |
| sigma | 11.886 | 0.008 | 0.425 | 2500 | 1 | 2531 | 2248 |
| mu_hypo | 68.975 | 0.016 | 0.605 | 1416 | 1 | 1424 | 1606 |
| tau | 0.055 | 0.001 | 0.019 | 1219 | 1 | 1259 | 1600 |
| ypred | 69.049 | 0.187 | 11.879 | 4096 | 1 | 4119 | 3605 |

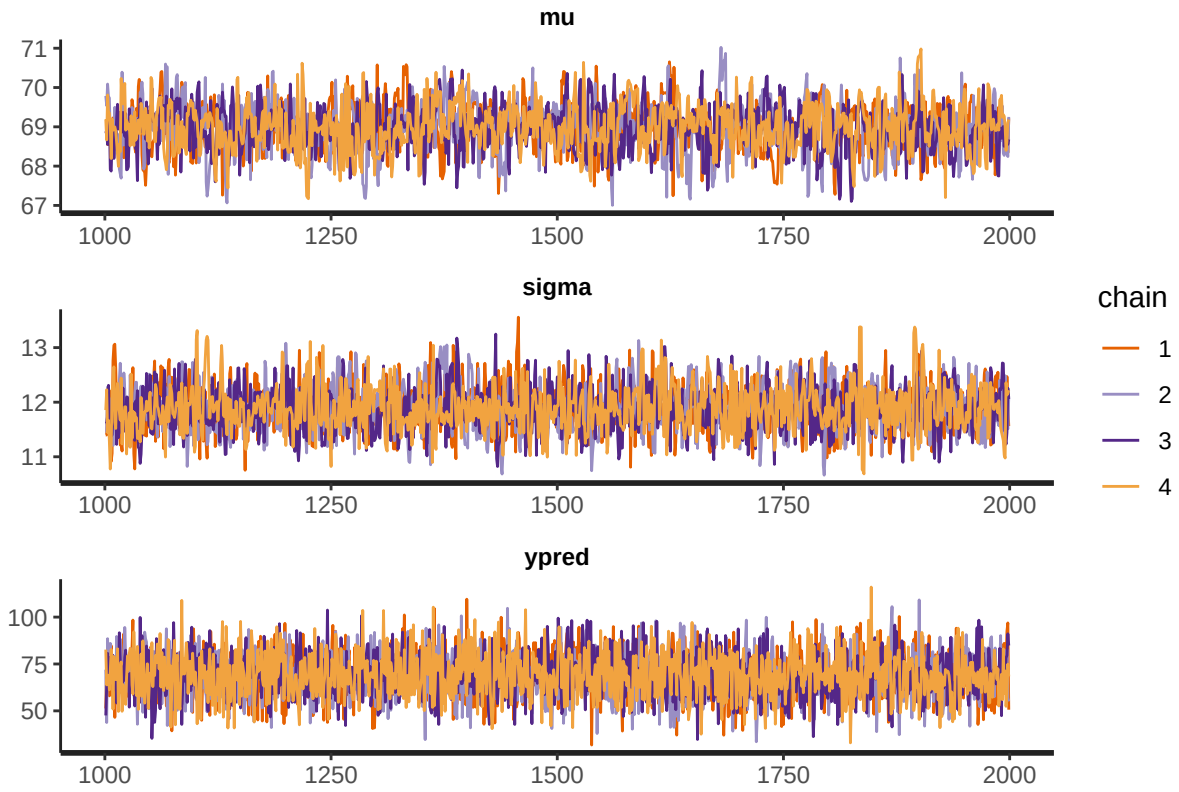


Figure 8: The trace plot of the hierarcical model for the young age group.

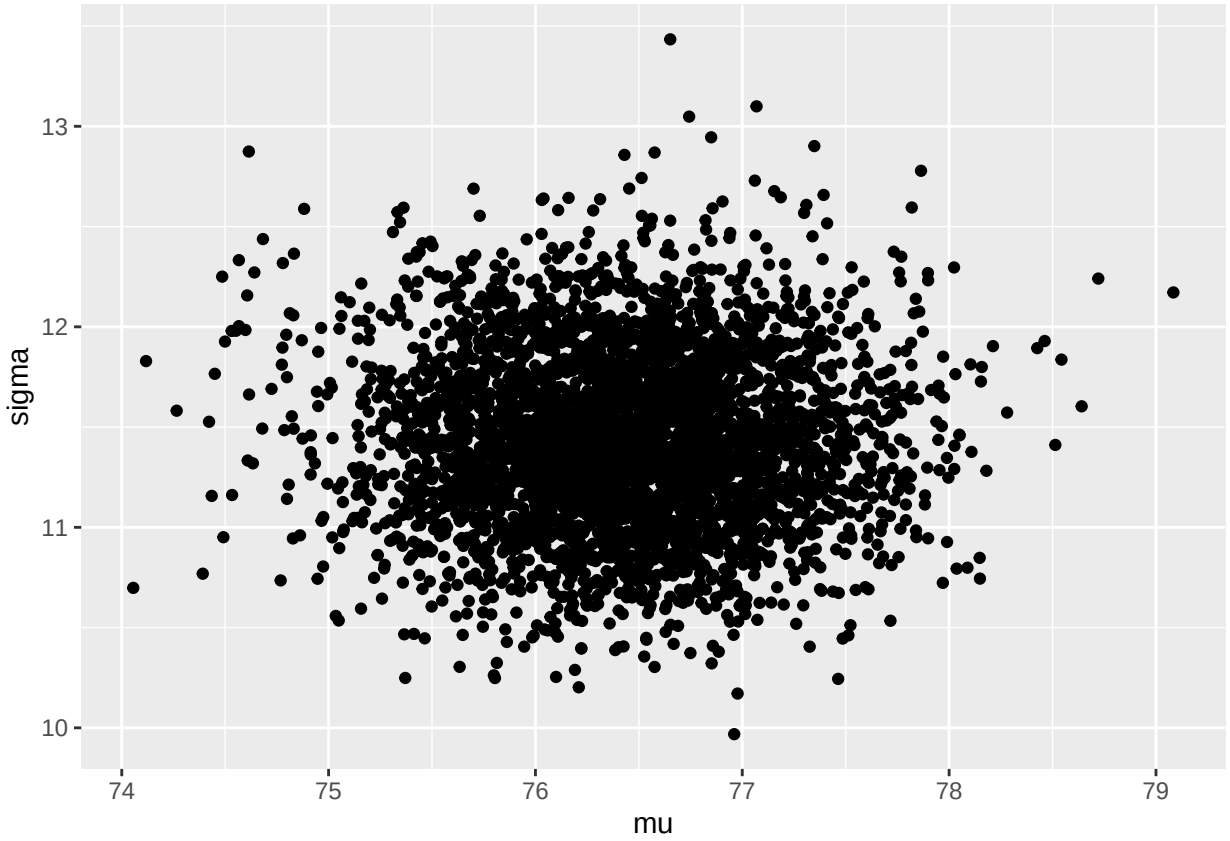


Figure 9: The paramters mu and sigma of the posterior scatter ploted.

## 8. Posterior predictive checks

We compare the posterior distribution with the original distribution visually. We can see that the posterior distribution is similar to the original distribution in all cases, indicating that the model is working as it should.

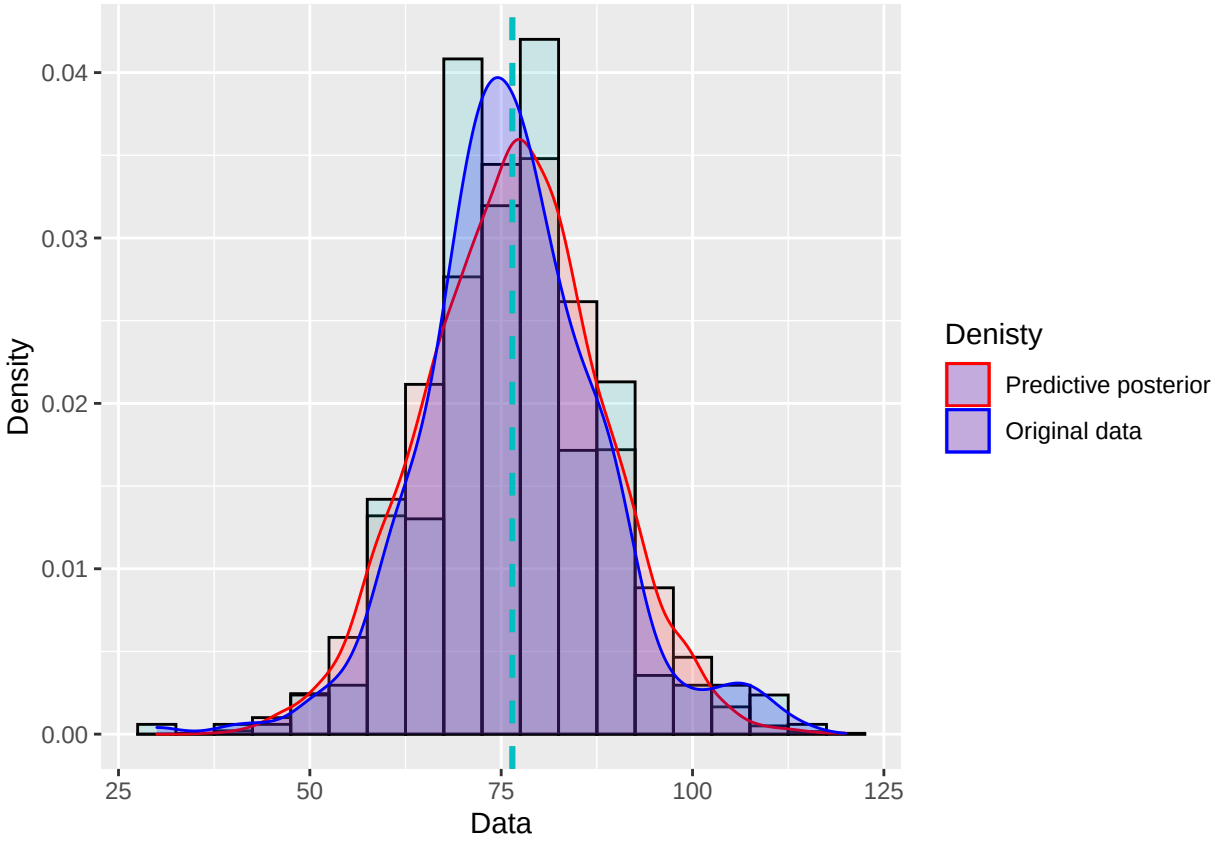


Figure 10: Hierarical posterior predictive vs original data for the old age group.

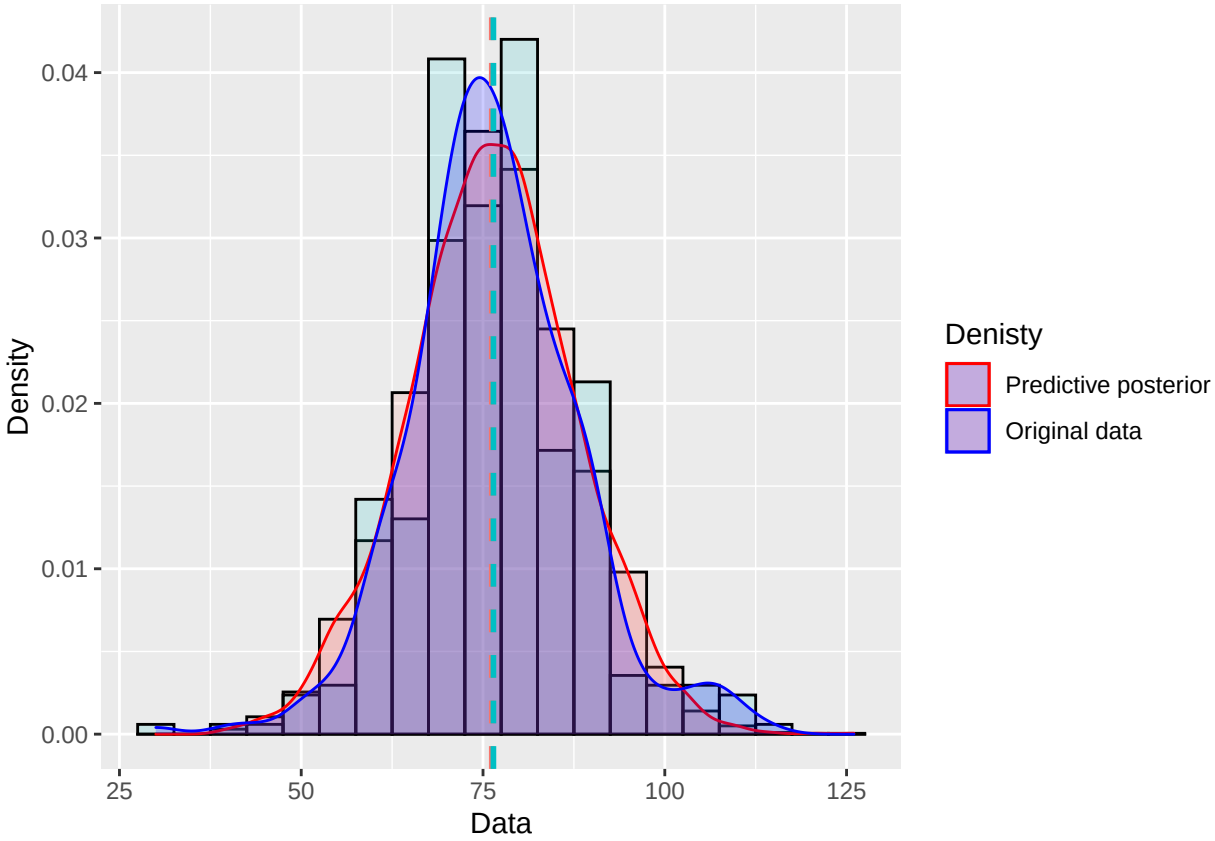


Figure 11: Hierarical posterior predictive vs original data for the old age group.

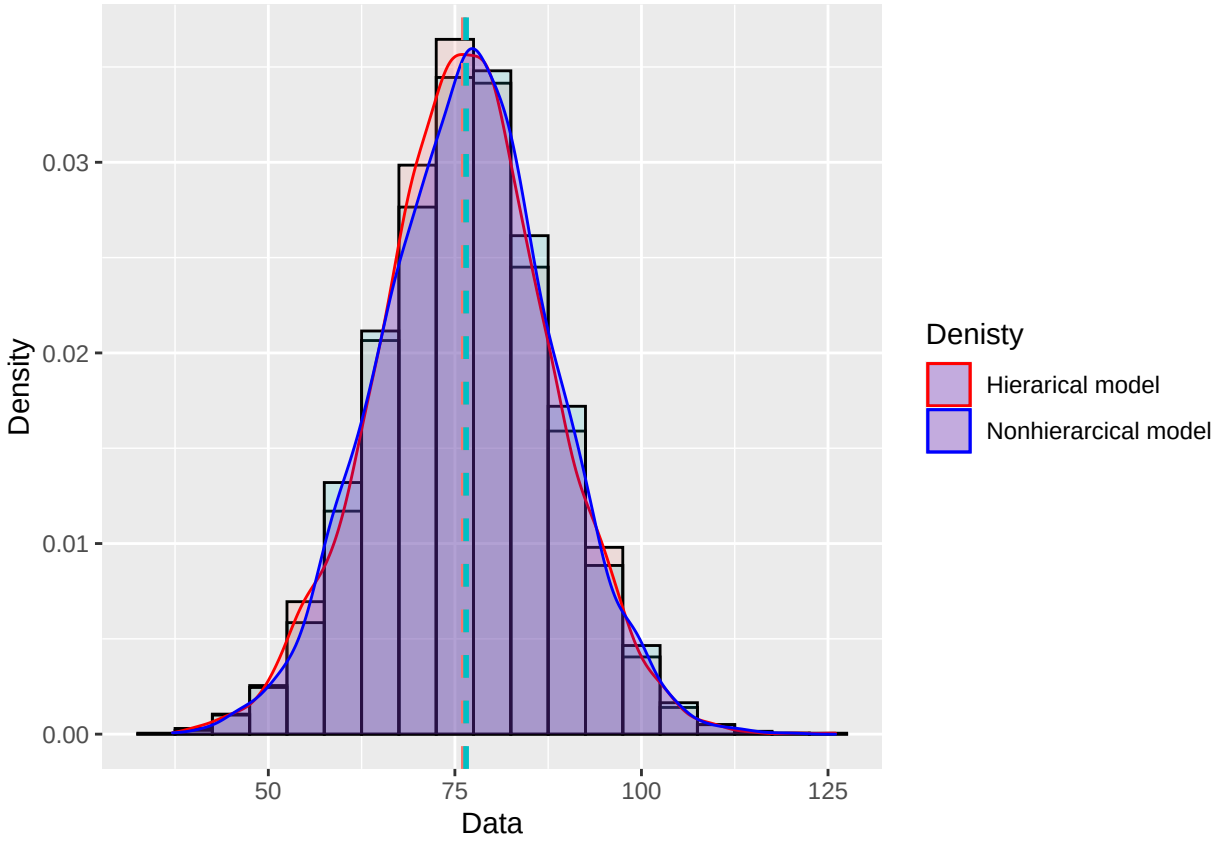


Figure 12: Posterior predictive distributions of the hierarcical versus non-hierarcical data for the old age group.

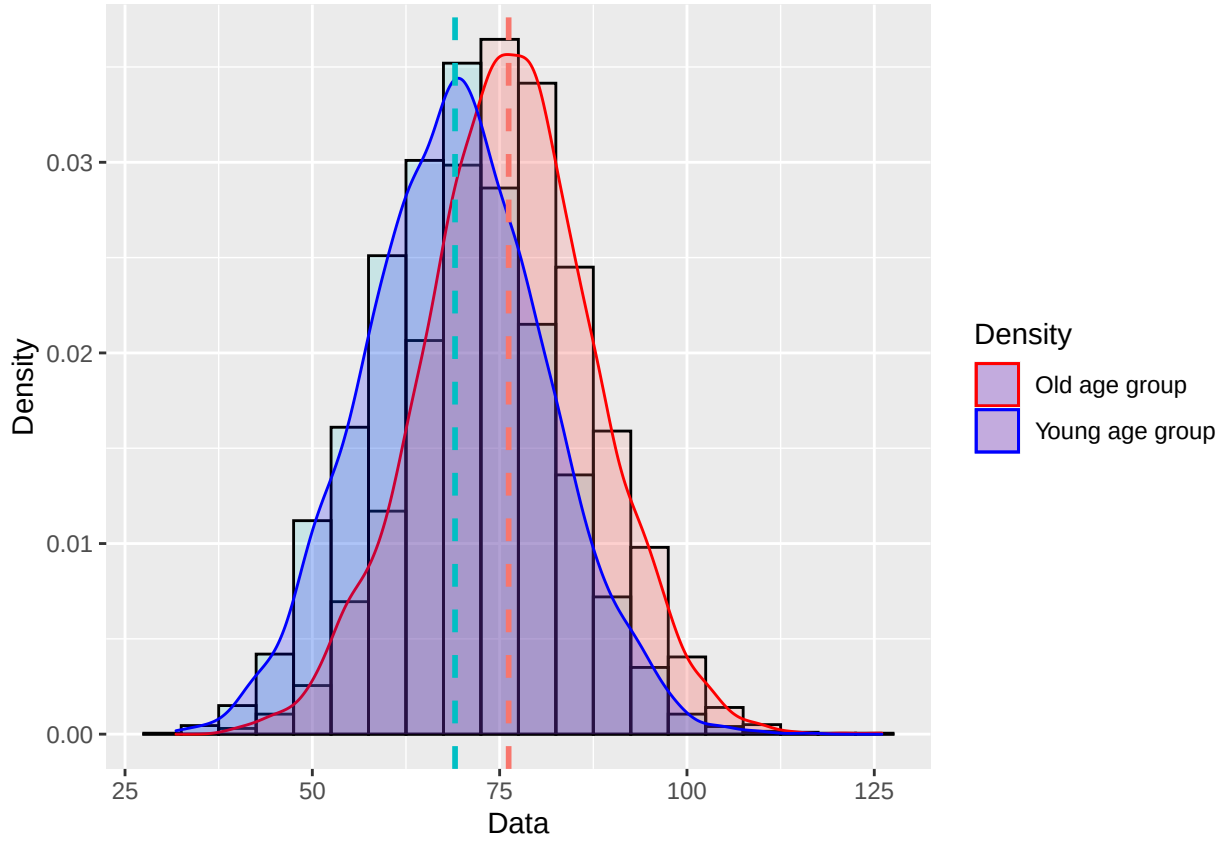


Figure 13: Posterior predictive distributions of the non hierarcical models for the old versus the young age group

## 9. Model comparison with LOO-CV

Next, we compare the nonhierarchical and the hierarchical model with LOO-CV. We can see that both models are equally good for the age groups.

```
##
## Computed from 4000 by 338 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1309.1 17.4
## p_loo         2.7  0.5
## looic      2618.2 34.8
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

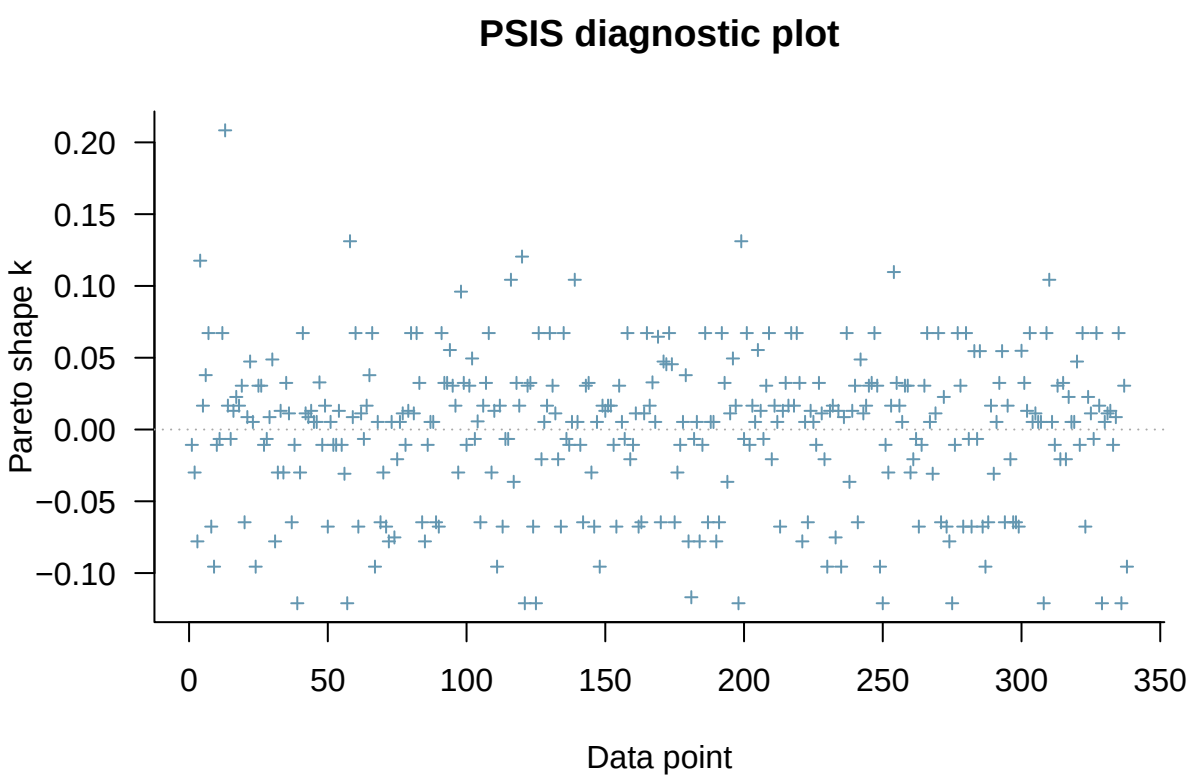


Figure 14: The pareto k values using leave-one-out method for the nonhierarchical model for the old age group.

```
##
## Computed from 4000 by 395 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1544.8 18.2
## p_loo         2.7  0.6
## looic      3089.6 36.4
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
##
## Computed from 4000 by 338 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1309.1 17.4
## p_loo         2.7  0.5
## looic      2618.2 34.8
## ------
## Monte Carlo SE of elpd_loo is 0.0.
```

```
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
##
## Computed from 4000 by 395 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1544.8 18.2
## p_loo         2.7  0.6
## looic      3089.6 36.4
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
print("The model for the old:")
```

```
## [1] "The model for the old:"
```

```
loo_compare(loo_nonhiera_old, loo_hiera_old)
```

```
##        elpd_diff se_diff
## model1 0.0       0.0
## model2 0.0       0.0
```

```
print("The model for the young:")
```

```
## [1] "The model for the young:"
```

```
loo_compare(loo_nonhiera_young, loo_hiera_young)
```

```
##        elpd_diff se_diff
## model1 0.0       0.0
## model2 0.0       0.0
```

# 10. Predictive performance assesment

We do not use our model to make predictions. Instead, we try to describe the blood pressure among the population and the effect of higher age on blood pressure. The predictive performance is not relevant as we do not predict anything with the current model.

# 11. Sensitivity analysis

To perform sensitivity analysis, we try a number of different priors to see if they change the posterior in a notable way. We try different combinations of the prior parameters, with $\mu$ values 0, 50, 100 and 1000, and $\sigma$ and Var values 1, 10, 100 and 1000. When we then plot the chains we can see that there are no large differences between how the posteriors behave, regardless of the priors. This indicates that the likelihood as a stronger influence than the prior, resulting in that the values of the prior parameters does not change the posterior.

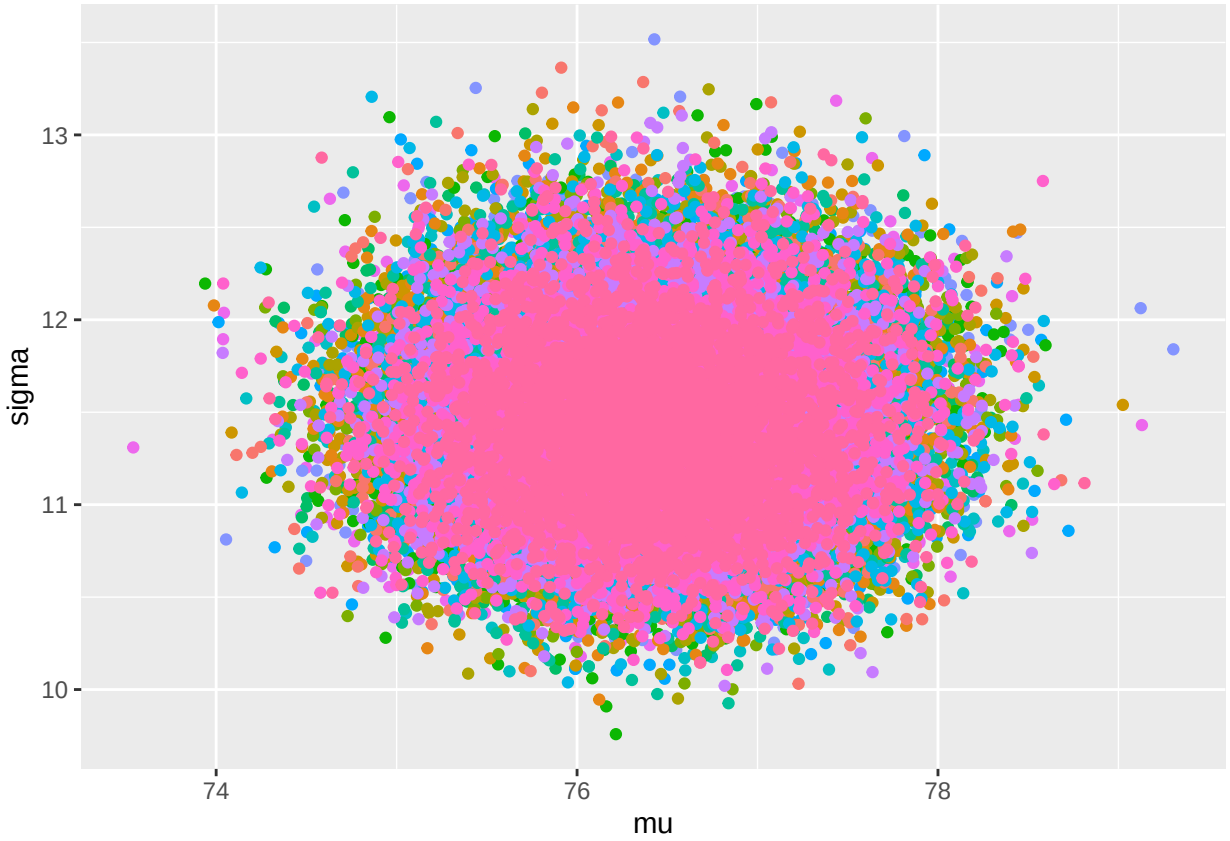


Figure 15: The complete chains, including warmup, of the sixteen combinations of prior values ploted.

## 12. Discussion

One issue we get when performing this analysis are that the LOO-values we get from both models are identical. For the old group it is -1309.1 both for separate and hierarchical. For the young group it is -1544.6 for both models. This causes that we cannot decide on which model performs better with respect to the data and priors.

We conclude that this is caused by our data of blood pressure being very close to normally distributed, hence easy to estimate. For improvement we could try our models on other datasets that are not as normally distributed to see how they perform.

## 13. Conclusion

We can conclude that blood pressure (at least with this dataset) is well described by the normal distribution. The analysis indicates that the the mean shifts a little bit with age. This corresponds well to the initial plots of the data, and also with real-world knowledge: older people tend to have a higher blood pressure due to e.g. heart disease, while younger people have healthier circulatory systems.

## 14. Self-reflection

While making this project, our group learned that we can get accurate estimates of parameters using Bayesian models. It is an efficient compliment to classical statistical inference methods. Our dataset was in the end quite easy to estimate and our models worked well. We also got a lot of training in smart ways to visualize data an producing coherent reports.

## References


Gurven, Michael, Aaron D. Blackwell, Daniel Eid Rodríguez, Jonathan Stieglitz, and Hillard Kaplan. 2012. "Does Blood Pressure Inevitably Rise with Age?" *Hypertension* 60 (1): 25–33. https://doi.org/10.1161/hypertensionaha.111.189100.

Mayo Clinic Staff. 2021. "Blood Pressure Chart: What Your Reading Means." 2021. https://www.mayoclinic.org/diseases-conditions/high-blood-pressure/in-depth/blood-pressure/art-20050982.