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.

Main modeling idea is to test with different Bayesian models how to get accurate estimates of the parameters that could describe blood pressure.

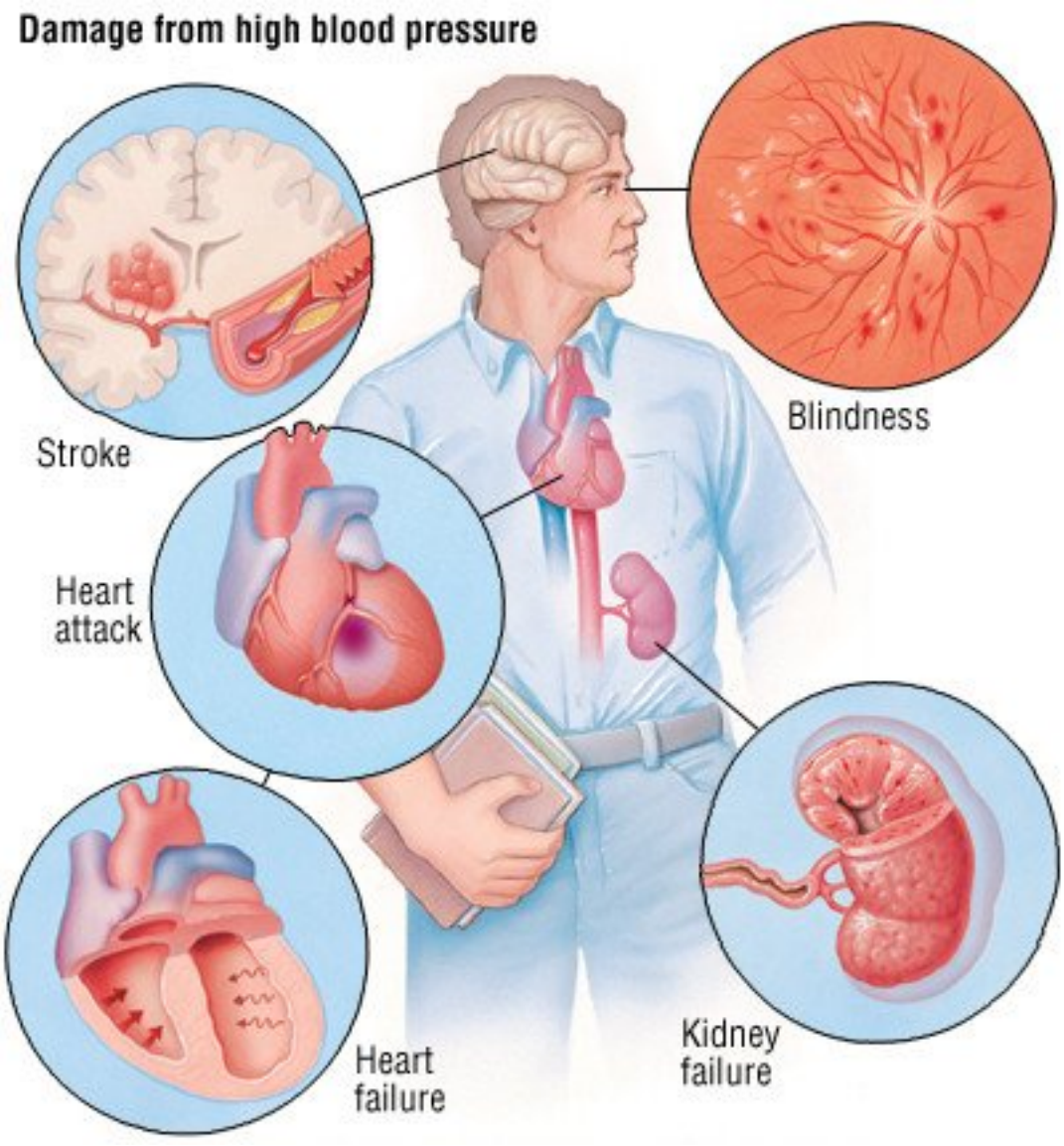
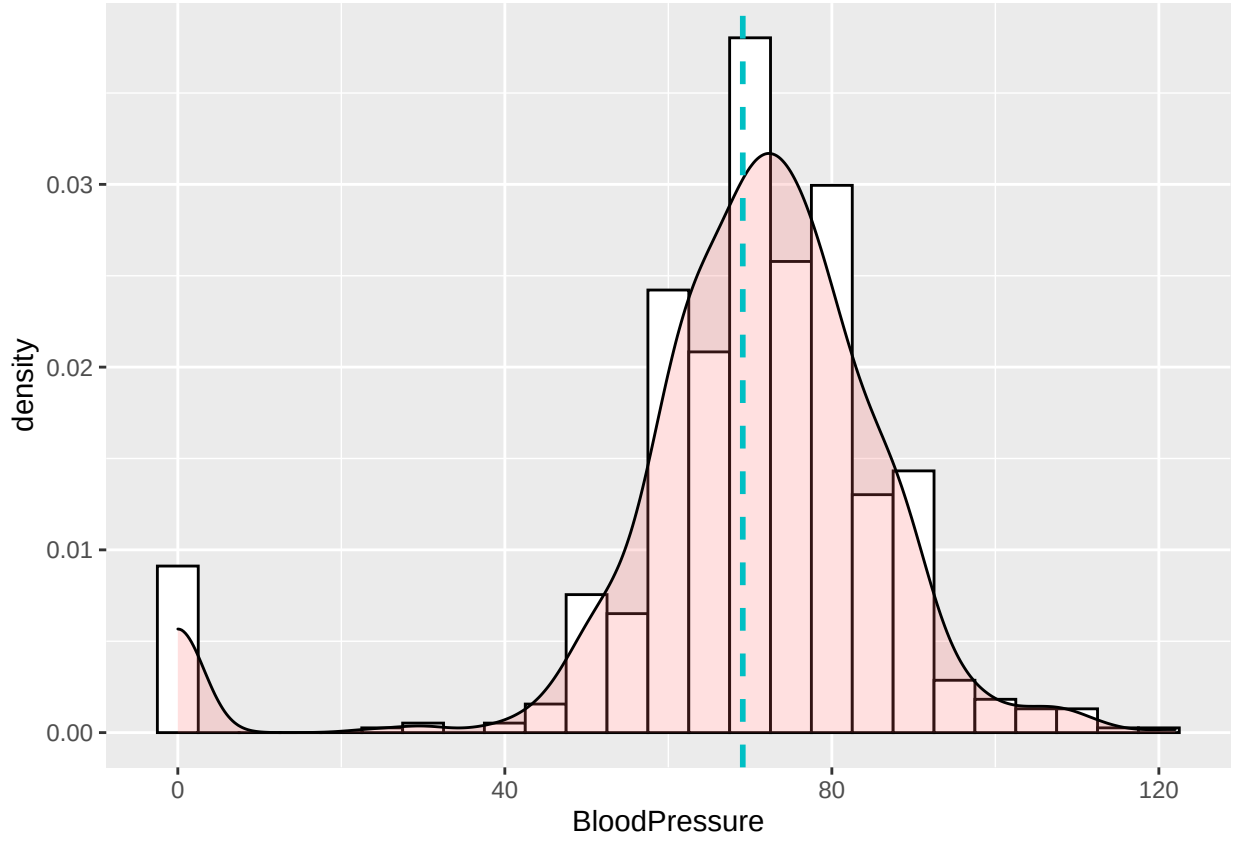


Figure 1: Damage that can be caused from high blood pressure

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

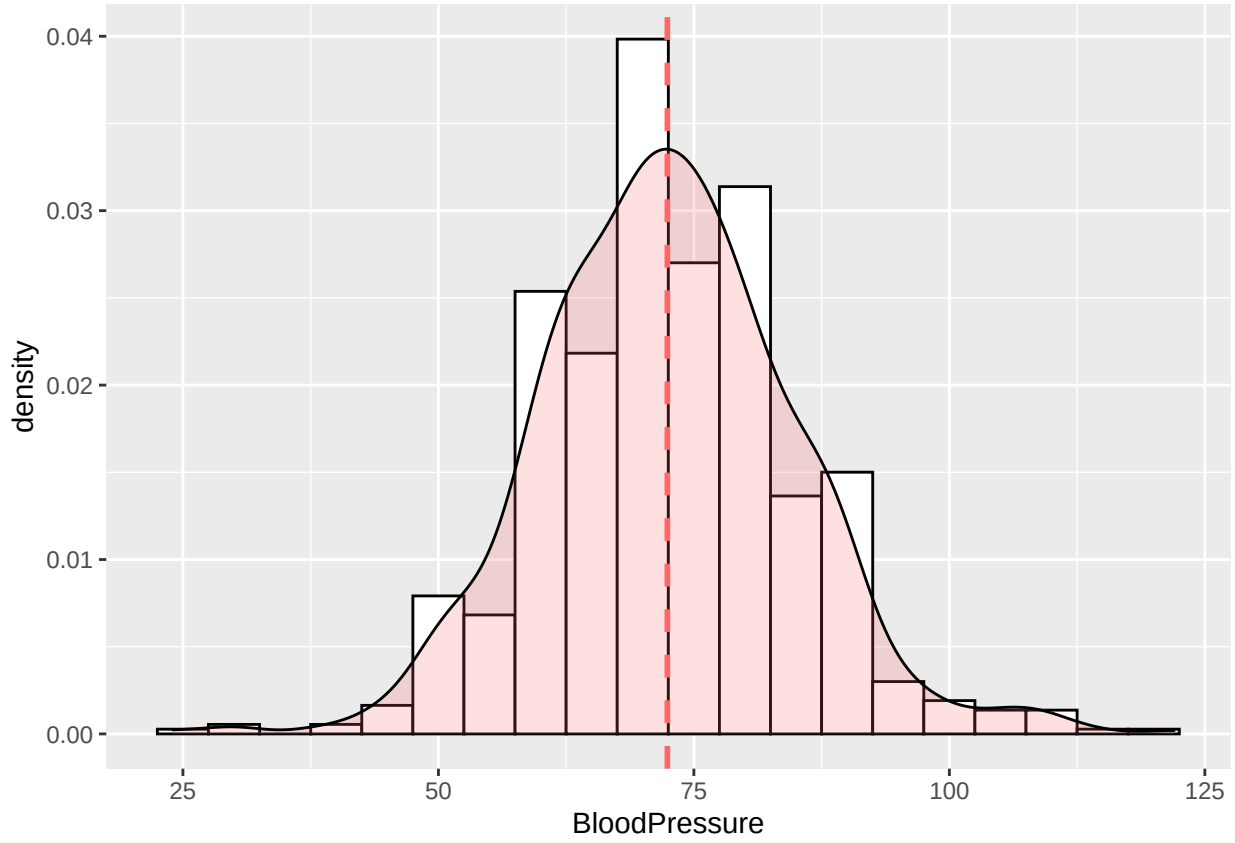


Figure 2: A histogram of the complete data set for the blood pressure. A density approximation plot and a mean line is included as well.

```
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)
```

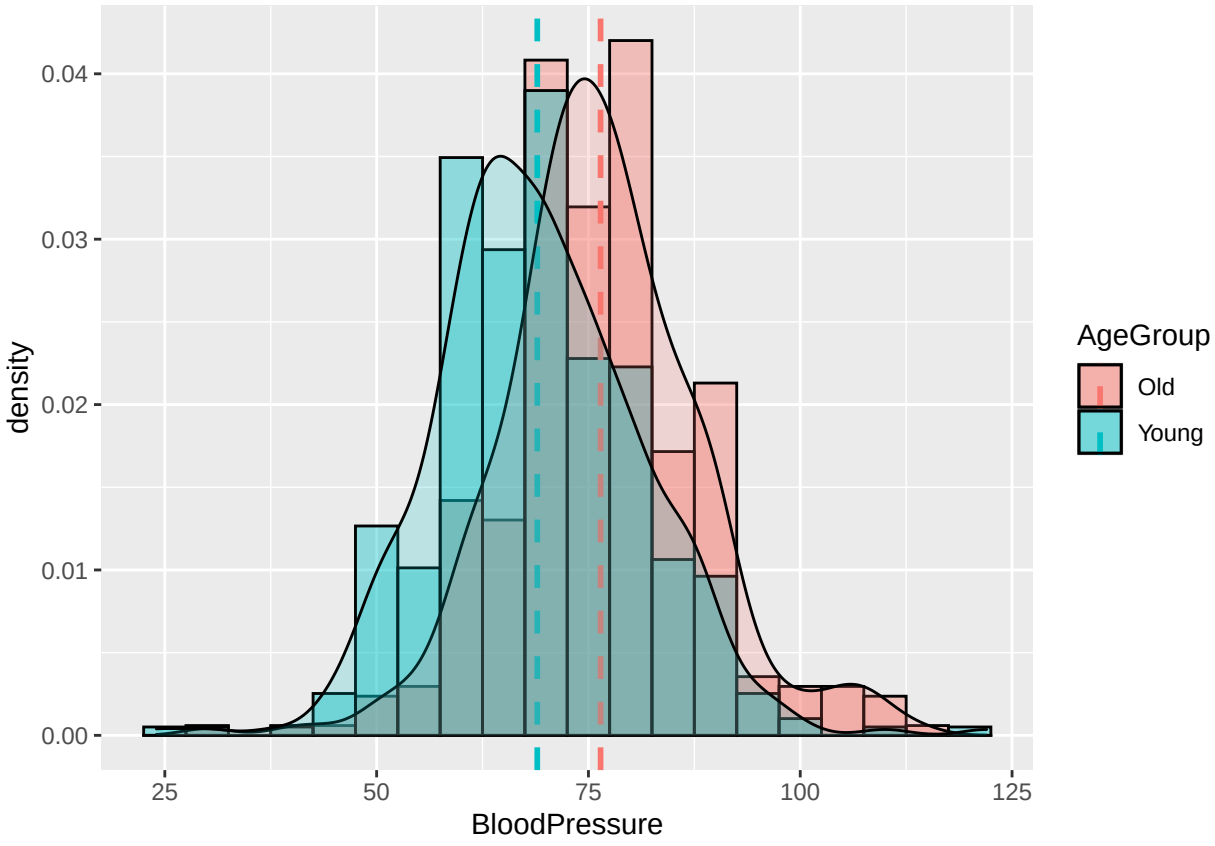


Figure 3: Histograms with density approximations and mean lines for both age groups.

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:

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

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

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

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

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

4. Priors

We use weekly informative priors for parameters μ, σ 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. μ, σ 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 μ 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 μ is higher, because it can be seen from the plot that the mean is higher, but other priors stay the same.

5. Stan code

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;
  real<lower=0> sigma;
}

model {
  //prior
  mu ~ normal(mean_mu_prior, mean_sigma_prior);
  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;                      //
```

```
  real mean_mu_prior;              //
  real<lower=0> mean_sigma_prior;  //
  real<lower=0> var_prior;         //
}

parameters {
  real mu;
  real<lower=0> sigma;
  real mu_hypo;
  real<lower=0> tau;
}

model {
  //hyperpriors
  mu_hypo ~ normal(mean_mu_prior, mean_sigma_prior);
  tau ~ inv_chi_square(var_prior);
  //prior
  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

Nonhierarchical model

Old group

```
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
)
```

```
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

```
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
)
```

```
fit_nonhiera_young = sampling(nonhieramodel,
  data = data_nonhiera_young,            # 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
  )
```

Hierarchical model

Old group

```
mean_mu_prior = 70
mean_sigma_prior = 10
var_prior = 20
data_hiera_old <- list(
  y = data_old$BloodPressure,
  N = length(data_old$BloodPressure),
  mean_mu_prior = mean_mu_prior,
  mean_sigma_prior = mean_sigma_prior_old,
  var_prior = var_prior
)
```

```
fit_hiera_old = sampling(hieramodel,
  data = data_hiera_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

```
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
)
```

```
fit_hiera_young = sampling(hieramodel,
  data = data_hiera_young,            # 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
  )
```

7. Convergence diagnostics

```
knitr::kable(head(monitor(fit_nonhiera_old, print = FALSE),3),
             caption = "The diagnostics of the nonhierarcical model for the old age group.")
```

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

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat	valid	Q5	Q50	Q95	MCSE_Q5	MCSE_Q25	MCSE_Q50	MCSE_Q75	MCSE_Q95	MCSE_SD	n_eff_Q5	n_eff_Q25	n_eff_Q50	n_eff_Q75	n_eff_Q95	n_eff_SD	ESS
mu	76.4	0.010	0.618	75.2	76.0	76.4	76.8	77.6	3639	1	1	75.4	76.4	77.4	0.026	0.013	0.012	0.018	0.024	0.0073657	2797						
sigma	11.4	0.007	0.433	10.6	11.1	11.4	11.7	12.3	3493	1	1	10.7	11.4	12.2	0.018	0.008	0.007	0.011	0.024	0.0053542	2604						
ypred	76.8	0.185	11.495	54.6	69.1	76.8	84.3	98.8	3831	1	1	57.9	76.8	96.0	0.492	0.278	0.184	0.220	0.508	0.1313875	3890						

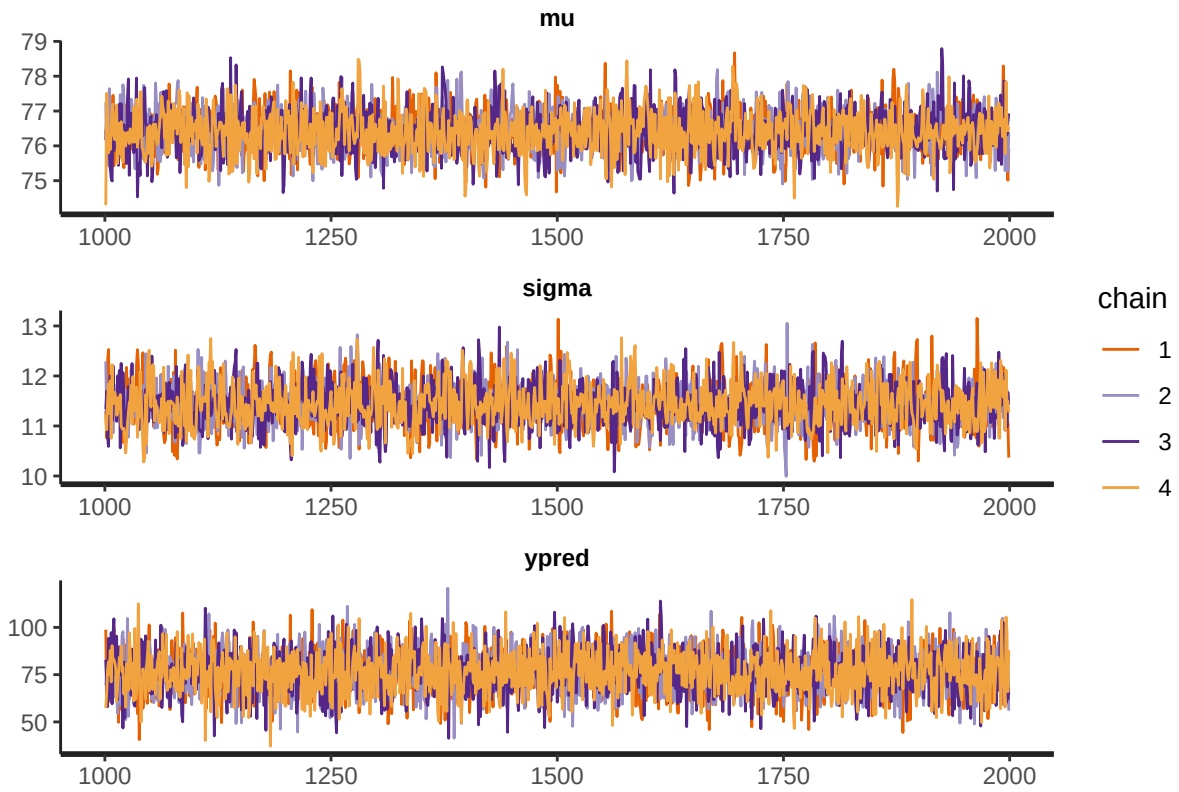


Figure 4: The trace plot of the nonhierarcical model for the old age group.

```
knitr::kable(head(monitor(fit_nonhiera_young, print = FALSE),3),
             caption = "The diagnostics of the nonhierarcical model for the young age group.")
```

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

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat	valid	Q5	Q50	Q95	MCSE	MCSE	MCSE	MCSE	MCSE	MCSE	MCSE	Bulk_ESS	Tail_ESS
mu	69.0	0.010	0.596	67.8	68.6	69.0	69.4	70.1	3701	1	1	68.0	69.0	69.9	0.032	0.012	0.011	0.013	0.039	0.007	3735	2633	
sigma	11.9	0.007	0.416	11.1	11.6	11.9	12.2	12.7	3464	1	1	11.2	11.9	12.6	0.017	0.009	0.009	0.012	0.032	0.005	3510	2492	
ypred	69.4	0.197	11.682	46.6	61.5	69.3	77.3	92.1	3473	1	1	50.3	69.3	88.5	0.523	0.256	0.247	0.284	0.433	0.139	3512	3971	

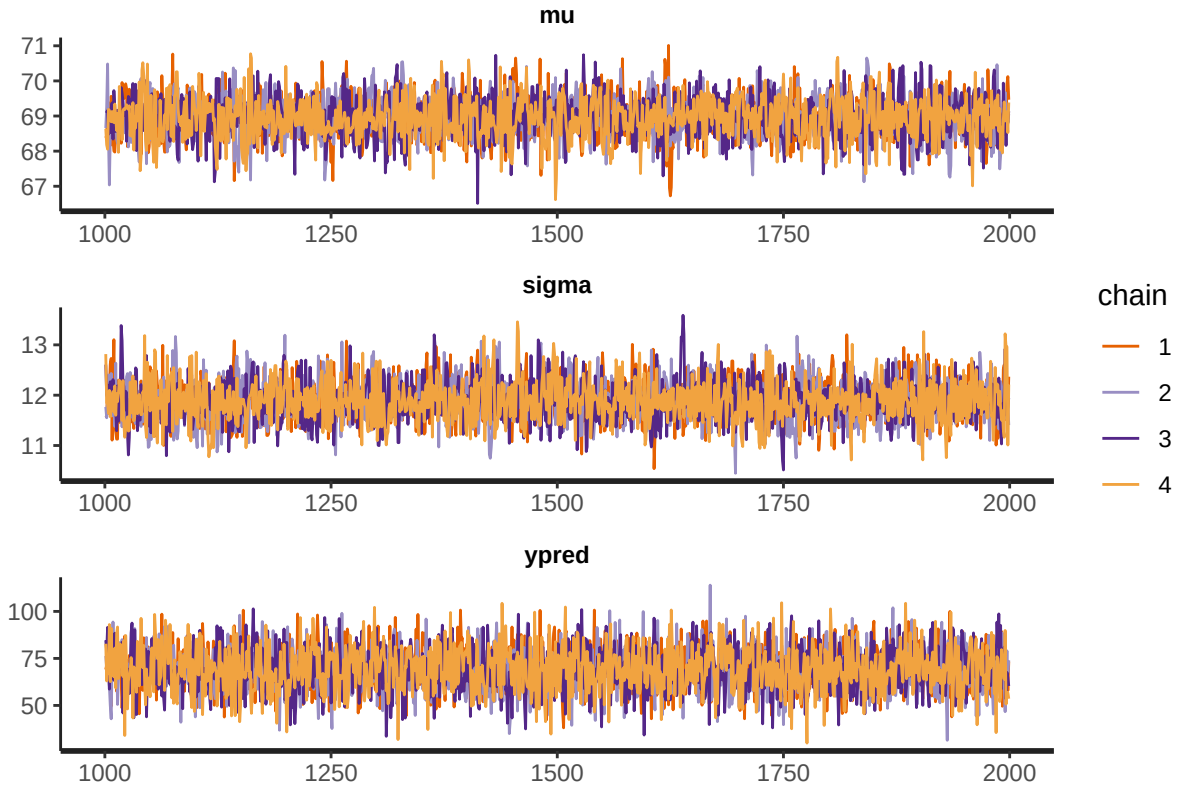


Figure 5: The trace plot of the nonhierarcical model for the young age group.

```
knitr::kable(head(monitor(fit_hiera_old, print = FALSE),3),
             caption = "The diagnostics of the hierarcical model for the old age group.")
```

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

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat	valid	Q5	Q50	Q95	MCSE	MCSE_Q2	MCSE_Q5	MCSE_Q10	MCSE_Q25	MCSE_Q50	MCSE_Q75	MCSE_Q90	MCSE_Q95	MCSE_Q97.5	MCSE_SD	ESS
mu	76.4	0.018	0.634	75.1	76.0	76.4	76.8	77.6	1291	1	1	75.4	76.4	77.4	0.053	0.018	0.020	0.020	0.042	0.0121287	1534					
sigma	11.4	0.008	0.431	10.6	11.1	11.4	11.7	12.3	2780	1	1	10.7	11.4	12.2	0.019	0.010	0.008	0.012	0.023	0.0062829	2024					
mu_hyper	76.4	0.018	0.636	75.1	76.0	76.4	76.8	77.6	1321	1	1	75.4	76.4	77.4	0.049	0.021	0.021	0.016	0.033	0.0121314	1446					

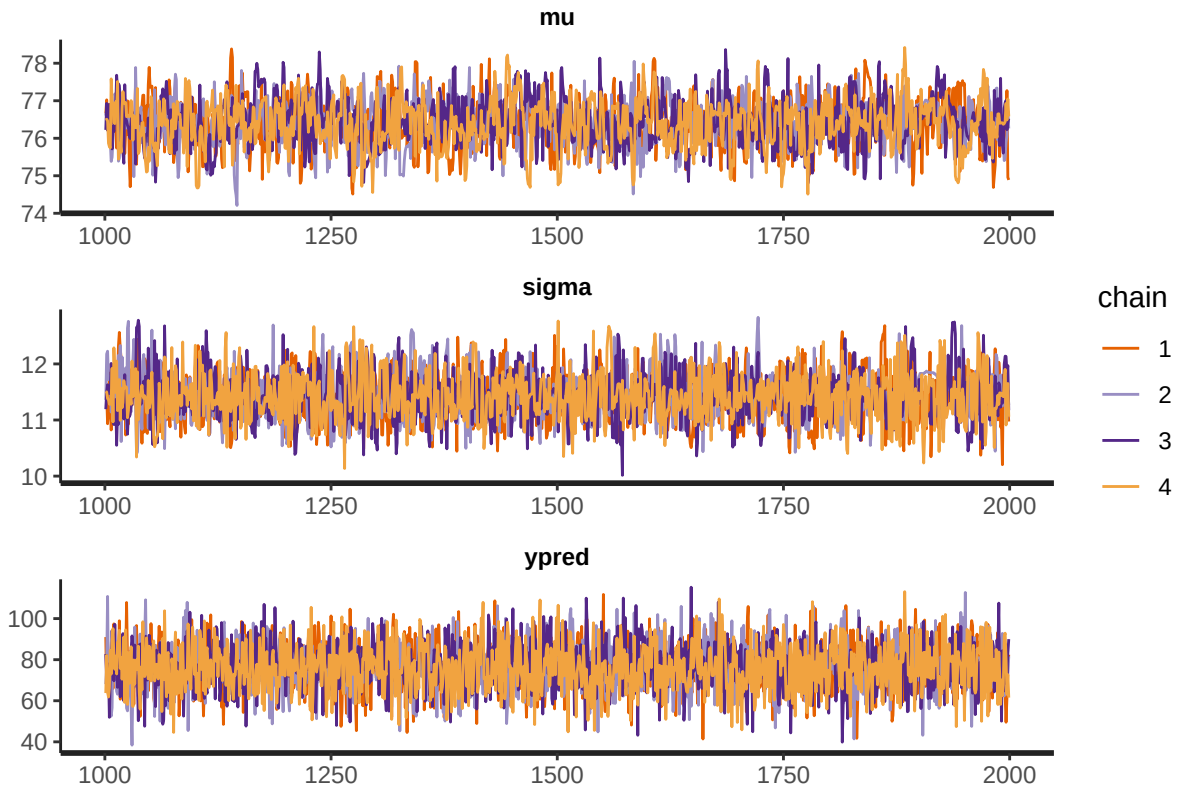


Figure 6: The trace plot of the hierarcical model for the old age group.

```
knitr::kable(head(monitor(fit_hiera_young, print = FALSE),3),
             caption = "The diagnostics of the hierarcical model for the young age group.")
```

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

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat	valid	Q5	Q50	Q95	MCSE	MCSE_Q2	MCSE_Q5	MCSE_Q10	MCSE_Q25	MCSE_Q50	MCSE_Q75	MCSE_Q90	MCSE_Q95	MCSE_Q97.5	MCSE_SD	ESS
mu	69.0	0.017	0.597	67.7	68.6	69.0	69.4	70.1	1222	1	1	67.9	69.0	69.9	0.093	0.025	0.018	0.018	0.036	0.0121242	461					
sigma	11.9	0.009	0.413	11.1	11.6	11.9	12.2	12.7	1958	1	1	11.2	11.9	12.6	0.023	0.011	0.010	0.014	0.019	0.0072001	2143					
mu_hyper	69.0	0.017	0.599	67.7	68.6	68.9	69.3	70.1	1234	1	1	67.9	68.9	69.9	0.087	0.022	0.018	0.016	0.035	0.0121267	451					

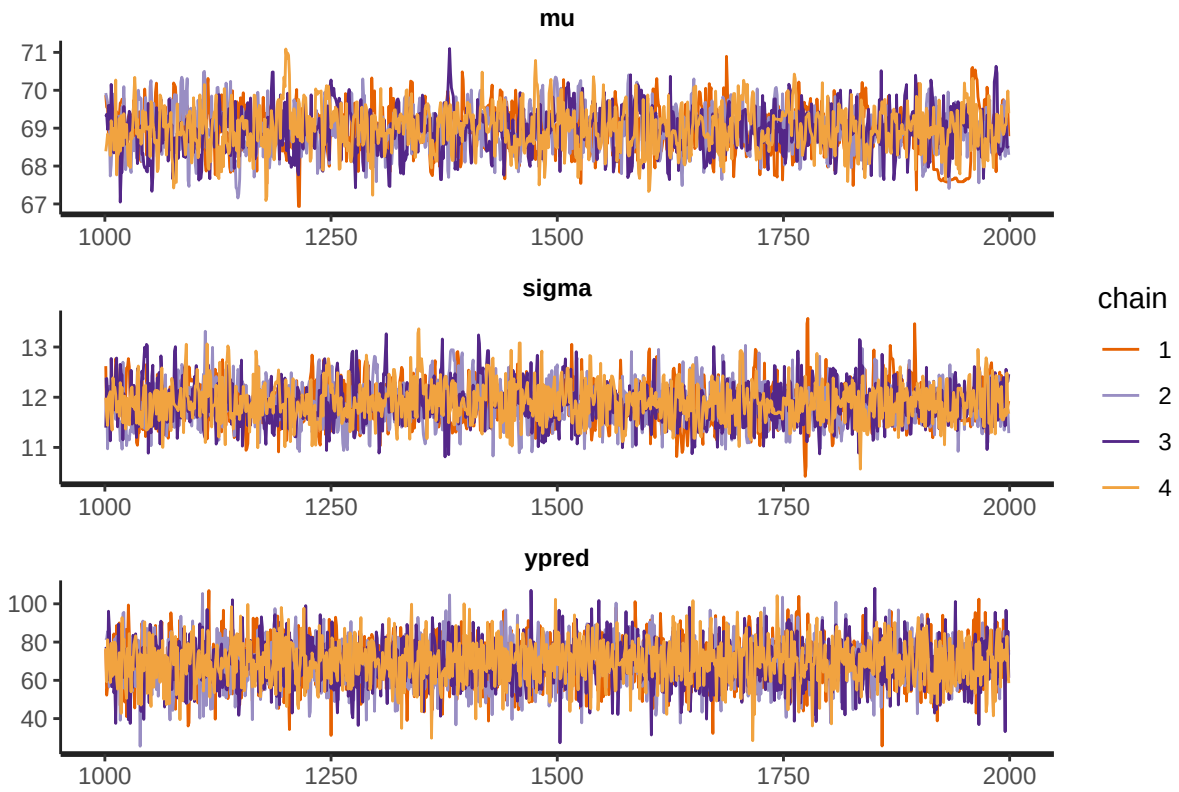
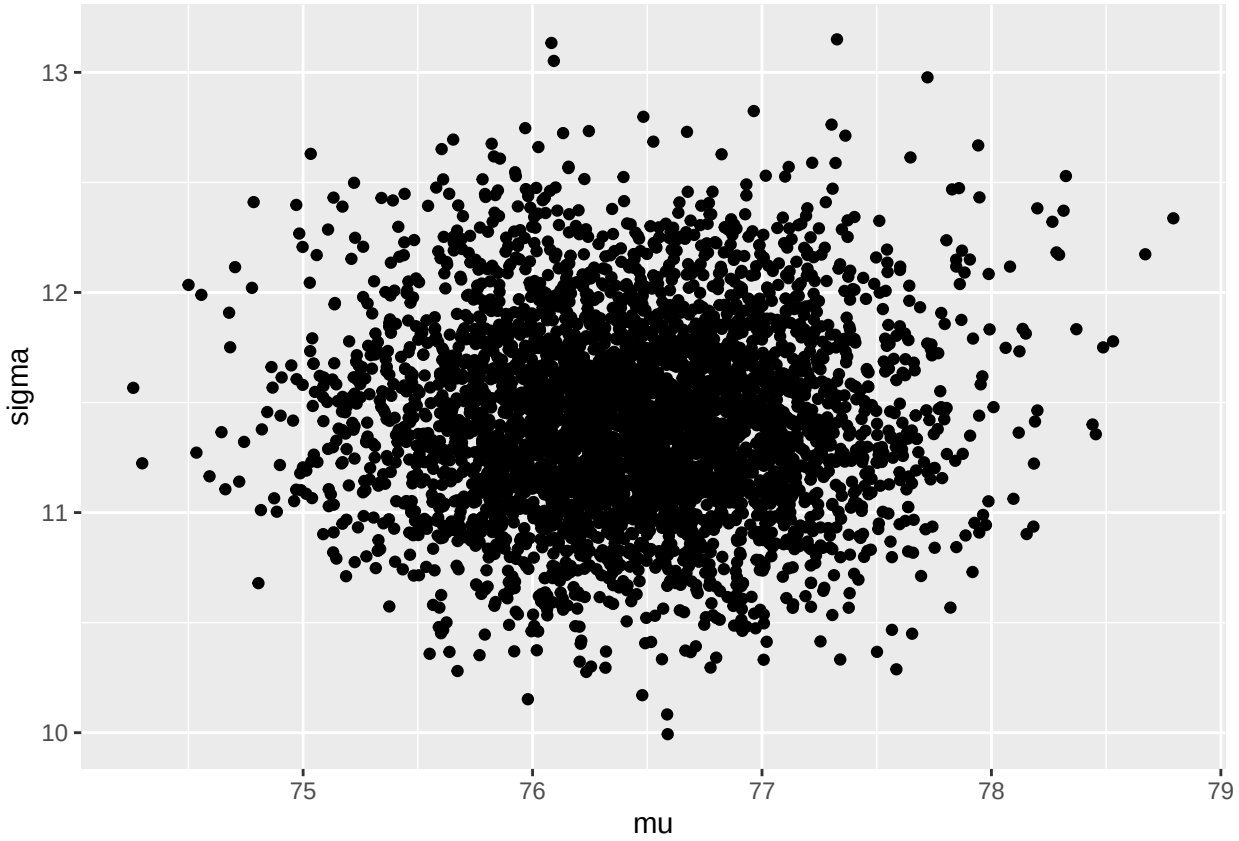


Figure 7: The trace plot of the hierarcical model for the young age group.



8. Posterior predictive checks

```
extract_hiera_old <- data.frame(extract(fit_hiera_old))
extract_nonhiera_old <- data.frame(extract(fit_nonhiera_old))
extract_hiera_young <- data.frame(extract(fit_hiera_young))
extract_nonhiera_young <- data.frame(extract(fit_nonhiera_young))
```

9. Model comparison with LOO-CV

```
loo_nonhiera_old <- loo(fit_nonhiera_old, pars="log_lik")
loo_nonhiera_old
```

```
##
## Computed from 4000 by 338 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1309.0 17.4
## p_loo         2.7  0.5
## looic      2618.1 34.8
```

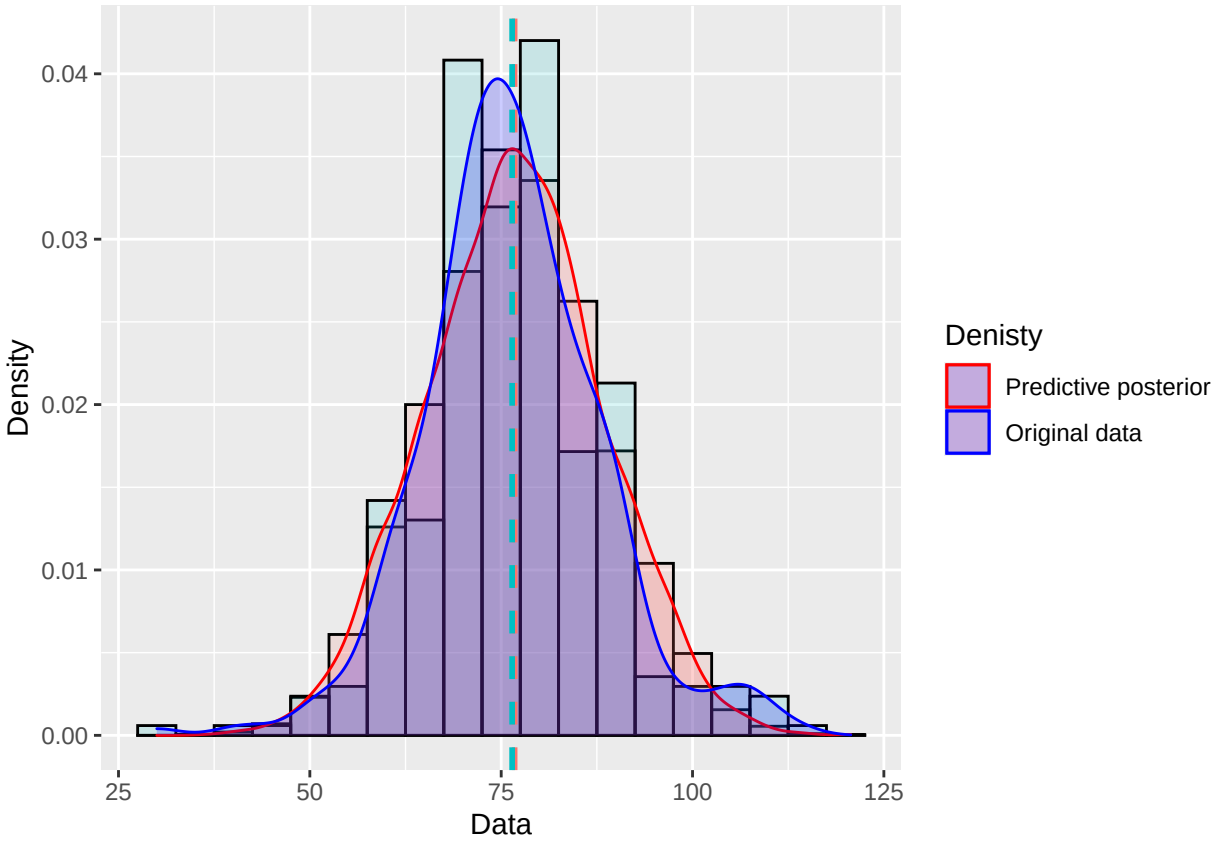


Figure 8: Hierarical posterior predictive vs original data for the old age group.

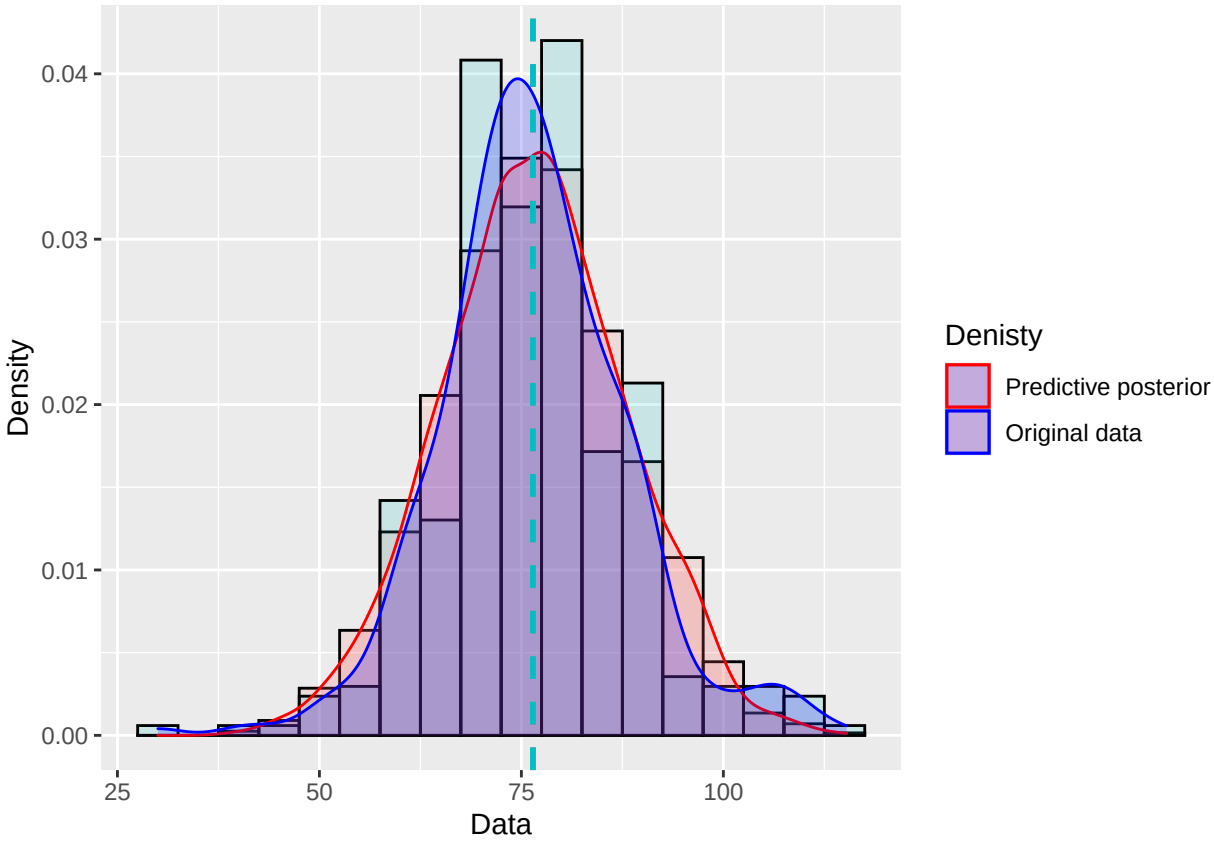


Figure 9: Hierarical posterior predictive vs original data for the old age group.

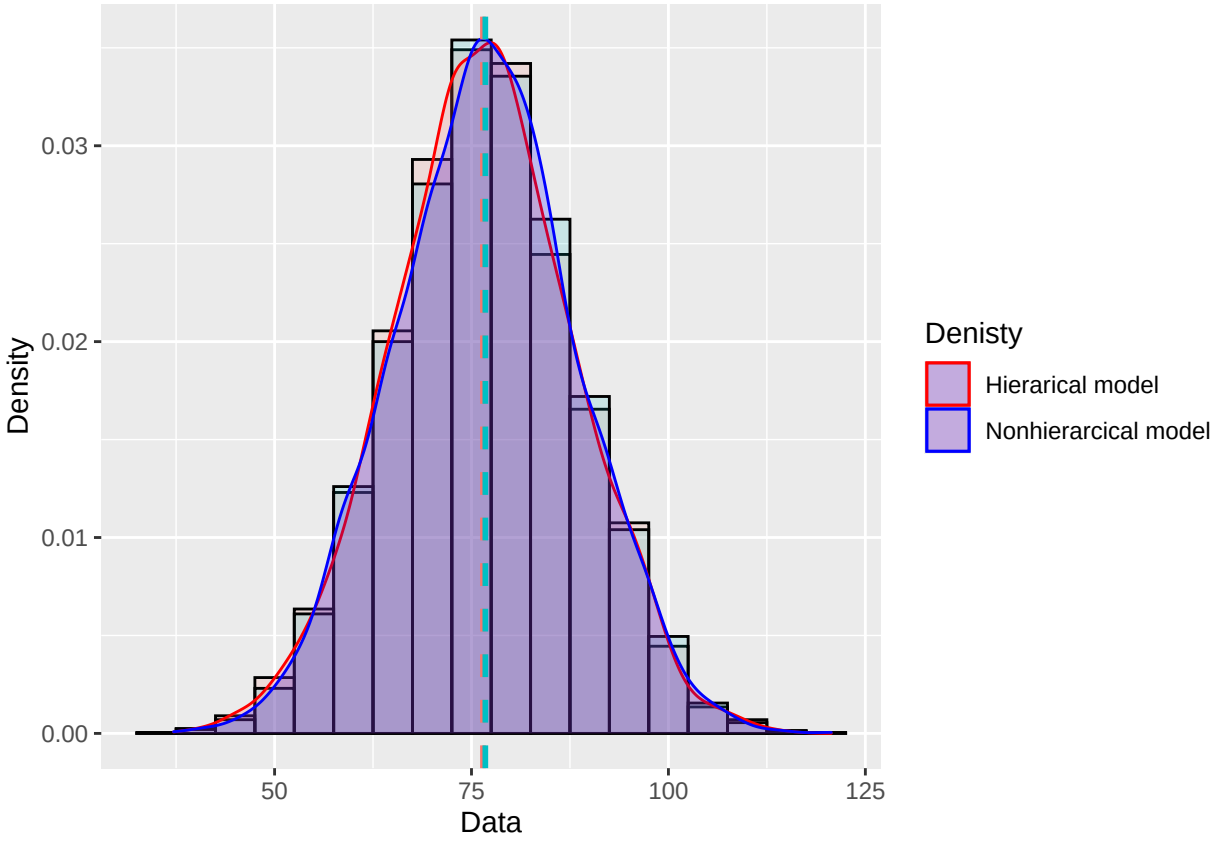


Figure 10: Posterior predictive distributions of the hierarcical versus non-hierarcical data for the old age group.

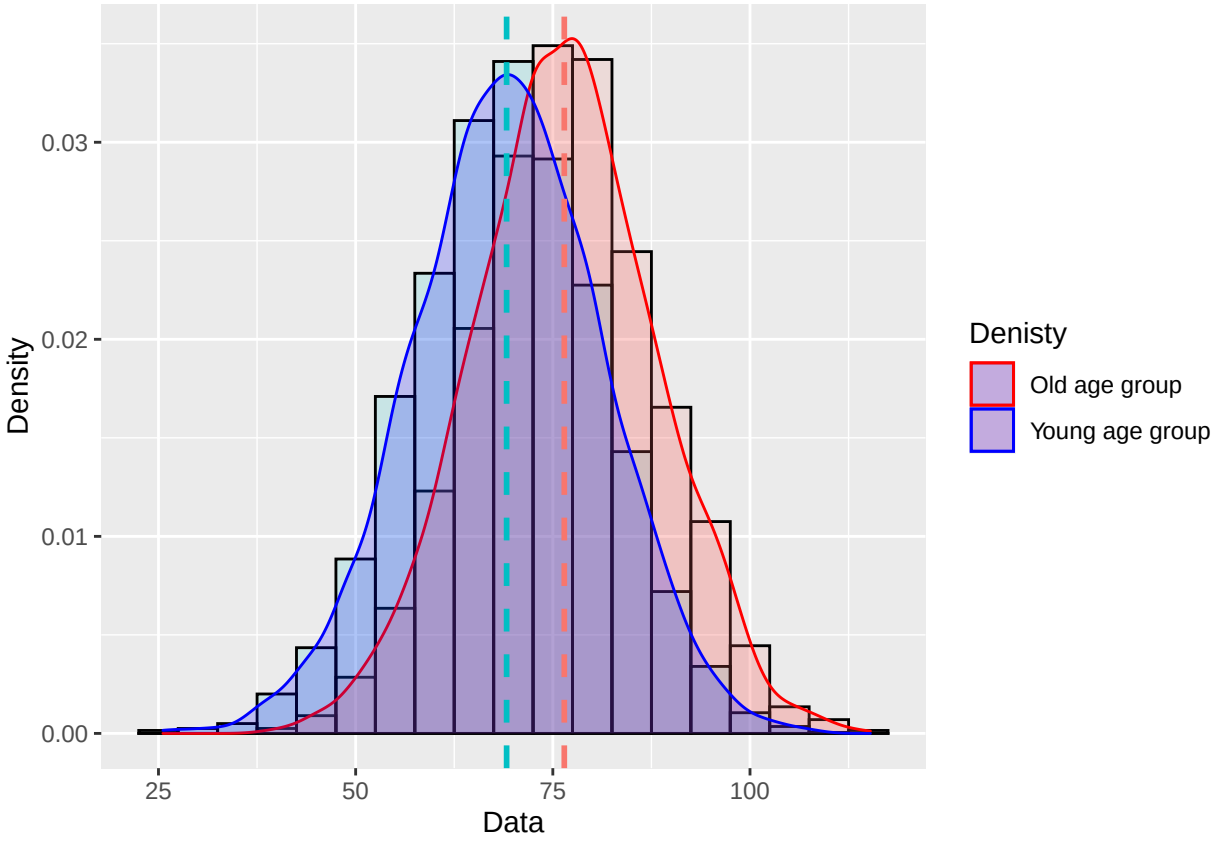


Figure 11: Posterior predictive distributions of the non hierarcical models for the old versus the young age group

```
## ------
## 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.
```

```
loo_nonhiera_young <- loo(fit_nonhiera_young, pars="log_lik")
loo_nonhiera_young
```

```
##
## Computed from 4000 by 395 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1544.6 18.2
## p_loo         2.6  0.6
## looic      3089.3 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.
```

```
loo_hiera_old <- loo(fit_nonhiera_old, pars="log_lik")
loo_hiera_old
```

```
##
## Computed from 4000 by 338 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1309.0 17.4
## p_loo         2.7  0.5
## looic      2618.1 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.
```

```
loo_hiera_young <- loo(fit_nonhiera_young, pars="log_lik")
loo_hiera_young
```

```
##
## Computed from 4000 by 395 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo  -1544.6 18.2
## p_loo         2.6  0.6
## looic      3089.3 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

```
mean_mu_prior_sensitivity = c(0, 50, 100, 1000)
mean_sigma_prior_old_sensitivity = c(1, 10, 100, 1000)
var_prior_old_sensitivity = c(1, 10, 100, 1000)
fit_sensitivity = c()
for (i in 1:length(mean_mu_prior_sensitivity)){
  for (j in 1:length(mean_sigma_prior_old_sensitivity)){
  data_sensitivity <- list(
    y = data_old$BloodPressure,
    N = length(data_old$BloodPressure),
    mean_mu_prior = mean_mu_prior_sensitivity[i],
    mean_sigma_prior = mean_sigma_prior_old_sensitivity[j],
    var_prior = var_prior_old_sensitivity[j]
  )

  fit_sensitivity = c(fit_sensitivity,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
    ))
  }
}
```

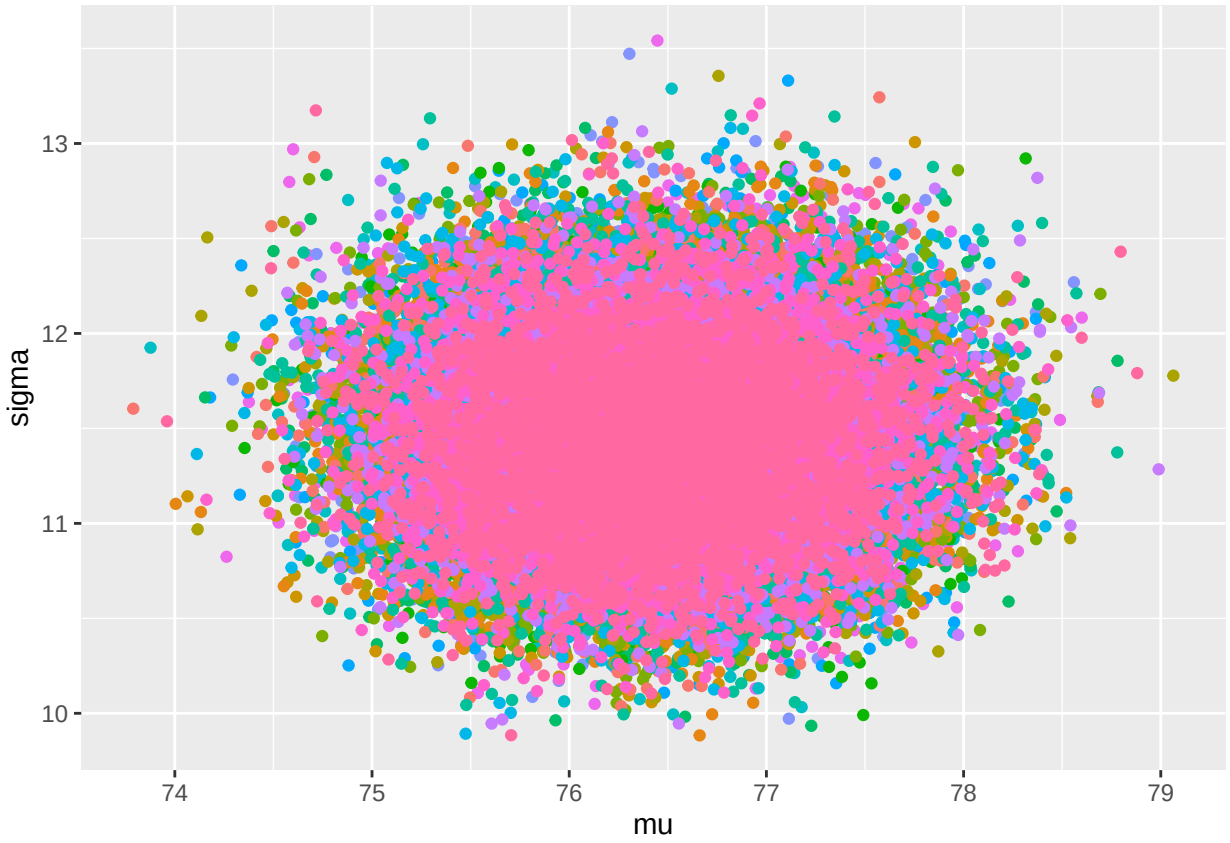


Figure 12: The complete chains, including warmup, of the sixteen combinations of prior values ploted.

12. Discussion

13. Conclusion

14. Self-reflection

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.