Bayesian Data Analysis Project

Distribution of children's heights in a family

Daniel Wohlrath

Winter 2021/2022

Contents

1 Introduction

1.1 Motivation

Various sport disciplines, overall appearance and even the ability of finding a life-long partner, in each of these situations one's height plays a role to a larger or smaller extent. If we could somehow predict heights of people before they become adults, we could prevent *problems* such as a sports-gifted children giving up on playing basketball, since they had not grown up sufficiently. Being able to predict that, the gifted child could have shifted his focus to another sport, where the height factor does not play a crucial part.

A natural approach to aforementioned predictions would be a linear or nonlinear regression based on child's parents' heights. Indeed, the scatter plots[1] shown in figure 1 clearly suggest a positive correlation between the height of a parent and height of his/her child, which is also ultimately asserted by using *corr* function in R library *corrgram*, see figure 2. As regression has been already extensively studied on this dataset, e.g. in [2], in this work we will focus more on the distribution of heights of children in a family.

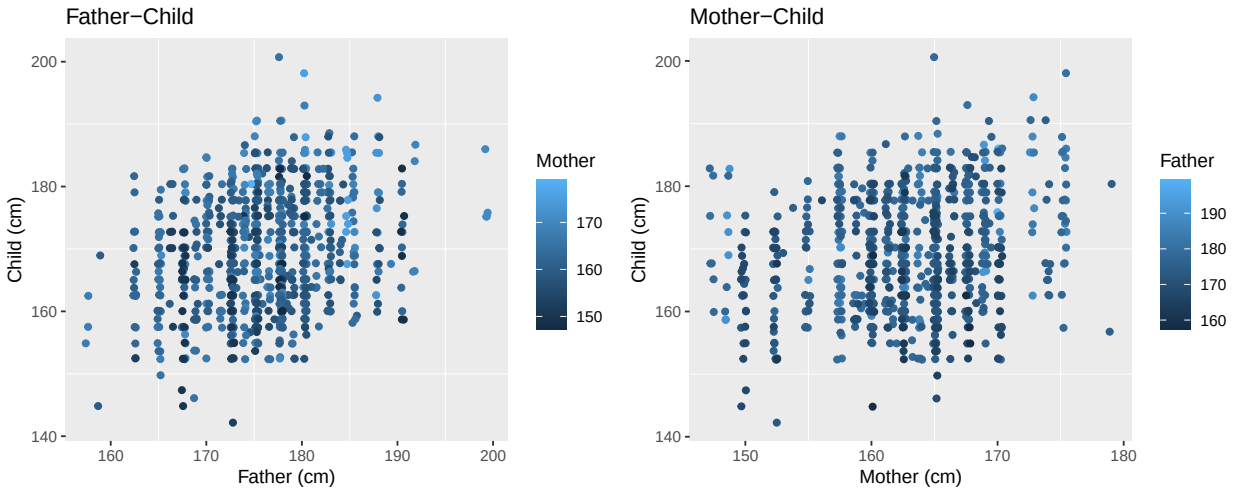

Figure 1: Clear positive correlation in height of parents and their child.

1.2 Problem

It has been shown in previous studies, see [2], [4], that height of a child can be predicted with heights of the child's parents by using a linear fit. However, the connection between heights of the siblings is not well understood. It is not clear whether there is a dependency on being a firstborn, a second born, or more generally a middle child. Other key aspect of different heights between siblings could be nutrition during the infant age and possibly also parents' age. Also note, that even identical twins do not have to be the same height when they grow up.

Clearly, the height of both parents have an effect of the child's height, mathematically

$$y = f(x_1, x_2, \dots)\,,$$

where x_1 and x_2 are father's and mother's heights, respectively, y is height of their child after it grows up and f is some unknown function. As it is clear by now, we cannot determine all the variables of such function f. Hence we must resort to some stochastic model. The main focus will be on the mean μ and standard deviation σ of the distribution of heights belonging to a family, i.e. a pair consisting of a father and a mother represented by their heights.

$$y \sim P(\Theta(\mathbf{x}))\,, \tag{1}$$

[1] created from Galton's Height Experiment

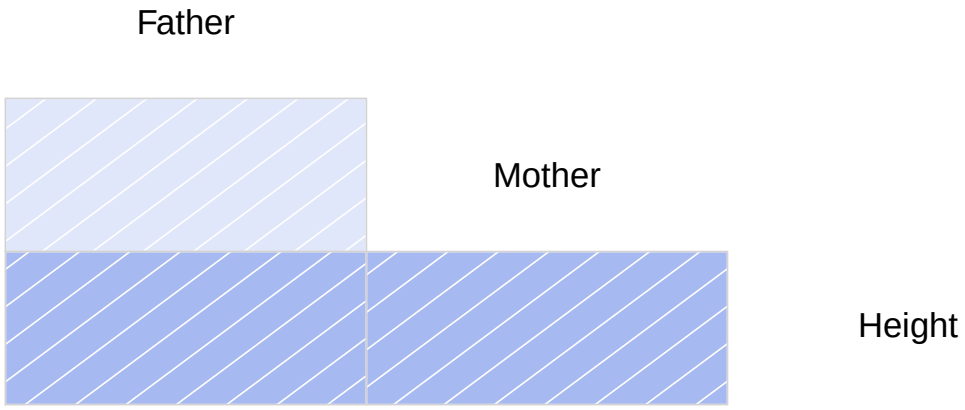

Figure 2: Proof of positive correlation in height of a parent and a child using *corr* function in R library *corrgram* .

where $\mathbf{x}$ denotes available data, Θ is a transformation of such data and P is an unknown probability distribution. The problem this project is attempting to study is illustrated in figure 3.

1.3 Modelling idea

For that purpose we follow the Bayesian approach. We use the Bayes' rule

$$p(\theta \,|\, y) \propto p(\theta) p(y \,|\, \theta)\,, \tag{2}$$

to obtain posterior distribution $p(\theta \,|\, y)$ of the parameter $\theta = (\mu,\, \sigma)$ from the likelihood $p(y \,|\, \theta)$ and assumed prior $p(\theta)$. Specifically, for generated θ_j from the prior of jth family, we compute the likelihood $p(y_{ij} \,|\, \theta_j)$ of the ith child of jth family to be born into that family. In this paper, we will consider normal distribution to be the candidate of P from the equation (1):

$$y_{ij} \sim \mathrm{N}(\mu_j, \sigma_j)\,, \tag{3}$$

with specific means μ_j and standard deviations σ_j.

The assumption of normal distribution is chosen since the behavior of the distribution is not yet well understood. The impression so far is that the distribution of heights of children is caused by "random noise" which is usually well described by normal distributions.

2 Data

2.1 Description of data

This project explores the distribution of heights of children in 197 different families that participated in famous Galton's height experiment in 1885. The dataset is available at [1] and for Galton's studies see [3].

The head of the original dataset is shown in figure 4. The columns of the dataset are:

- Family: the family that the child belongs to, labeled from 1 to 205,

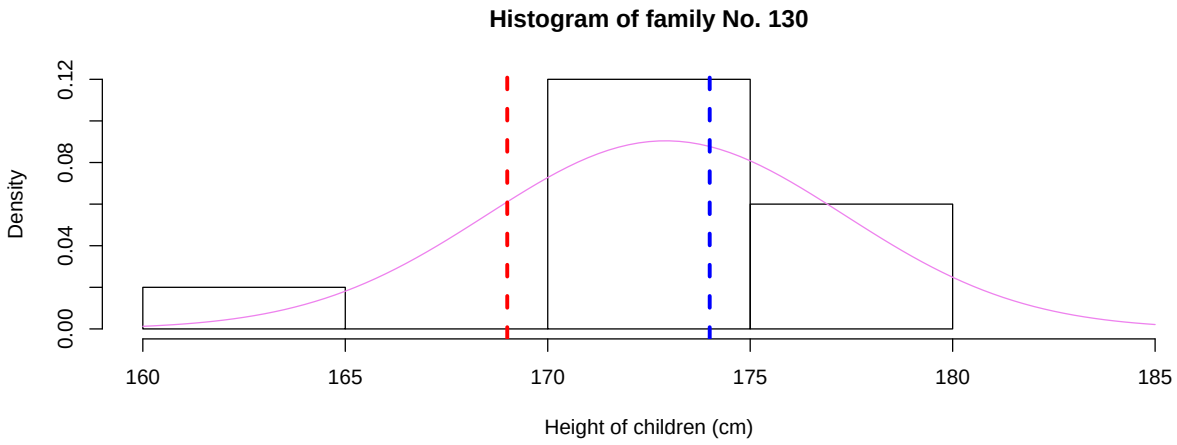

Figure 3: Fitted normal distribution to family No.130, which has 11 children. Blue and red vertical dotted lines represent height of the father and mother, respectively.

	Father (cm)	Mother (cm)	Child (cm)
Min.	162.6	148.6	142.2
1st Qu.	172.7	160.0	162.6
Median	175.3	163.8	167.6
Mean	175.3	163.2	168.7
3rd Qu.	178.6	166.4	175.3
Max.	190.5	175.3	200.7

Table 1: Summary statistics of the chosen subset of original dataset

- Father: the father's height, in inches,
- Mother: The mothers's height, in inches,
- Gender: the gender of the child, male (M) or female (F),
- Height: the height of the child, in inches,
- Kids: total number of kids in the corresponding family.

Not all of these listed features will concern us. First, to put the measurements into a more convenient format[2], we scale the heights appropriately so we obtain measurements in centimeters.

Second, we will not take into account the gender of the child. Even though men are taller then women on average, this is done purely for simplicity and intelligibility of our study.

Lastly, it is obvious that it is very easy to fit a normal distribution to a dataset of size 1 or 2. In the special case of sample size being equal to one, the fitted normal distribution is the degenerated Dirac's distribution. Hence, we will discard all data corresponding to families with less than 6 children. After that, we are left with data of 66 families with more than 5 children, aggregating to a total of 504 observations[3]. The statistics of our chosen subset of original dataset are summarized in table 1.

[2]the use of the word "convenient" is subjective to one's nationality

[3]some families have more than 6 children

Family	Father	Mother	Gender	Height	Kids
1	78.5	67.0	M	73.2	4
1	78.5	67.0	F	69.2	4
1	78.5	67.0	F	69.0	4
1	78.5	67.0	F	69.0	4
2	75.5	66.5	M	73.5	4
2	75.5	66.5	M	72.5	4
2	75.5	66.5	F	65.5	4
2	75.5	66.5	F	65.5	4
3	75.0	64.0	M	71.0	2
3	75.0	64.0	F	68.0	2
4	75.0	64.0	M	70.5	5
4	75.0	64.0	M	68.5	5

Figure 4: First rows of the original Galton's dataset

2.2 Using the data

Before we proceeded to statistical inference, the data was scraped in the following way. For more precise information see listing 1.

1. We permute the rows of the dataframe so any possible structures[4] are deleted
2. We scaled the original data by 2.54 in order to convert measurements in inches to measurements in centimeters.
3. We selected families with 6 children or more for meaningfulness of our study.
4. We created an array *dataset* which stores for each family first six height measurements[5] of children belonging to that family.
5. We created vectors *fathers* and *mothers* in which we store the corresponding heights of parents.

```
#Loading data
height <- read_csv("height.csv")

#Permuting rows so there is no hidden structure
height <- height[sample(nrow(height)),]

#Units conversion
height$Father = height$Father*2.54
height$Mother = height$Mother*2.54
height$Height = height$Height*2.54

#Lower limit of kids for family to be studied
numkids = 6

#filter for families with more kids than numkids-1
data = height[height$Kids>=numkids,]

#getting number of families in my data
numfams = length(unique(data$Family))
```

[4] e.g. descending order of heights
[5] these measurements are actually random after we perform permuting of rows

```
#creating dataset which will be put into STAN
#Each column dataset[,j] represents measurements of jth family
#Each row dataset[i,] represents ith measurements in all families
dataset<- matrix(nrow=numkids,ncol=numfams )
j=1
for (k in unique(data$Family)){
  for(i in 1:numkids){
    dataset[i,j] = data$Height[data$Family==k][i]
  }
  j=j+1
}

#Getting height of fathers and mothers in families
fathers <- c()
mothers <- c()
m = 1
for (j in unique(data$Family)){
  fathers[m] = unique(data$Father[data$Family==j])
  mothers[m] = unique(data$Mother[data$Family==j])
  m=m+1
}
```

Listing 1: Preparation of data

The output dataset is prepared to be converted into the data frame class and be used as an input for Stan model.

Although we are discarding some information for families with more than 6 children, it is very desirable to use the *loo* function in rstan package for model comparison. Unfortunately, loo function is not yet available for ragged arrays, see more in section 9.

3 Models

As mentioned earlier in section 1, it is likely that height of parents has an influence on the height of their child. Although it is not precisely known what kind of influence, we can propose different models based on what specific influence is present..

3.1 Pooled model

The pooled model is simple and yields fast simulation draws, since the number of parameters is not extremely high. The possible weakness of this model is however that it treats every family as the same, with identical prior mean and standard deviation distribution. Therefore it generally cannot fit large datasets accurately.

A mathematical description of the implemented pooled model is

$$\begin{aligned} y_{ij} &\sim \mathrm{N}(\mu, \sigma)\,, \\ \mu &\sim \mathrm{N}(170, 10)\,, \\ \sigma &\sim \mathrm{Gamma}(7.5, 1)\,, \end{aligned} \tag{4}$$

where y_{ij} are the height measurements and μ and σ are the common mean and standard deviation of the normal distribution, respectively. Index j denotes the index of the family and i the index of the child in that family, i.e. $i < 7$, since we chose only families with equal to or more than 6 children. Note that we are using the following notation for Gamma distribution: first argument of $\mathrm{Gamma}(\cdot, \cdot)$ is the shape parameter $\alpha > 0$ and the second argument is the rate parameter $\beta > 0$.

3.2 Hierarchical model

In hierarchical models we use a prior distribution in which the $\theta_j = (\mu_j, \sigma_j)$'s are viewed as a sample from a common population distribution. Such application enables us to estimate aspects of the population distribution of the θ_j's even though the values of θ_j are not themselves observed.

A mathematical description for our hiearchical model with common variance can be given as

$$\begin{aligned} y_{ij} &\sim \mathrm{N}(\mu_j,\, \sigma)\,, \\ \mu_j &\sim \mathrm{N}(\mu_0, \sigma_0)\,, \\ \sigma &\sim \mathrm{Gamma}(7.5, 1)\,, \\ \mu_0 &\sim \mathrm{N}(170, 10) \\ \sigma_0 &\sim \mathrm{Gamma}(7.5, 1)\,, \end{aligned} \tag{5}$$

where in addition to model (4), we have hyperparameters μ_0, σ_0 and their corresponding hyperpriors.

3.3 Hierarchical model with parental prior mean

We also propose a hierarchical model with prior means dependent on the heights of fathers $(f_j)_{j\in J}$ and mothers $(m_j)_{j\in J}$, where J denotes the set of all families.

A mathematical description of this hierarchical model with common variance can be given as

$$\begin{aligned} y_{ij} &\sim \mathrm{N}(\mu_j,\, \sigma)\,, \\ \mu_j &\sim \mathrm{N}(0.5 f_j + 0.5 m_j, \sigma_0)\,, \\ \sigma &\sim \mathrm{Gamma}(7.5, 1)\,, \\ \sigma_0 &\sim \mathrm{Gamma}(7.5, 1)\,. \end{aligned} \tag{6}$$

The difference between models (5) and (6) is that the prior means in (6) are related to heights of the corresponding parents. The motivation for this was discussed briefly in section 1.

4 Priors

In the pooled model (4), a normal distribution was chosen as the prior for the common mean μ and gamma distribution was selected to be the prior for the common standard deviation σ. These are commonly used weakly informative priors in cases where the parameters are the mean and standard deviation of a normal distribution, which justification is stated in subsection 1.3. We set the mean and the standard deviation of the prior for μ as 170 and 10,respectively, since they are roughly the sample mean and sample standard deviation of all heights in the considered dataset. Similarly, we chose the shape and rate parameters for σ prior as $\alpha = 7.5$, $\beta = 1$ in order to have $\sigma > 0$ and its mean value α/β not too high or too low to affect the posterior draws drastically.

In the hierarchical model (5) we use the same justification of priors and hyperpriors as in the pooled model (4).

In the hierarchical model with parental prior mean (6) we take into account the heights of parents. The weight coefficients for father's height and mother's height are the same, $w_f = w_m = 0.5$. That corresponds to the fact that we take into account the affect of parents' heights but also neglect any possible genetic dominance in both genders. Other specific values in this model are justified analogically to the previous two models.

The parameter choices for the prior distributions are explored further in section 11.

5 Stan code

This section provides the code of the Stan models used in this paper. As described in section 3, we studied three different models, one pooled and two hierarchical. The implementation of the pooled model is presented in the subsection 5.1, the first hierarchical version is shown in 5.2 and the second in 5.3.

Variable names in the code chunks shown below are following the convention established so far in this paper. It is worth noting that Stan models also include posterior predictive samples *ypred* and generated quantity *log_lik* which are essential in the analysis discussed in section 8 and model comparison in section 9, respectively.

5.1 Pooled model

The Stan code for the pooled is shown in listing 2. Further information on the model is provided in the comments of the code.

```
# The input data is a vector 'y' of length 'N'.
data {
   int<lower=0> N; #number of children considered
   int<lower=0> J; #number of families
   vector[J] y[N]; #matrix of measurements
}

# The parameters accepted by the model. Our model
# accepts two parameters 'mu' and 'sigma'.
parameters {
  real mu;
  real<lower=0> sigma;
}

# The model to be estimated. We model the output
# 'y' to be normally distributed with common mean 'mu'
# and common standard deviation 'sigma'.
model {
  #PRIORS
  mu ~ normal(170,10);
  sigma ~ gamma(7.5,1);
  #Likelihood
  for (i in 1:N){
    for (j in 1:J){
      y[i,j] ~ normal(mu,sigma);
    }
  }
}

# Generated quantities used for model comparison
# and predictive performance
generated quantities{
    vector[J] ypred;
    vector[J] log_lik[N];
    for (j in 1:J){
      ypred[j] = normal_rng(mu,sigma);
    }
    for (i in 1:N){
      for (j in 1:J){
        log_lik[i,j] = normal_lpdf(y[i,j] | mu, sigma);
      }
    }
}
```

Listing 2: Implementation of pooled model in Stan

5.2 Hierarchical model

The implementation of the hierarchical model in Stan is shown below in listing 3. The comments within the code should be helpful.

```
# The input data is a vector 'y' of length 'N'.
data {
  int<lower=0> N; #number of children considered
  int<lower=0> J; #number of families
  vector[J] y[N]; #matrix of measurements
}

# The parameters accepted by the model. Our model
# accepts 4 parameters 'mu', 'sigma', 'sigma0' and 'mu0'.
parameters {
  vector[J] mu;
```

```
  real<lower=0> sigma;
  real<lower=0> sigma0;
  real mu0;
}

model {
  #PRIORS
  mu0 ~ normal(170,10);# weakly informative prior
  sigma0 ~ gamma(7.5,1);# weakly informative prior
  sigma ~ gamma(7.5,1);# weakly informative prior
  for (j in 1:J){
    mu[j] ~ normal(mu0,sigma0);# population prior with unknown parameters
  }
  #LIKELIHOOD
  for (i in 1:N){
    for (j in 1:J){
      y[i,j] ~ normal(mu[j], sigma);
    }
  }
}
# Generated quantities used for model comparison
# and predictive performance
  generated quantities {
    vector[J] ypred;
    vector[J] log_lik[N];
    for (j in 1:J){
      ypred[j] = normal_rng(mu[j],sigma);
    }
    for (i in 1:N){
      for (j in 1:J){
        log_lik[i,j] = normal_lpdf(y[i,j] | mu[j], sigma);
      }
    }

}
```

Listing 3: Implementation of hierarchical model in Stan

5.3 Hierarchical model with parental prior mean

The implementation of the hierarchical model with parental prior mean in Stan is shown below in listing 4.

```

// The input data is a vector 'y' of length 'N'.
data {
  int<lower=0> N; #number of children considered
  int<lower=0> J; #number of families
  vector[J] y[N]; #matrix of measurements
  vector[J] f; #height of father
  vector[J] m; #height of mother
}

# The parameters accepted by the model. Our model
# accepts 3 parameters 'mu', 'sigma' and 'sigma0'.
parameters {
  vector[J] mu;
  real<lower=0> sigma;
  real<lower=0> sigma0;
}

#PRIORS
  sigma ~ gamma(7.5,1);
  sigma0 ~ gamma(7.5,1);
  for (j in 1:J){
    mu[j] ~ normal(0.5*f[j]+0.5*m[j],sigma0);
  }
  #LIKELIHOOD
```

```
  for (i in 1:N){
    for (j in 1:J){
      y[i,j] ~ normal(mu[j], sigma);
    }
  }
}
# Generated quantities used for model comparison
# and predictive performance
  generated quantities {
    vector[J] ypred;
    vector[J] log_lik[N];
    for (j in 1:J){
      ypred[j] = normal_rng(mu[j],sigma);
    }
    for (i in 1:N){
      for (j in 1:J){
        log_lik[i,j] = normal_lpdf(y[i,j] | mu[j], sigma);
      }
    }

}
```

Listing 4: Implementation of hierarchical model with parental prior mean in Stan

6 Options used for simulation

All models were run using a sampler with 4 chains and warm-up length was 2000. For each chain a total of 4000 samples were drawn so there were 2000 post warm-up samples per chain. Since no divergent transitions were observed, we also let the default settings for *max_treedepth* and initial point selector.

The codes for sampling from models are shown in listings 5, 6 and 7 for pooled, hierarchical and hierarchical with parental prior mean model, respectively. Note that all models require data of the height observations, number of children considered and total number of families. In addition, the hierarchical model with parental prior mean requires also vectors of heights of fathers and mothers for each family, as it is clear from mathematical description of the model (6) and its implementation in Stan 4.

```
#Input for STAN pooled model
pool_data =list(N=numkids,
                J=numfams,
                y=data.frame(dataset))

#Compiling STAN
modelp = stan_model("pool.stan",model_name="pooled")

#SAMPLING from STAN
extract_pool= rstan::sampling(modelp, data=pool_data,iter=4000)

#EXTRACTING INFO from STAN
la_pool = extract(extract_pool)
```

Listing 5: Sampling from the pooled model

```
#Input for STAN hierarchical model
hier_data =list(N=numkids,
                J=numfams,
                y=data.frame(dataset))

#Compiling STAN
model = stan_model("hier.stan",
                   model_name="hier")

#SAMPLING from STAN
extract_hier= rstan::sampling(model,data=hier_data,iter=4000)

#EXTRACTING INFO from STAN
```

```
la_hier = extract(extract_hier)
```

Listing 6: Sampling from the hierarchical model

```
#Input for STAN hierarchical model with parental prior mean
hier2_data =list(N=numkids,
                 J=numfams,
                 y=data.frame(dataset),
                 f=fathers,m=mothers)

#Compiling STAN
modelh2 = stan_model("hier2.stan",
                     model_name="hier2")

#SAMPLING from STAN
extract_hier2 = rstan::sampling(modelh2, data=hier2_data,iter=4000)

#EXTRACTING INFO from STAN
la_hier2 = extract(extract_hier2)
```

Listing 7: Sampling from the hierarchical model with parental prior mean

More analysis on the convergence is discussed in section 7 and the final evaluation is given in section 10.

7 Convergence diagnostics

We are able to diagnose the convergence of sampling with two quantitative values: the $\hat{R}$-value and the minimum effective sample size ESS.

$\hat{R}$ convergence diagnostic compares the between- and within-chain estimates for model parameters and other univariate quantities of interest. If chains have not mixed well (i.e., the between- and within-chain estimates do not agree), $\hat{R}$ is larger than 1. It is recommended in [Aki] to use the sample if $\hat{R}$ is less than 1.05. Stan reports $\hat{R}$ which is the maximum of rank normalized split-$\hat{R}$ and rank normalized folded-split-$\hat{R}$.

On the other hand, effective sample size for each parameter plays the role in the Markov chain Monte Carlo central limit theorem (MCMC CLT) as the number of independent draws plays in the standard central limit theorem (CLT). The amount by which autocorrelation within the chains increases uncertainty in estimates can be measured by ESS. In other words, ESS reflects the number of independent samples that are required for estimation as well as the autocorrelated samples of the sampling chains. Large values indicate convergence of chains.

The code that extracts information about $\hat{R}$ values and ESSs from each model is shown in listing 8.

```
#getting Rhat and ESS values for Pooled model
summpool = summary(extract_pool)$summary
rhat_pool = summpool[,10]
ess_pool = summpool[,9]
max(rhat_pool)
min(ess_pool)

#getting Rhat and ESS values for Hierarchical model
summhier = summary(extract_hier)$summary
rhat_hier = summhier[,10]
ess_hier = summhier[,9]
max(rhat_hier)
min(ess_hier)

#getting Rhat and ESS values for Hierarchical model
#with parental prior mean
summhier2 = summary(extract_hier2)$summary
rhat_hier2 = summhier2[,10]
ess_hier2 = summhier2[,9]
max(rhat_hier2)
min(ess_hier2)
```

Listing 8: Computing $\hat{R}$ values and ESSs

Model	max $\hat{R}$	min ESS
Pooled	1.000475	3782.62
Hierarchical	1.001515	2271.10
Hierarchical with parental mean prior	1.002099	801.80

Table 2: Maximum $\hat{R}$ values and minimum ESS values when sampling from models

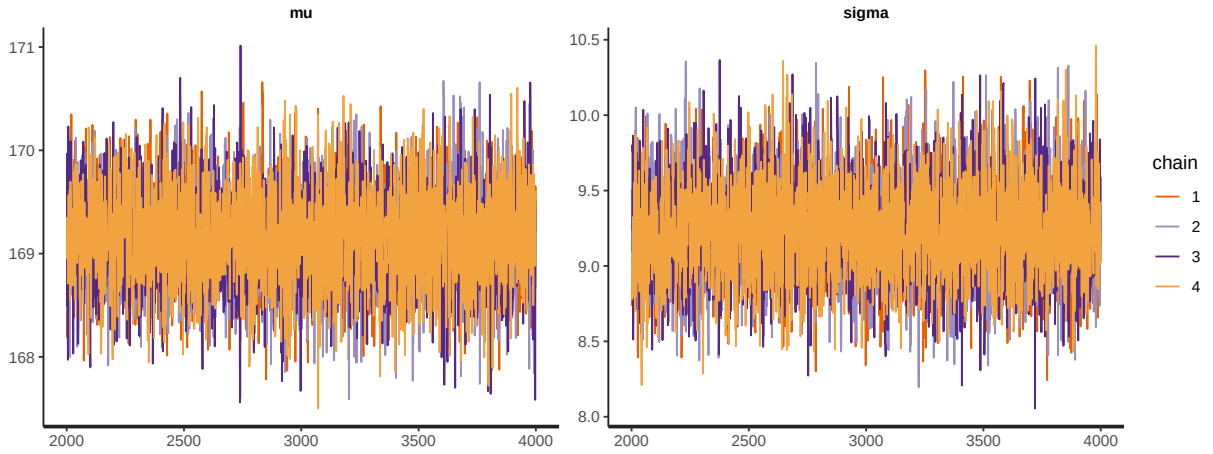

Figure 5: Mixing of chains when sampling from the pooled model

Maximal $\hat{R}$ values and minimal ESS values obtained from sampling from the models are shown in the table 2. All of $\hat{R}$ values are smaller than 0.01 which implies that chains have mixed well and the estimates are likely to be reliable. Mixing of chains for different models are also shown in figures 5, 6, 7. The chains for each parameter look like a "fat hairy caterpillar".

Minimum ESS values for pooled and hierarchical models are larger than 2000, which was the actual sample size. This concludes a highly succesful convergence rate. However, for the hierarchical model with parental prior mean, the minimum ESS was around 800.

8 Posterior predictive checks

The goal of posterior predictive check is to find out whether the model fits the data.We will draw samples from the posterior of the parameters using the Stan models implemented in section 5. We wish to sample new simulated heights of children from the posterior predictive distribution

$$p(\tilde{y} \,|\, y) = \int p(\tilde{y} \,|\, \theta) p(\theta \,|\, y) \, \mathrm{d}\theta = \int p(\tilde{y}, \theta \,|\, y) \, \mathrm{d}\theta \,,$$

where $\tilde{y}$ is a simulated height that could have been observed and y is the actual data.

8.1 Quantitative summaries

We can for example check summaries of estimated parameters, such as means, quantiles and standard deviation. Tables 3, 4 and 5 shows such comparison for hierarchical models in families $1, 2, 3$.

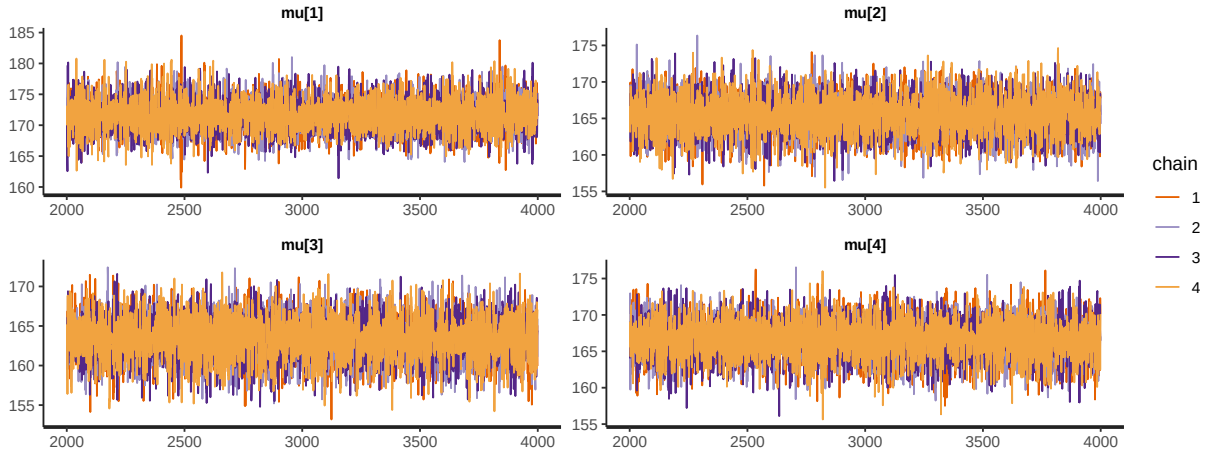

Figure 6: Mixing of chains when sampling from the hierarchical model

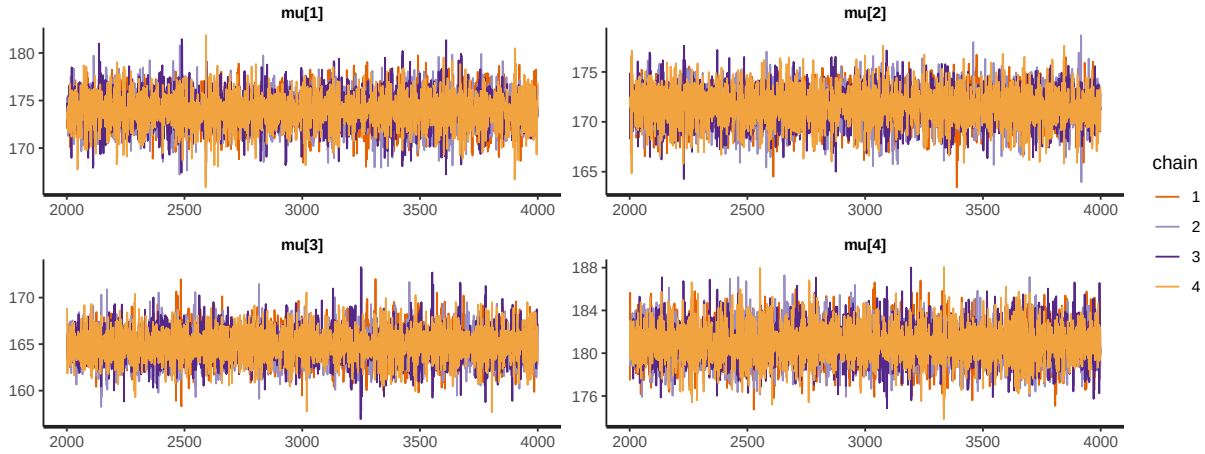

Figure 7: Mixing of chains when sampling from the hierarchical model with parental prior mean

	Family 1	Hier. pred. 1	Hier2. pred. 1
Mean	174.20	171.72	174.48
Standard deviation	10.42	2.71	1.90
25%	165.42	169.89	173.24
75%	179.71	173.52	175.72

Table 3: Summary of predictions and comparison with sample data for hierarchical model and hierarchical model with parental prior mean for family No.1

	Family 2	Hier. pred. 2	Hier2. pred. 2
Mean	164.67	165.47	168.43
Standard deviation	6.75	2.68	1.98
25%	161.61	163.71	167.15
75%	168.91	167.27	169.76

Table 4: Summary of predictions and comparison with sample data for hierarchical model and hierarchical model with parental prior mean for family No.2

	Family 3	Hier. pred. 3	Hier2. pred. 3
Mean	168.783	163.37	174.07
Standard deviation	7.71	2.75	1.99
25%	163.83	161.52	172.82
75%	174.05	165.24	175.44

Table 5: Summary of predictions and comparison with sample data for hierarchical model and hierarchical model with parental prior mean for family No.3

8.2 Graphical comparison

Here we use the R function *ppc_dens_overlay* to compare density estimate of observed data to density estimates of predicted data by our Stan models.

In figures 8, 9 and 10 we plotted density estimates of family No.50 together with 200 density estimates using predicted data from each of the Stan models.

If we would like to compare estimated means and standard deviations visually, cf. 8.1, we can compare the pair of sample mean and sample standard deviation to the predicted means and standard deviations from our Stan models. Such comparison is shown in figures

9 Model comparison

We use the log-likelihoods, in R code denoted *log_lik*, obtained from sampling the Stan models in order to compute the Pareto $\hat{k}$ values and PSIS expected log-predictive density values (elpd). The PSIS elpd value is an unbiased estimate of log posterior predictive density for new observations. Hence, high PSIS elpd value indicates that observed data are well described by the model. Pareto $\hat{k}$ values serve as a measure of evaluation of reliability of PSIS elpd values. We also compute the number of efficient parameters p_{eff} for

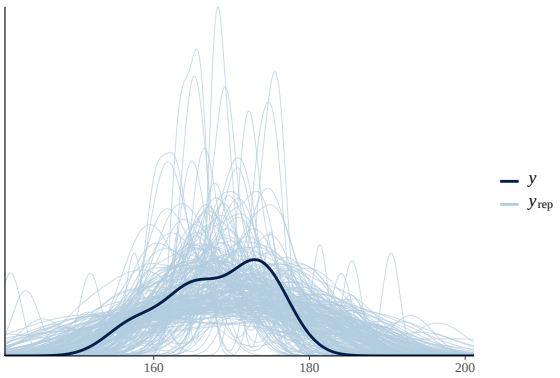

Figure 8: Comparison of density estimations in pooled model for family No.50

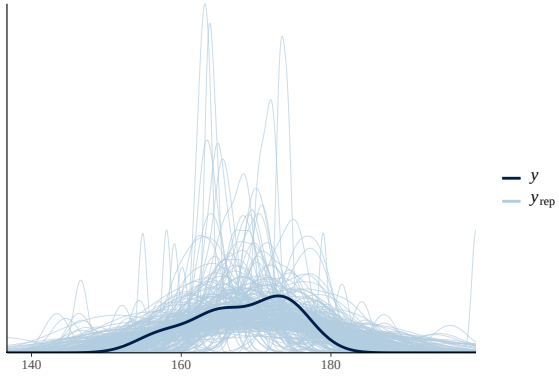

Figure 9: Comparison of density estimations in hierarchical model for family No.50

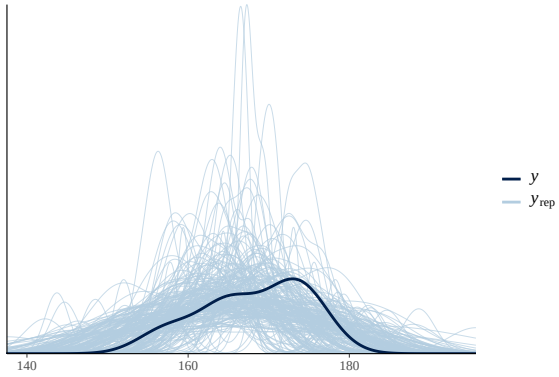

Figure 10: Comparison of density estimations in hierarchical model with parental prior mean for family No.50

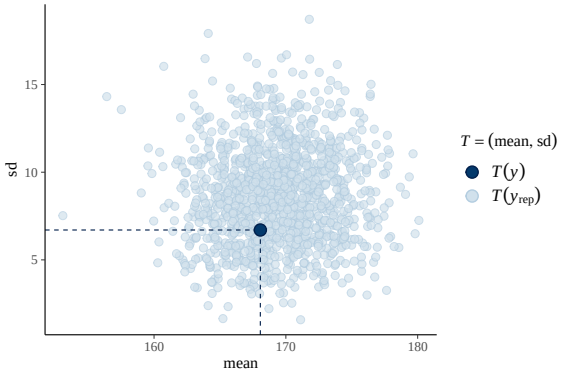

Figure 11: Comparison of sample mean and standard deviation and predicted values in pooled model for family No.50

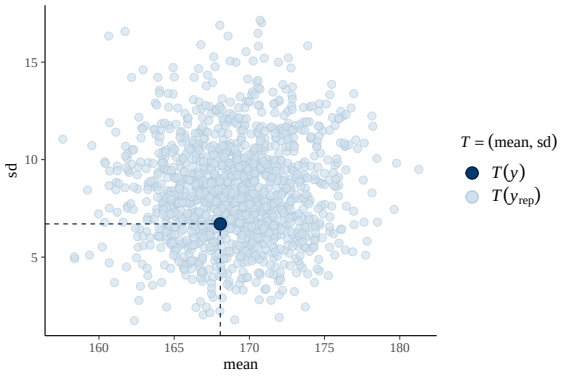

Figure 12: Comparison of sample mean and standard deviation and predicted values in hierarchical model for family No.50

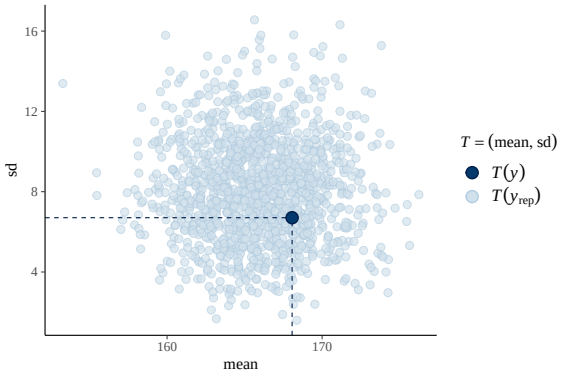

Figure 13: Comparison of sample mean and standard deviation and predicted values in hierarchical model with parental prior mean for family No.50

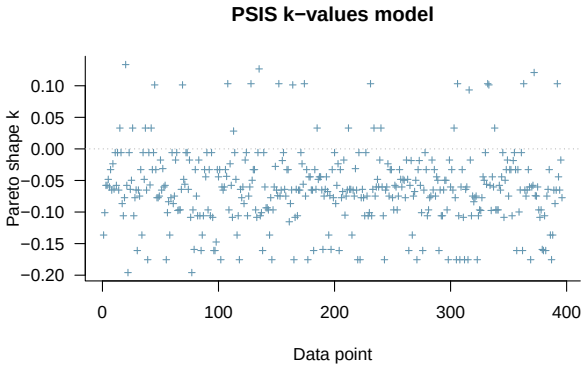

Figure 14: Pareto $\hat{k}$ values for pooled model

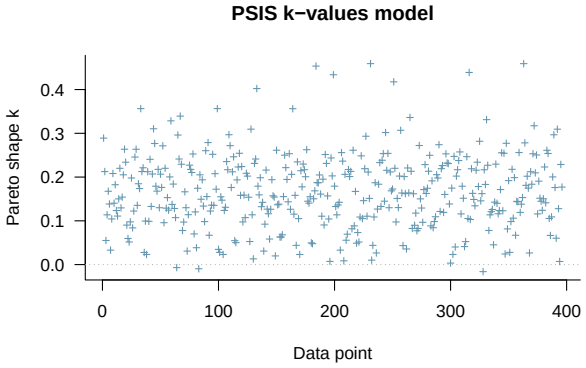

Figure 15: Pareto $\hat{k}$ values for hierarchical model

different models. p_{eff} serves well to address possible flaws in the model when Pareto $\hat{k}$ values are too high.

Obtained Pareto $\hat{k}$ values are shown in figures 14, 15 and 16 for pooled, hierarchical and hierarchical with parental prior mean, respectively. The reliability of PSIS elpd values can be assessed using Pareto $\hat{k}$ values in the following way. If $\hat{k} < 0.5$, then the PSIS-estimates are reliable. From the shown figures it is clear, that all Pareto $\hat{k}$ values yield this property.

Obtained PSIS elpd values and p_{eff} are shown in table 6. Based on the PSIS elpd values, it seems that the pooled model has the worst performance. All of the models have almost the same PSIS elpd value, so we cannot choose one to be certainly the best. However, based on PSIS elpd values, the best performing model is the Hierarchical model with parental mean prior.

Number of efficient parameters p_{eff} for pooled model is 1.85 and the actual number of model parameters is 2. The actual number of parameters for hierarchical model is $66 + 1 + 2 = 69$ (number of families $+\sigma + \sigma_0 + \mu_0$). The actual number of parameters for hierarchical model with parental mean prior is 68.

10 Evaluation

The motivation of this work was to confirm that the distribution of heights in between siblings is normal. For this purpose, we developed 3 mathematical models, from which one is a pooled model and two are hierarchical.

For atleast somewhat sensible results, we needed to "restrict" our point of view only to families with

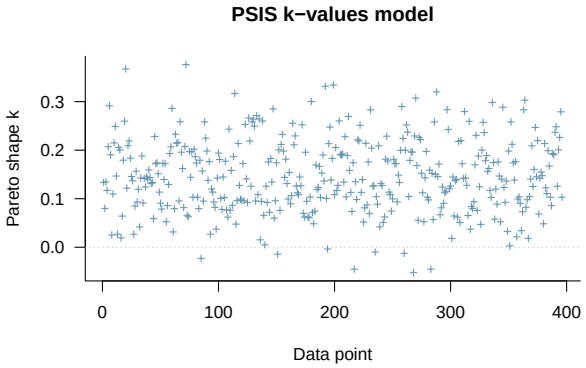

Figure 16: Pareto $\hat{k}$ values for hierarchical model with parental mean prior

Model	PSIS elpd value	p_{eff}
Pooled	-1443.84	1.85
Hierarchical	-1426.53	39.20
Hierarchical with parental mean prior	-1413.38	23.32

Table 6: PSIS elpd values and p_{eff} for different Stan models

more than 5 kids. The Monte Carlo chains converted sufficiently, as $\hat{R}$ values were very close to 1 and none of the parameters in any model had $\hat{R}$ value above 1.01.

The posterior predictive analysis has shed light on some shortcomings of each of the models. There are way too many "outlying" families, that have all kids roughly the same height. This fact results in a very low estimation of the standard deviation and eventually ruining the overall fit of the model.

Comparison of the posterior distributions for family No.1 are shown in figure 17. Mean, standard deviation and .25 and .75 quartiles corresponding to family No.1 are shown in table 3. Whilst the posteriors for both hierarchical models are almost identical, the pooled model posterior stays the same for every family. Similar comparison for family No.2 is shown in figure 18.

According to PSIS elpd values, the pooled model is the worst option. It is quite intuitive, since it has no reflection on the differences between families. Based on PSIS elpd values, the hierarchical model with parental mean prior should be chosen as the best option. It was shown in section 8, that this hierarchical model fits the data quite well, however there is not a significant difference, when compared to the other hierarchical model.

11 Sensitivity analysis

We perform the sensitivity analysis with respect to prior choices, in orded to conclude whether the prior distributions have an undesired influence on the posterior distributions. If the results do not change drastically over varying priors, the model is reliable.

11.1 Pooled sensitivity

The priors for pooled model are

$$\mu \sim \mathrm{N}(\mu_{prior}, \sigma_{prior})\,,$$
$$\sigma \sim \mathrm{Gamma}(\alpha_{prior}, \beta_{prior})\,,$$

where $\mu_{prior}, \sigma_{prior}$, $\alpha_{prior}, \beta_{prior}$ denote the prior parameters that may vary in the prior sensitivity analysis.

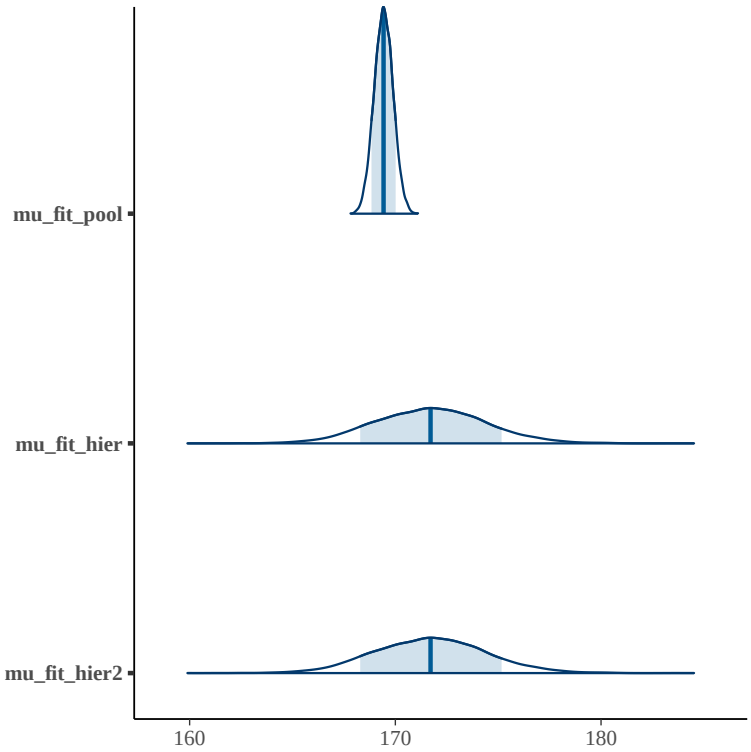

Figure 17: Comparison of posterior distributions of different models belonging to family No.1

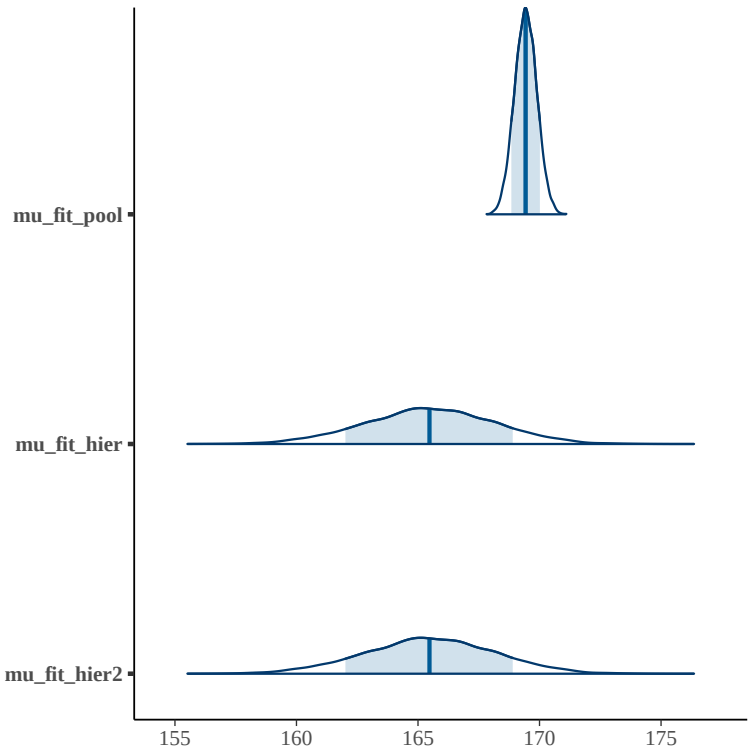

Figure 18: Comparison of posterior distributions of different models belonging to family No.2

(a) New priors (b) Old priors

Figure 19: Comparison of posterior predictive distribution for pooled models with different priors.

(a) New priors (b) Old priors

Figure 20: Comparison of posterior predictive distribution for hierarchical models with different priors.

We tried the following values of the parameters:

$$\mu_{prior} = 180, \quad \sigma_{prior} = 20,$$
$$\alpha_{prior} = 10, \quad \beta_{prior} = 1\,.$$

The comparison of posterior predictive distributions with old and new priors is shown in figure 19. Note that Stan is a stochastic program and the samples are never the same. It looks with new priors, the pooled model predictive posterior is more precise when we relaxed the priors a little bit. After all, we enlarged the standard variations in both prior distributions. Maybe the original priors were too informative. On the other hand, the PSIS elpd value got even lower with the new prior choice.

11.2 Hierarchical sensitivity

The priors for hierarchical model are

$$\sigma \sim \text{Gamma}(\alpha_{prior}, \beta_{prior})\,,$$
$$\sigma_0 \sim \text{Gamma}(\alpha_{prior}, \beta_{prior})\,,$$

and weight coefficients for f_j and m_j, where the shape and scale parameters for gamma distribution could be different in general, however, there is no reason to distinguish between them.

We tried the following values of the parameters:

$$\alpha_{prior} = 10, \quad \beta_{prior} = 1,$$
$$\text{w}_f = 0.7, \quad \text{w}_m = 0.3\,,$$

putting more emphasis on the height of the father and at the same time increasing the total variance by increasing α_{prior}.

The comparison of posterior predictive distributions with old and new priors is shown in figure 20. Again, after relaxing the priors a little, the new posterior predictive distribution fits the data better.

12 Discussion

One of the main issues was the initial choice of too informative priors in section 3. As was discussed in section 11, less informative priors would be probably more suitable for the established problem.

One of the possible future improvements would be definitely more insightful priors, chosen with respect to the already exposed positive correlation between the heights of child's parents and child's own height. Our lack of knowledge in this field may have compromised the results heavily.

Also, the PSIS-LOO analysis should be available for ragged arrays in 2022. With such tool, we would not have to restrict ourselves only to families with more than 5 children and we could conduct the analysis with less stochasticity that was incorporated during the data scraping (see section 2).

13 Conclusion

It is very likely, that the distribution of heights in between siblings follow the normal distribution. We formulated 3 different models, from which the results are not clearly decisive. Considering the PSIS elpd values, the hierarchical model with parental prior mean is the most suitable one. Probably because there is some information of the parents incorporated in the model. However, the initial priors were too restrictive/informative and that have maybe compromised to possible outcomes of our analysis.

During the analysis, we discovered that many families might have their children's height distributed normally with very low variance. This phenomenon is then hardly modelable with our predictive posterior, which is affected heavily with the too informative prior distribution.

References

[1] Galton's height data. https://www.randomservices.org/random/data/Galton.html. Accessed: 2022-02-01.

[2] Simple linear regression with galton. https://pygot.wordpress.com/2017/03/25/simple-linear-regression-with-galton/. Accessed: 2022-02-01.

[3] Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 1886.

[4] Hao Han, Yeming ma, and Wei Zhu. Galton's family heights data revisited. 08 2015.