The data is a poly-pharmacy dataset which is available from “polypharm” in the R package “aplore3”. The set contains data on 500 subjects studied over 7 years. The response is whether the subject is taking drugs from 3 or more different groups. We consider the covariates, Gender = 1 if male and 0 if female, Race = 0 if subject is white and 1 otherwise, Age, and the following binary indicators for the number of outpatient mental health visits, $\text{MHV}_1 = 1$ if $1 \leq \text{MHV} \leq 5$, $\text{MHV}_2 = 1$ if $6 \leq \text{MHV} \leq 14$ and $\text{MHV}_3 = 1$ if $\text{MHV} \geq 15$. Let INPTMHV = 0 if there were no inpatient mental health visits and 1 otherwise. We consider a logistic random intercept model of the form:

$$\begin{aligned}\text{logit}(\mu_{ij}) = \;& \beta_0 + \beta_1 \text{Gender}_i + \beta_2 \text{Race}_i + \beta_3 \text{Age}_{ij} + \\ & \beta_4 \text{MHV}_{1ij} + \beta_5 \text{MHV}_{2ij} + \beta_6 \text{MHV}_{3ij} + \beta_7 \text{INPTMHV}_{ij} + u_i,\end{aligned}$$

for $i = 1, \ldots, 500,\; j = 1, \ldots, 7$, where $u_i \sim N(0, e^{2\xi})$. The priors can be $\beta_j \sim N(0, \sigma_\beta^2)$ and $\xi \sim N(0, \sigma_\xi^2)$ with $\sigma_\beta^2 = \sigma_\xi^2 = 100$.

1 Random intercept model

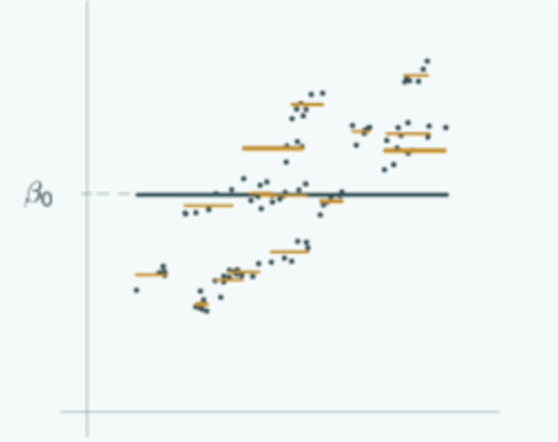
The random intercept model allows for some correlation between explanatory variables, as our dataset collects 7 years of data from 500 patients, and the data from the same patients are correlated.

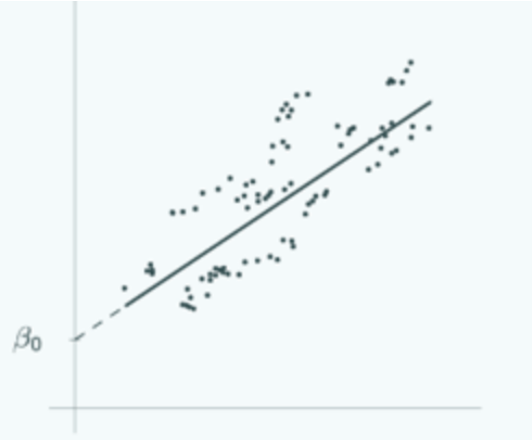
One-way ANOVA:

$$y_{ij} = \beta_0 + u_j + e_{ij} \qquad e_{ij} \sim N(0, \sigma_e^2) \qquad u_j \sim N(0, \sigma_u^2)$$

Regression model:

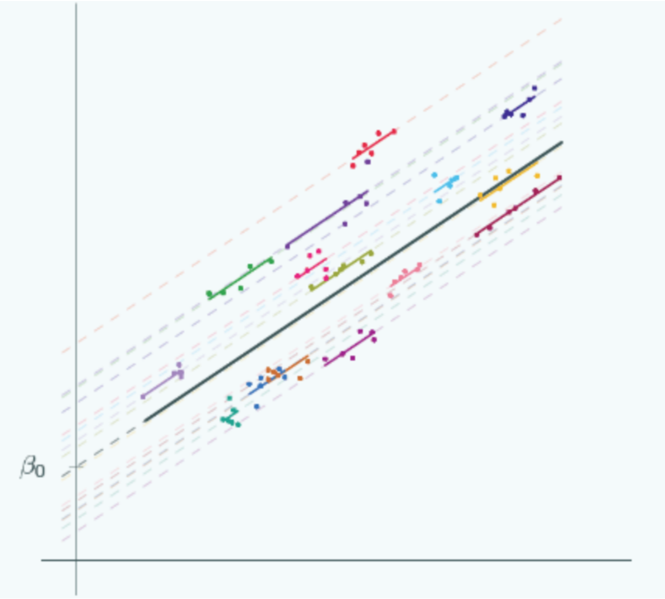
$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad e_i \sim N(0, \sigma^2)$$





Random intercept model: (ANCOVA)

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij} \qquad e_{ij} \sim N(0, \sigma_e^2) \qquad u_j \sim N(0, \sigma_u^2)$$



In the fitting results, the slope of each group is the same, and the intercept is different.

2 Problem setting

What we concern in this problem is whether the subject is taking drugs from 3 or more different groups, define this as random variable z_{ij}, and define μ_{ij} as the probability of $z_{ij} = 1$.

$$z_{ij}|\mu_{ij} \sim Bernoulli(\mu_{ij}) \tag{1}$$

Choosing logit function as the link regression and assume:

$$\begin{aligned} logit(\mu_{ij}) &= \beta_0 + \beta_1 Gender_{ij} + \beta_2 Race_{ij} + \beta_3 Age_{ij} + \beta_4 MHV_{1ij} \\ &+ \beta_5 MHV_{2ij} + \beta_6 MHV_{3ij} + \beta_7 INPTMHV_{ij} + u_i \\ &:= x_{ij}\beta + u_i \qquad for i = 1, \cdots, 500;\ j = 1, \cdots, 7, \end{aligned} \tag{2}$$

where $u_i \sim \mathcal{N}(0, e^{2\xi})$ with priors $\beta_j \sim \mathcal{N}(0, 100), \xi \sim \mathcal{N}(0, 100)$.

Note that there is no residual in the above random intercept model, there is no definition of the residual for the generalized regression model.

3 Theoretical Derivation

To derive the posterior distribution for the parameters.

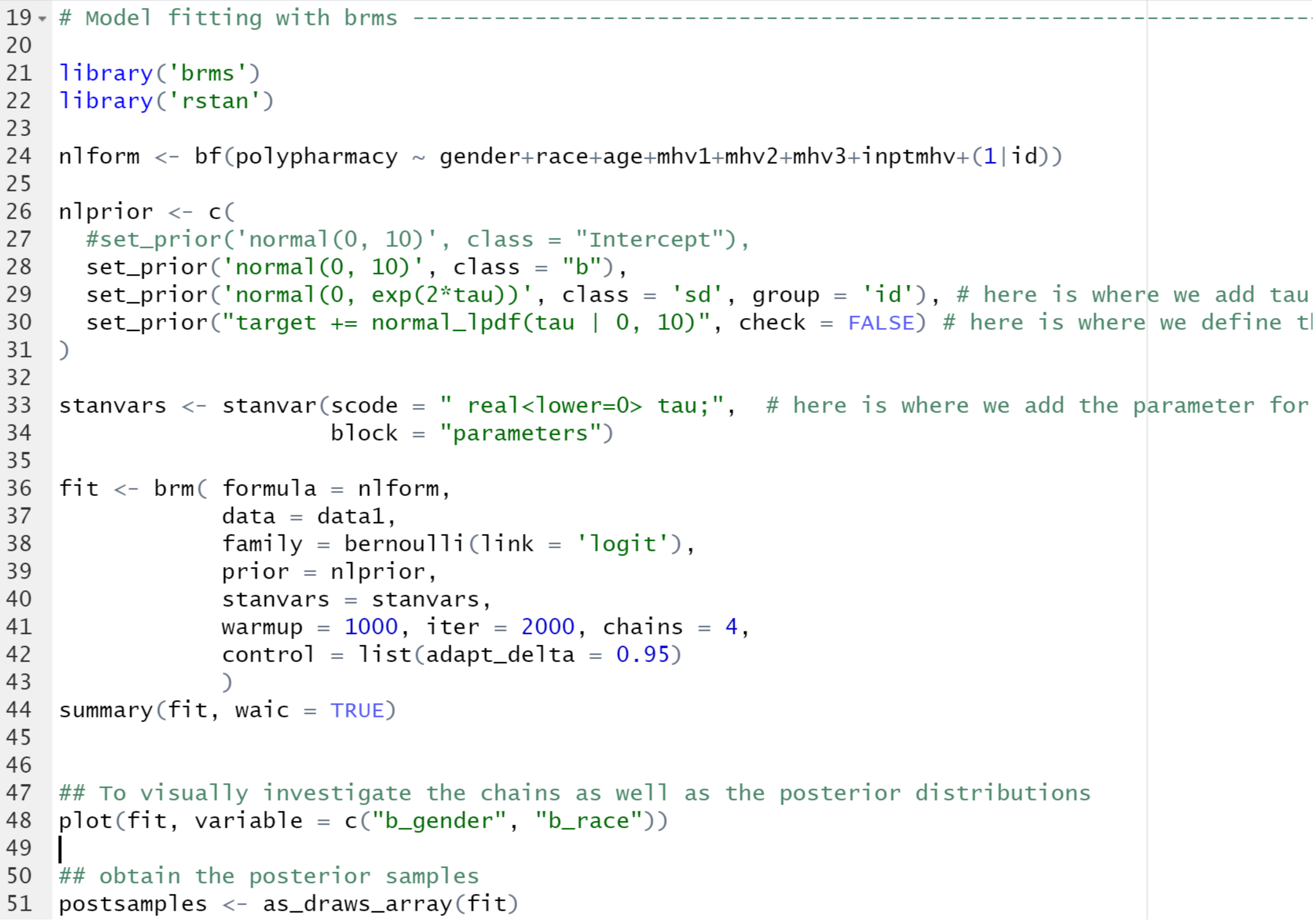
$$\begin{aligned} p(\beta, u_i, \xi, |z_{ij}) &\propto \prod_{i=1}^{500}\prod_{j=1}^{7} p(z_{ij}|\mu_{ij}) \cdot p(\beta) \cdot p(u_i|\xi) \cdot p(\xi) \\ &= \left(\prod_{i=1}^{500}\prod_{j=1}^{7} \mu_{ij}^{z_{ij}}(1-\mu_{ij})^{1-z_{ij}}\right) \cdot \left(\prod_{k=0}^{7} \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{\beta_k^2}{2\sigma^2}\}\right) \\ &\cdot \left(\prod_{i=1}^{500} \frac{1}{\sqrt{2\pi}\sigma_u} \cdot \exp\{-\frac{u_i^2}{2\sigma_u^2}\}\right) \cdot \frac{1}{\sqrt{2\pi}\sigma} \exp\{-\frac{\xi^2}{2\sigma^2}\}, \end{aligned} \tag{3}$$

where $\mu_{ij} = logit^{-1}(x_{ij}\beta + u_i)$, $\sigma = 10$ and $\sigma_u = e^{\xi}$.

We have $8 + 500 + 1 = 509$ unknown parameters.

4 Sampling

Using the 'brms' package in R to sample the posterior samples:

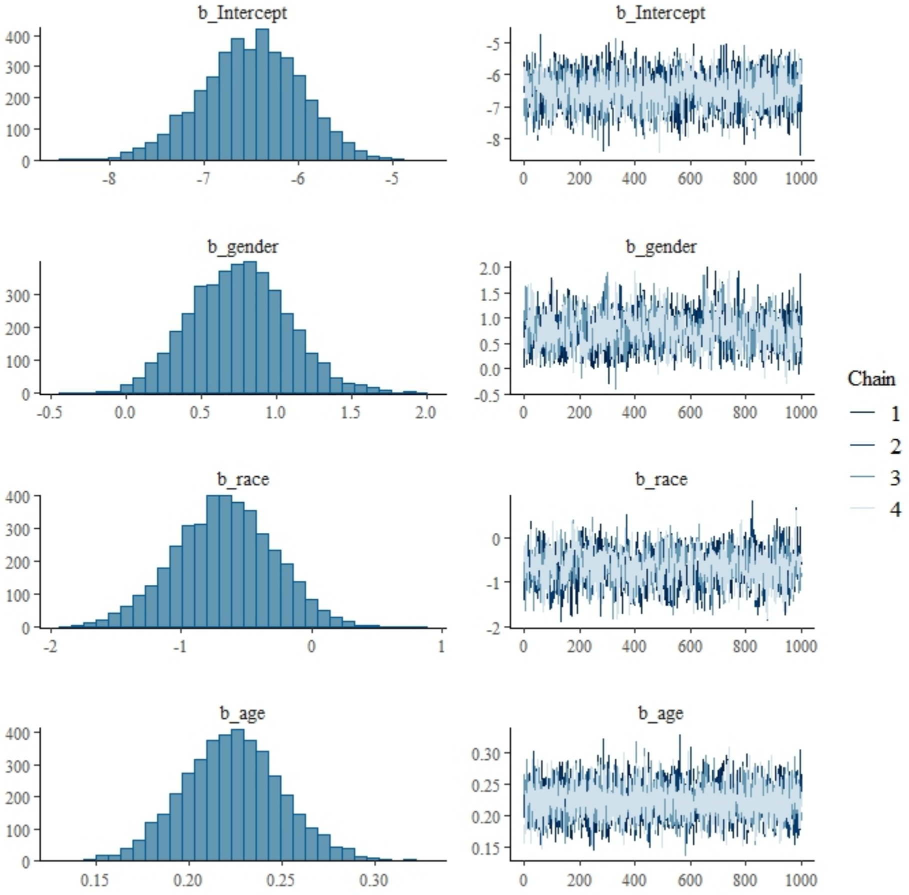


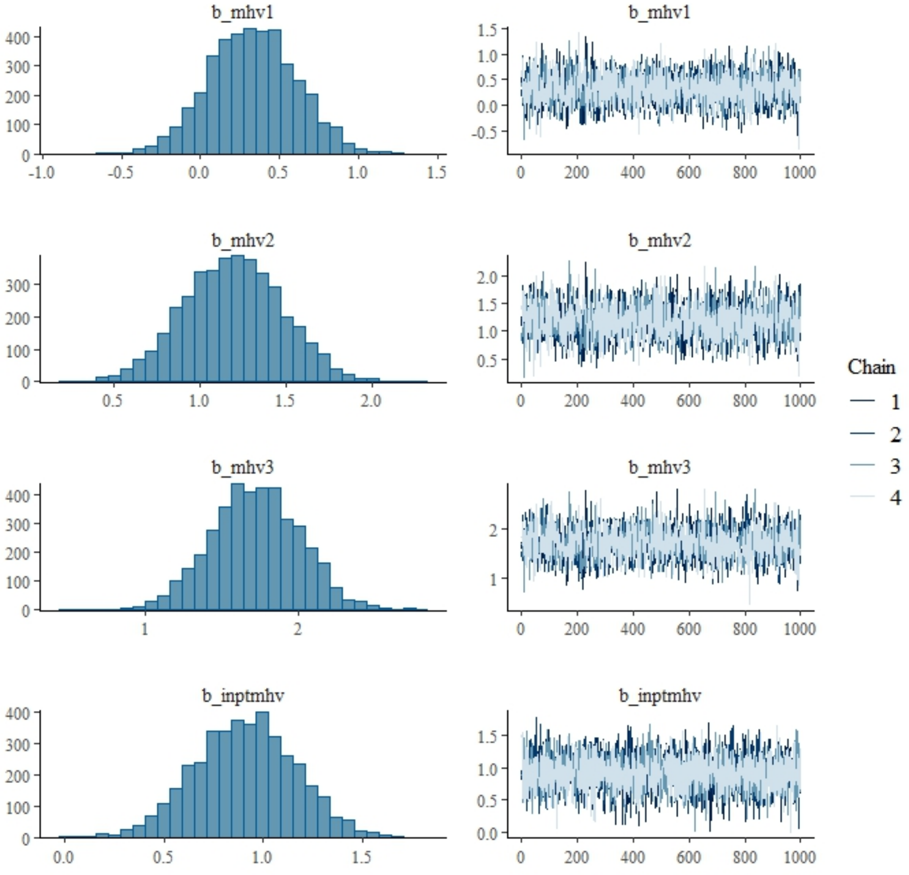
```r
# Model fitting with brms ------------------------------------------------------------------

library('brms')
library('rstan')

nlform <- bf(polypharmacy ~ gender+race+age+mhv1+mhv2+mhv3+inptmhv+(1|id))

nlprior <- c(
  #set_prior('normal(0, 10)', class = "Intercept"),
  set_prior('normal(0, 10)', class = "b"),
  set_prior('normal(0, exp(2*tau))', class = 'sd', group = 'id'), # here is where we add tau
  set_prior("target += normal_lpdf(tau | 0, 10)", check = FALSE) # here is where we define t
)

stanvars <- stanvar(scode = " real<lower=0> tau;",  # here is where we add the parameter for
                    block = "parameters")

fit <- brm( formula = nlform,
            data = data1,
            family = bernoulli(link = 'logit'),
            prior = nlprior,
            stanvars = stanvars,
            warmup = 1000, iter = 2000, chains = 4,
            control = list(adapt_delta = 0.95)
            )
summary(fit, waic = TRUE)


## To visually investigate the chains as well as the posterior distributions
plot(fit, variable = c("b_gender", "b_race"))

## obtain the posterior samples
postsamples <- as_draws_array(fit)
```

Draws were sampled using sampling(NUTS). For each parameter, Bulk ESS and Tail ESS are effective sample size measures, and $\hat{R}$ is the potential scale reduction factor on split chains (at convergence, $\hat{R} = 1$). Use the posterior mean as the estimation of parameters.

	Estimate	Error	$\hat{R}$	Bulk-ESS	Tail- ESS
Intercept	-6.53	0.53	1.00	2080	2405
gender	0.76	0.34	1.00	1220	2431
race	-0.67	0.39	1.00	1079	1884
age	0.22	0.03	1.00	3733	2895
mhv1	0.32	0.29	1.00	2306	2672
mhv2	1.18	0.30	1.00	1911	2517
mhv3	1.71	0.30	1.00	1943	2357
inptmhv	0.91	0.26	1.00	5387	2817
sd(Intercept)	2.48	0.17	1.00	995	1870

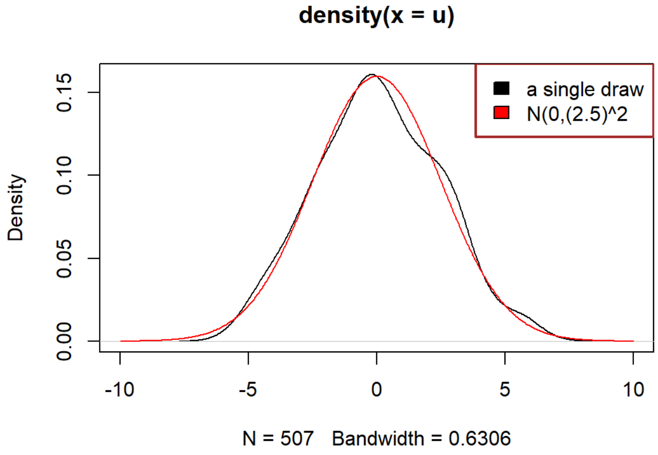
From the trace plot, we can see that the four chains are mixed and obtain stationary.


b_Intercept
400
300
200
100
0
-8
-7
-6
-5
b_Intercept
-5
-6
-7
-8
0
200
400
600
800
1000
b_gender
300
200
100
0
-0.5
0.0
0.5
1.0
1.5
2.0
b_gender
2.0
1.5
1.0
0.5
0.0
-0.5
0
200
400
600
800
1000
Chain
1
2
3
4
b_race
400
300
200
100
0
-2
-1
0
1
b_race
0
-1
-2
0
200
400
600
800
1000
b_age
400
300
200
100
0
0.15
0.20
0.25
0.30
b_age
0.30
0.25
0.20
0.15
0
200
400
600
800
1000


b_mhv1
400
300
200
100
0
-1.0
-0.5
0.0
0.5
1.0
1.5
b_mhv1
1.5
1.0
0.5
0.0
-0.5
0
200
400
600
800
1000
b_mhv2
300
200
100
0
0.5
1.0
1.5
2.0
b_mhv2
2.0
1.5
1.0
0.5
0
200
400
600
800
1000
Chain
1
2
3
4
b_mhv3
400
300
200
100
0
1
2
b_mhv3
2
1
0
200
400
600
800
1000
b_inptmhv
400
300
200
100
0
0.0
0.5
1.0
1.5
b_inptmhv
1.5
1.0
0.5
0.0
0
200
400
600
800
1000

5 Model checking for prior

If the model were true, we would expect any single simulation draw of the vectors of u_i (random intercept) parameters to look like independent draws from the its prior distribution. (for detailed information, see pdf 'model checking for prior')



The result suggests that the prior may follow a normal distribution with mean 0 and sd 2.5. So we refit the model.

```
nlform <- bf(polypharmacy ~ gender+race+age+mhv1+mhv2+mhv3+inptmhv+(1|id))

nlprior <- c(
  set_prior('normal(0, 10)', class = "Intercept"),
  set_prior('normal(0, 10)', class = "b"),
  set_prior('normal(0, 2.5)', class = 'sd', group = 'id')
)

fit <- brm( formula = nlform,
            data = data1,
            family = bernoulli(link = 'logit'),
            prior = nlprior,
            warmup = 1000, iter = 2000, chains = 4
)
summary(fit, waic = TRUE)
```

The results are shown below:

	Estimate	Error	$\hat{R}$	Bulk-ESS	Tail- ESS
Intercept	-6.52	0.53	1.00	1528	2190
gender	0.75	0.34	1.00	851	1655
race	-0.66	0.38	1.00	943	1524
age	0.22	0.03	1.00	3112	3032
mhv1	0.32	0.29	1.00	1482	2475
mhv2	1.19	0.29	1.00	1391	2094
mhv3	1.72	0.30	1.00	1324	1713
inptmhv	0.92	0.26	1.00	3803	3197
sd(Intercept)	2.49	0.16	1.00	988	1555

The estimation of the parameters are very close for the two model. But the ESS are lower for the reduced model, since the reduced model does not have hyper-parameter, its parameter's structure is simpler.