Visualization in Bayesian Data Analysis

Data Visualization Assignment

Luong Thi My Hanh

Referrences

- Visualization in Bayesian Data Analysis (III.16), Kerman et al. (book)
- Visualization in Bayesian workflow, Gabry et al. Royal statistical society, 2019.
- https://www.monicaalexander.com/posts/2020-28-02-bayes_viz/ (practice)

Workflow

(a) exploratory data analysis -> initial models

(b) computational model checks using fake data simulation and the prior predictive distribution

(c) computational checks for the inference algorithms

(d) posterior predictive checks

(e) model comparison

Data and aims of data analysis

- a sample of available dataset of all births, US, 2017

```
> nrow(d)
[1] 3864754
> ncol(d)
[1] 240
> d <- d %>% select(mager, mracehisp, meduc, bmi, sex, combgest, dbwt, ilive)
> head(d)
# A tibble: 6 × 8
  mager mracehisp meduc   bmi sex   combgest  dbwt ilive
  <dbl>     <dbl> <dbl> <dbl> <chr>    <dbl> <dbl> <chr>
1    31         2     5  32.5 F           40  3653 Y
2    33         1     8  20.2 M           38  2987 Y
3    36         1     1  28.3 M           37  3445 Y
4    26         5     3  24.5 M           39  3645 Y
5    19         3     2  15.6 F           24   860 Y
6    20         7     3  20.5 F           37  2875 Y
```

Focus on: birth weight and gestational age

birthweight = dbwt;
gest=combgest=combined gestation

sex: gender of babies
ilive: infant living at time of report
mager: age of mother
meduc: mother's education
mracehisp: mother's race/hispanic

```
> ds <- d[sample(1:nrow(d), nrow(d)*0.001),]
> nrow(ds)
[1] 3864
> ds <- ds %>% mutate(dbwt= dbwt/1000)
write_rds(ds, path = "births_2017_sample.RDS")
Warning message:
The `path` argument of `write_rds()` is deprecated as of readr 1.4.0.
Please use the `file` argument instead.
This warning is displayed once every 8 hours.
Call `lifecycle::last_lifecycle_warnings()` to see where this warning was generated.
> head(ds)
# A tibble: 6 × 8
  mager mracehisp meduc   bmi sex   combgest  dbwt ilive
  <dbl>     <dbl> <dbl> <dbl> <chr>    <dbl> <dbl> <chr>
1    36         1     3  19.1 M           43  3.77 Y
2    25         1     4  22.5 F           40  2.89 Y
3    40         7     3  26.6 F           39  2.97 Y
4    39         7     7  24.1 F           39  2.55 Y
5    38         8     9  31   M           35  2.76 Y
6    29         1     4  27.9 M           39  3.5  Y
```

```
> summary(ds$combgest)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  20.00   38.00   39.00   38.63   40.00   99.00
> summary(ds$dbwt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.227   2.950   3.291   3.271   3.629   9.999
>
```

```
> ds <- read_rds("births_2017_sample.RDS")
> ds <- ds %>% rename(birthweight = dbwt, gest = combgest) %>% mutate(preterm = ifelse(gest<3
2, "Y", "N")) %>% filter(ilive=="Y",gest< 99, birthweight<9.999)
> head(ds)
# A tibble: 6 × 9
  mager mracehisp meduc   bmi sex    gest birthweight ilive preterm
  <dbl>     <dbl> <dbl> <dbl> <chr> <dbl>       <dbl> <chr> <chr>
1    36         1     3  19.1 M        43        3.77 Y     N
2    25         1     4  22.5 F        40        2.89 Y     N
3    40         7     3  26.6 F        39        2.97 Y     N
4    39         7     7  24.1 F        39        2.55 Y     N
5    38         8     9  31   M        35        2.76 Y     N
6    29         1     4  27.9 M        39        3.5  Y     N
>
```

observations	birthweight
gestational age +	+
preterm	-
sex male	-
others: height of parents, educations,....	

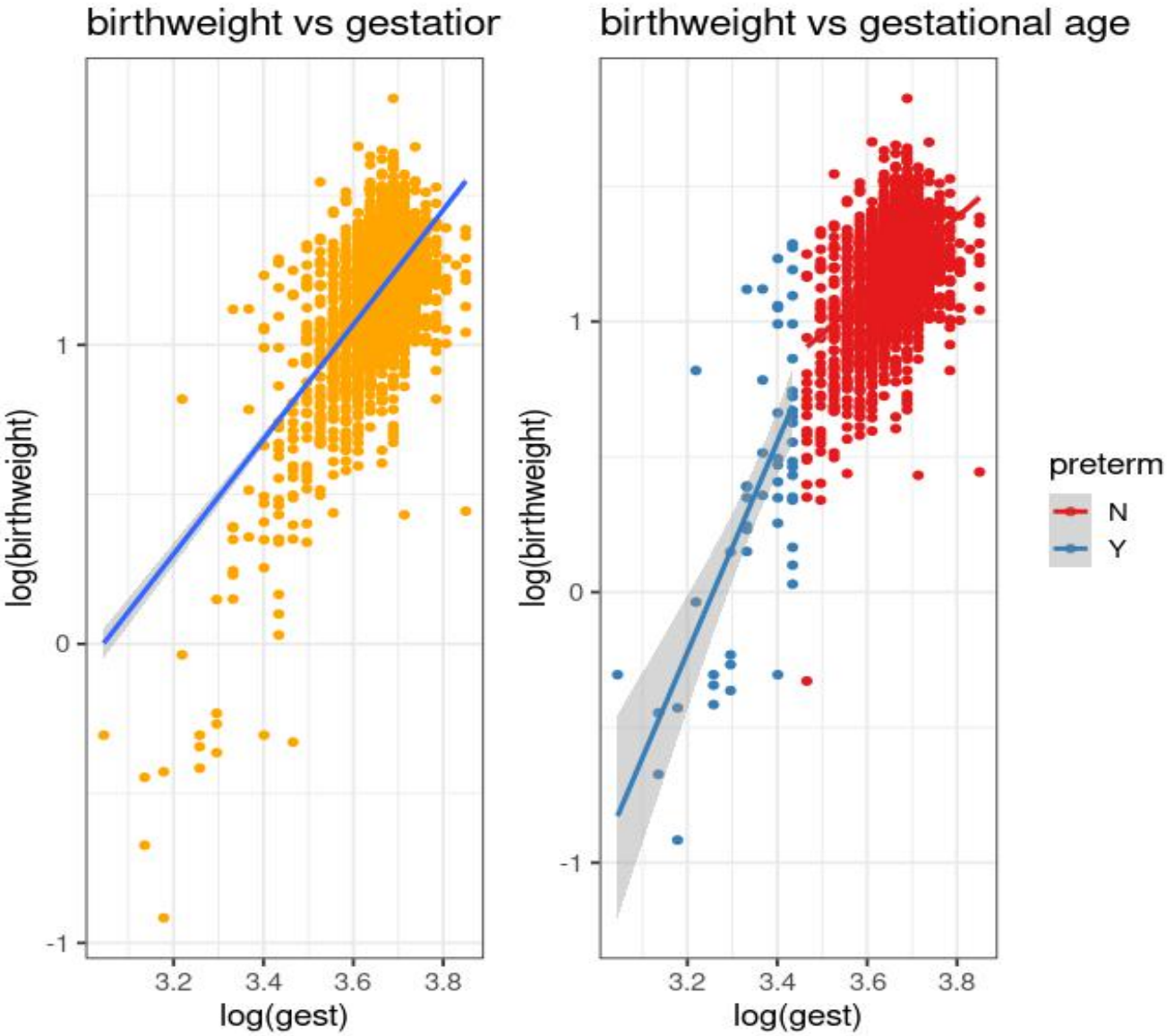
Aim of analysis: build a predictive model of birthweight.
Aim of class: Focus on three simple models and to show how visualization can be used to help to construct, sense-check, compute and evaluate these models (rstan and brms packages)

(a)Exploratory data analysis

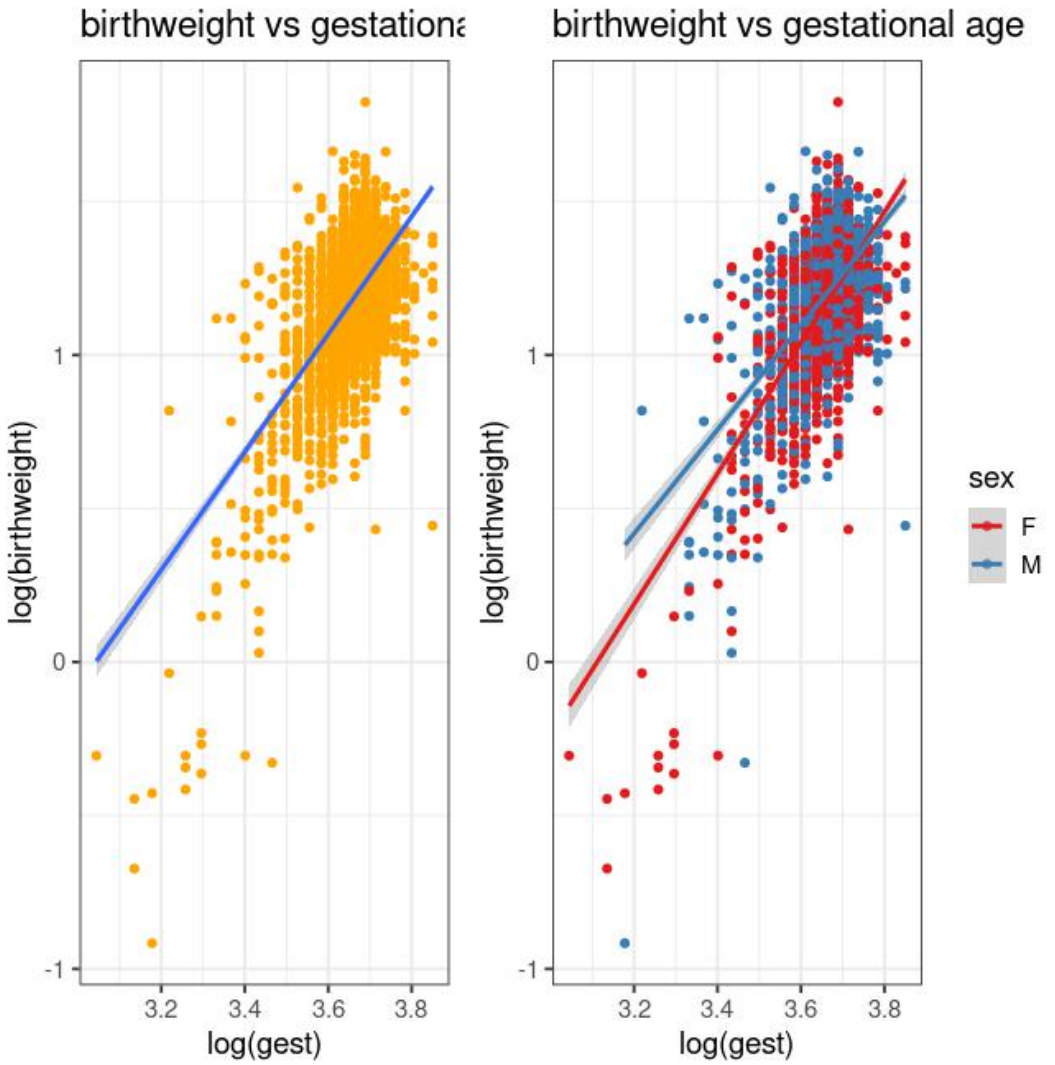
- Important first step to model building
- information of potential variables or functional forms of models

Plotting gestational age, preterm, sex and birthweight

```
> p1=ds %>% ggplot(aes(log(gest), log(birthweight))) + geom_point(col="orange") + geom_smooth
(method="lm") + scale_color_brewer(palette="Set1") + theme_bw(base_size=14) +  ggtitle("birth
weight vs gestational age")
p2=ds %>% ggplot(aes(log(gest), log(birthweight), color=preterm)) + geom_point() + geom_smoot
h(method="lm") + scale_color_brewer(palette="Set1") + theme_bw(base_size=14) +  ggtitle("birt
hweight vs gestational age")
p1+p2
```



```
> p1=ds %>% ggplot(aes(log(gest), log(birthweight))) + geom_point(col="orange") + geom_smooth
(method="lm") + scale_color_brewer(palette="Set1") + theme_bw(base_size=14) +  ggtitle("birth
weight vs gestational age")
p2=ds %>% ggplot(aes(log(gest), log(birthweight), color=sex)) + geom_point() + geom_smooth(me
thod="lm") + scale_color_brewer(palette="Set1") + theme_bw(base_size=14) +  ggtitle("birthwei
ght vs gestational age")
p1+p2
```



Some simple candidate models

Model 1:

$$\log(y_i) \sim N(\beta_0 + \beta_1 \log(x_i), \sigma^2)$$

Model 2 and 3:

$$\log(y_i) \sim N(\beta_0 + \beta_1 \log(x_i) + \gamma_0 z_i + \gamma_1 \log(x_i) z_i, \sigma^2)$$

- y_i is weight in kg
- x_i is gestational age in weeks
- z_i is preterm (0 or 1, if gestational age is less than 32 weeks)

or is fetal gender (0 if male, 1 is female)

in stan: mod1, mod2, mod3
in brms: mod1b, mod2b, mod2c

(b)Simulation and prior predictive checks

- Situation: no available data on birthweights
- what distribution of weights are implied by the choice of priors and likelihood?

=> simulate from the priors and likelihood, and plot the resulting distribution

$$\mathrm{P}(\theta \mid \mathrm{data}) \propto \mathrm{P}(\theta) \times \mathrm{P}(\mathrm{data} \mid \theta)$$

Posterior probablity | **Prior** probablity | **Likelihood**

Prior Choice Recommendations

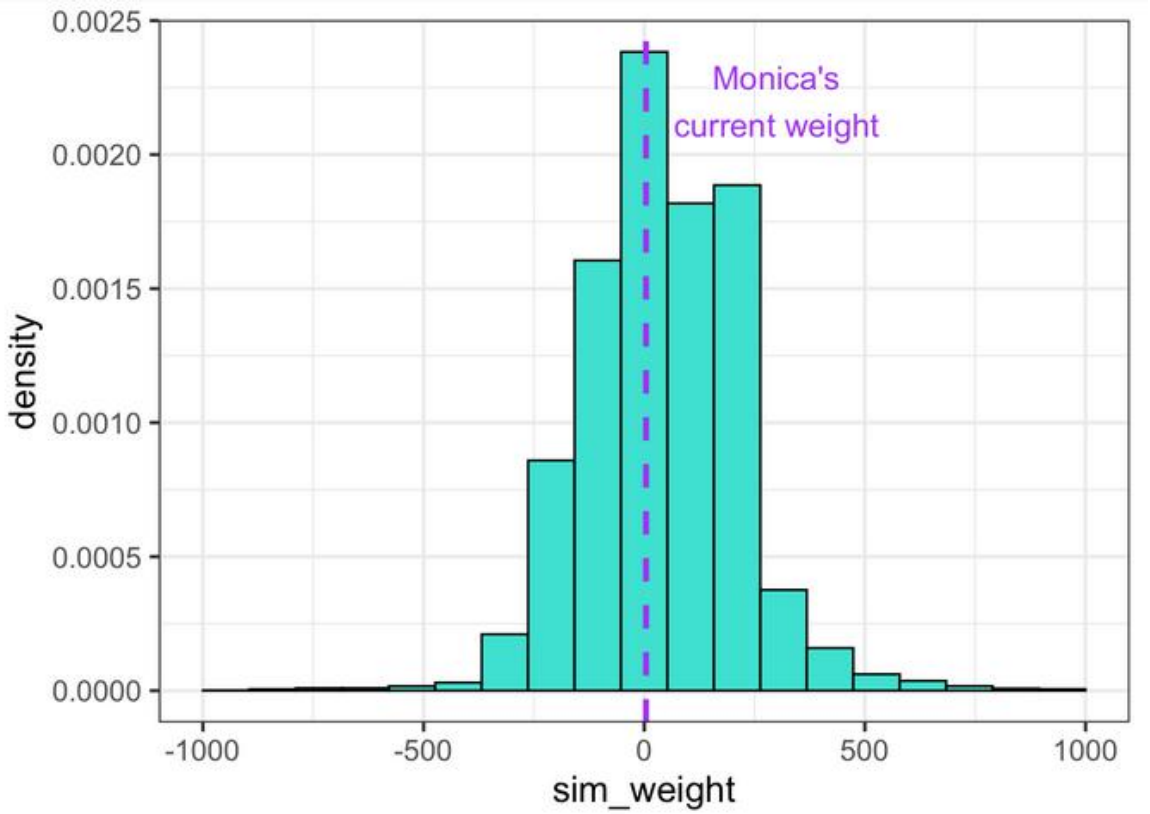
5 levels of priors

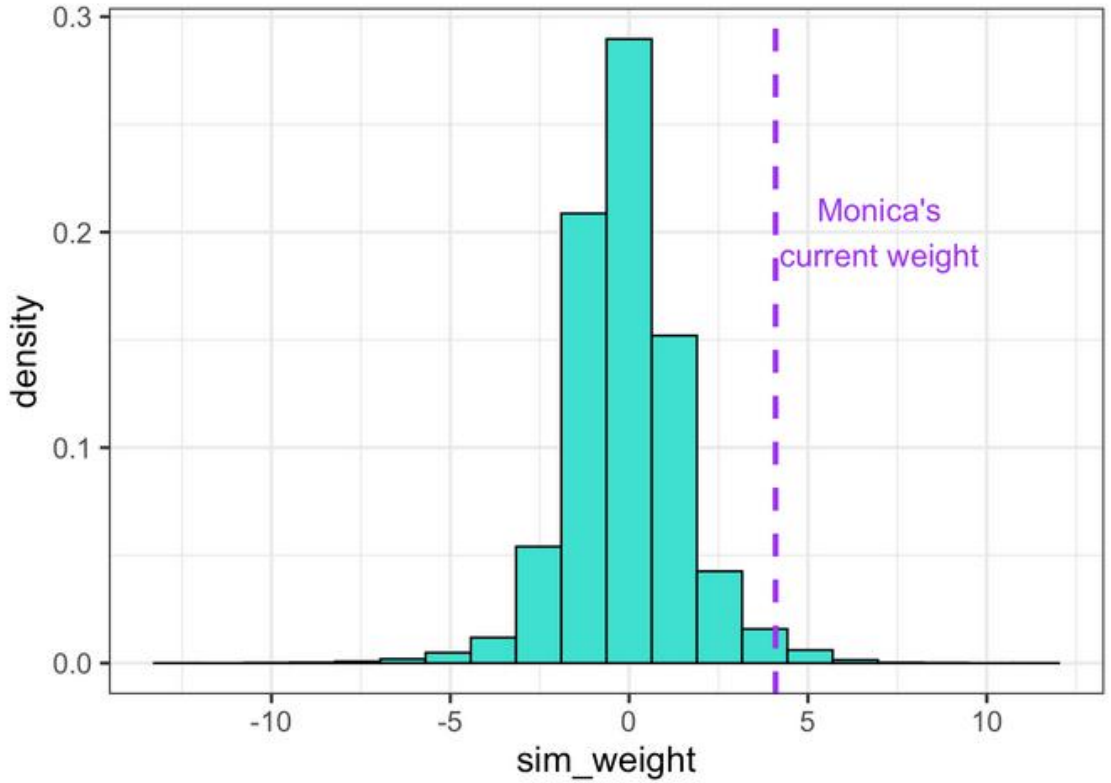
- Flat prior (not usually recommended);
- Super-vague but proper prior: normal(0, 1e6) (not usually recommended);
- Weakly informative prior, very weak: normal(0, 10);
- Generic weakly informative prior: normal(0, 1);
- Specific informative prior: normal(0.4, 0.2) or whatever. Sometimes this can be expressed as a scaling followed by a generic prior: theta = 0.4 + 0.2*z; z ~ normal(0, 1);

https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations

Vague and weakly informative priors

Model 1: $$\log(y_i) \sim N(\beta_0 + \beta_1 \log(x_i), \sigma^2)$$ Monica's weight=60 kg





vague: beta0 <- rnorm(nsims, 0, 100)
beta1 <- rnorm(nsims, 0, 100)

weakly: beta0 <- rnorm(nsims, 0, 1)
beta1 <- rnorm(nsims, 0, 1)

(c) computational checks

run the models with stan and brms

- mod1, mod1b: birthweight ~gestational age
- mod2, mod2b: birthweight ~ gestational age + preterm
- mod3, mod2c: birthweight ~ gestational age + sex

```
summary(mod1)[["summary"]][c(paste0("beta[",1:2, "]"), "sigma"),]

           mean      se_mean          sd      2.5%       25%       50%
beta[1] 1.1695800 7.474798e-05 0.002661259 1.1641541 1.1678824 1.1695891
beta[2] 0.1197083 7.647743e-05 0.002454188 0.1151330 0.1179753 0.1196762
sigma   0.1611435 1.020006e-04 0.001801512 0.1576861 0.1599025 0.1611106
              75%     97.5%     n_eff     Rhat
beta[1] 1.1713305 1.1748537 1267.5798 1.001259
beta[2] 0.1214014 0.1242364 1029.7908 1.000467
sigma   0.1622649 0.1649477  311.9382 1.007719
```

rstan

1 standard deviation increase in the gestation weeks leads to a 0.12 increase in birth weight (log).

```
#####
Running models with brms

mod1b <- brm(log_weight~log_gest_c, data = ds)

> summary(mod1b)
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: log_weight ~ log_gest_c
   Data: ds (Number of observations: 3842)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept      1.17      0.00     1.16     1.17 1.00     4465     2936
log_gest_c     0.12      0.00     0.11     0.12 1.00     4965     3134

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.16      0.00     0.16     0.16 1.00     1759     1786

Draws were sampled using sampling(NUTS). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

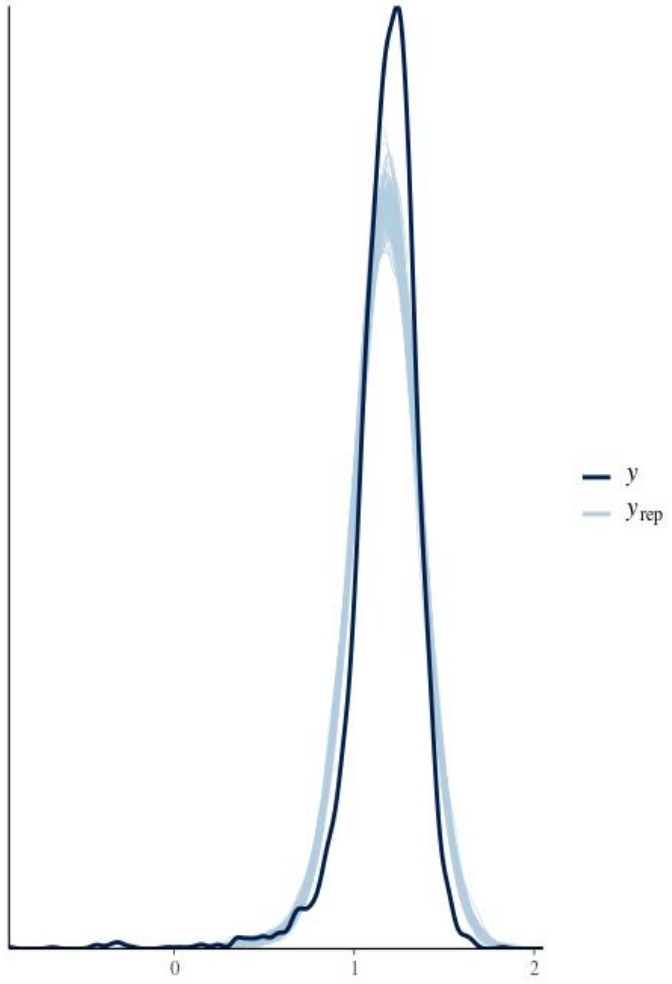
brms

(d)Posterior predictive checks: stan

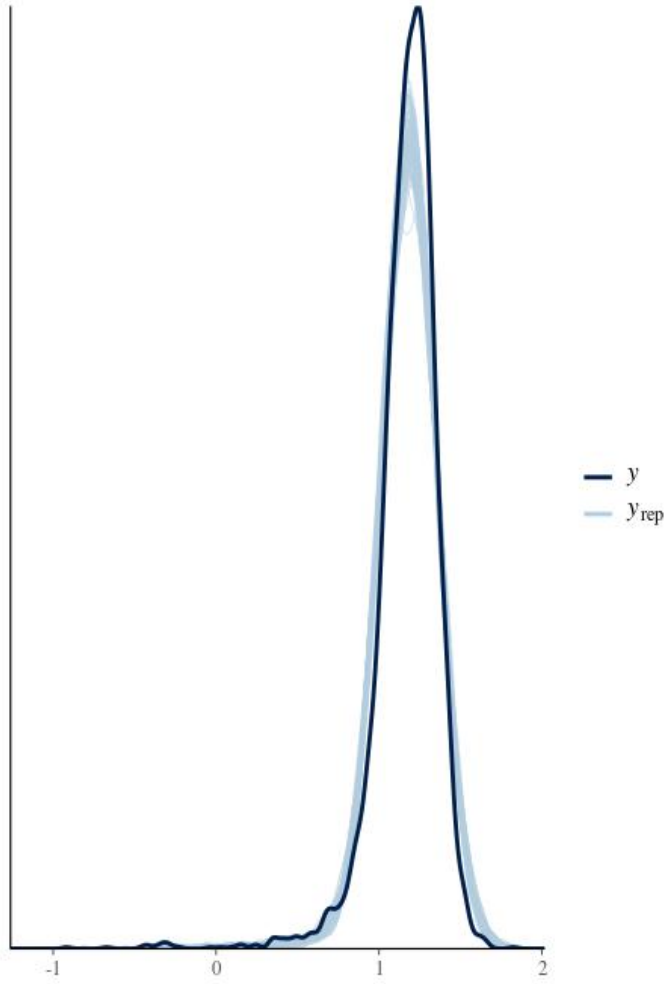
replicated datasets vs observed dataset

log_weight

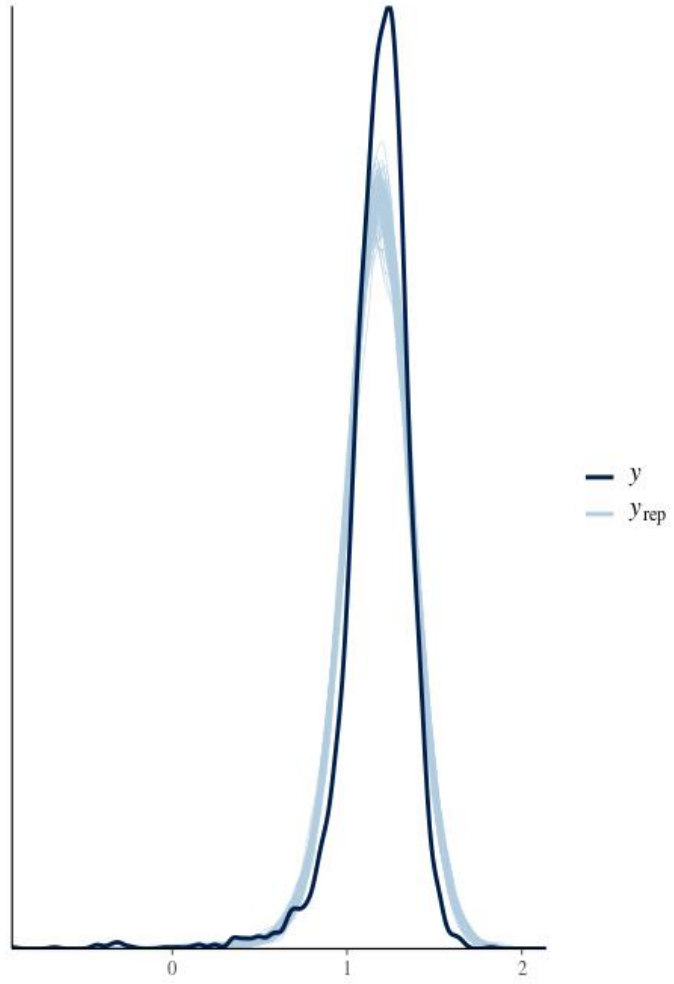
stan
Density plot



mod1



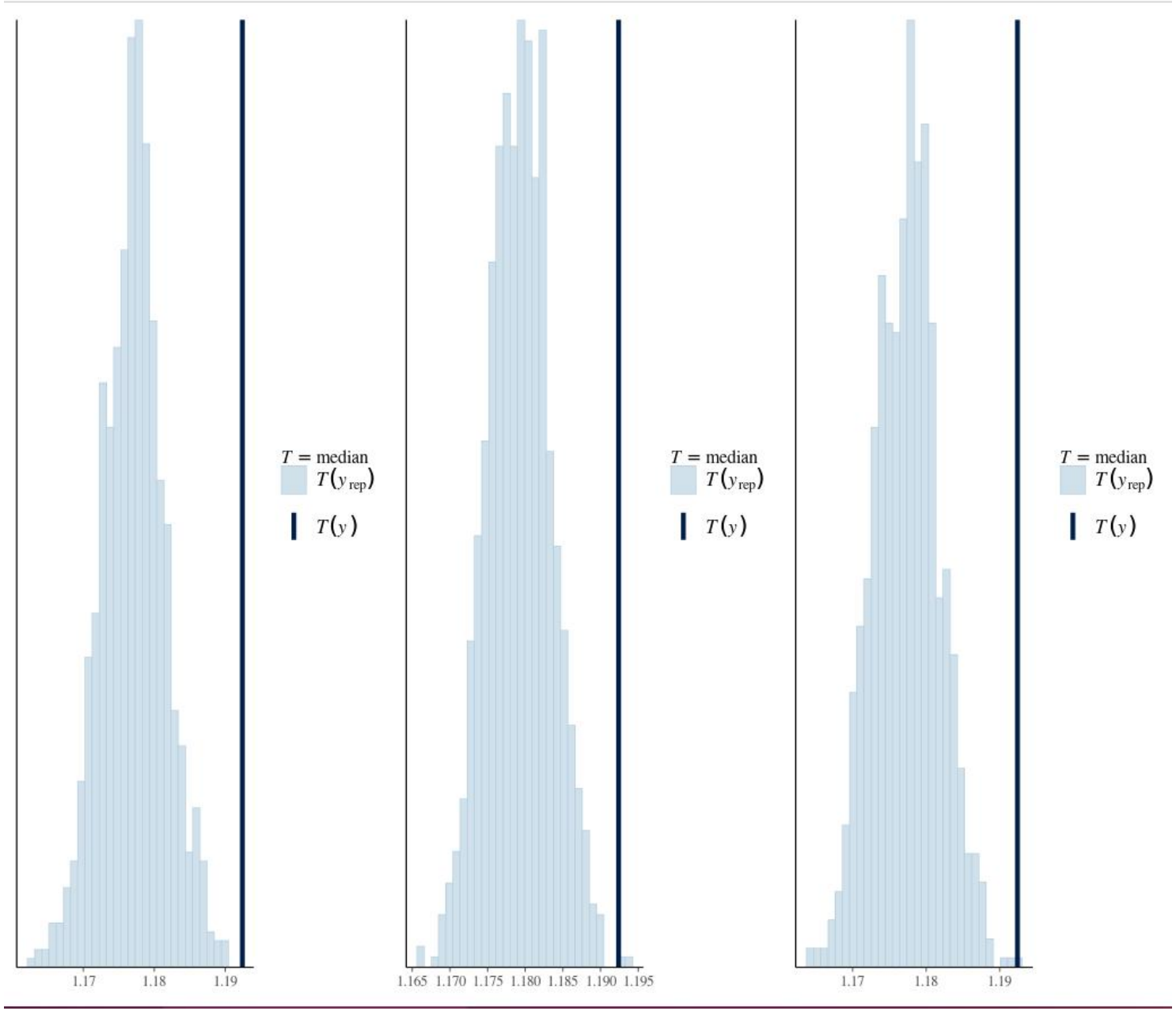
mod 2

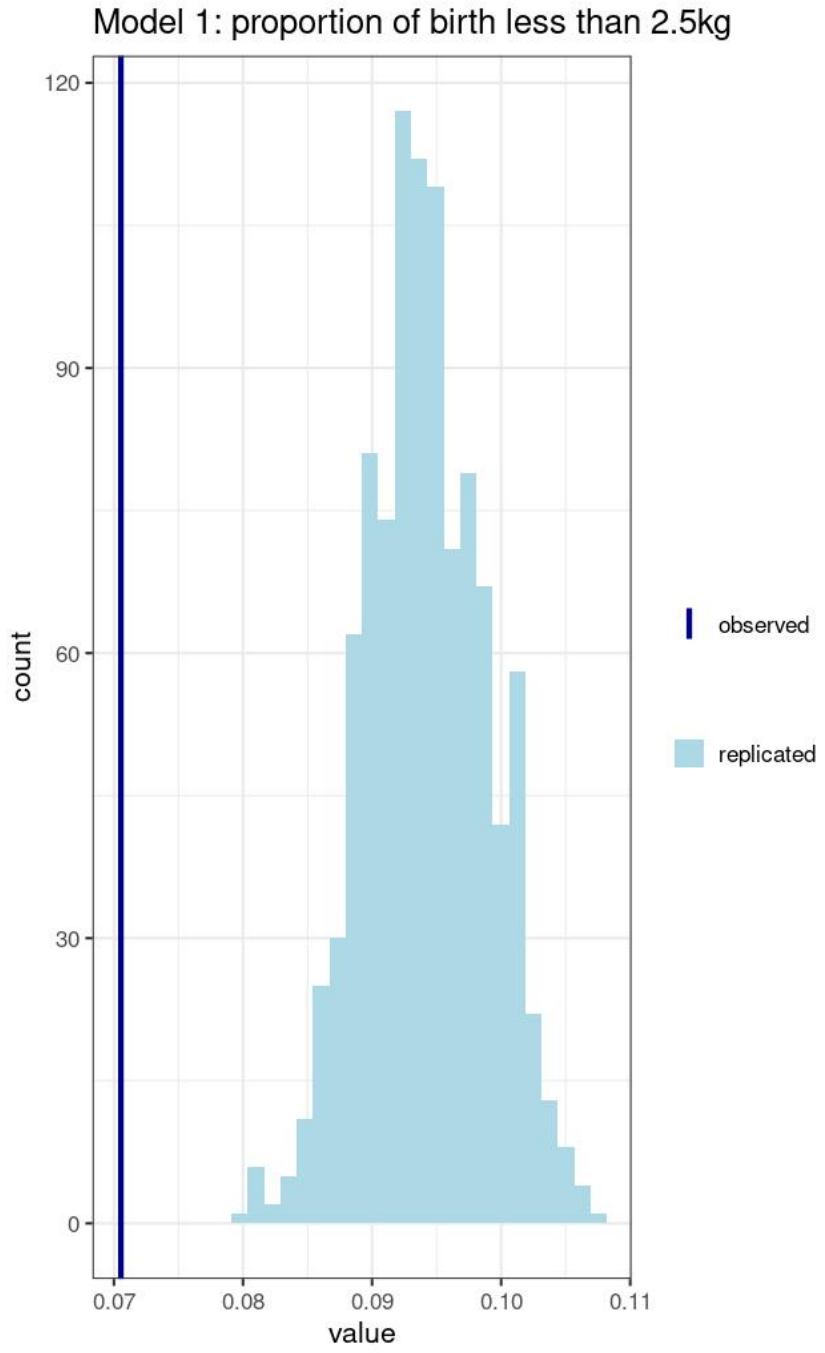


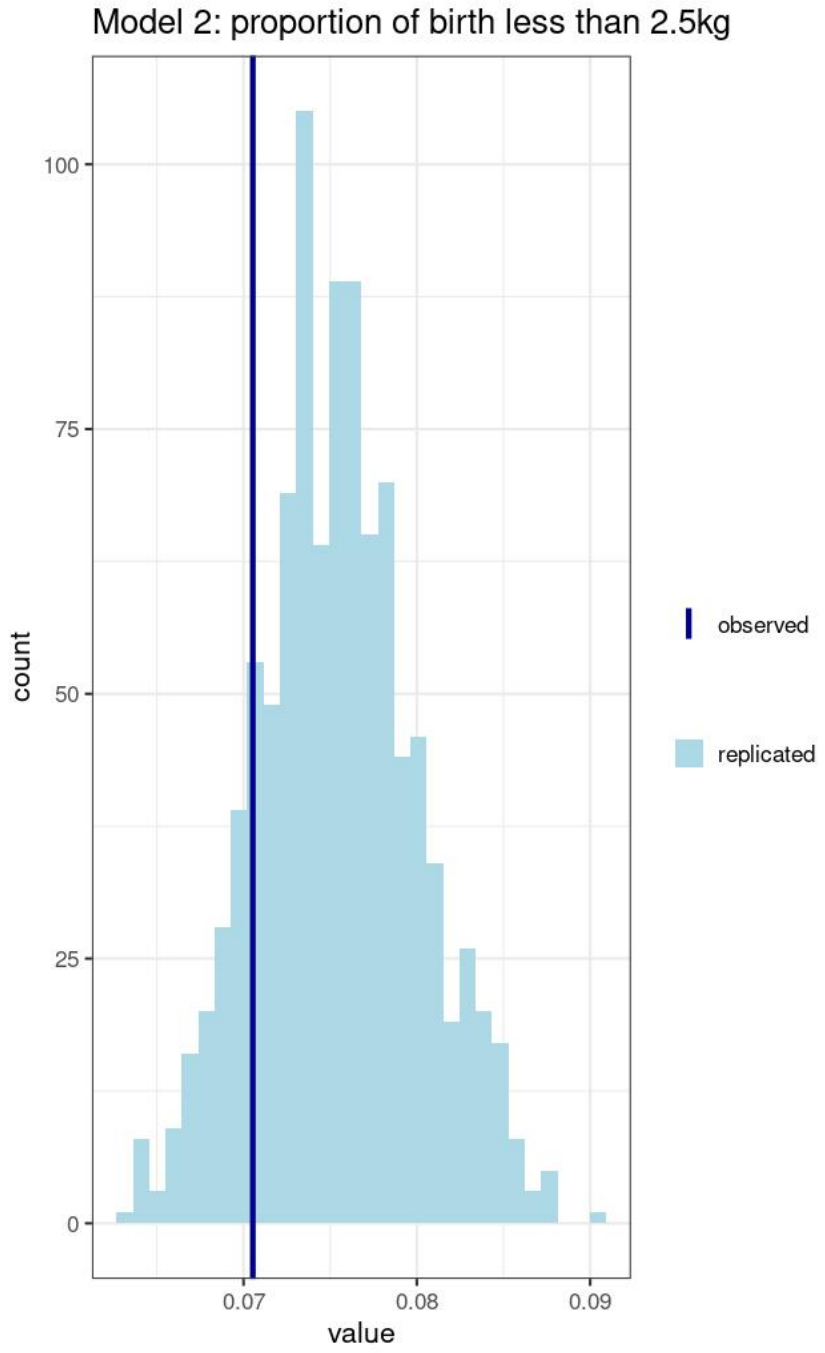
mod3

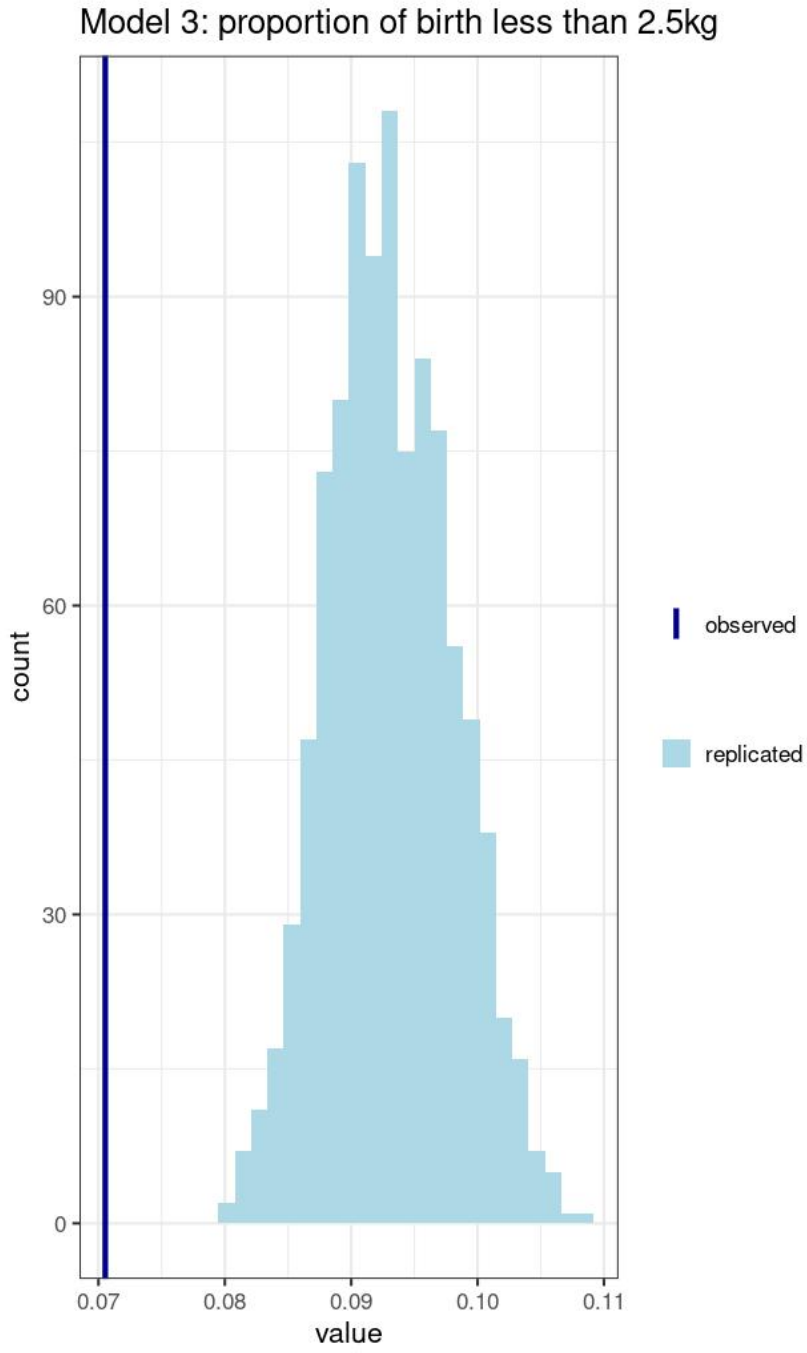
Test Statistics

median of log_weight



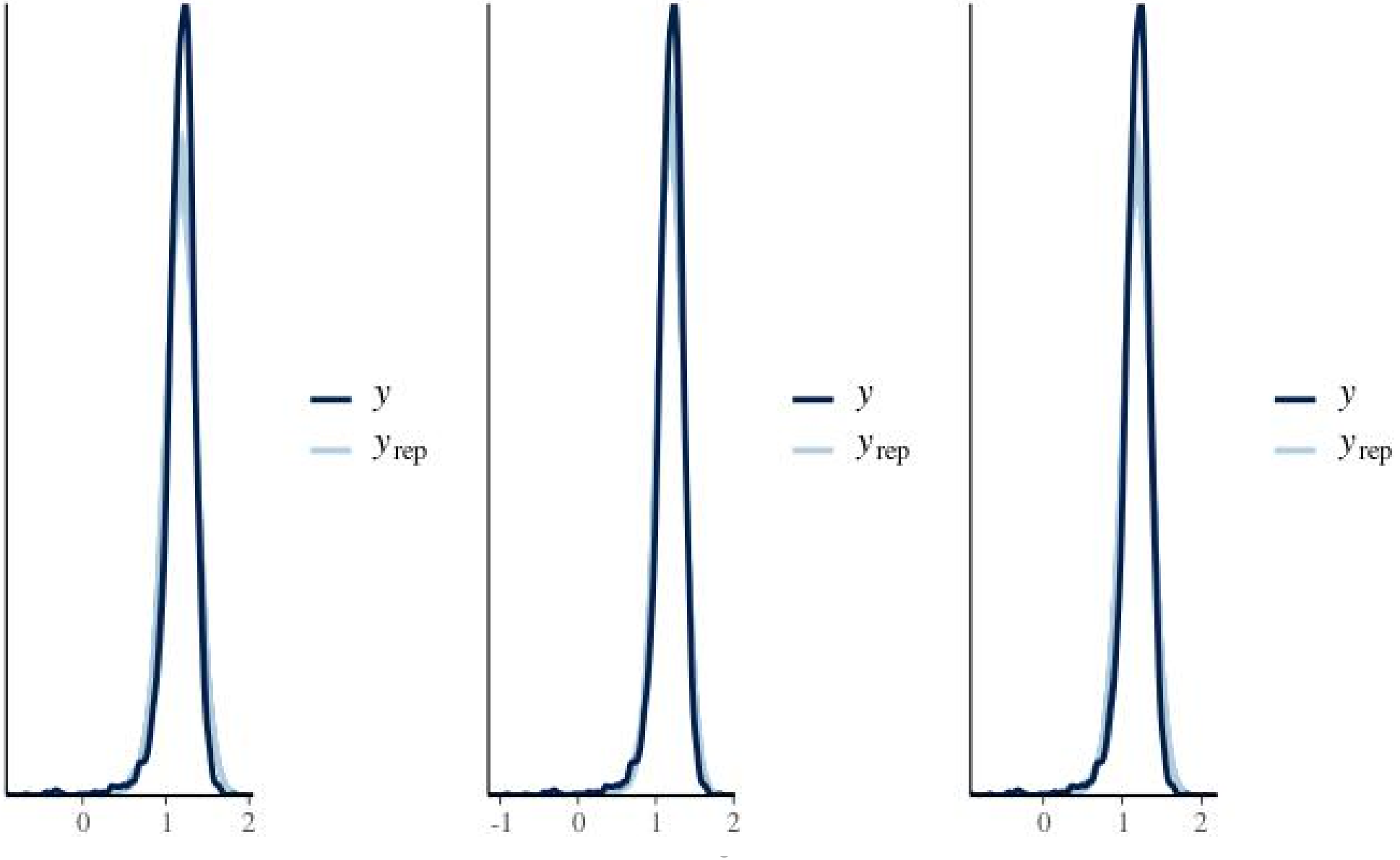

Model 1: proportion of birth less than 2.5kg
count
value
observed
replicated
0
30
60
90
120
0.07
0.08
0.09
0.10
0.11


Model 2: proportion of birth less than 2.5kg
count
value
observed
replicated
0
25
50
75
100
0.07
0.08
0.09

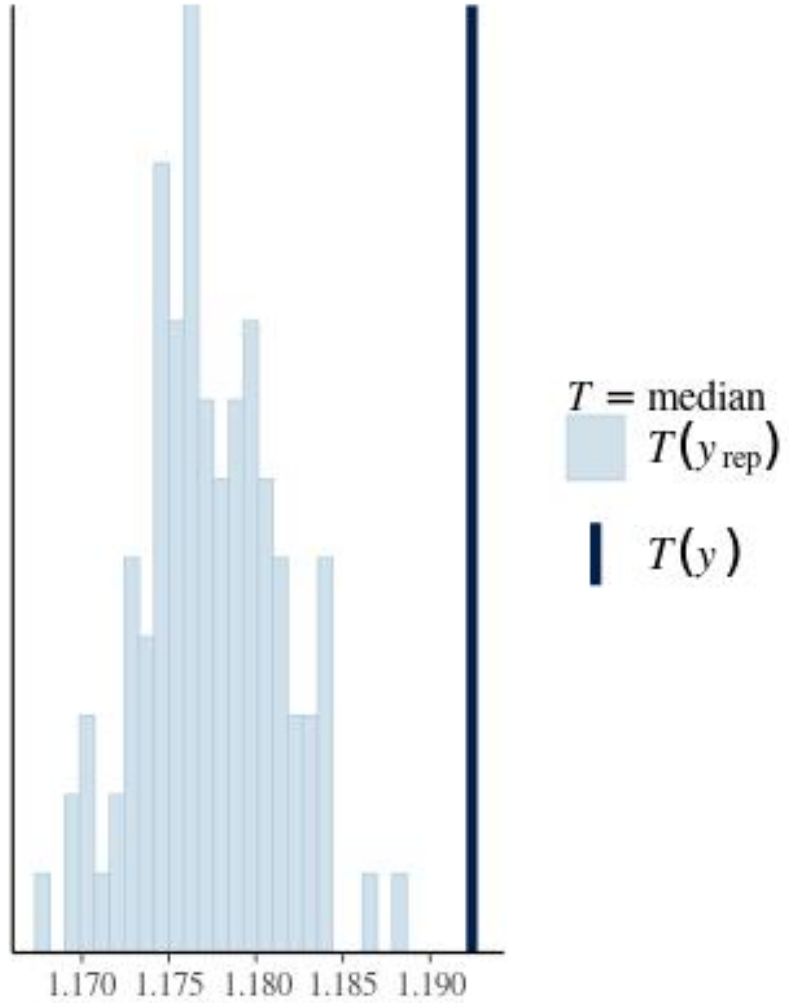

Model 3: proportion of birth less than 2.5kg
count
value
observed
replicated
0
30
60
90
0.07
0.08
0.09
0.10
0.11

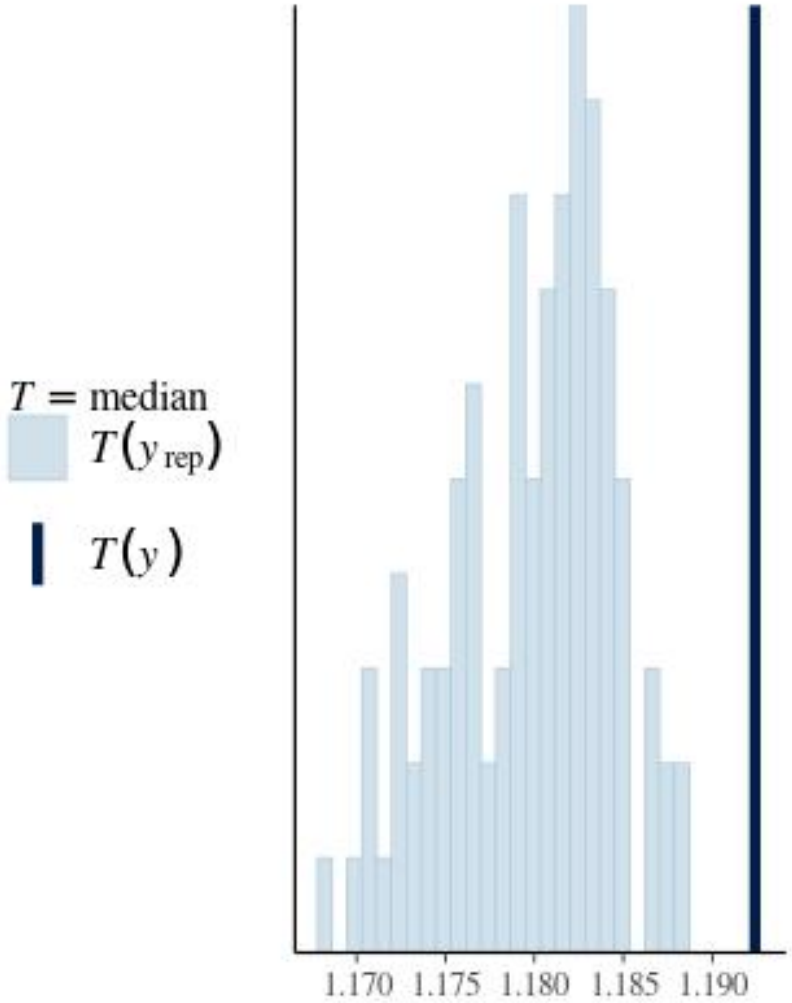
Posterior predictive checks: brms

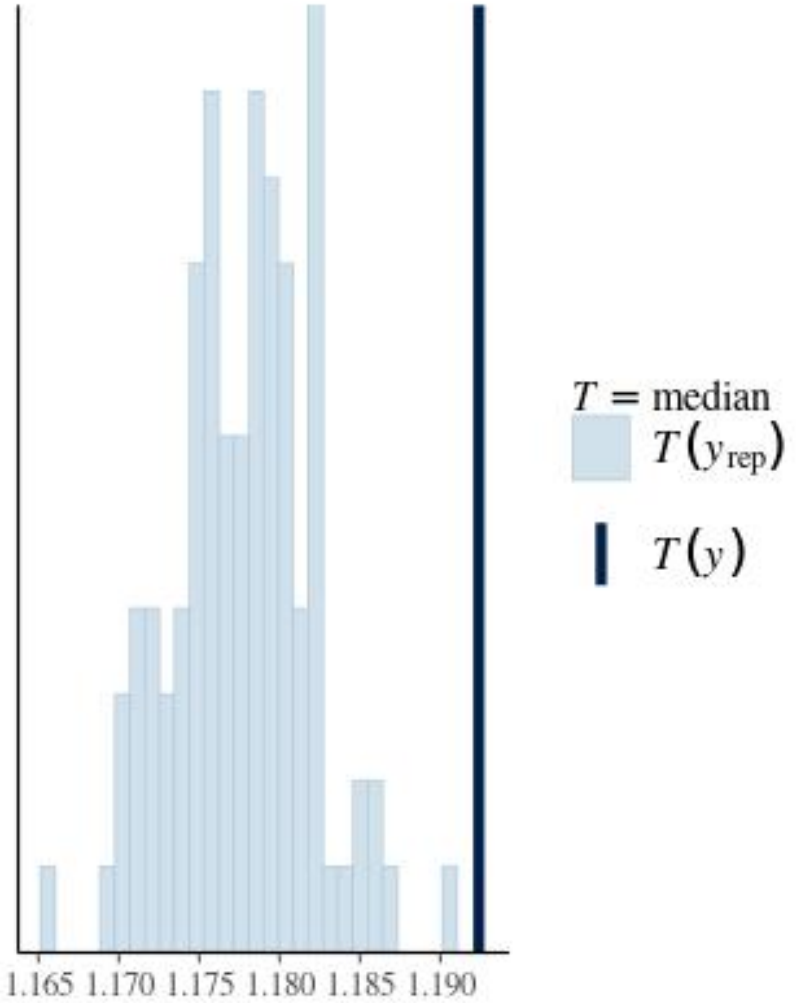
density plot



test statistics







(e) model comparison

- comparing models using leave-one-out cross validation (LOO-CV)
- LOO posterior predictive densities using Pareto Smoothed Importance Sampling (PSIS)
- LOO probability integral transform (LOO-PIT) (bayesplot)

stan

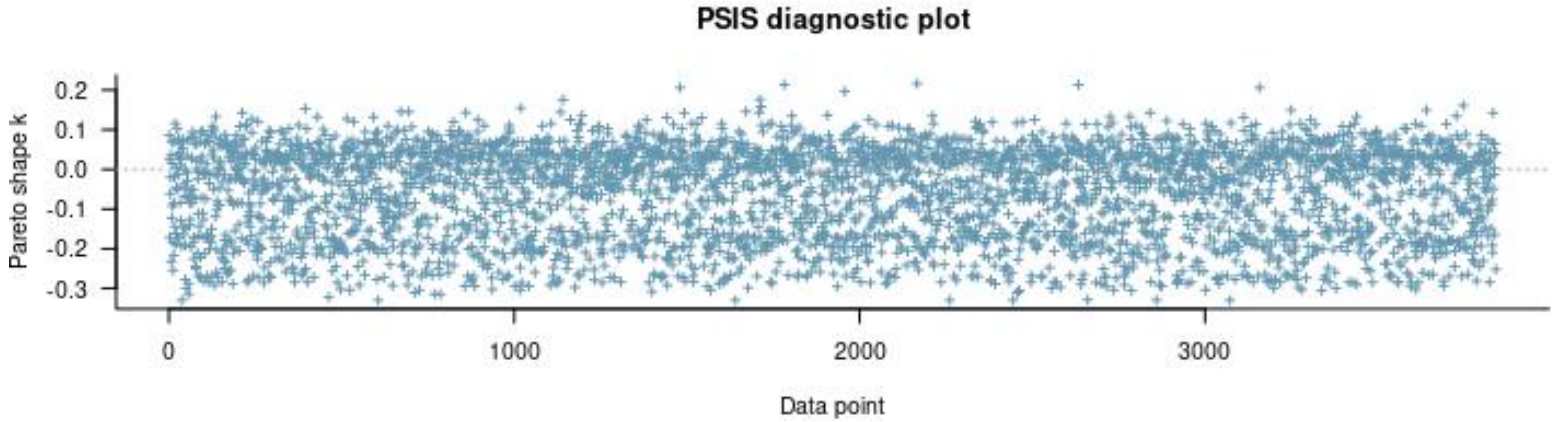
LOO-CV

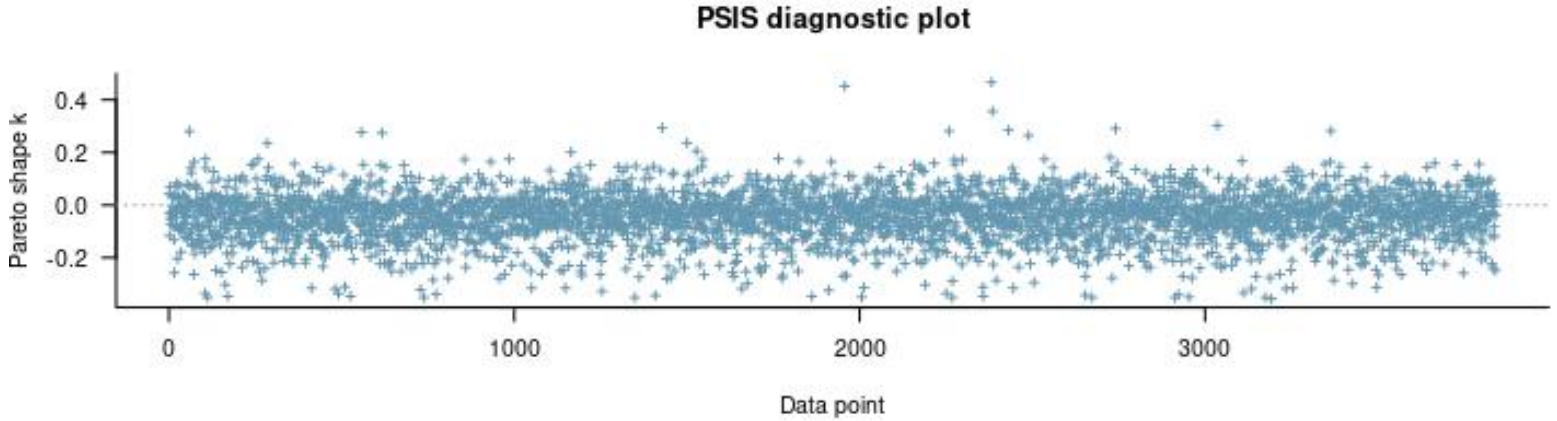
	loo1	loo2	loo3
elpd	1556.8	**1684.7**	1586.8

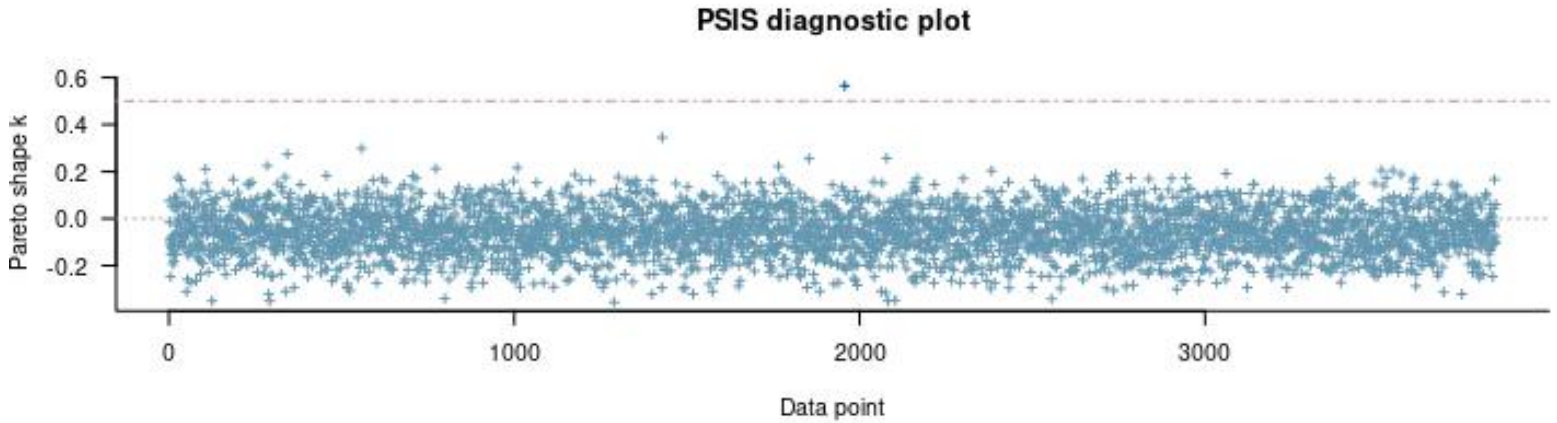
The elpdLOO is higher for Model 2, so it is preferred.

All Pareto k estimates are good (k < 0.5) (not good if k>0.7)

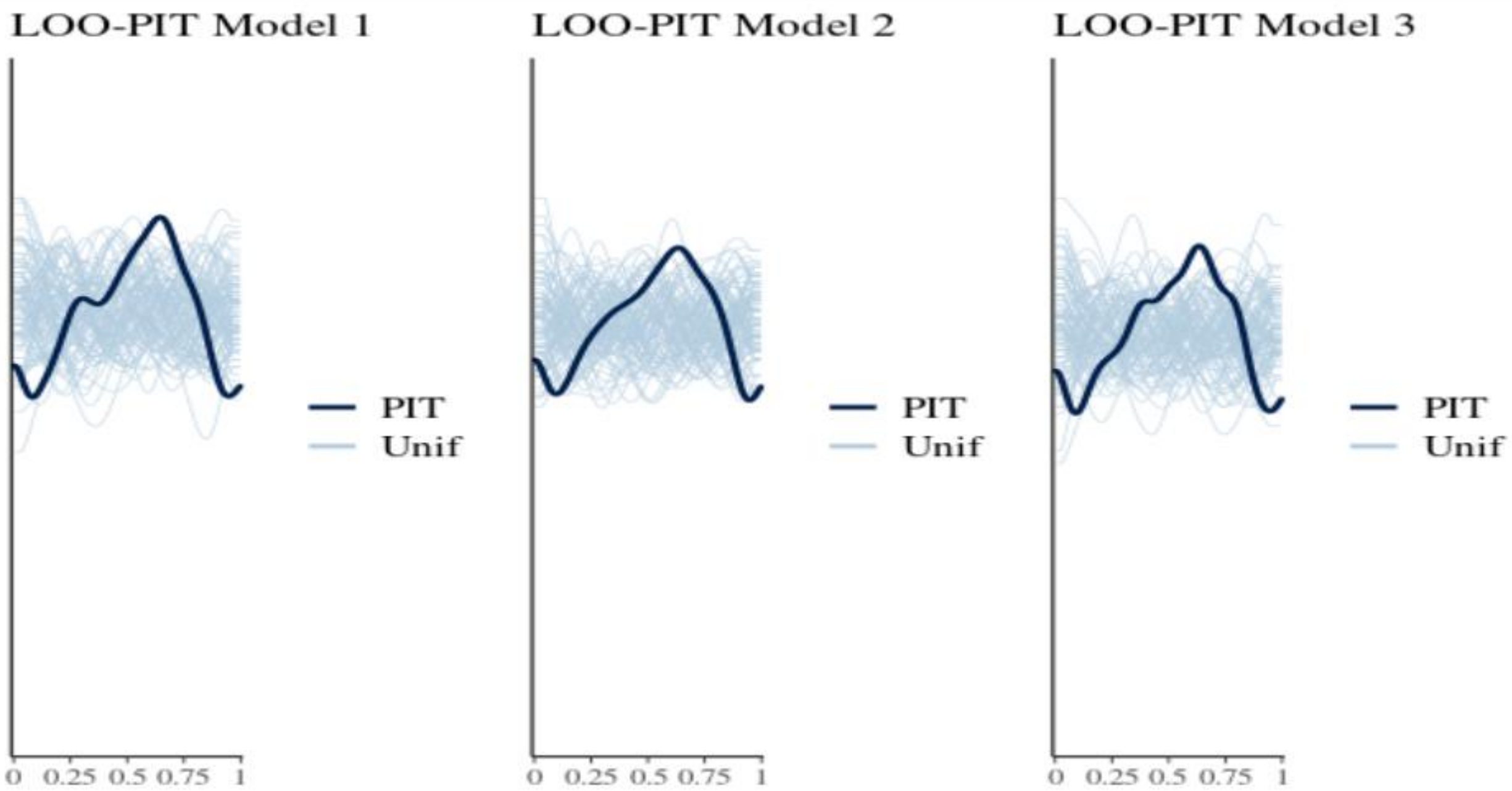
PSIS


PSIS diagnostic plot
Pareto shape k
0.2
0.1
0.0
-0.1
-0.2
-0.3
0
1000
2000
3000
Data point


PSIS diagnostic plot
Pareto shape k
0.4
0.2
0.0
-0.2
0
1000
2000
3000
Data point


PSIS diagnostic plot
Pareto shape k
0.6
0.4
0.2
0.0
-0.2
0
1000
2000
3000
Data point

LOO-PIT


LOO-PIT Model 1
PIT
Unif
0 0.25 0.5 0.75 1
LOO-PIT Model 2
PIT
Unif
0 0.25 0.5 0.75 1
LOO-PIT Model 3
PIT
Unif
0 0.25 0.5 0.75 1

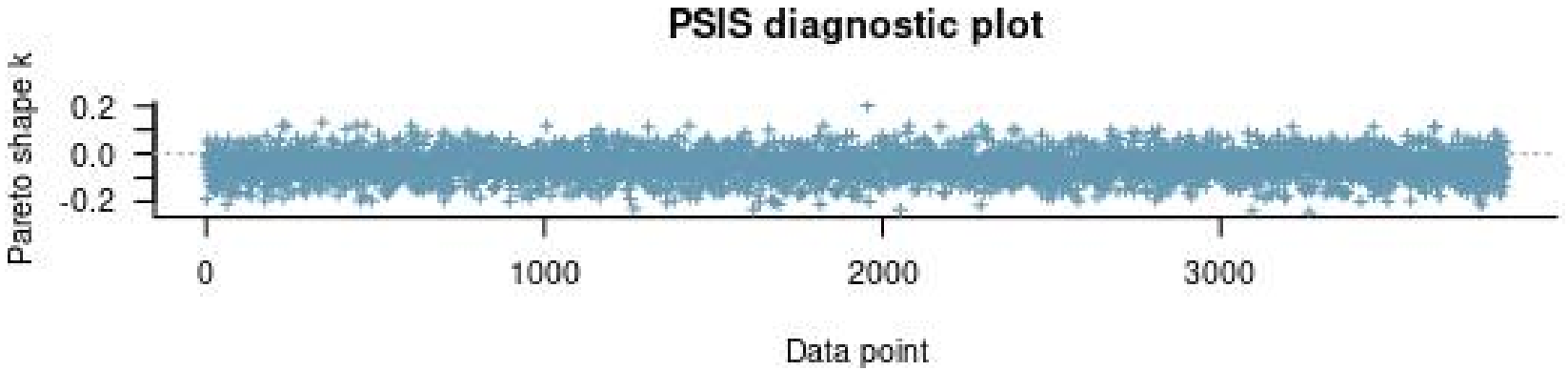
brms

LOO-CV

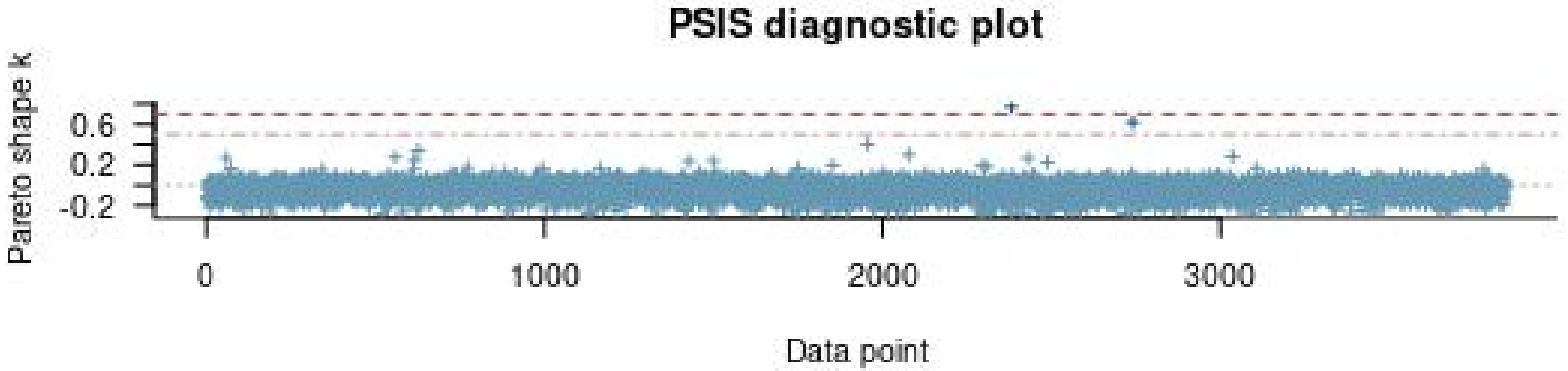
	loo1	loo2	loo3
elpd	1556.8	**1684.7**	1586.8

	loo1b	loo2b	loo2c
elpd	1556.1	**1684.4**	1586.2

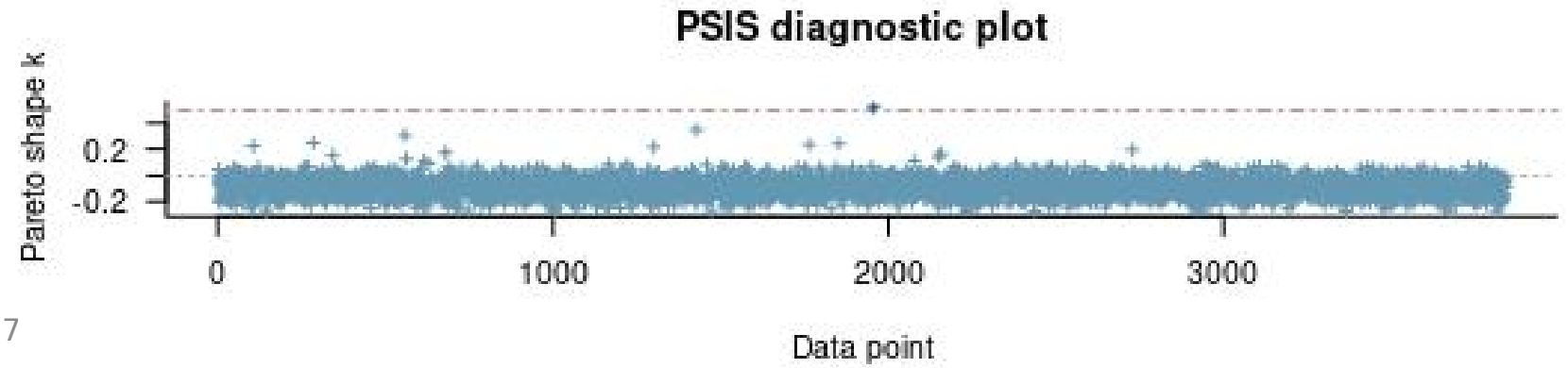
PSIS

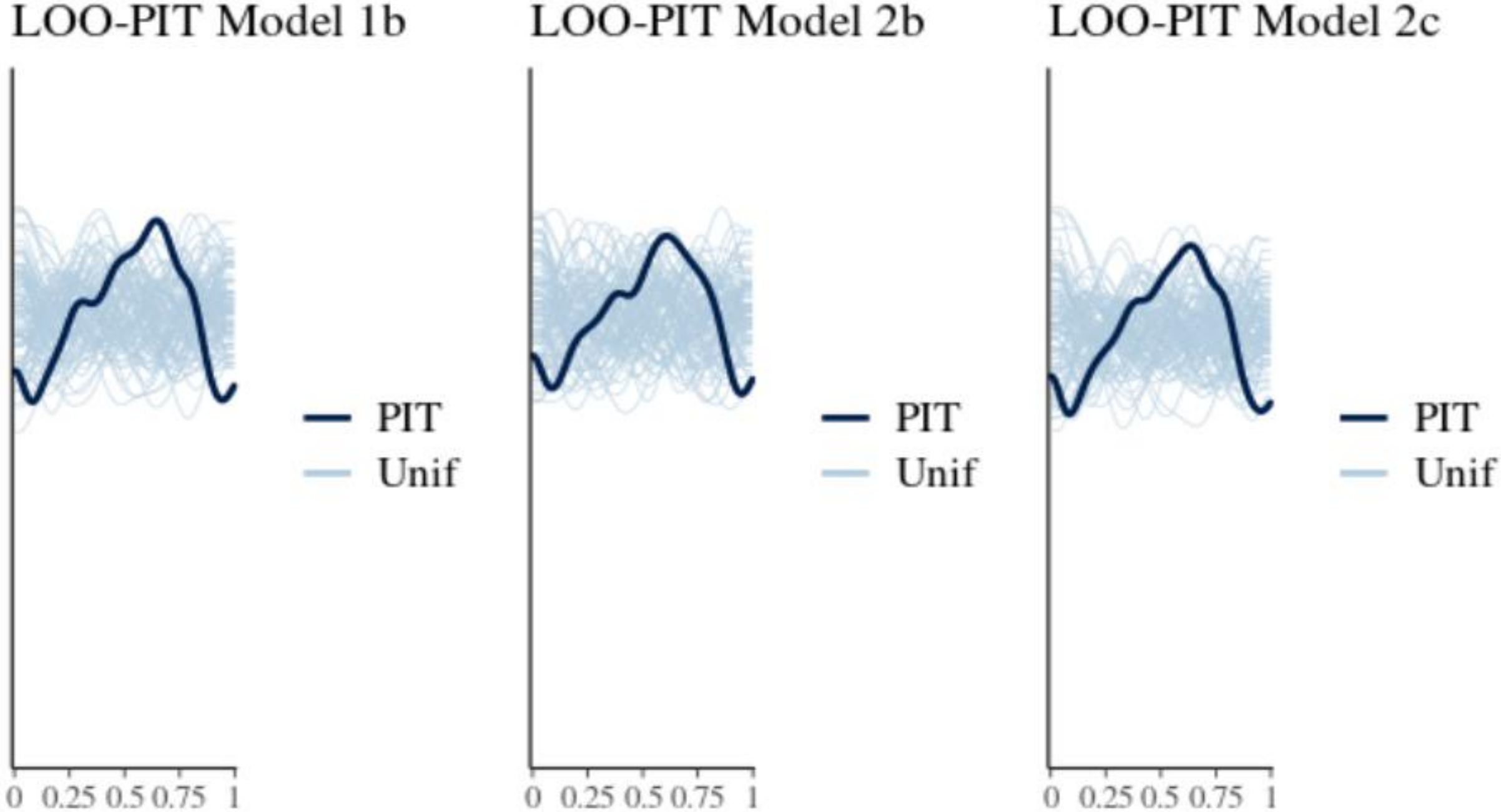


brms



K in mod2b, 2c are still good, but not good as in stan




LOO-PIT Model 1b
LOO-PIT Model 2b
LOO-PIT Model 2c
PIT
Unif
PIT
Unif
PIT
Unif
0 0.25 0.5 0.75 1
0 0.25 0.5 0.75 1
0 0.25 0.5 0.75 1

Practice