The adventures of Bayesian programming in Stan

Samuel Hudec

Matej Bel University
Faculty of Natural Sciences
Department of Mathematics

9. 2. 2018

Topics

- Introduction
- Bayesian approach
- MCMC
- Stan enviroment
- First model in Stan
- Regression model in Stan
- Basic Hierarchical model in Stan
- Conclusion

Introduction

http://mc-stan.org/

Stan is a state-of-the-art platform for statistical modeling and high-performance statistical computation. Thousands of users rely on Stan for statistical modeling, data analysis, and prediction in the social, biological, and physical sciences, engineering, and business.

Introduction

http://mc-stan.org/

Stan is a state-of-the-art platform for statistical modeling and high-performance statistical computation. Thousands of users rely on Stan for statistical modeling, data analysis, and prediction in the social, biological, and physical sciences, engineering, and business.

What you can get?

- full Bayesian statistical inference with MCMC sampling
- approximate Bayesian inference with variational inference
- penalized maximum likelihood estimation with optimization

Introduction (Why Stan?)

Introduction (Why Stan?)

Named by **Stanislaw Ulam**.

Who was a scientist in the fields of mathematics and nuclear physics, also he invated the Monte Carlo method to computation, ...

Introduction (Stanislaw Ulam)

Introduction (Stan interfaces)

- RStan (R)
- PyStan (Phyton)
- CmdStan (shell, command-line terminal)
- MatlabStan (MATLAB)
- StataStan (Stata)
- MathematicaStan (Mathemtatica)

Introduction (Stan interfaces)

- RStan (R)
- PyStan (Phyton)
- CmdStan (shell, command-line terminal)
- MatlabStan (MATLAB)
- StataStan (Stata)
- MathematicaStan (Mathemtatica)

Higher-Level Interface RStanArm provides an R formula interface for Bayesian regression modeling.
etc.

Bayesian approach

Based on Bayesian theorem (1763)

$$\boldsymbol{p(\theta|D)} = \frac{\boldsymbol{p(D|\theta)}p(\theta)}{\int_\Theta \boldsymbol{p(D|\theta)p(\theta)}\,\mathrm{d}\,\boldsymbol{\theta}}$$

Model (likelihood) $= p(D|\theta)$
Prior $= p(\theta)$ (conjugate, noninformative...)
evidence $= \int_\Theta p(D|\theta)p(\theta)\,\mathrm{d}\,\theta \triangleq p(D)$
Posterior $= p(\theta|D)$

Bayesian approach

Based on Bayesian theorem (1763)

$$\boldsymbol{p(\theta|D)} = \frac{\boldsymbol{p(D|\theta)}p(\theta)}{\int_\Theta \boldsymbol{p(D|\theta)}\boldsymbol{p(\theta)}\,\mathrm{d}\,\boldsymbol{\theta}}$$

Model (likelihood) $= p(D|\theta)$
Prior $= p(\theta)$ (conjugate, noninformative...)
evidence $= \int_\Theta p(D|\theta)p(\theta)\,\mathrm{d}\,\theta \triangleq p(D)$
Posterior $= p(\theta|D)$

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

Bayesian approach

Based on Bayesian theorem (1763)

$$\boldsymbol{p(\theta|D)} = \frac{\boldsymbol{p(D|\theta)}p(\theta)}{\int_\Theta \boldsymbol{p(D|\theta)p(\theta)}\,\mathrm{d}\,\boldsymbol{\theta}}$$

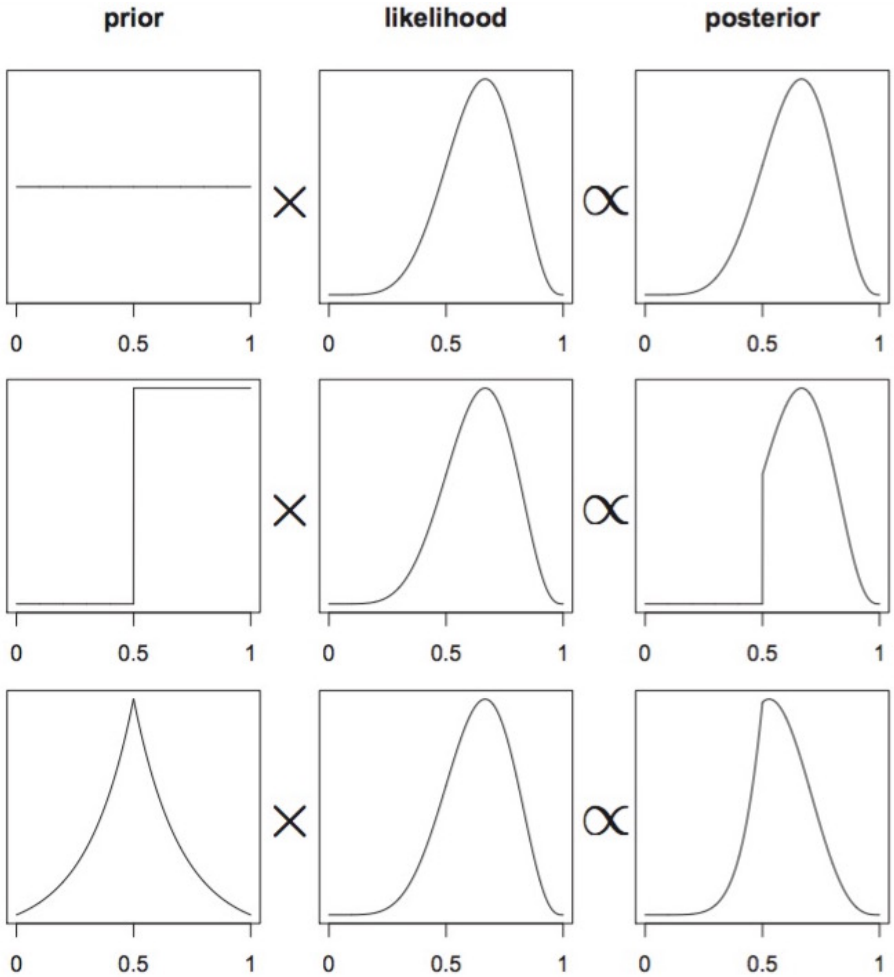
Model (likelihood) $= p(D|\theta)$
Prior $= p(\theta)$ (conjugate, noninformative...)
evidence $= \int_\Theta p(D|\theta)p(\theta)\,\mathrm{d}\,\theta \triangleq p(D)$
Posterior $= p(\theta|D)$

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

Posterior predictive density (prediction, validation,...)

$$p(\tilde{D}|D) = \int_\Theta p(\tilde{D}|\theta)p(\theta|D)\,\mathrm{d}\,\theta$$

Bayesian approach



source: McElreath, R., *Statistical Rethinking* CRC Press 2016.

Bayesian approach

$$\frac{p(D|\theta)p(\theta)}{\int_{\Theta} p(D|\theta)p(\theta)\,\mathrm{d}\,\theta} \xrightarrow{???} p(\theta|D)$$

Bayesian approach

$$\frac{p(D|\theta)p(\theta)}{\int_{\Theta} p(D|\theta)p(\theta)\,\mathrm{d}\,\theta}$$

$$p(\theta|D)$$

MCMC

"Markov Chain Monte Carlo methods are a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after a number of steps is then used as a sample of the desired distribution. The quality of the sample improves as a function of the number of steps."

MCMC

"Markov Chain Monte Carlo methods are a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after a number of steps is then used as a sample of the desired distribution. The quality of the sample improves as a function of the number of steps."

- How many samples do you need?
- How many chains we need?

MCMC

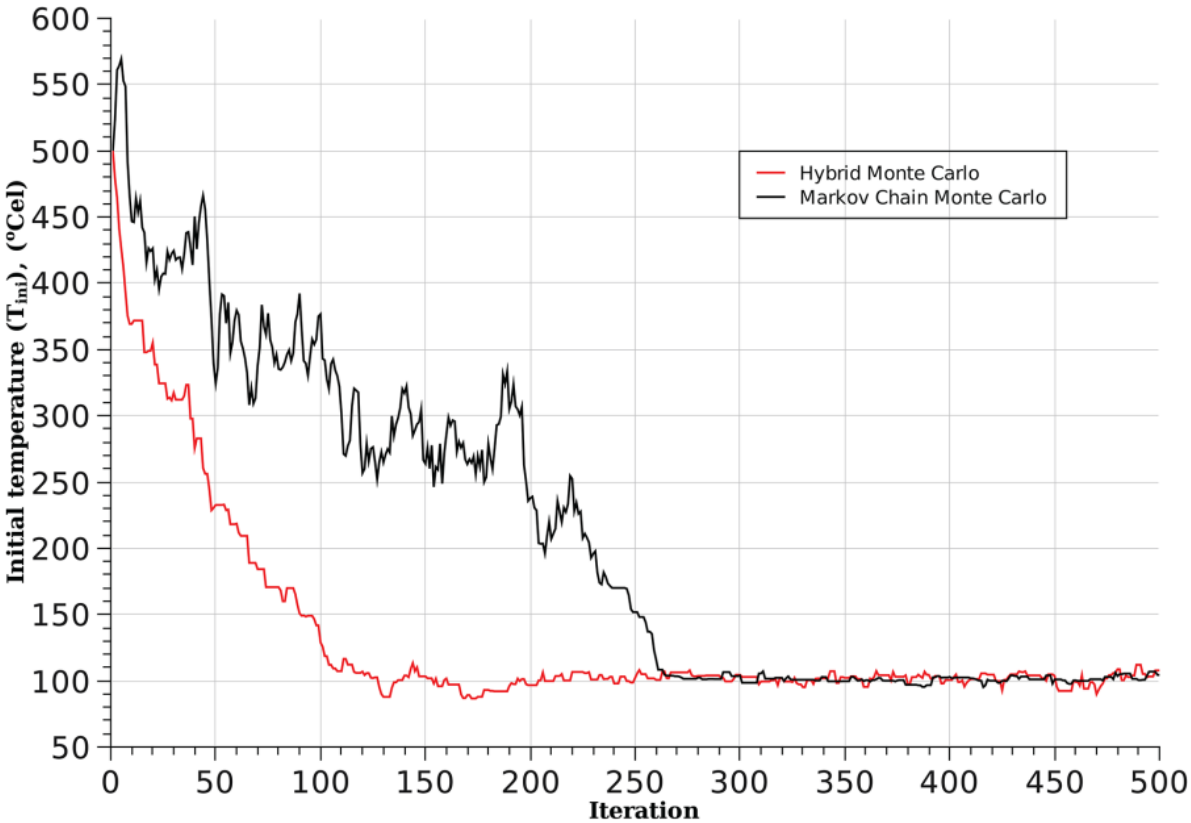
"Markov Chain Monte Carlo methods are a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its equilibrium distribution. The state of the chain after a number of steps is then used as a sample of the desired distribution. The quality of the sample improves as a function of the number of steps."

- How many samples do you need?
- How many chains we need?

for reggresion motto

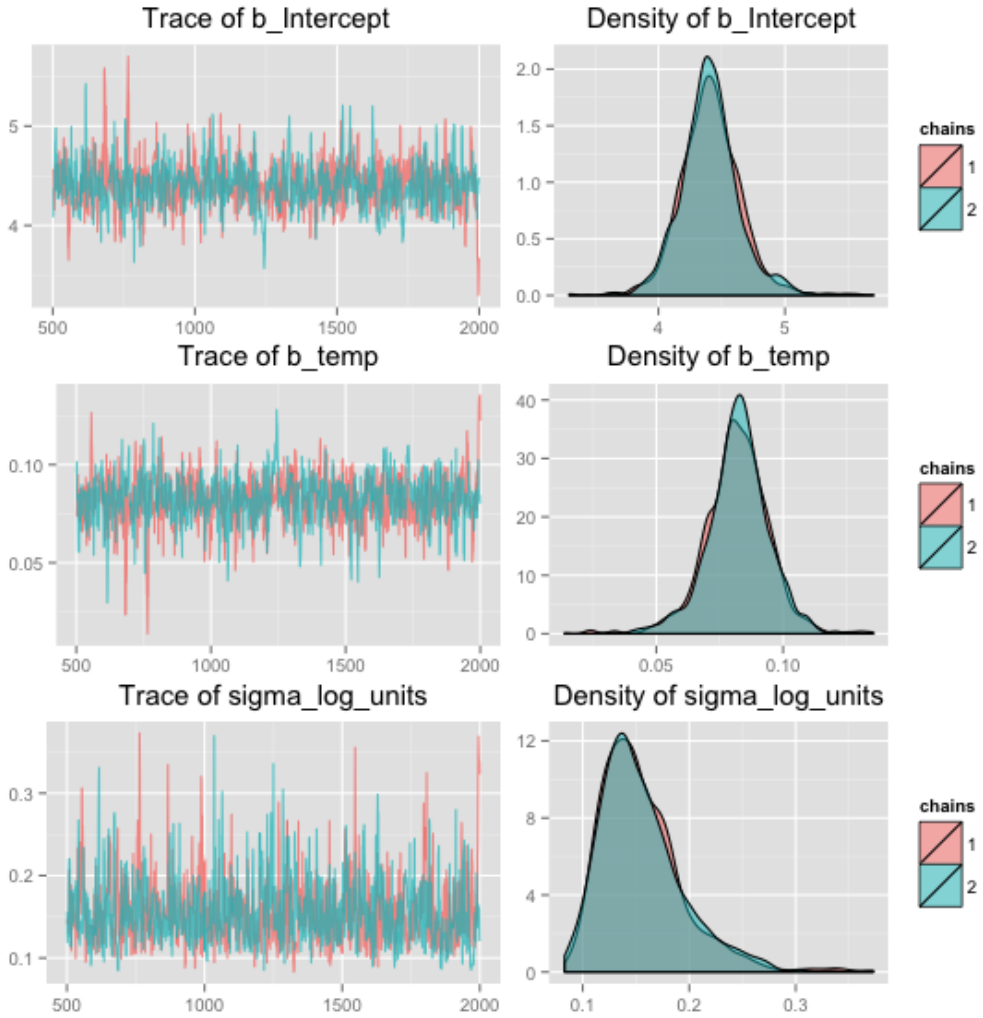
"four short chains to check, one long chain for inference"

MCMC



sourse: http://heattransfer.asmedigitalcollection.asme.org/

MCMC


Trace of b_Intercept
Density of b_Intercept
Trace of b_temp
Density of b_temp
Trace of sigma_log_units
Density of sigma_log_units
chains
1
2

MCMC (diagnostic)

The default diadnostic output from Stan includes two metrics

- `n_eff` measure of the effective number of samples
- `Rhat` $\hat{R}$ Gelman-Rubin convergence diagnostic (above 1 usually indicates that the chain has not yet converged)

MCMC (diagnostic)

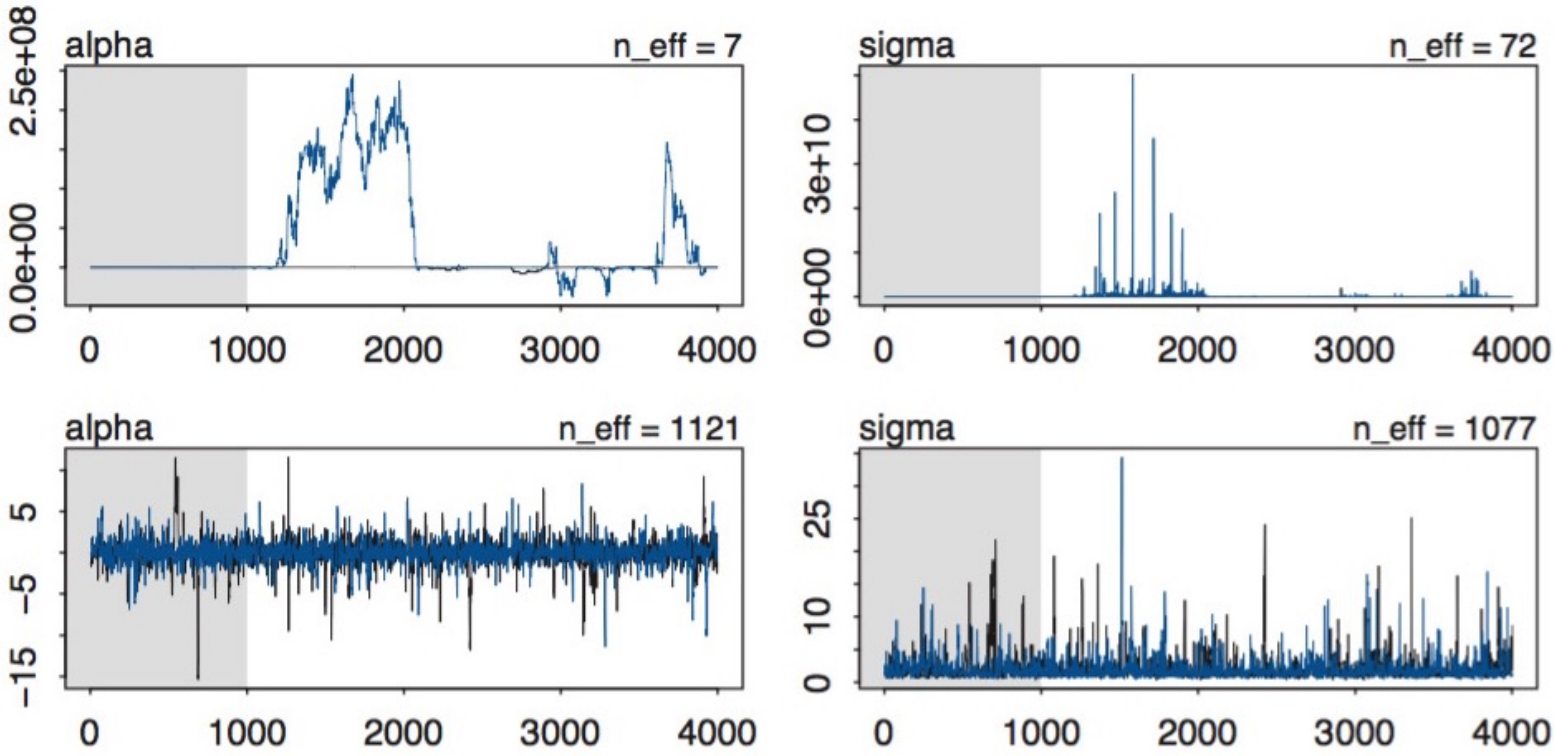
The default diadnostic output from Stan includes two metrics

- `n_eff` measure of the effective number of samples
- `Rhat` $\hat{R}$ Gelman-Rubin convergence diagnostic (above 1 usually indicates that the chain has not yet converged)

Common problems with unstacionarity or "Timing a wild chain"

- broad, tlat regions of the posterior density
- non-identifiable parameters

MCMC (diagnostic)



source: McElreath, R., *Statistical Rethinking* CRC Press 2016.

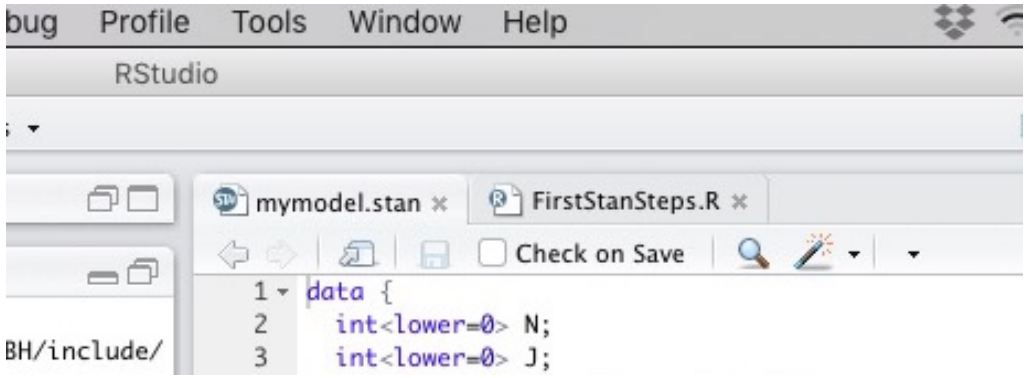
Stan enviroment (before we start)

- build model in Stan file `mymodel.stan`
- use R (or another enviroment) for load data set, runnig Stan model, print fit, diagnostics, plotting, ...

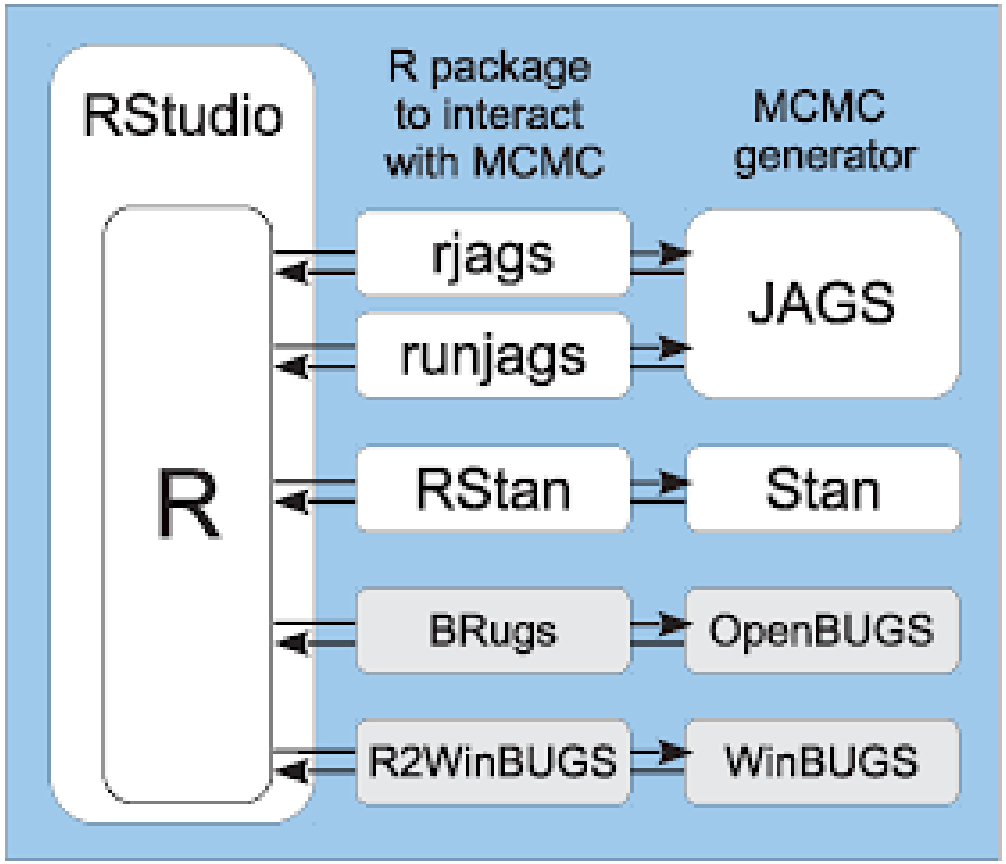
Stan enviroment (before we start)

- build model in Stan file `mymodel.stan`
- use R (or another enviroment) for load data set, runnig Stan model, print fit, diagnostics, plotting, ...

strongly recommend installing RStudio which has basic support for .stan file types and syntax highlighting for Stan



Stan enviroment



source: Kruschke, J., K., *Doing Bayesian Data Analysis* Elsevier Inc. 2015

Stan enviroment (General structure of Stan model specification)

```
data {
... declarations ...
}
```

Stan enviroment (General structure of Stan model specification)

```
data {
... declarations ...
}
transformed data {
... declarations ... statements ...
}
```

Stan enviroment (General structure of Stan model specification)

```
data {
... declarations ...
}
transformed data {
... declarations ... statements ...
}
parameters {
... declarations ...
}
```

Stan enviroment (General structure of Stan model specification)

```
data {
... declarations ...
}
transformed data {
... declarations ... statements ...
}
parameters {
... declarations ...
}
transformed parameters {
... declarations ... statements ...
}
```

Stan enviroment (General structure of Stan model specification)

```
data {
... declarations ...
}
transformed data {
... declarations ... statements ...
}
parameters {
... declarations ...
}
transformed parameters {
... declarations ... statements ...
}
model {
... declarations ... statements ...
}
```

Stan enviroment (General structure of Stan model specification)

```
data {
... declarations ...
}
transformed data {
... declarations ... statements ...
}
parameters {
... declarations ...
}
transformed parameters {
... declarations ... statements ...
}
model {
... declarations ... statements ...
}
generated quantities{
... declarations ... statements ...
}
```

First model in Stan

Let start with model for estimating a Bernoulli parameter (tossing coin example)

- $y|\theta \sim bernoulli(\theta)$ $\quad$ $likelihood$ $\quad$ $(model)$
- $\theta \sim beta(?, ?)$ $\quad$ $prior$

First model in Stan

Let start with model for estimating a Bernoulli parameter (tossing coin example)

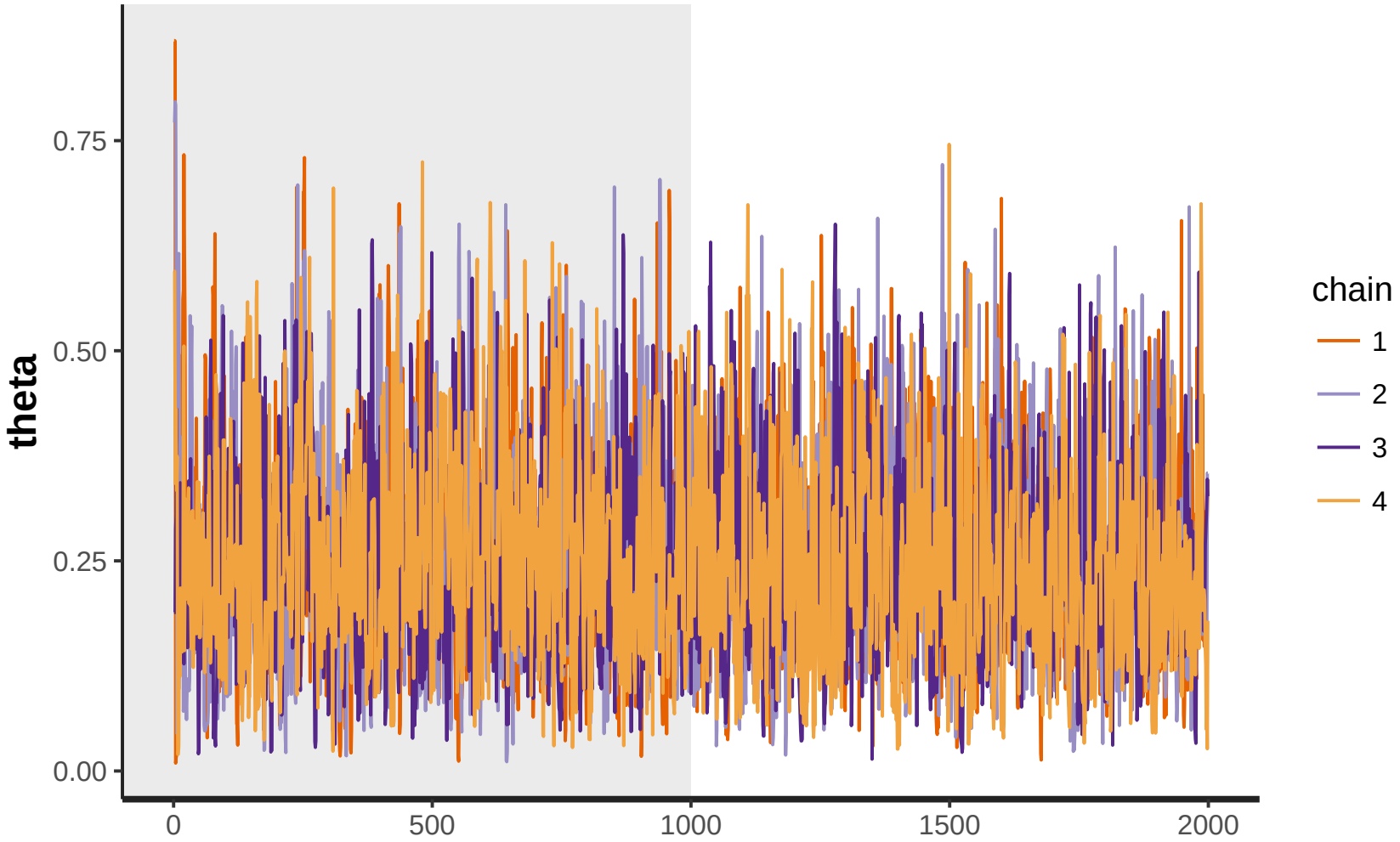
- $y|\theta \sim bernoulli(\theta)$ *likelihood* *(model)*
- $\theta \sim beta(?, ?)$ *prior*

```
data {
  int<lower=0> N; // N >= 0
  int<lower=0, upper=1> y[N]; // y[n] in [0, 1]
}
parameters {
  real<lower=0, upper=1> theta; // theta in [0, 1]
}
model {
  theta ~ beta(1, 1); // prior
  y ~ bernoulli(theta); // likelihood
}
```

First model in R

```
> N <- 10
> y <- c(0, 1, 0, 0, 0, 0, 0, 0, 0, 1)
> berdat <- list(N=N,y=y)
> fit<-stan(file="mymodel.stan",data=berdat)
> print(fit)

Inference for Stan model: mymodel.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

       mean se_mean sd 2.5% 25% 50% 75% 97.5% n_eff Rhat
theta 0.25 0.00 0.12 0.06 0.16 0.23 0.33 0.51 1311 1
lp__ -7.27 0.02 0.73 -9.35 -7.46 -7.00 -6.80 -6.75 1370 1

> traceplot(fit,inc_warmup=T)
```

First model in Stan

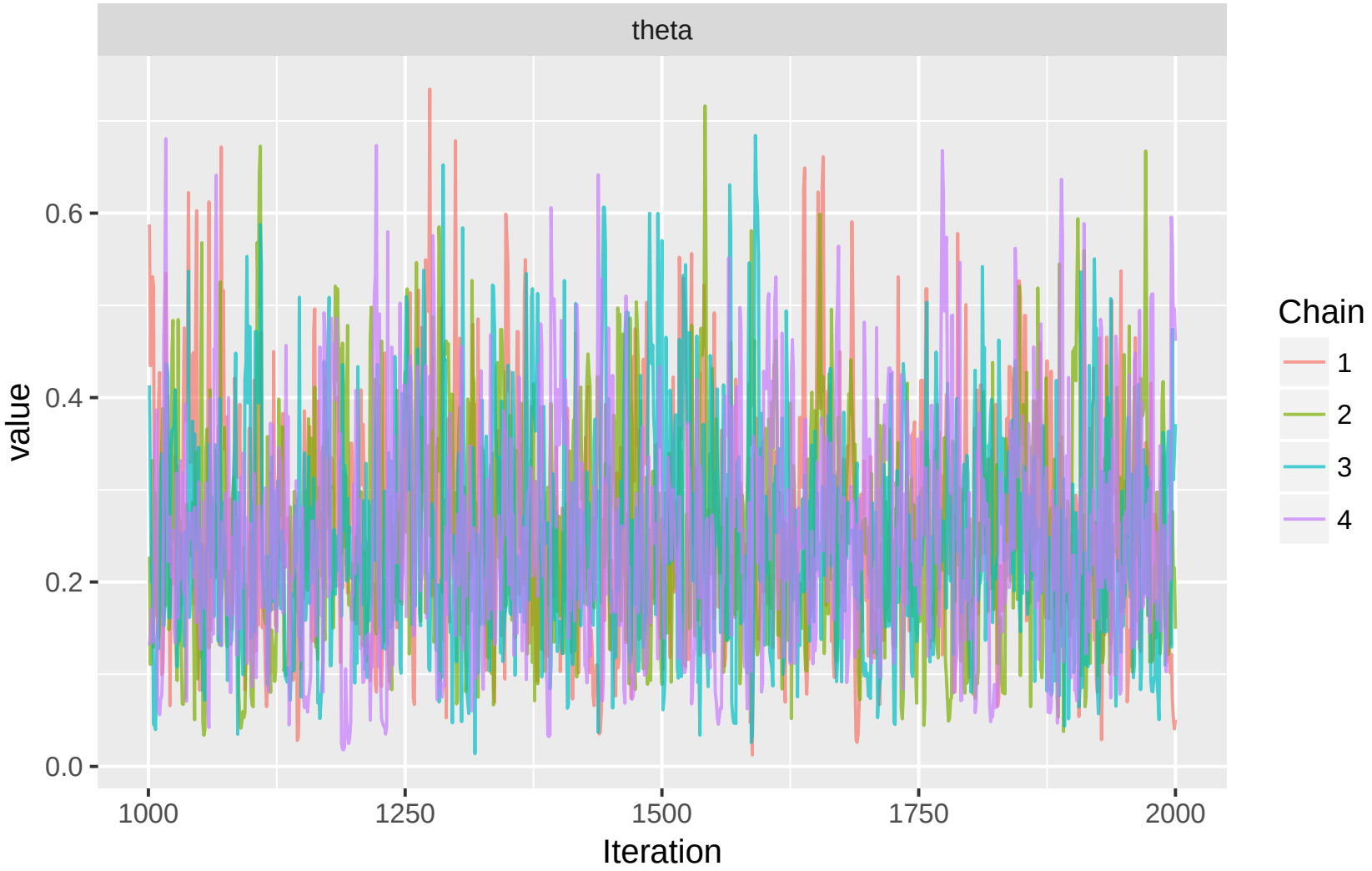


First model in Stan

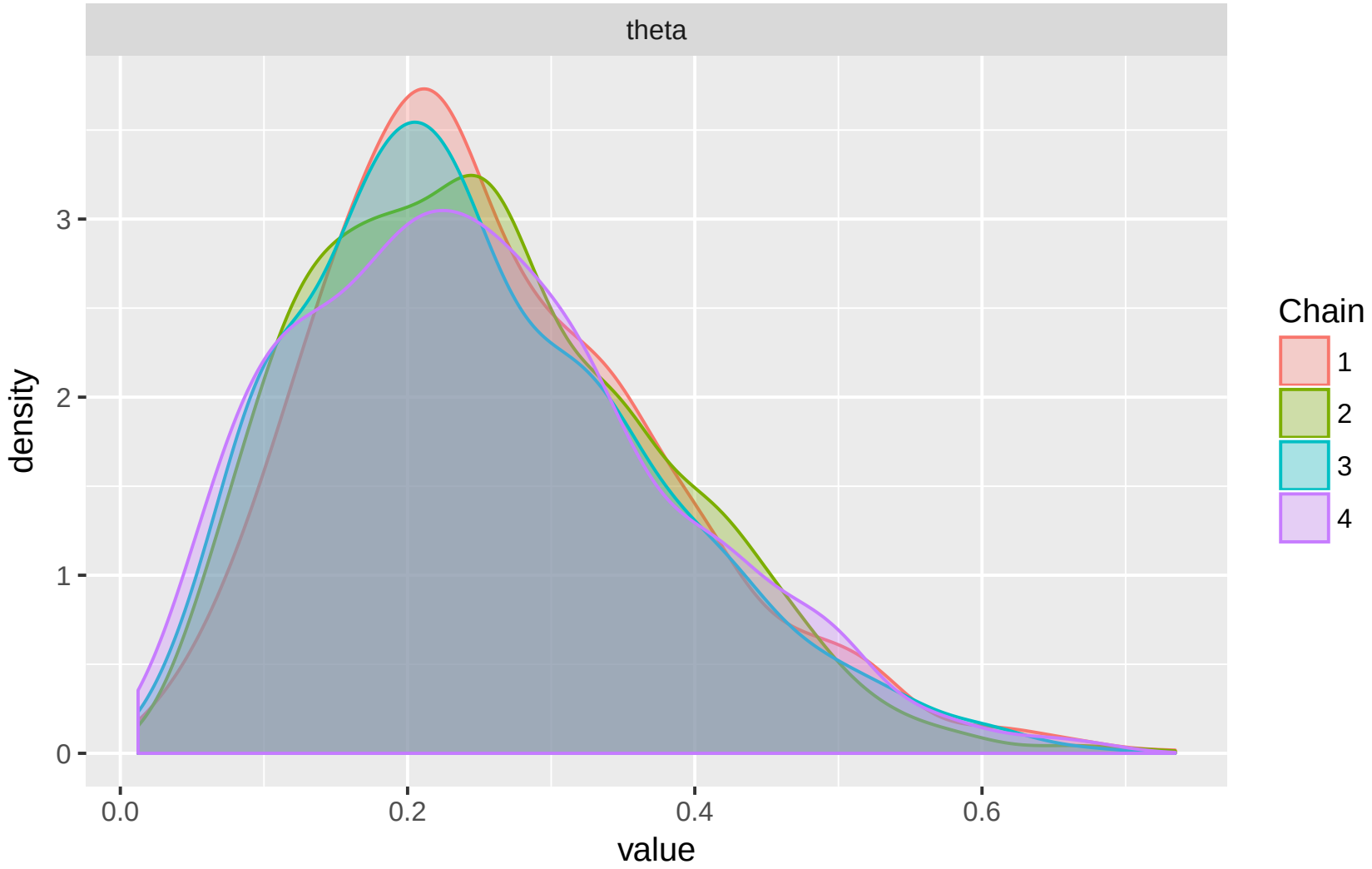
Advanced plotting with `library(ggmcmc)`

```
> S <- ggs(fit)
> ggs_histogram(S)
> ggs_traceplot(S)
> ggs_density(S)
> ggs_compare_partial(S)
> ggs_running(S)
> ggs_autocorrelation(S)
```

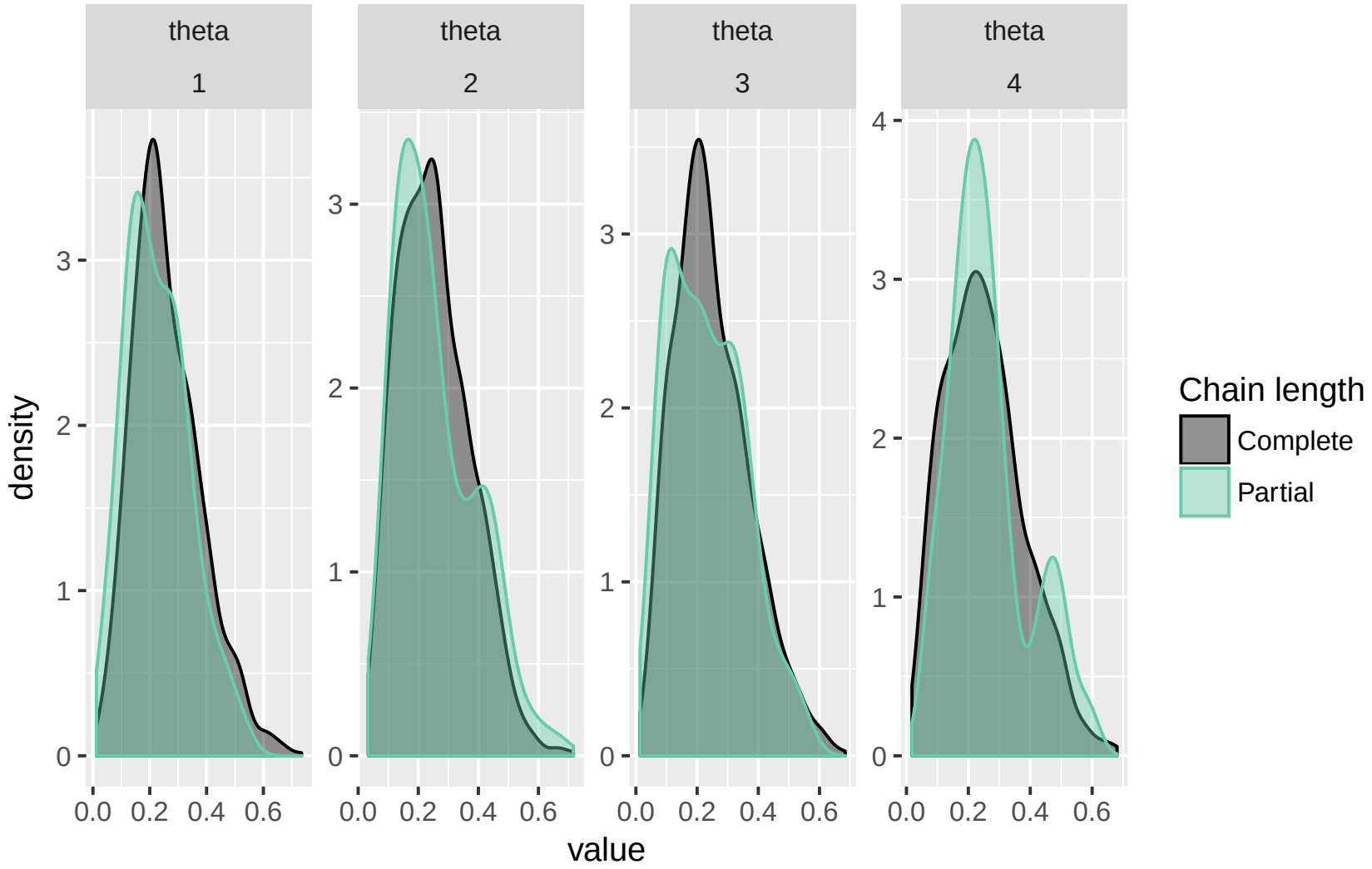
First model in Stan



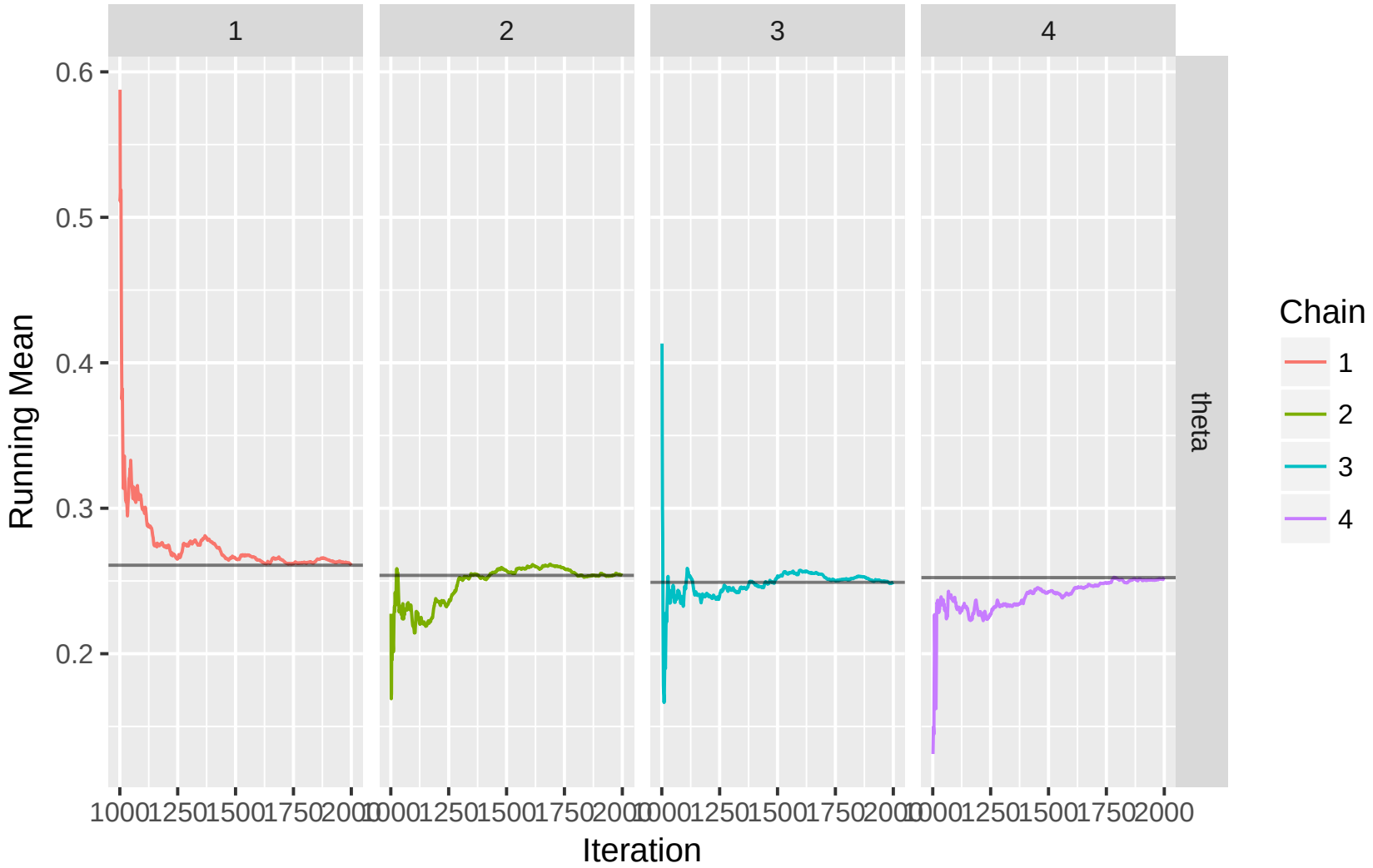
First model in Stan



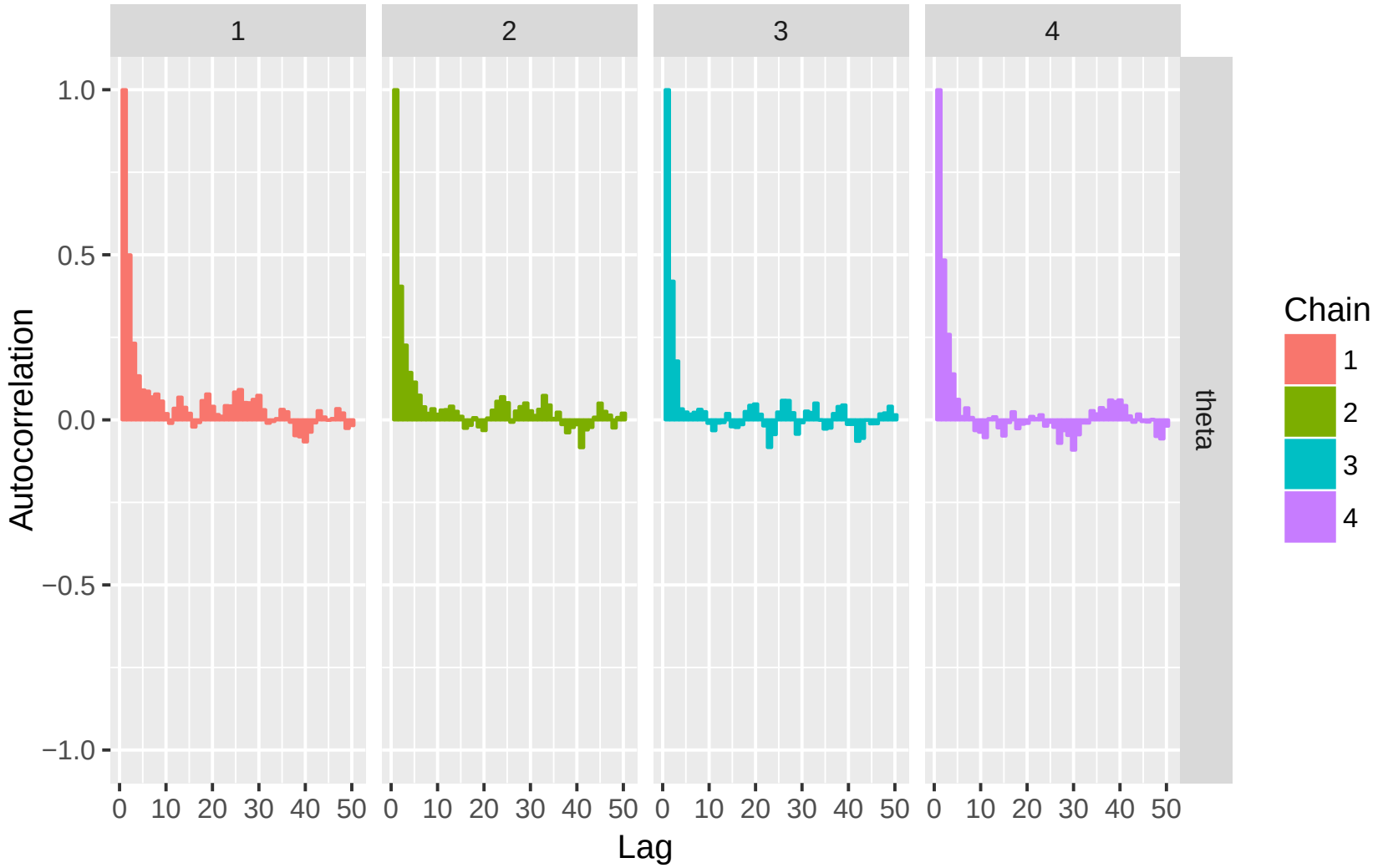
First model in Stan



First model in Stan



First model in Stan



Regression model in Stan

Frequentist:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$$

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$$

- $\hat{\boldsymbol{\beta}}_{ML} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$
- $\hat{\sigma}^2_{ML} = \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n}$
- $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{ML}$

$$\hat{\boldsymbol{\beta}}_{ML} \sim N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$$

$$\hat{\sigma}^2_{ML} \sim Gamma\left(\frac{n-k}{2}, \frac{2\sigma^2}{n}\right)$$

Regression model in Stan

Formal bayesian justification:

$$p(\boldsymbol{\beta}, \sigma | y, \boldsymbol{X}) \propto p(y | \boldsymbol{X}, \boldsymbol{\beta}, \sigma) p(\boldsymbol{\beta}, \sigma)$$

Regression model in Stan

Formal bayesian justification:

$$p(\boldsymbol{\beta}, \sigma | y, \boldsymbol{X}) \propto p(y | \boldsymbol{X}, \boldsymbol{\beta}, \sigma) p(\boldsymbol{\beta}, \sigma)$$

Bayesians:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\boldsymbol{y} | \boldsymbol{\beta}, \sigma, \boldsymbol{X} \sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$ *likelihood*
- $\boldsymbol{\beta} \sim$??? *prior*
- $\sigma \sim$??? *prior*

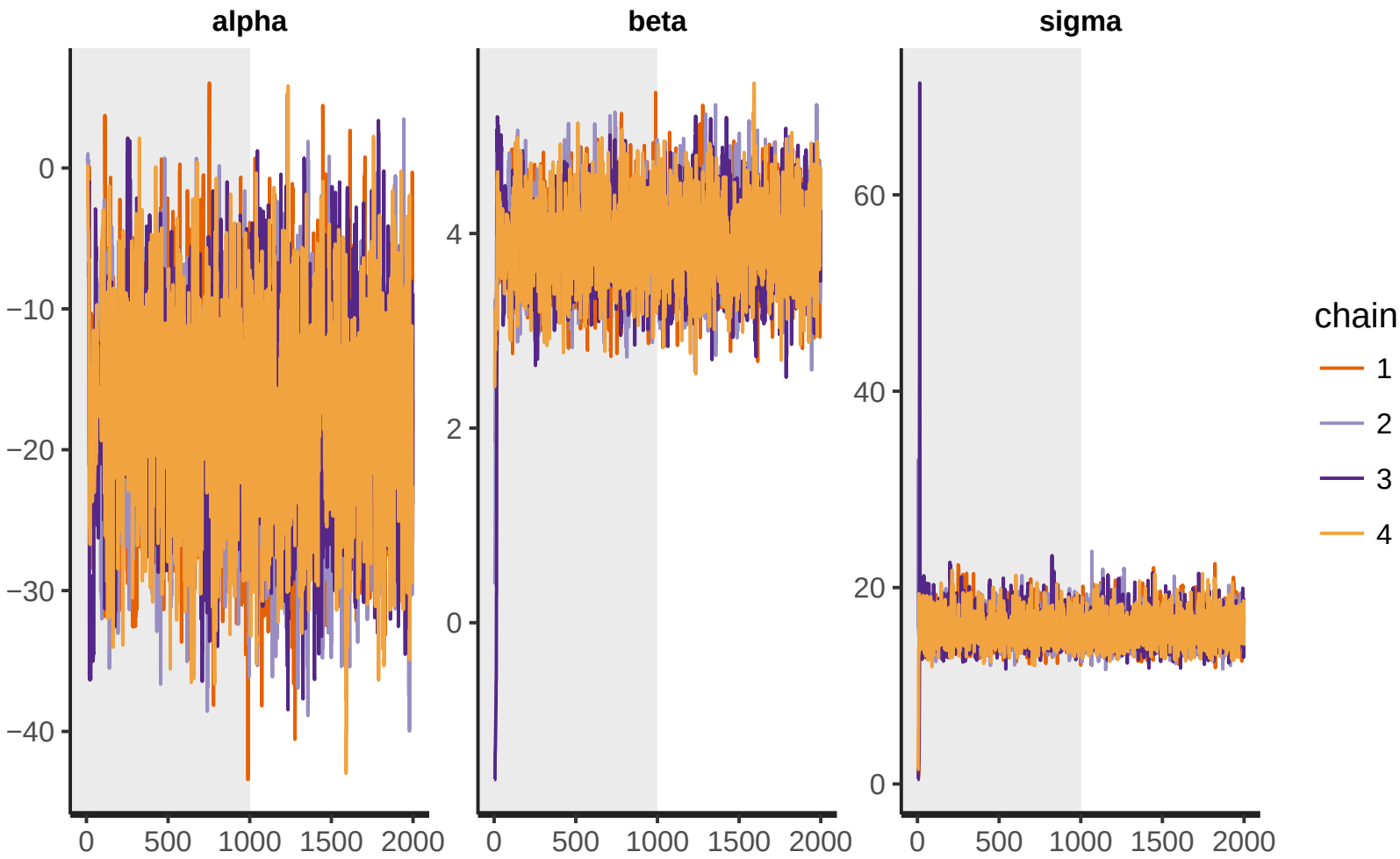
Regression model in Stan

```
data {
int<lower=1> N;
vector[N] x;
vector[N] y;
}
parameters {
real alpha;
real beta;
real<lower=0> sigma;
}
model {
// alpha ~ cauchy(0,10);
// beta ~ cauchy(0,2.5);
// sigma ~ cauchy(0, 2.5);
y ~ normal(beta*x + alpha, sigma); // likelihood
}
```

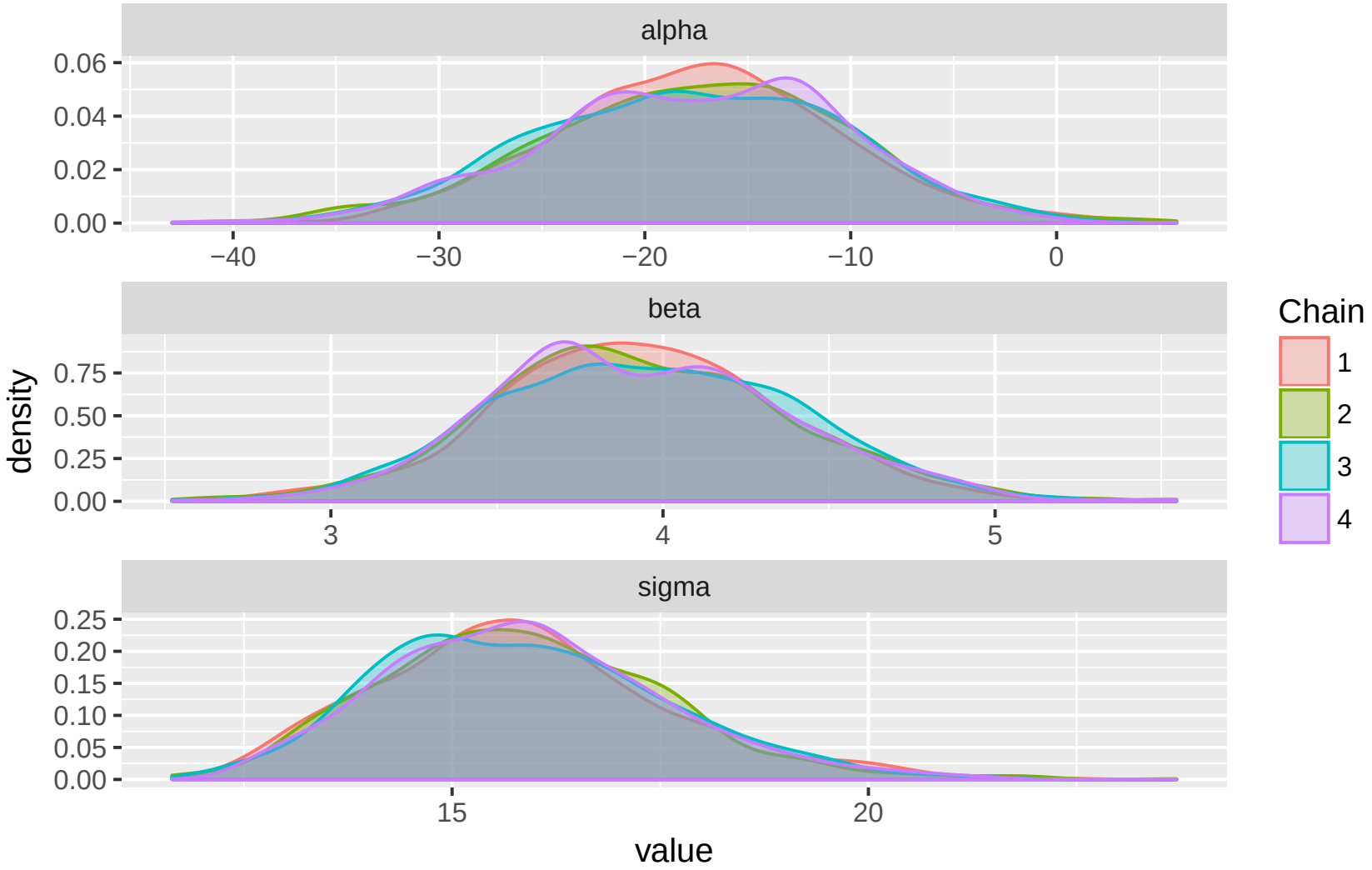
Regression model in Stan

```
> data(cars)
> carr <- list(N=50,x=cars[,1],y=cars[,2])
> fitstan<-stan(file="regression.stan",data=carr)

# No priors
      mean se_mean sd 2.5% 25% 50% 75% 97.5% n_eff Rhat
alpha -17.87 0.20 6.95 -31.14 -22.53 -18.04 -13.22 -3.90 1232 1
beta 3.95 0.01 0.43 3.10 3.67 3.97 4.24 4.77 1288 1
sigma 15.83 0.04 1.63 13.08 14.67 15.70 16.82 19.42 1758 1


# Priors
      mean se_mean sd 2.5% 25% 50% 75% 97.5% n_eff Rhat
alpha -12.43 0.19 6.76 -26.02 -16.93 -12.23 -7.71 0.43 1276 1
beta 3.62 0.01 0.42 2.83 3.34 3.62 3.90 4.46 1319 1
sigma 15.56 0.03 1.61 12.80 14.41 15.38 16.56 19.15 2258 1
```

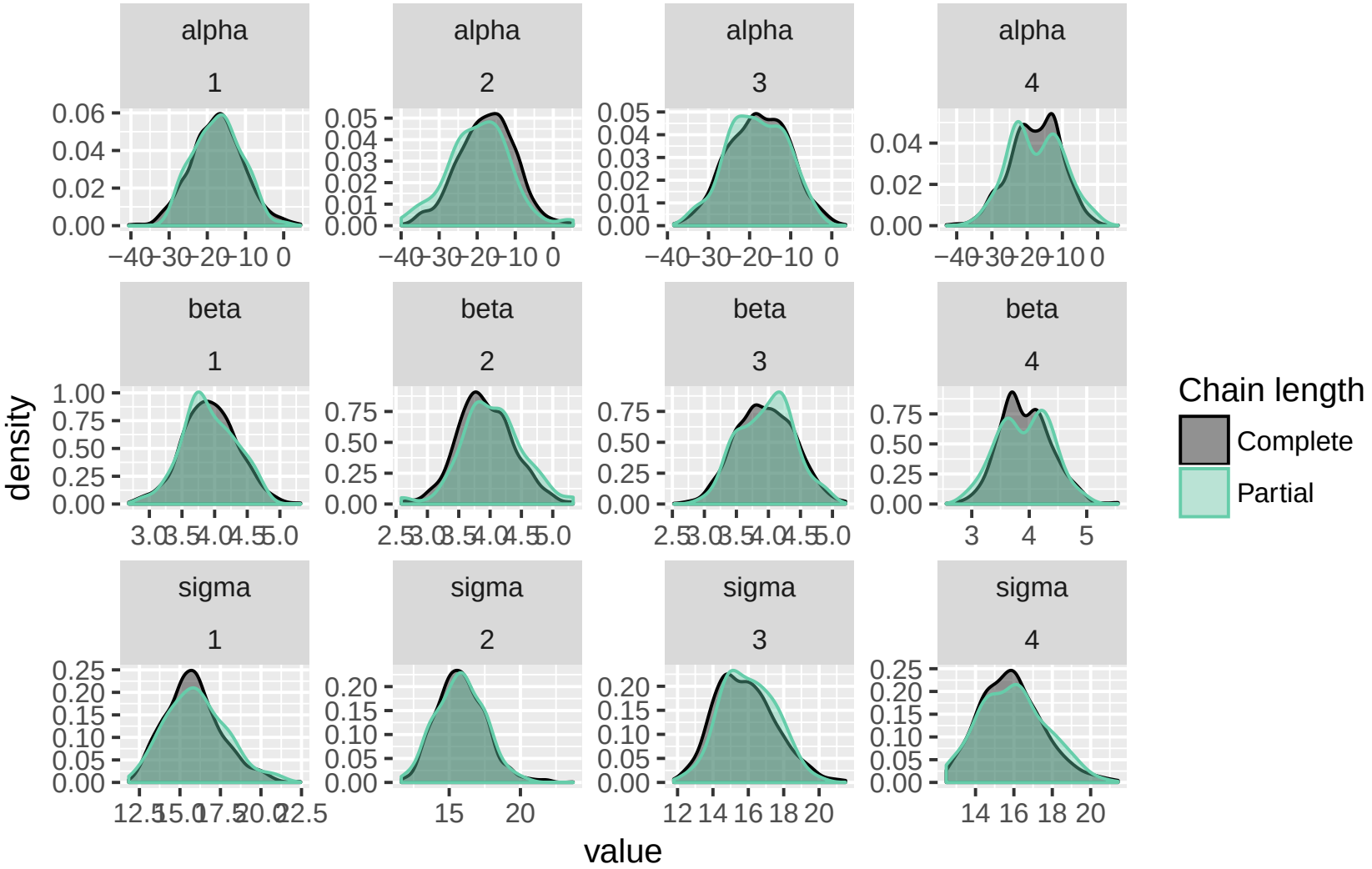
Regression model in Stan



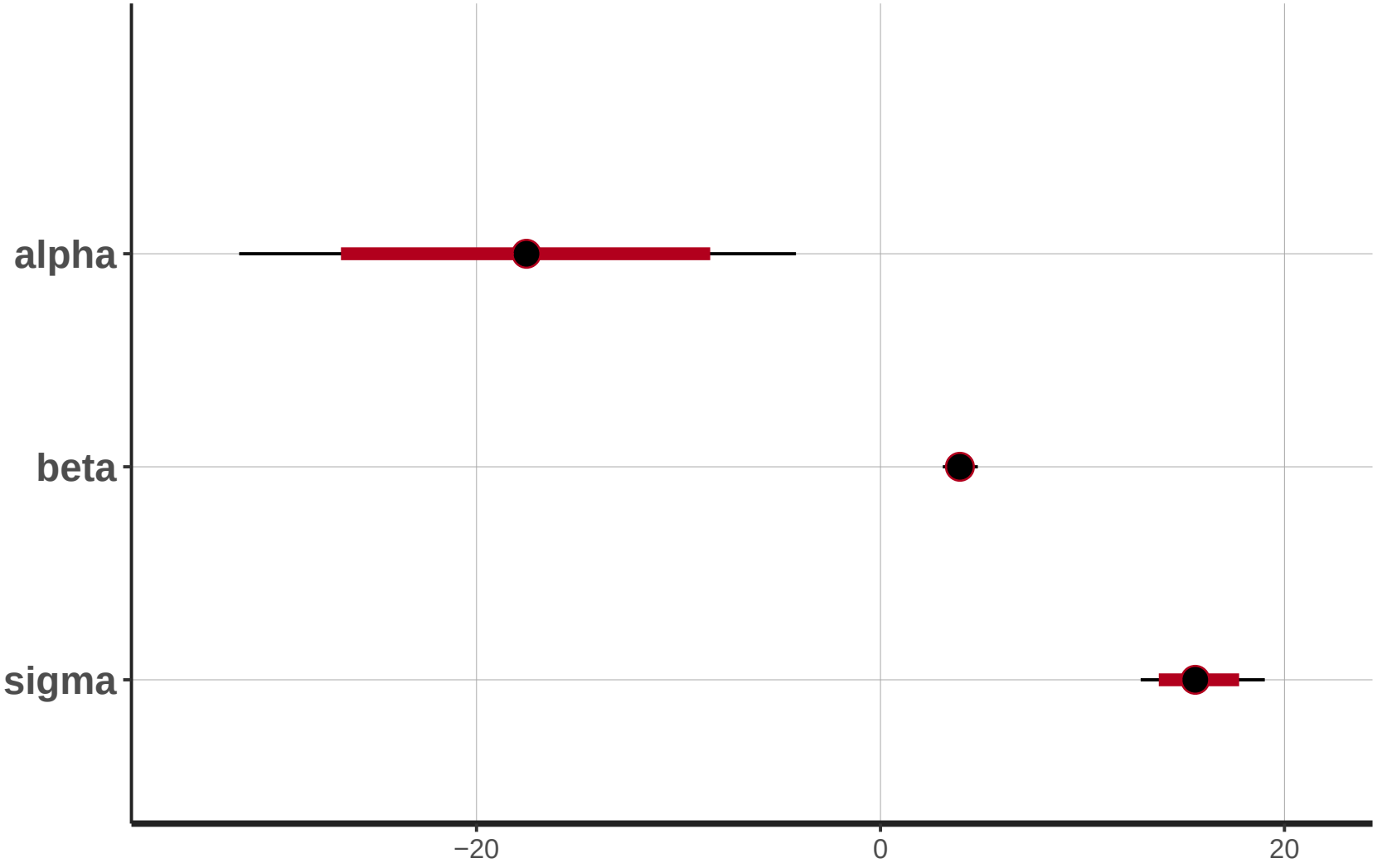
Regression model in Stan



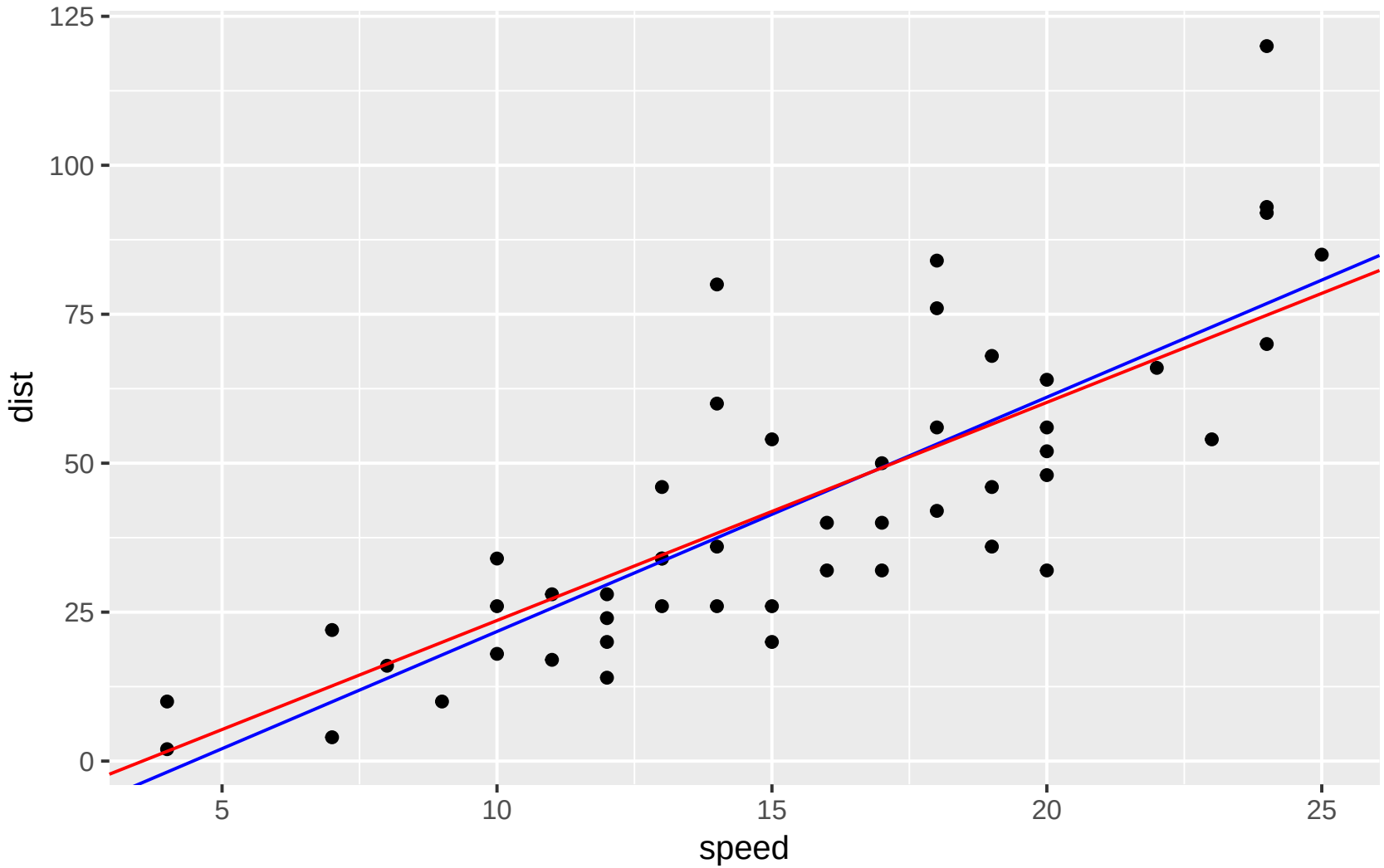
Regression model in Stan



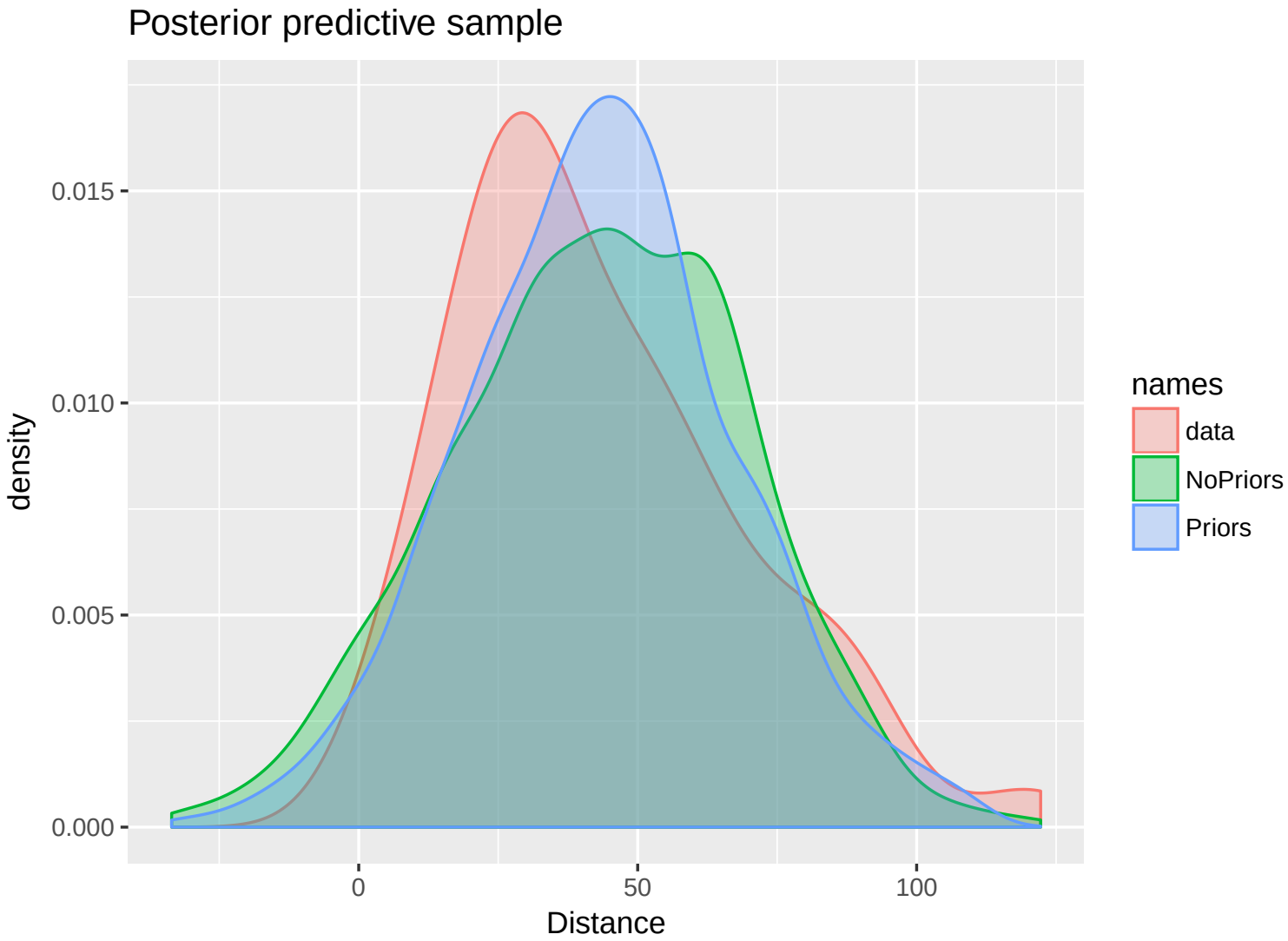
Regression model in Stan



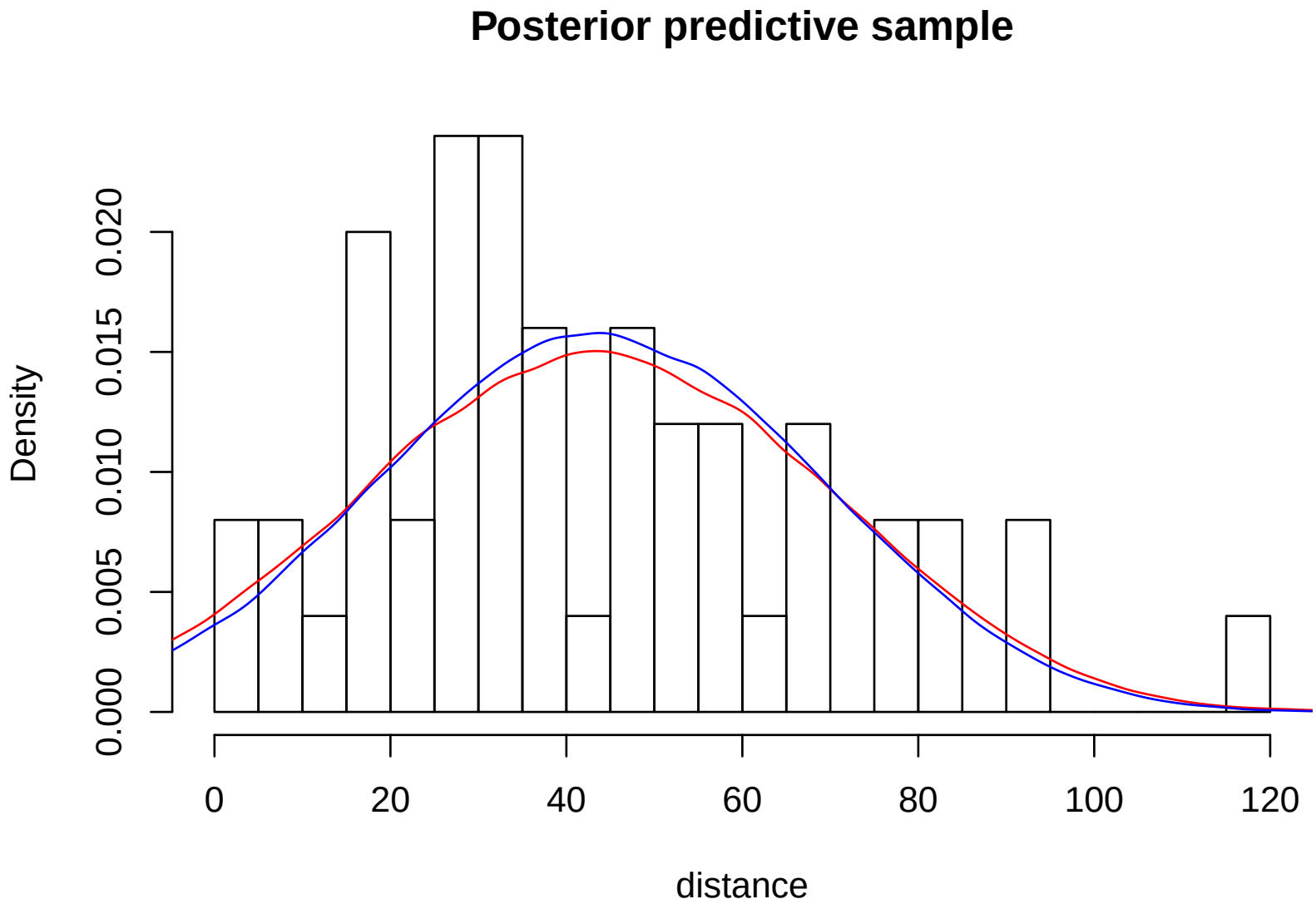
Regression model in Stan



Regression model in Stan



Regression model in Stan



Basic hierarchical model in Stan

Parallel experiments in eight schools

A study was performed for the Educational Testing Service to analyze the effects of special coaching programs on thest scores.

School	Estimated treatment	Standard error of estimate
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

Basic hierarchical model in Stan

Formal bayesian justification:

$$p(\boldsymbol{\theta}, \mu, \tau|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mu, \tau)p(\mu, \tau)$$

Normal model with exchangeable parameters:

- $\boldsymbol{y}|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ *likelihood*
- $\boldsymbol{\theta}|\mu, \tau \sim N(\mu, \tau^2)$ *prior*
- $\mu, \tau \sim ???$ *prior*

Basic hierarchical model in Stan

Formal bayesian justification:

$$p(\boldsymbol{\theta}, \mu, \tau|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mu, \tau)p(\mu, \tau)$$

Normal model with exchangeable parameters:

- $\boldsymbol{y}|\boldsymbol{\theta} \sim N(\boldsymbol{\theta}, \boldsymbol{\sigma}^2)$ *likelihood*
- $\boldsymbol{\theta}|\mu, \tau \sim N(\mu, \tau^2)$ *prior*
- $\mu, \tau \sim$??? *prior*

Frequentist:

$$\boldsymbol{y} = \overbrace{\mu + \boldsymbol{\eta}}^{\boldsymbol{\theta}} + \boldsymbol{\epsilon}$$

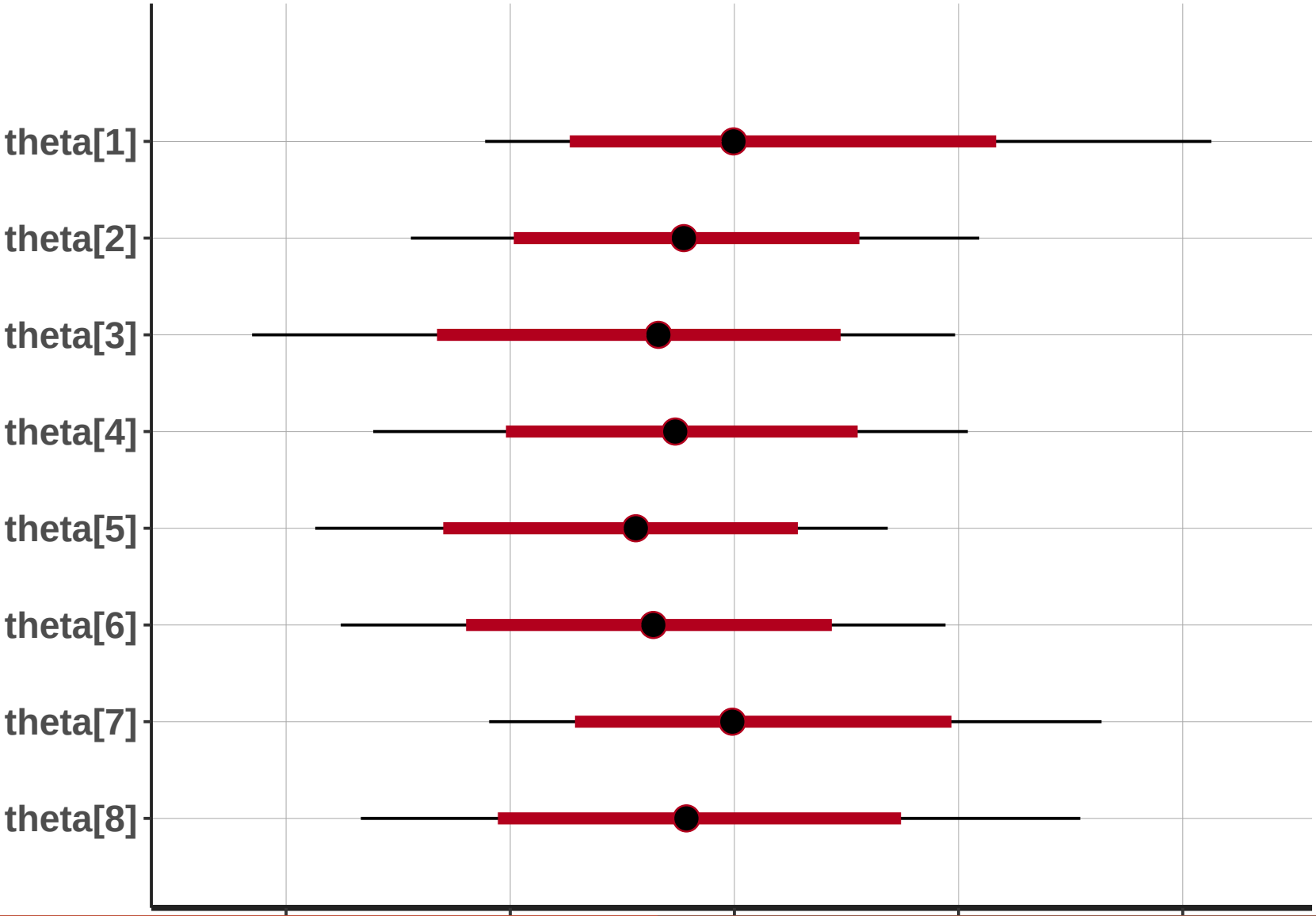
Basic hierarchical model in Stan

```
data {
int<lower=0> J; // number of schools
real y[J]; // estimated treatment effects
real<lower=0> sigma[J]; // s.e. of effect estimates
}
parameters {
real mu; // population mean
real<lower=0> tau; // population sd
vector[J] eta; // school-level errors
}
transformed parameters {
vector[J] theta; // school effects
theta = mu + tau*eta;
}
model {
eta ~ normal(0, 1); // prior
y ~ normal(theta, sigma); // likelihood
}
```

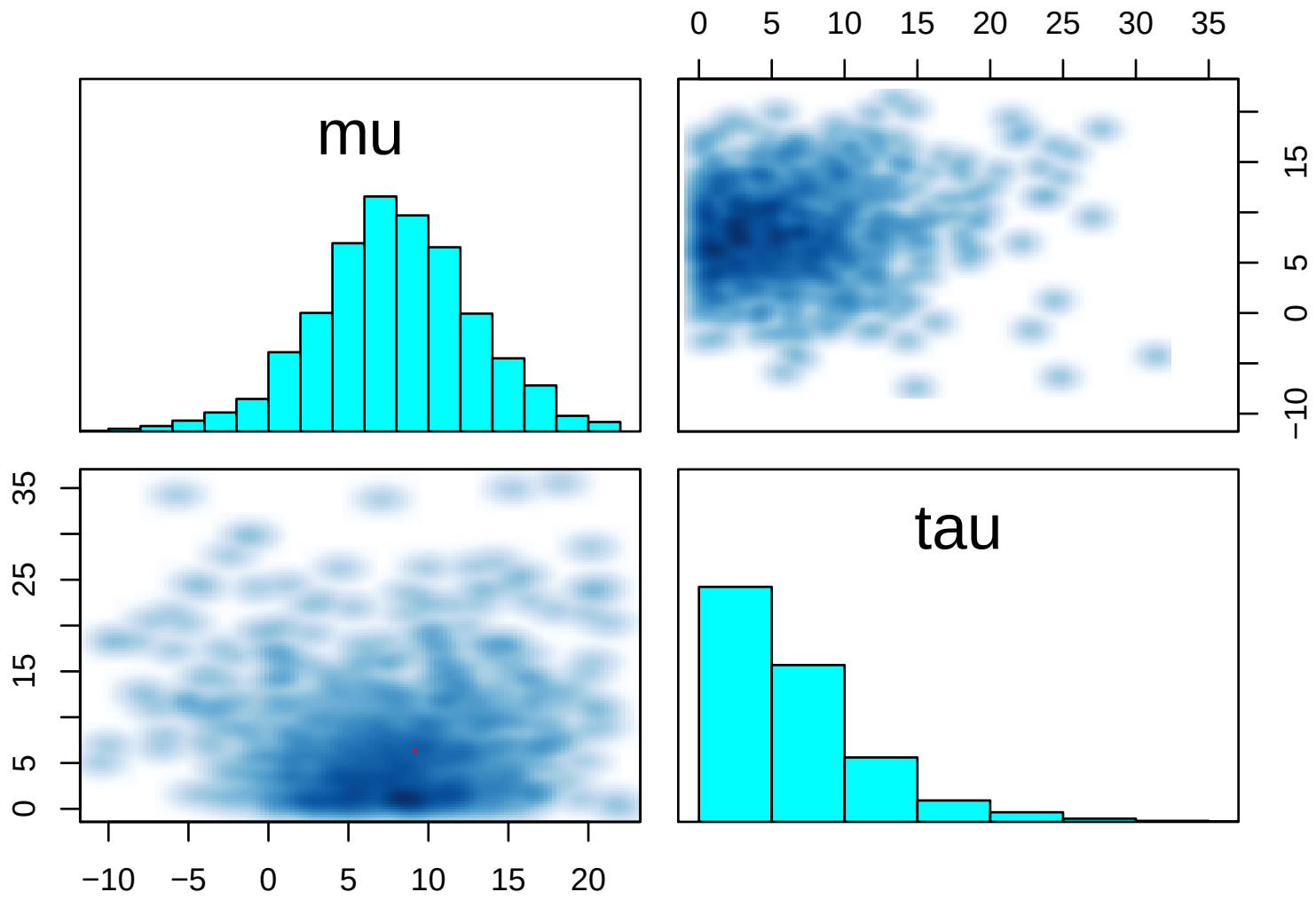
Basic hierarchical model in Stan

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
mu	7.88	0.15	4.95	-2.23	4.69	7.91	11.14	17.35	1119	1.00
tau	6.44	0.21	5.46	0.18	2.32	5.27	9.11	21.38	655	1.01
eta[1]	0.40	0.02	0.92	-1.44	-0.23	0.42	1.03	2.19	1853	1.00
eta[2]	0.00	0.02	0.87	-1.69	-0.58	0.00	0.58	1.69	1560	1.00
eta[3]	-0.18	0.02	0.94	-2.04	-0.80	-0.17	0.45	1.68	2000	1.00
eta[4]	-0.04	0.02	0.89	-1.79	-0.61	-0.08	0.56	1.75	1680	1.00
eta[5]	-0.32	0.02	0.90	-2.08	-0.90	-0.33	0.23	1.49	1662	1.00
eta[6]	-0.22	0.02	0.95	-2.06	-0.84	-0.23	0.40	1.70	1910	1.00
eta[7]	0.35	0.02	0.89	-1.47	-0.22	0.36	0.92	2.11	2000	1.00
eta[8]	0.04	0.02	0.93	-1.80	-0.54	0.03	0.66	1.93	2000	1.00
theta[1]	11.31	0.23	8.14	-1.13	5.96	9.96	15.46	31.27	1252	1.00
theta[2]	7.79	0.14	6.29	-4.43	3.83	7.74	11.67	20.92	2000	1.00
theta[3]	6.06	0.17	7.74	-11.52	2.17	6.61	10.72	19.83	2000	1.00
theta[4]	7.47	0.15	6.57	-6.10	3.62	7.35	11.49	20.41	2000	1.00
theta[5]	5.29	0.15	6.49	-8.70	1.48	5.60	9.80	16.84	2000	1.00
theta[6]	6.26	0.15	6.65	-7.56	2.55	6.38	10.41	19.41	2000	1.00
theta[7]	10.72	0.16	6.94	-0.94	6.10	9.91	14.62	26.37	2000	1.00
theta[8]	8.30	0.17	7.70	-6.65	3.87	7.86	12.58	25.42	2000	1.00
lp__	-5.02	0.12	2.71	-11.10	-6.64	-4.73	-3.05	-0.63	545	1.01

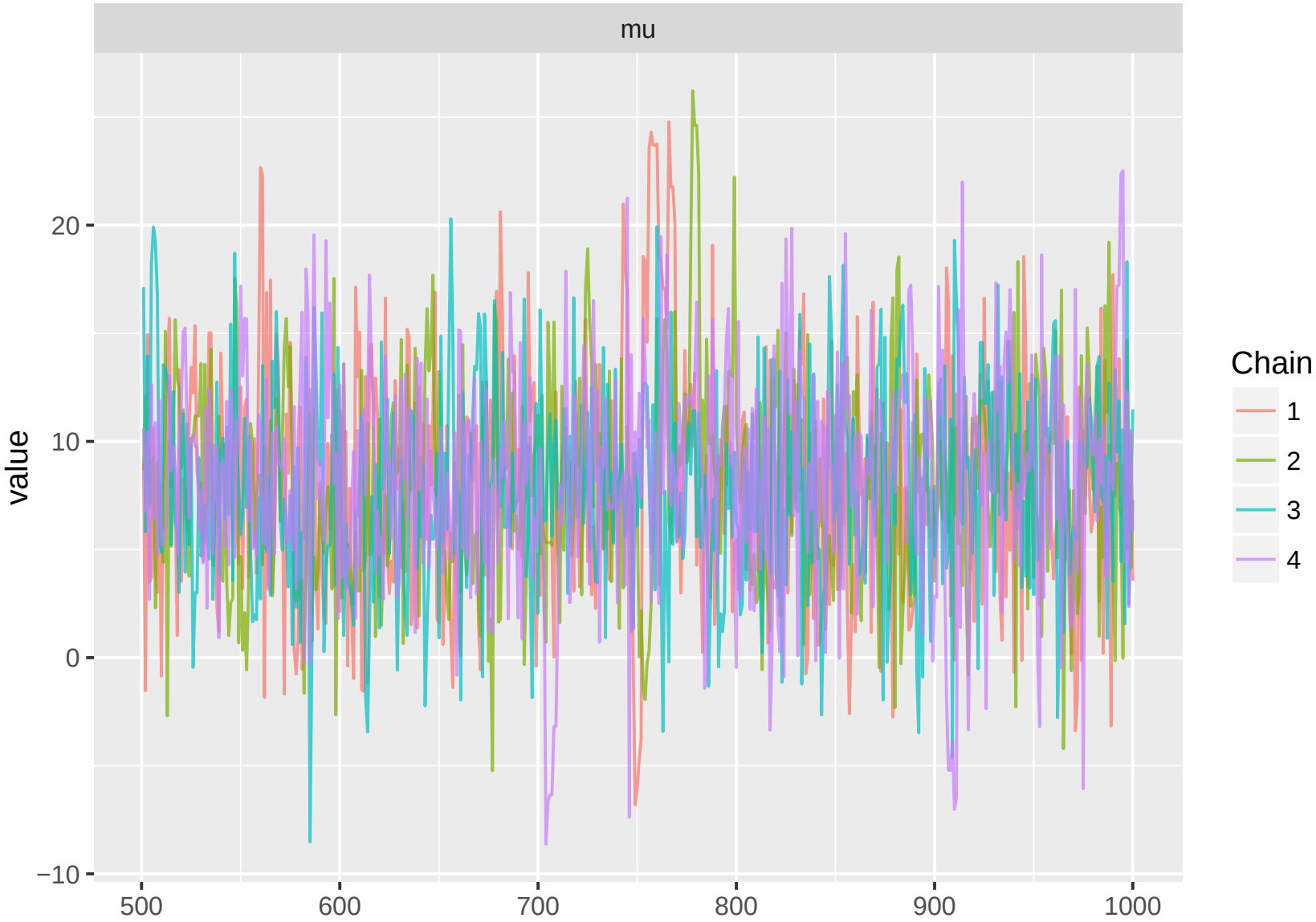
Basic hierarchical model in Stan



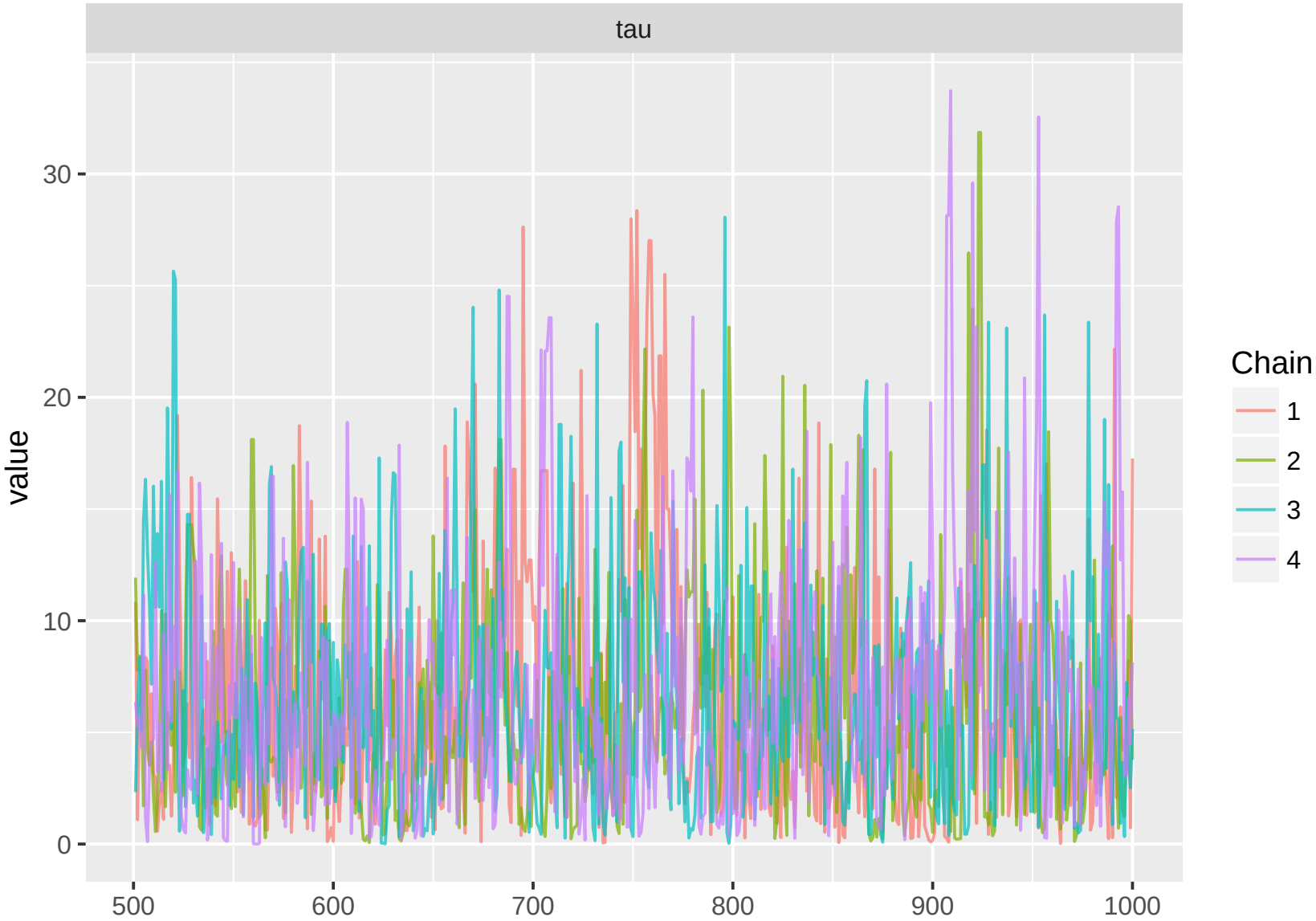
Basic hierarchical model in Stan



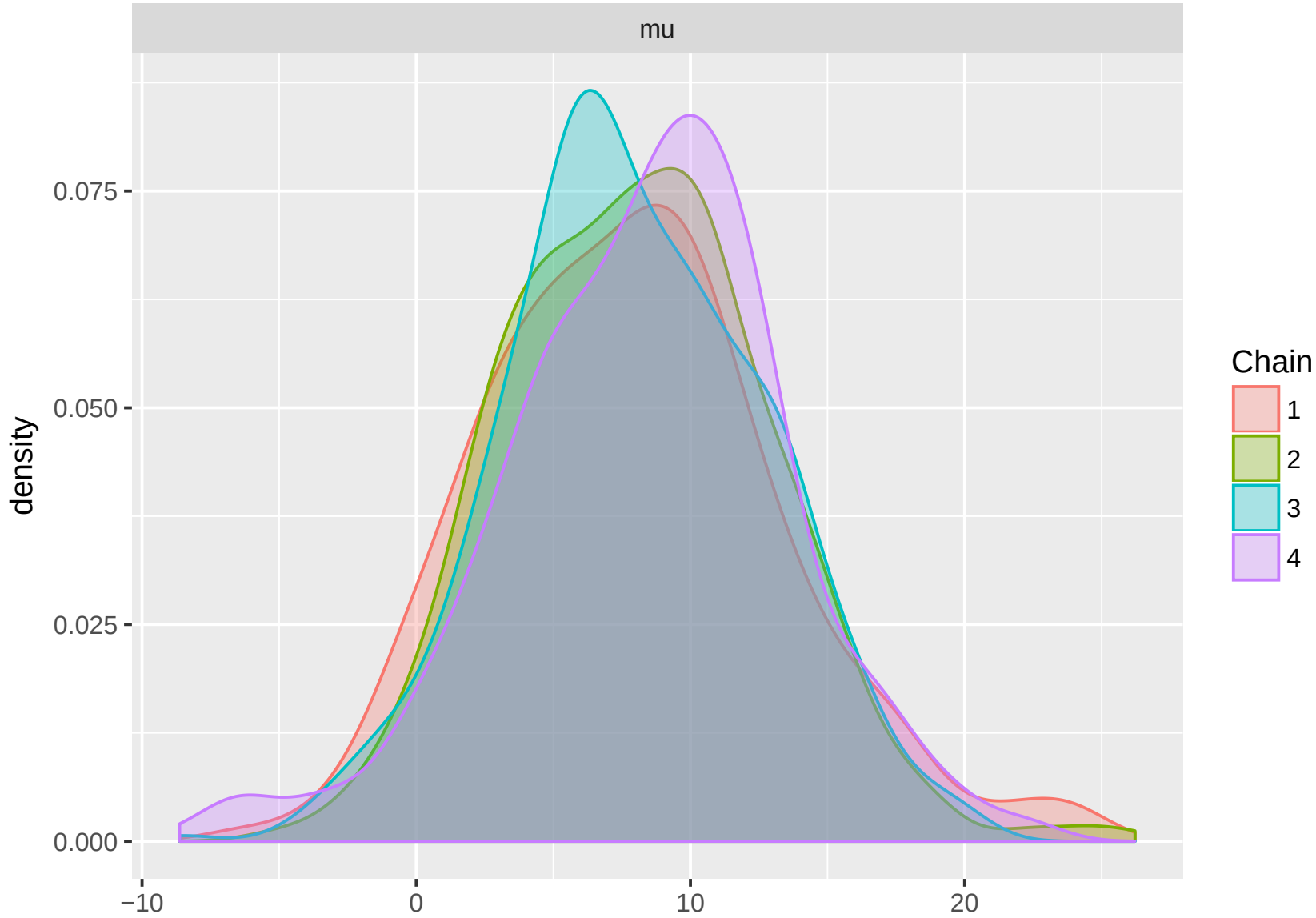
Basic hierarchical model in Stan



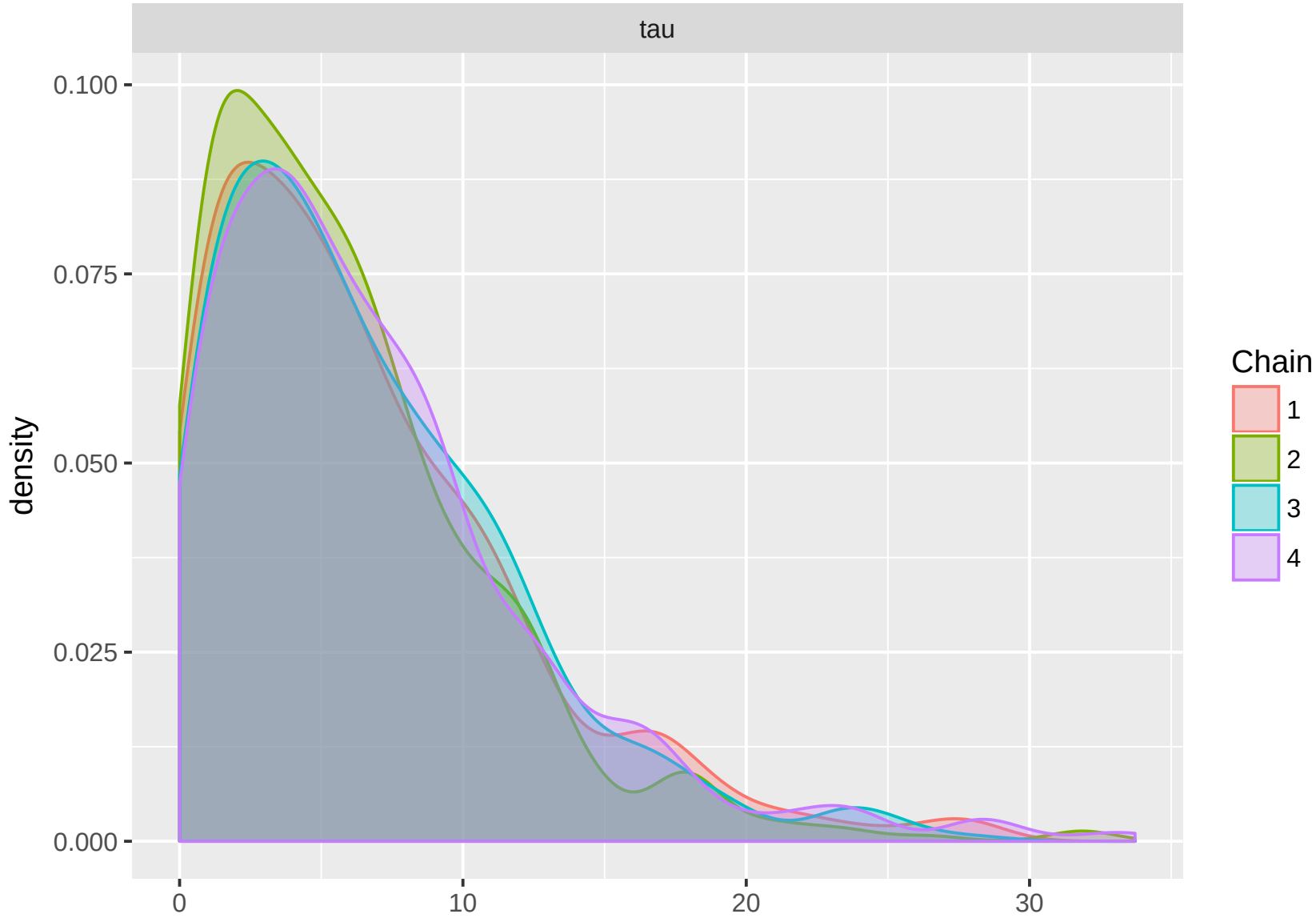
Basic hierarchical model in Stan



Basic hierarchical model in Stan



Basic hierarchical model in Stan



Conclusion (Bayesian data analysis)

Gelmans viewpoint

1. Setting up a full probability model (teoretical)
2. Calculating and interpreting the appropriate posterior distribution (continuing on observed data)
3. Evaluating the fit of the model and the implications of the resulting posterior distribution

Conclusion

- This is the key different between Baysian and Frequentist:
  - The frequentist get the data and try to generalize that data to the whole population, it's very human nature. You see all your neighbors driving pinky cars, you end up the whole city love pinky cars.
  - Instead, Bayesian thinks about some parameters affected the choice of your neighbors. You have strong evidence that all your neighbors are beatiful young girls. So that why most of the cars are pink.

Conclusion

- This is the key different between Baysian and Frequentist:
  - The frequentist get the data and try to generalize that data to the whole population, it's very human nature. You see all your neighbors driving pinky cars, you end up the whole city love pinky cars.
  - Instead, Bayesian thinks about some parameters affected the choice of your neighbors. You have strong evidence that all your neighbors are beatiful young girls. So that why most of the cars are pink.

  But what if some guys love driving pinky cars also ? This is the risk of Bayesian, if your prior is wrong and the data you have is not big enough, everything goes wrong!

Conclusion

- This is the key different between Baysian and Frequentist:
  - The frequentist get the data and try to generalize that data to the whole population, it's very human nature. You see all your neighbors driving pinky cars, you end up the whole city love pinky cars.
  - Instead, Bayesian thinks about some parameters affected the choice of your neighbors. You have strong evidence that all your neighbors are beatiful young girls. So that why most of the cars are pink.

  But what if some guys love driving pinky cars also ? This is the risk of Bayesian, if your prior is wrong and the data you have is not big enough, everything goes wrong!
- Good for longitudial study, mixed effects, models which requires strong expert knowledge...

Conclusion

- This is the key different between Baysian and Frequentist:
  - The frequentist get the data and try to generalize that data to the whole population, it's very human nature. You see all your neighbors driving pinky cars, you end up the whole city love pinky cars.
  - Instead, Bayesian thinks about some parameters affected the choice of your neighbors. You have strong evidence that all your neighbors are beatiful young girls. So that why most of the cars are pink.

  But what if some guys love driving pinky cars also ? This is the risk of Bayesian, if your prior is wrong and the data you have is not big enough, everything goes wrong!
- Good for longitudial study, mixed effects, models which requires strong expert knowledge...
- Stan vs. JAGS, BUGS

Conclusion (to be continue)

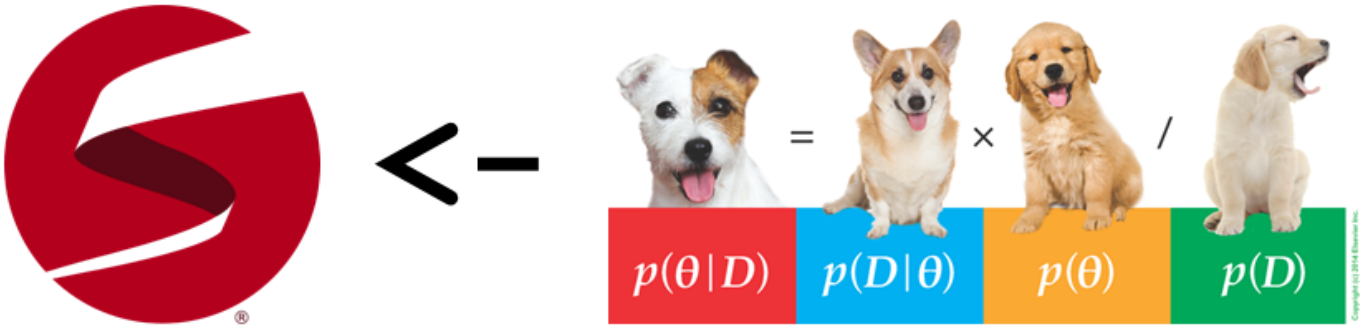
Suggested ideas

- Faraway: STAN for linear mixed models,
- Use Stan for solve your own model
- or ....

Bibliography

- Gelman, A., Carlin, J., B., Stern, H., S., Dunson, D., B., Vehtari, A., Rubin, D., B., *Bayesian Data Analysis* Third Edition CRC Press 2013.
- Kruschke, J., K., *Doing Bayesian Data Analysis* Academic Press 2015.
- McElreath, R., *Statistical Rethinking* CRC Press 2016.
- Faraway, J., J., *Extenging the Linear Model with R* CRC Press 2016.
- Carpenter, G., Gelman, A,, et al. *Stan: A Probabilistic Programming Language* Journal of Statistical Software, January 2017, Volume 76, Issue 1.

Thank you for your attention



doingbayesiandataanalysis.blogspot.com