AN INTRODUCTION TO BAYESIAN ANALYSIS AND BAYESIAN REGRESSION IN R

Michael Crawford & Michael Cooney

Barnett Waddingham

BAYESIAN STATISTICS IN 5 MINUTES

WHICH IS IMPOSSIBLE

SO LET'S RATHER TRY TO...

LEARN ABOUT BAYES RULE IN 5 MINUTES

SO WHAT IS BAYES RULE?

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

HOW DO WE USE THIS?

YOU USE BAYES RULE ALL THE TIME

YOU JUST DON'T KNOW IT

BAYES RULE

SHOWS HOW YOUR BELIEFS CHANGE

WHEN YOU GET NEW INFORMATION

LET'S START WITH SOMETHING FAMILIAR

TOSSING A COIN

LET'S TOSS THE COIN A FEW TIMES

AFTER EACH TOSS

TELL ME IF THE COIN IS FAIR

AT OUTSET IT'S REASONABLE TO

ASSUME THE COIN IS FAIR

THIS IS YOUR INITIAL BELIEF

First toss: H	Fair? Yes
Second toss: H	Fair? Yes
Third toss: H	Fair? Yes
Fourth toss: H	Fair? Yes doubts set in
Fifth toss: H	Fair? Possibly
Sixth - Tenth toss: H	Fair? Probably not
Eleventh - Twentieth toss: H	Fair: Defintely not
Chances of this are:	1 in 1,048,576

YOU HAVE JUST INTUITIVELY APPLIED BAYES RULE WITHOUT REALISING IT

AS YOU GOT MORE INFORMATION

YOU ALTERED YOUR BELIEF

THAT THE COIN IS FAIR

NOW LET'S LOOK AT BAYES RULE AGAIN

BAYES RULE ALLOWS YOU QUANTIFY THIS QUALITATIVE PROCESS

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}$$

WE CAN SIMPLIFY THIS

$$P(A \mid B) \propto P(B \mid A)\,P(A)$$

$$P(A)$$

IS THE PRIOR DISTRIBUTION

AND REPRESENTS OUR INITIAL BELIEF

$$P(B \mid A)$$

IS THE LIKELIHOOD MODEL

AND UPDATES AS THE DATA ARRIVES

$$P(A \mid B)$$

IS THE POSTERIOR DISTRIBUTION

OUR UPDATED BELIEFS FOR A

Prior Distribution Parameters

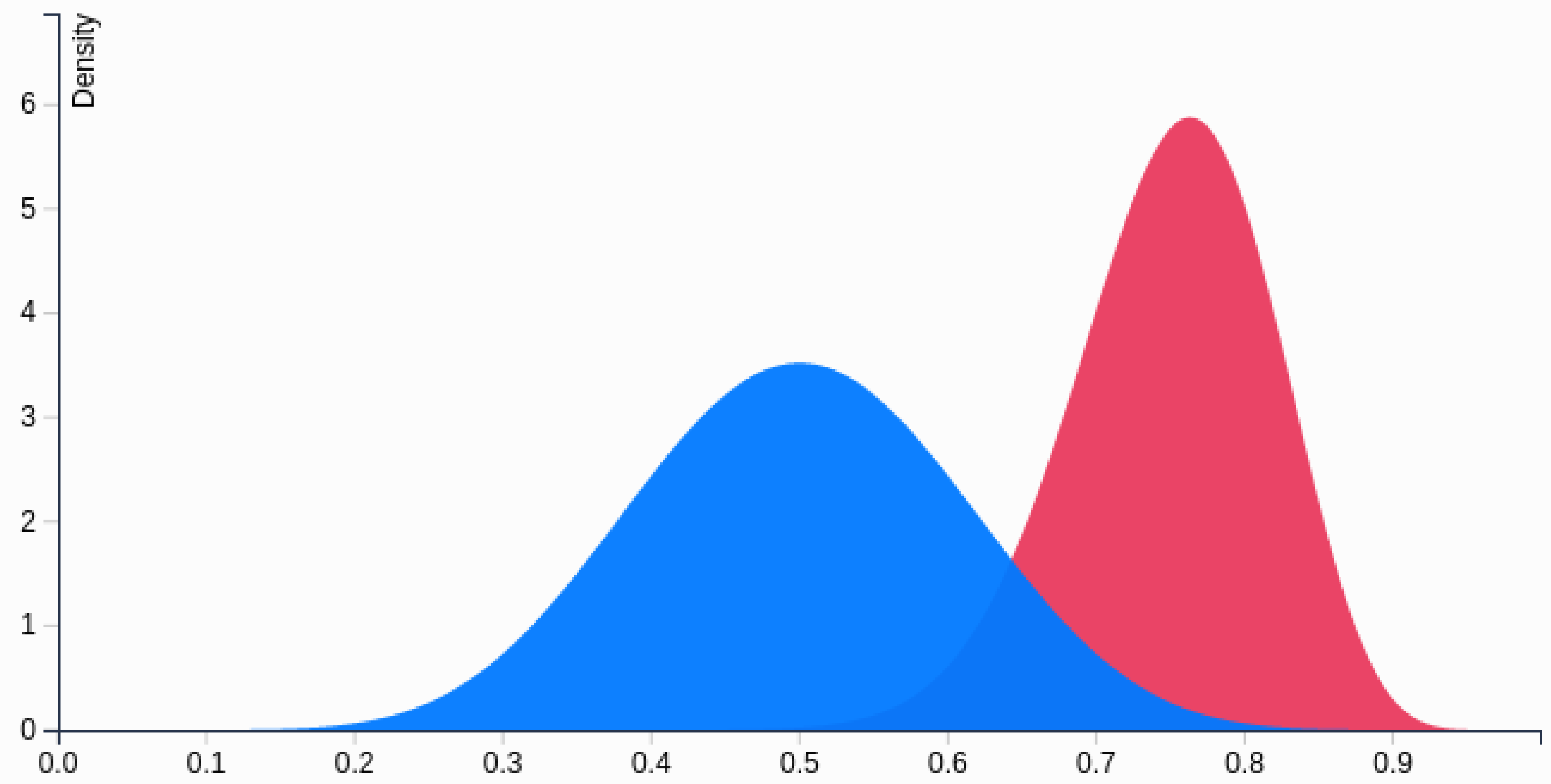
Alpha	10
Beta	10

Coin Toss Results

Heads	20
Tails	0

Calculate Reset

Prior Probability Density Function



USING THESE SIMPLE BUILDING BLOCKS

WE CAN BUILD POWERFUL MODELS

WHY BAYESIAN?

INTERPRETABILITY

Natural interpretation of output

Credibility region vs Confidence interval

SPARSE DATA PROBLEMS

Claims reserving

Pricing

Account segmentation

CLAIMS RESERVING

Changes in reserving philosophy

Assessing multiple insurers

THE BAYESIAN APPROACH

POSTERIOR DISTRIBUTION

$$p(\theta \,|D) = \frac{\int p(D \mid \theta)\, p(\theta)}{\int p(D)}$$

$$\begin{aligned} p(\theta) &= \text{prior distribution of } \theta \\ p(\theta \,|D) &= \text{posterior distribution of } \theta \text{ given } D \end{aligned}$$

$$p(\theta \,|D) \propto \int p(D \mid \theta)\, p(\theta)$$

How do we calculate this integral?

STAN

MCMC via HMC

Probabilistic Programming Language

C++ backend

Excellent online community

Why bother?

Captures uncertainty

Easy to iterate and improve

Allows generative modelling

Hierarchical modelling for sparse data

PITFALLS

Learning ‘cliff’

Requires aspects of physics, computation, statistics

Can seem overwhelming

Start with simple linear model

LINEAR MODELS

GETTING STARTED

Ordinary Least Squares (OLS)

Input variables X, parameters β

$$y = \beta X + \epsilon,$$
$$\epsilon \sim \mathcal{N}(0, \sigma)$$

Constant variance σ.

Rethink linear models in Bayesian language

Need probability model

BASIC ASSUMPTIONS

Data distributed as Normal

Mean for each point is linear function of X, βX

$$y \sim \mathcal{N}(\beta X, \sigma)$$

SIMPLE CLAIMS MODEL

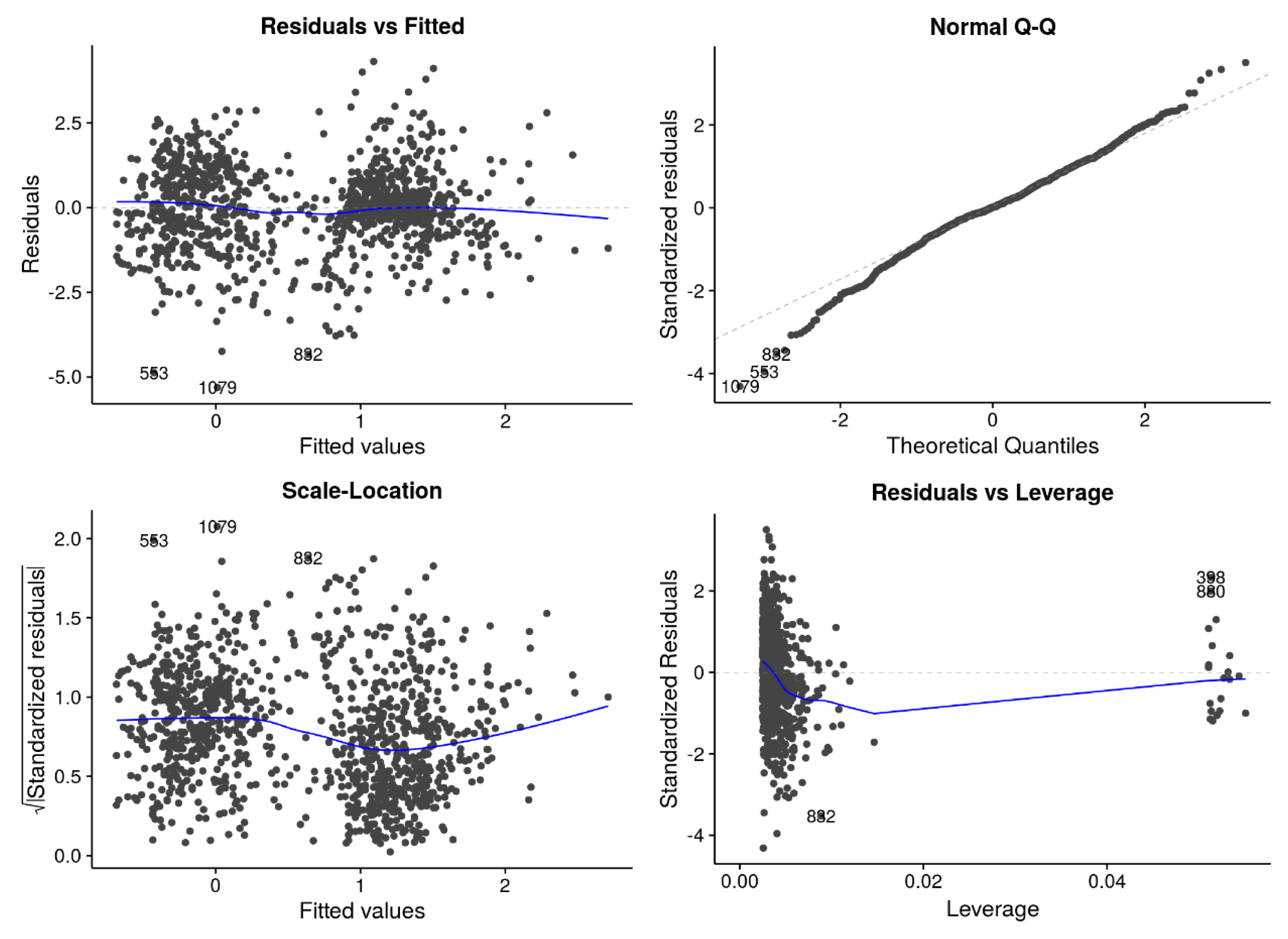
```
  log_loss lawyer gender seatbelt age
1   3.553632    yes   male      yes  50
2   2.388029     no female      yes  28
3  -1.108663     no   male      yes   5
4   2.401253    yes   male       no  32
5  -1.980502     no   male      yes  30
6  -1.174414    yes female      yes  35
7   1.263562    yes   male      yes  19
```

$$\text{log\_loss} \sim \text{lawyer} + \text{seatbelt} + \text{gender} + \text{age}$$

‘Formula notation’

MLE MODEL (IN R)

```
model_lm <- lm(log_loss ~ lawyer + seatbelt + gender + age
               ,data = modeldata_tbl)
```


Residuals vs Fitted
Residuals
Fitted values
553
1079
882
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Residuals vs Leverage
Standardized Residuals
Leverage
398
880

RSTANARM PACKAGE

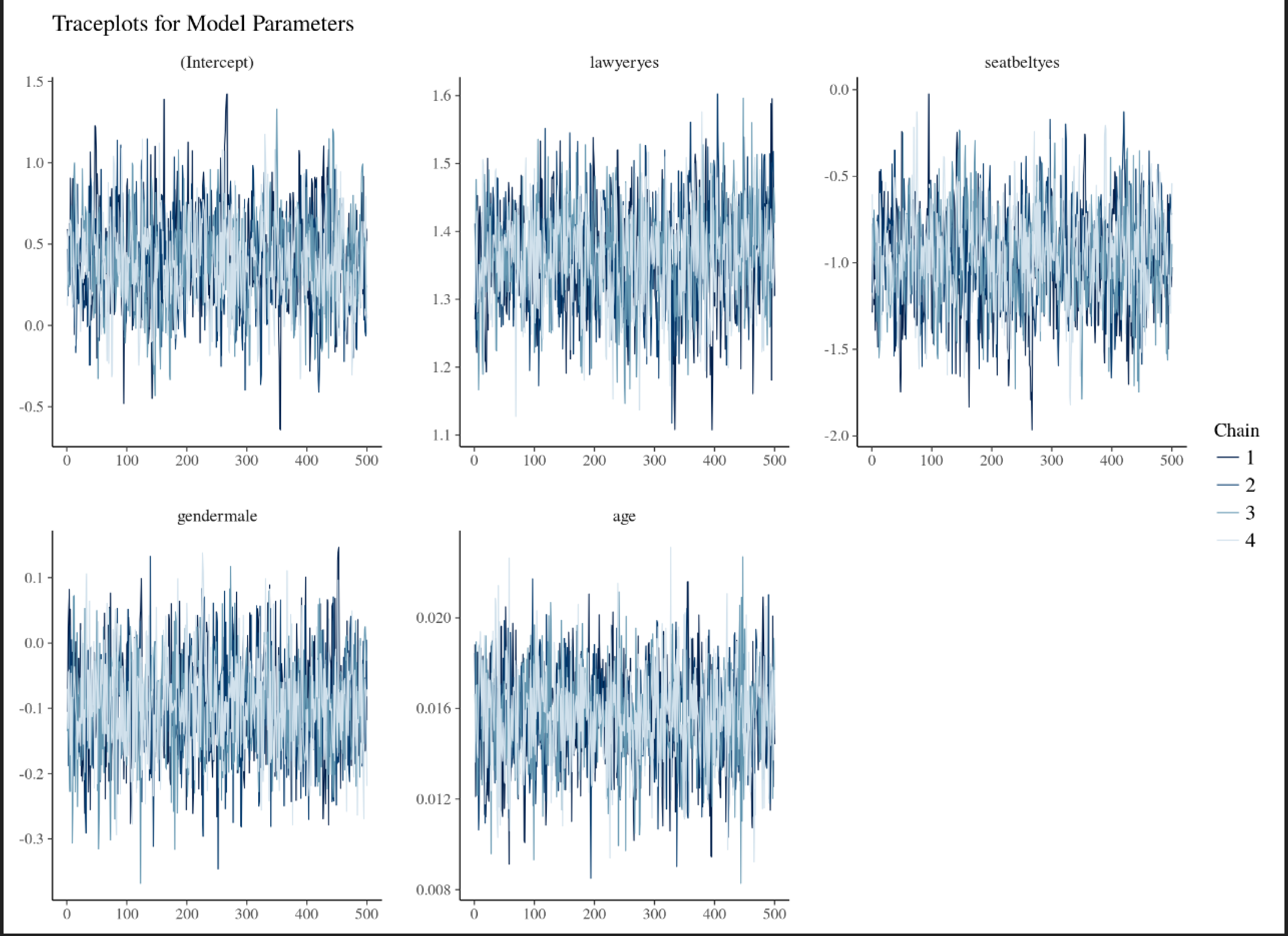
Pre-built models

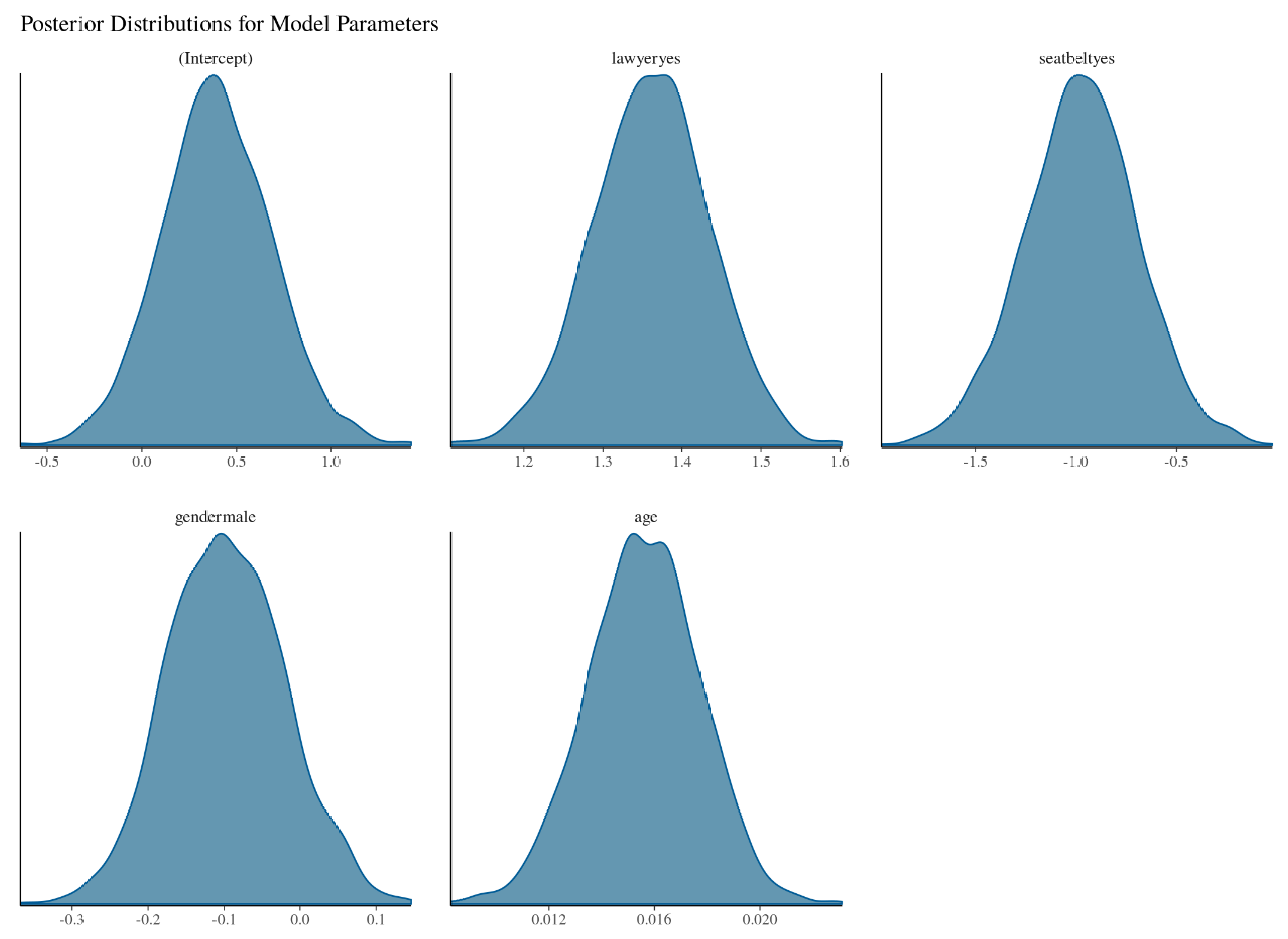
Linear models, GLMs, ANOVA, etc.

Built for ease of use

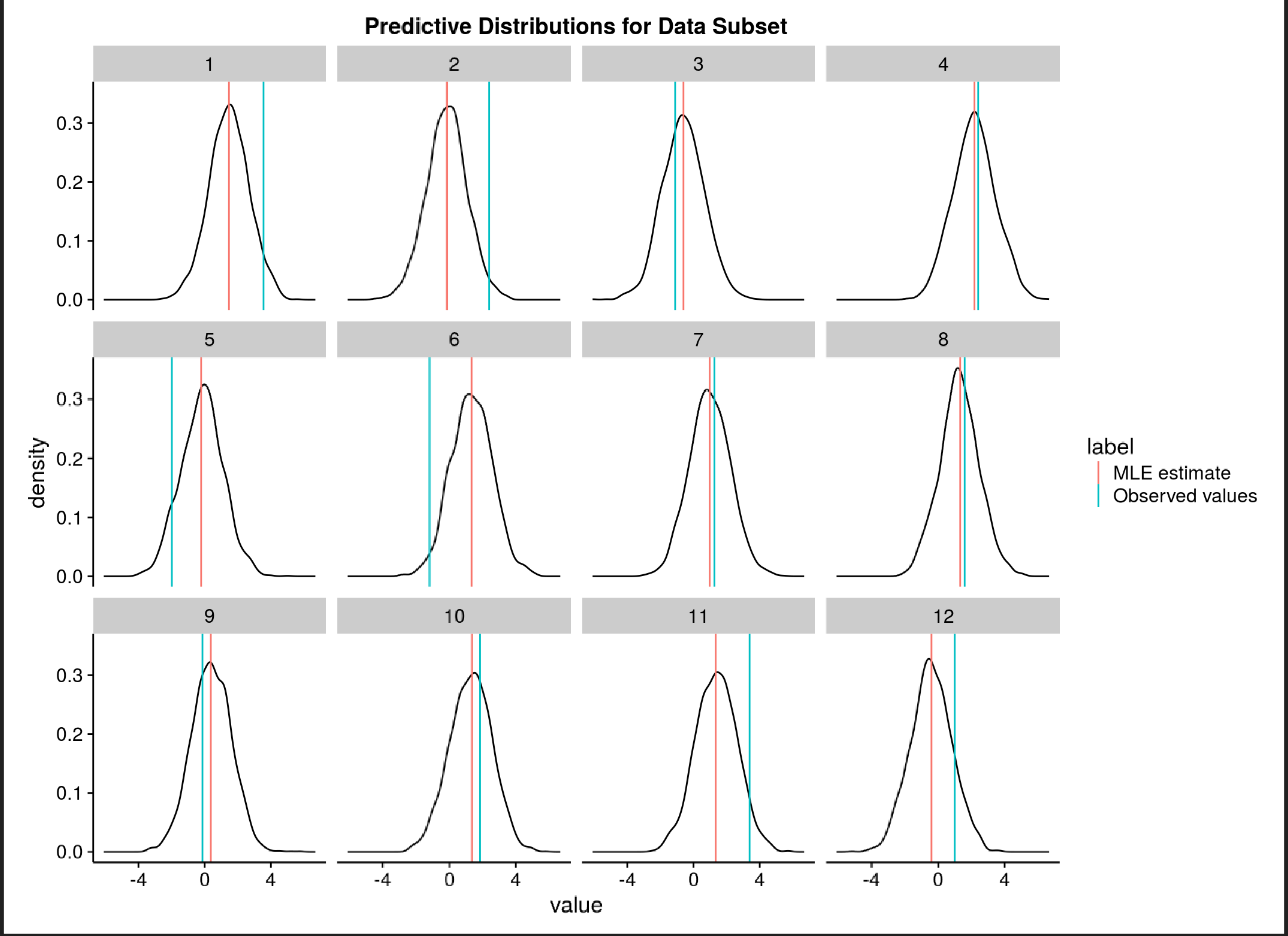
RSTANARM VERSION

```
model_stanlm <- stan_lm(log_loss ~ lawyer + seatbelt + gender + age
                       ,prior  = R2(location = 0.8)
                       ,data   = modeldata_tbl)
```


Traceplots for Model Parameters
(Intercept)
lawyeryes
seatbeltyes
gendermale
age
Chain
1
2
3
4



Posterior Distributions for Model Parameters
(Intercept)
-0.5
0.0
0.5
1.0
lawyeryes
1.2
1.3
1.4
1.5
1.6
seatbeltyes
-1.5
-1.0
-0.5
gendermale
-0.3
-0.2
-0.1
0.0
0.1
age
0.012
0.016
0.020


Predictive Distributions for Data Subset
1
2
3
4
5
6
7
8
9
10
11
12
density
0.0
0.1
0.2
0.3
value
-4
0
4
label
MLE estimate
Observed values

SUMMARY

Bayesian output captures uncertainty

More and more common

Learning curve

FURTHER READING

Stan Documentation/Vignettes/Case Studies, *Stan Core Team et al.*
http://www.mc-stan.org

Data Analysis Using Regression and Multilevel/Hierarchical Models, *Gelman and Hill*
http://www.stat.columbia.edu/~gelman/arm/

Statistical Rethinking, *McElreath*
http://xcelab.net/rm/statistical-rethinking/

Doing Bayesian Data Analysis, *Kruschke*
https://sites.google.com/site/doingbayesiandataanalysis/

QUESTIONS?

Nah, we're running outta time. Seriously.

michael.crawford@barnett-waddingham.co.uk

michael.cooney@barnett-waddingham.co.uk

THANKS FOR LISTENING

The R code is available on request