Bayesian Logistic Regression

Laurent Smeets

22 juni 2019

packages

```
library(tidyverse)
library(brms)
library(ggridges)
```

data

```
head(dat)
```

```
##   n_datapoints mean_accuracy ratio_urban_area classified2 classified
## 1         1114     10.883152        0.5323160           1    correct
## 2          671      7.592801        0.0000000           0  incorrect
## 3          355     97.586193        0.5690141           1    correct
## 4          539     12.411861        0.3896104           1    correct
## 5          549     17.015165        0.3952641           1    correct
## 6          649     15.557957        0.1325116           1    correct
##            mode3
## 1          Train
## 2 BikeNonElectric
## 3          Train
## 4            Car
## 5            Car
## 6            Car
```

priors

Generic weakly informative priors were chosen for both the intercept - student_t(3, 0, 10) - and the coefficient for all the features - N(0,10) (Gelman, 2019; Gelman, Jakulin, Pittau, & Su, 2008; Ghosh, Li, & Mitra, 2017).

```
priors <- c(
  prior(normal(0, 10),       class = "b"),
  prior(student_t(3, 0, 10),   class = "b", coef = "Intercept")
)
```

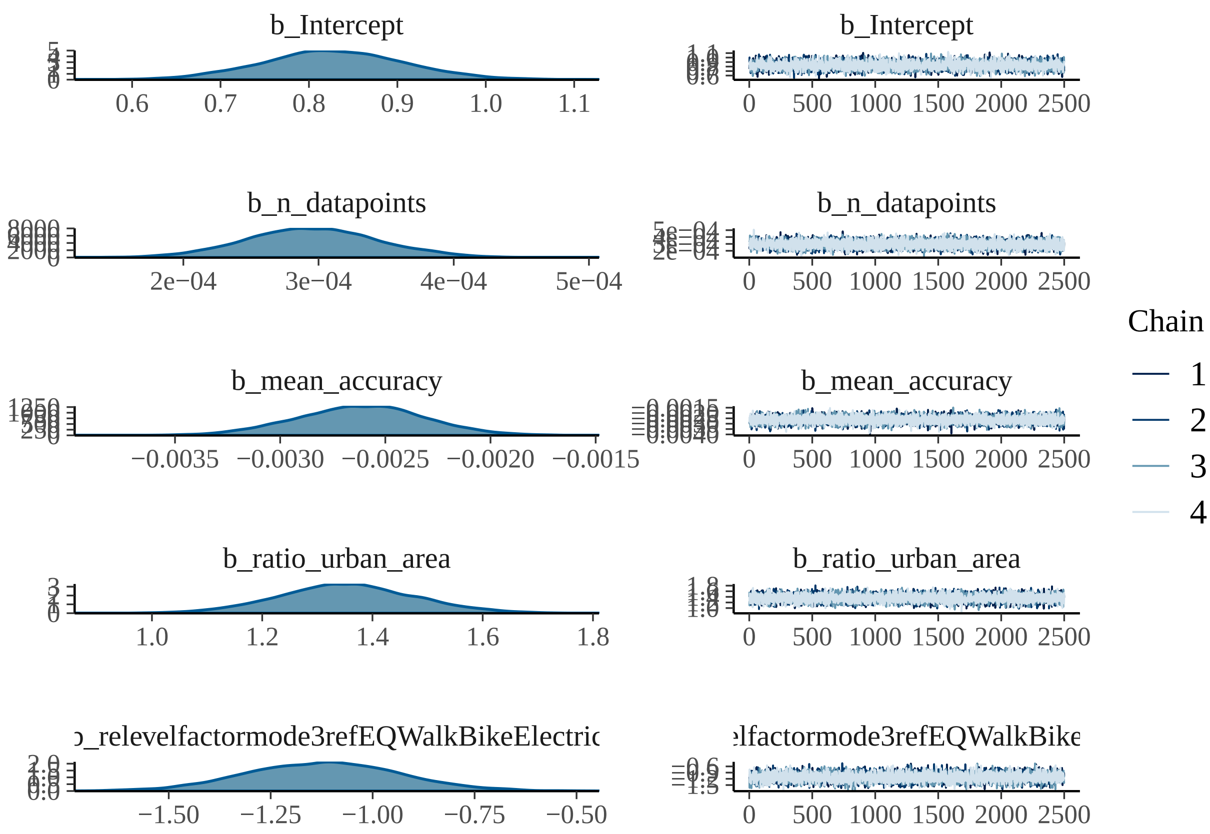
model

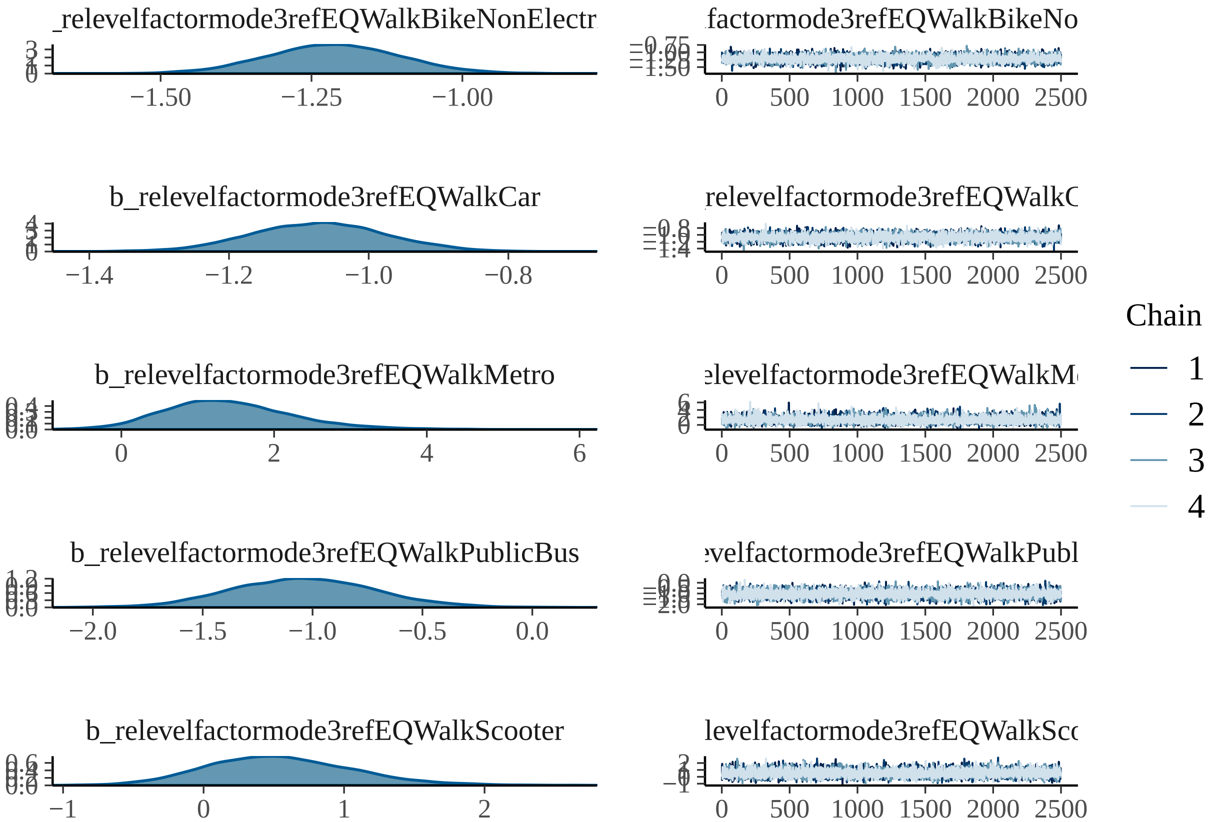
```
model <- brm(classified2 ~ 0 + Intercept + n_datapoints + mean_accuracy + ratio_urban_area+ relevel(fac
          data    = dat,
          family = "bernoulli",
```

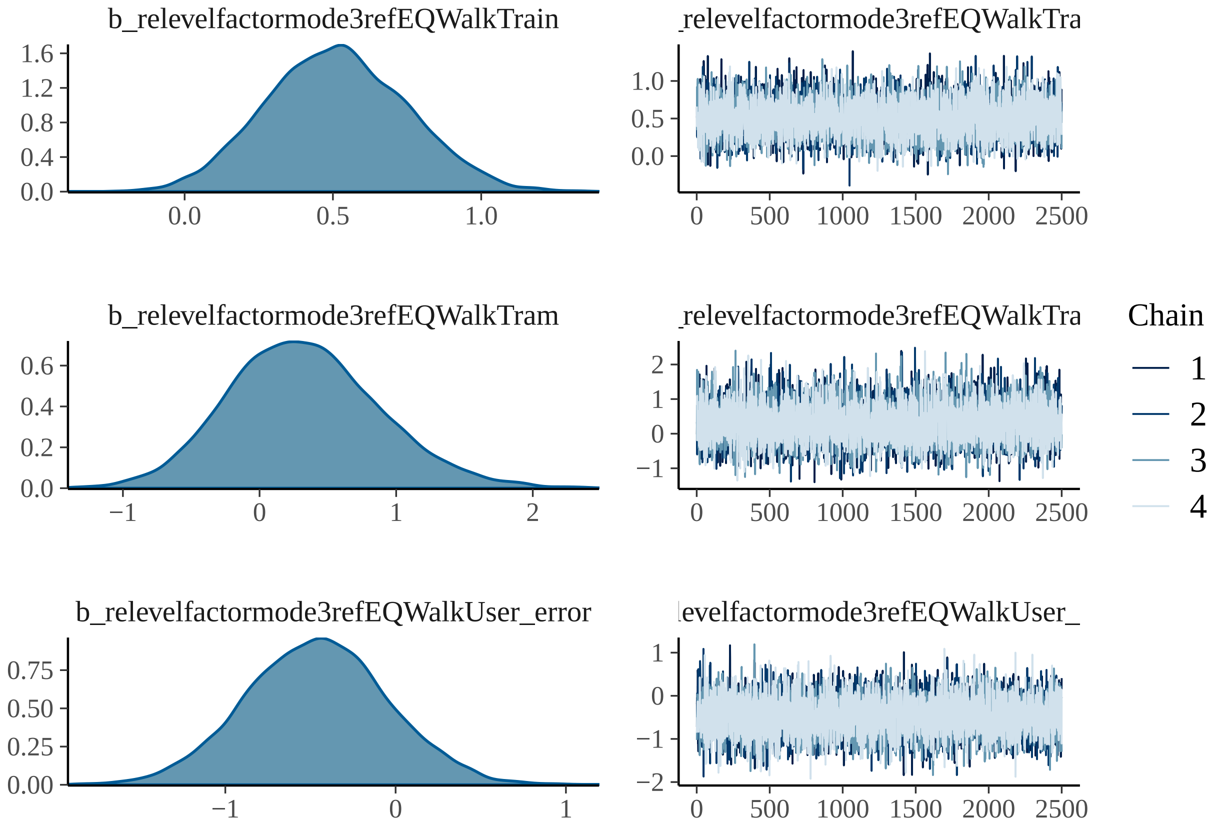
```
           iter    = 5000,
           seed    = 123,
           cores   = 4,
           prior   = priors)
```

Check for convergence and posterior

```
plot(model)
```




_relevelfactormode3refEQWalkBikeNonElectr
−1.50 −1.25 −1.00
factormode3refEQWalkBikeNo
b_relevelfactormode3refEQWalkCar
−1.4 −1.2 −1.0 −0.8
relevelfactormode3refEQWalkC
b_relevelfactormode3refEQWalkMetro
0 2 4 6
b_relevelfactormode3refEQWalkPublicBus
−2.0 −1.5 −1.0 −0.5 0.0
b_relevelfactormode3refEQWalkScooter
−1 0 1 2
0 500 1000 1500 2000 2500
Chain
1
2
3
4



Posteriors

```r
round(posterior_summary(model), 5)
```

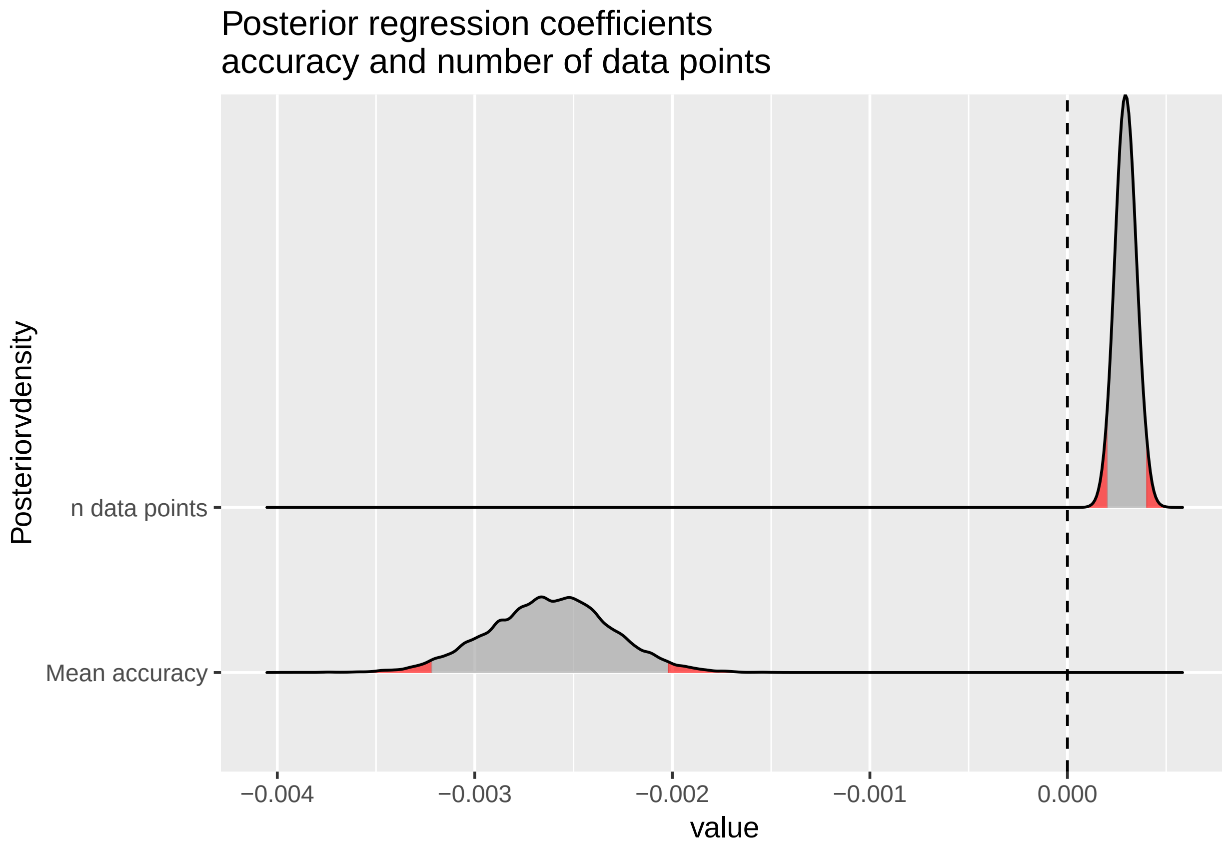
```
##                                             Estimate Est.Error
## b_Intercept                                  0.82793   0.08092
## b_n_datapoints                               0.00030   0.00005
## b_mean_accuracy                             -0.00262   0.00031
## b_ratio_urban_area                           1.35434   0.12316
## b_relevelfactormode3refEQWalkBikeElectric   -1.12093   0.18849
## b_relevelfactormode3refEQWalkBikeNonElectric -1.21311  0.11054
## b_relevelfactormode3refEQWalkCar            -1.07092   0.09662
## b_relevelfactormode3refEQWalkMetro           1.40921   0.85113
## b_relevelfactormode3refEQWalkPublicBus      -1.03862   0.32743
## b_relevelfactormode3refEQWalkScooter         0.54807   0.52310
## b_relevelfactormode3refEQWalkTrain           0.51436   0.23695
## b_relevelfactormode3refEQWalkTram            0.32814   0.55618
## b_relevelfactormode3refEQWalkUser_error     -0.46115   0.41842
## lp__                                      -2661.35584  2.54961
##                                                 Q2.5     Q97.5
## b_Intercept                                  0.67175   0.98801
## b_n_datapoints                               0.00020   0.00039
## b_mean_accuracy                             -0.00323  -0.00202
```

```
## b_ratio_urban_area                                 1.11457     1.60179
## b_relevelfactormode3refEQWalkBikeElectric        -1.48545    -0.74895
## b_relevelfactormode3refEQWalkBikeNonElectric     -1.43079    -0.99427
## b_relevelfactormode3refEQWalkCar                 -1.25808    -0.88267
## b_relevelfactormode3refEQWalkMetro               -0.06156     3.30597
## b_relevelfactormode3refEQWalkPublicBus           -1.66828    -0.38096
## b_relevelfactormode3refEQWalkScooter             -0.40187     1.65916
## b_relevelfactormode3refEQWalkTrain                0.06863     0.98842
## b_relevelfactormode3refEQWalkTram                -0.68604     1.50644
## b_relevelfactormode3refEQWalkUser_error          -1.27993     0.36891
## lp__                                          -2667.22276 -2657.33952
```

Plot Posteriors

```r
posterior_samples(model, pars = c("mean_accuracy", "n_datapoints")) %>%
  gather() %>%
  ggplot(aes(x    = value,
             y    =  key,
             fill = factor(..quantile..)))+
  stat_density_ridges(geom = "density_ridges_gradient",
                      calc_ecdf = TRUE,
                      quantiles = c(0.025, 0.975),
                      scale     = 2.5)+
    scale_fill_manual(
      name   = "Probability",
      values = c("#FF0000A0", "#A0A0A0A0", "#FF0000A0"),
      labels = c("(0, 0.025]", "(0.025, 0.975]", "(0.975, 1]"))+
    geom_vline(xintercept = 0,
               linetype = "dashed")+
    #xlim(-8,9.5)+
    theme(legend.position = "none")+
    ylab("Posteriorvdensity")+
     scale_y_discrete(labels = c("Mean accuracy",
                                 "n data points"))+
  labs(title    = "Posterior regression coefficients\naccuracy and number of data points")
```

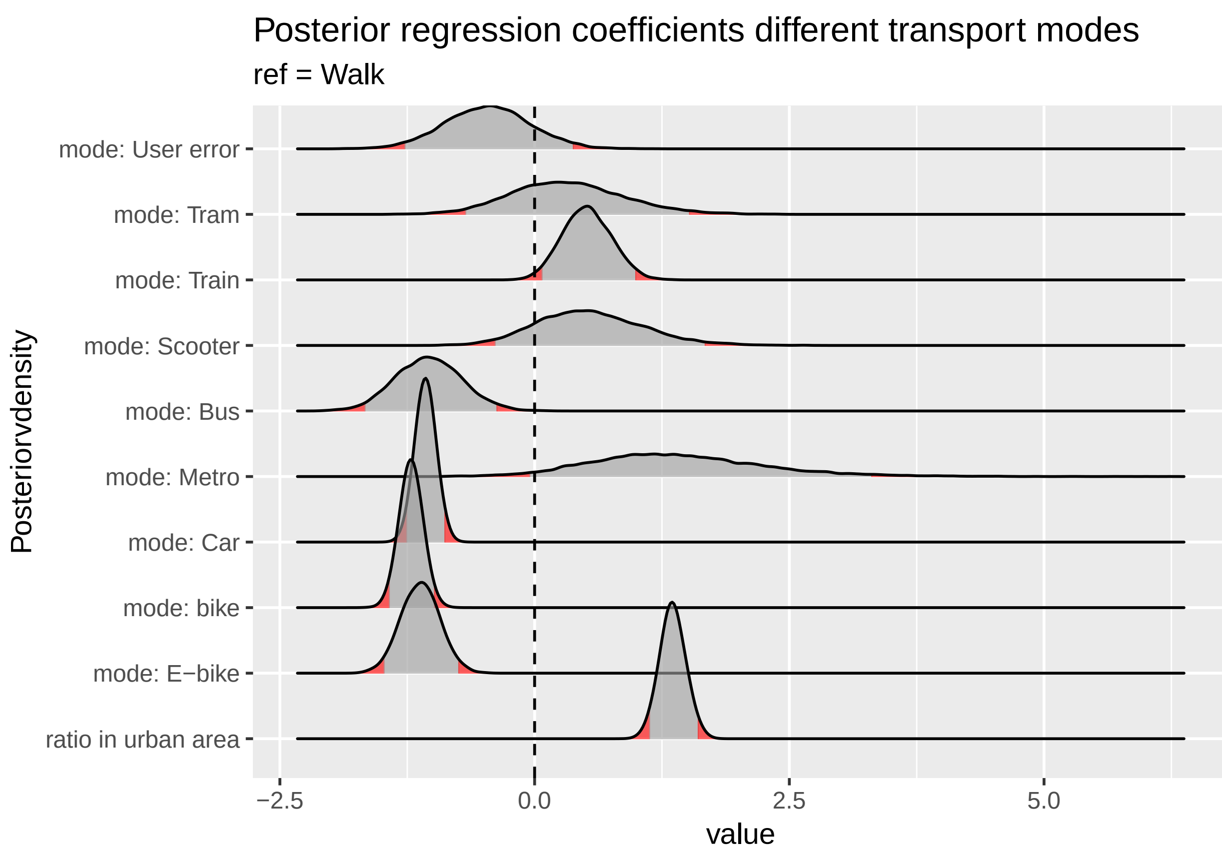
```
## Picking joint bandwidth of 2.52e-05
```




```
posterior_samples(model, pars = c("mode3", "ratio_urban_area")) %>%
  gather() %>%
  ggplot(aes(x    = value,
             y    =  key,
             fill = factor(..quantile..)))+
  stat_density_ridges(geom = "density_ridges_gradient",
                      calc_ecdf = TRUE,
                      quantiles = c(0.025, 0.975),
                      scale     = 2.5)+
  scale_fill_manual(
    name   = "Probability",
    values = c("#FF0000A0", "#A0A0A0A0", "#FF0000A0"),
    labels = c("(0, 0.025]", "(0.025, 0.975]", "(0.975, 1]"))+
  geom_vline(xintercept = 0,
             linetype = "dashed")+
  theme(legend.position = "none")+
  ylab("Posteriorvdensity")+
   scale_y_discrete(labels = c("ratio in urban area",
                               "mode: E-bike",
                               "mode: bike",
                               "mode: Car",
                               "mode: Metro",
                               "mode: Bus",
                               "mode: Scooter",
                               "mode: Train",
                               "mode: Tram",
```

```
                                "mode: User error"))+
  labs(title    = "Posterior regression coefficients different transport modes",
       subtitle = "ref = Walk")
```

```
## Picking joint bandwidth of 0.0483
```



References

- Gelman, A. (2019, May 2nd). Prior choice recommendations. Retrieved from https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations
- Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. (2008). A weakly informative default prior distribution for logistic and other regression models. The Annals of Applied Statistics, 2, 1360-1383. Doi: 10.1214/08-AOAS191
- Ghosh, J., Yi, L., Mitra, R. (2017). On the use of Cauchy prior distributions for bayesian logistic regression. The Annals of Applied Statistics