# BDA - Project brms library test

Arsi Ikäheimonen

## Contents

Load packages

```
library(aaltobda)
library(cmdstanr)
library(brms)
library(ggplot2)
library(gridExtra)
library(bayesplot)
library(ggdist)
theme_set(bayesplot::theme_default(base_family = "sans"))
library(rprojroot)
library(brms)
library(caret)
library(corrplot)
library(dplyr)
library(crosstable)
SEED <- 614273
```

Load data

```
data <- read.csv('Machine-Learning-with-R-datasets/insurance.csv')
head(data)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

Typecasting

```
data$region <- as.factor(data$region)
data$sex <- as.factor(data$sex)
data$smoker <- as.factor(data$smoker)
data$children <- as.integer(data$children)
data$region = recode(data$region, "southeast" = "south", "southwest" = "south", "northeast" = "north",
head(data)
```

```
##   age    sex    bmi children smoker region   charges
## 1  19 female 27.900        0    yes  south 16884.924
## 2  18   male 33.770        1     no  south  1725.552
## 3  28   male 33.000        3     no  south  4449.462
## 4  33   male 22.705        0     no  north 21984.471
## 5  32   male 28.880        0     no  north  3866.855
## 6  31 female 25.740        0     no  south  3756.622
```

Scaler functions

```
min_max_scaler <- function(values){
  scaled_data = (values - min(values)) / (max(values) - min(values))
  return(scaled_data)
}

descaler <- function(values, max, min){
  descaled_data = values*(max-min) + min
  return(descaled_data)
}
```

Scale the data

```
data$scaled_charges = min_max_scaler(data$charges)
data$scaled_age = min_max_scaler(data$age)
data$scaled_bmi = min_max_scaler(data$bmi)
data$scaled_children = min_max_scaler(data$children)
head(data)
```

```
##   age    sex    bmi children smoker region   charges scaled_charges scaled_age
## 1  19 female 27.900        0    yes  south 16884.924    0.251610757 0.02173913
## 2  18   male 33.770        1     no  south  1725.552    0.009635951 0.00000000
## 3  28   male 33.000        3     no  south  4449.462    0.053115162 0.21739130
## 4  33   male 22.705        0     no  north 21984.471    0.333010027 0.32608696
## 5  32   male 28.880        0     no  north  3866.855    0.043815557 0.30434783
## 6  31 female 25.740        0     no  south  3756.622    0.042056002 0.28260870
##   scaled_bmi scaled_children
## 1  0.3212268             0.0
## 2  0.4791499             0.2
## 3  0.4584342             0.6
## 4  0.1814635             0.0
## 5  0.3475921             0.0
## 6  0.2631154             0.0
```

Train / test data

```
inTrain <- createDataPartition(
  y = data$smoker,
  ## the outcome data are needed
  p = .75,
  ## The percentage of data in the
  ## training set
  list = FALSE
)

train_data <- data[inTrain,]
test_data <- data[-inTrain,]
```

Basic frequentist linear model

```
basic_model = lm(charges~age+sex+bmi+children+smoker+region, data = data) #Create the linear regression
summary(basic_model) #Review the results
```

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
##     region, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11281.1  -2825.0   -988.2   1336.0  29949.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12093.76     949.71 -12.734  < 2e-16 ***
## age            256.95      11.89  21.616  < 2e-16 ***
## sexmale       -130.29     332.76  -0.392 0.695459
## bmi            338.38      28.17  12.013  < 2e-16 ***
## children       473.12     137.61   3.438 0.000604 ***
## smokeryes    23851.76     411.95  57.899  < 2e-16 ***
## regionsouth   -820.68     341.26  -2.405 0.016317 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6059 on 1331 degrees of freedom
## Multiple R-squared:  0.7508, Adjusted R-squared:  0.7497
## F-statistic: 668.4 on 6 and 1331 DF,  p-value: < 2.2e-16
```

Check the data balance

```
crosstable(data, c(smoker,), by=region)
```

```
## # A tibble: 2 x 5
##   .id    label  variable north        south
##   <chr>  <chr>  <chr>    <chr>        <chr>
## 1 smoker smoker no       524 (49.25%) 540 (50.75%)
## 2 smoker smoker yes      125 (45.62%) 149 (54.38%)
```

PRIORS:

```
# Prior choice (used for all the models except the hierarchical model)
pr = prior(normal(0,1), class = 'b')

pr2 = (prior(normal(0,1), class = "b", coef = "scaled_age") +
            prior(student_t(3,0,0.1), class="sd", group="smoker"))

pr3 = (prior(normal(0,1), class = "b", coef = "scaled_age") +
       prior(normal(0,1), class="b", coef ="scaled_bmi") +
       prior(student_t(3,0,0.1), class="sd", group="smoker"))

pr4 = (prior(normal(0,1), class = "b", coef = "scaled_age") +
       prior(normal(0,1), class="b", coef ="scaled_bmi") +
       prior(student_t(3,0,0.1), class="sd", group="smoker") +
       prior(student_t(3,0,0.1), class="sd", group="region"))

pr5 = (prior(normal(0,1), class = "b", coef = "scaled_age") +
      prior(normal(0,1), class = "b", coef = "sexmale") +
      prior(normal(0,1), class="b", coef ="scaled_bmi") +
      prior(normal(0,1), class="b", coef ="scaled_children") +
      prior(student_t(3,0,0.1), class="sd", group="region") +
      prior(student_t(3,0,0.1), class="sd", group="smoker"))
```

MODELS:

```
# Baseline model
model_baseline = brm(
  scaled_charges ~ scaled_age,
  data  = train_data,
  prior = pr,
  cores = 4
)

# 2 effect model with BMI
```

```r
model_2_test1 = brm(
  scaled_charges ~ scaled_age + scaled_bmi,
  data  = train_data,
  prior = pr,
  cores = 4
)

# 2 effect non-hierarchical model with smoker
model_2_test2 = brm(
  scaled_charges ~ scaled_age + smoker,
  data  = train_data,
  prior = pr,
  cores = 4
)

# 2 effect hierarchical model with smoker
model_2 = brm(
  scaled_charges ~ scaled_age + (1|smoker),
  data  = train_data,
  prior = pr2,
  cores = 4
)
```

```r
# 3 effect hierarchical model
model_3 = brm(
  scaled_charges ~ scaled_age + scaled_bmi + (1|smoker),
  data  = train_data,
  prior = pr3,
  cores = 4
)

# 4 effect hierarchical model
model_4 = brm(
  scaled_charges ~ scaled_age + scaled_bmi + (1|smoker) + (1|region),
  data  = train_data,
  prior = pr4,
  cores = 4
)

# 5 effect non-hierarchical model
model_5_non_hier = brm(
  scaled_charges ~ scaled_age + sex + scaled_bmi + scaled_children + region + smoker,
  data  = train_data,
  prior = pr,
  cores = 4
)

# 5 effect hierarchical model
model_5 = brm(
  scaled_charges ~ scaled_age + sex + scaled_bmi + scaled_children + (1|region) + (1|smoker),
  data  = train_data,
  prior = pr5,
  cores = 4,
  control = list(adapt_delta = 0.9)
)
```

```
summary(model_baseline)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: scaled_charges ~ scaled_age
##    Data: train_data (Number of observations: 1004)
##   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      0.10      0.01     0.08     0.12 1.00     3966     3113
## scaled_age     0.20      0.02     0.16     0.23 1.00     3616     2924
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.18      0.00     0.18     0.19 1.00     3258     2749
##
## 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).
```

```
loo_baseline = loo(model_baseline)
loo_baseline
```

```
##
## Computed from 4000 by 1004 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo    272.6 31.3
## p_loo         4.2  0.4
## looic      -545.1 62.7
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
summary(model_2_test1)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: scaled_charges ~ scaled_age + scaled_bmi
##    Data: train_data (Number of observations: 1004)
##   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      0.03      0.02    -0.00     0.06 1.00     5075     3300
## scaled_age     0.18      0.02     0.15     0.22 1.00     3662     2808
## scaled_bmi     0.20      0.04     0.13     0.27 1.00     4341     3399
##
```

```
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.18      0.00     0.17     0.19 1.00     3580     2658
##
## 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).
```

```
loo_2_test1 = loo(model_2_test1)
loo_2_test1
```

```
##
## Computed from 4000 by 1004 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo    286.7 28.8
## p_loo         4.9  0.4
## looic      -573.3 57.6
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
summary(model_2_test2)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: scaled_charges ~ scaled_age + smoker
##    Data: train_data (Number of observations: 1004)
##   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      0.02      0.01     0.01     0.03 1.00     3997     3083
## scaled_age     0.20      0.01     0.18     0.22 1.00     3916     3045
## smokeryes      0.38      0.01     0.36     0.39 1.00     3923     3004
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.10      0.00     0.10     0.11 1.00     3287     2772
##
## 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).
```

```
loo_2_test2 = loo(model_2_test2)
loo_2_test2
```

```
##
## Computed from 4000 by 1004 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo    865.5 34.1
## p_loo         6.4  0.6
## looic     -1730.9 68.3
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
summary(model_2)
```

```
## Warning: There were 66 divergent transitions after warmup. Increasing
## adapt_delta above 0.8 may help. See http://mc-stan.org/misc/
## warnings.html#divergent-transitions-after-warmup
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: scaled_charges ~ scaled_age + (1 | smoker)
##    Data: train_data (Number of observations: 1004)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~smoker (Number of levels: 2)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.22      0.10     0.10     0.49 1.00      722      271
##
## Population-Level Effects:
##            Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept      0.21      0.16    -0.13     0.56 1.00      981      939
## scaled_age     0.20      0.01     0.18     0.22 1.00     2884     2391
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.10      0.00     0.10     0.11 1.00     2310     2405
##
## 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).
```

```
loo_2 = loo(model_2)
loo_2
```

```
##
## Computed from 4000 by 1004 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo    865.3 34.1
## p_loo         6.6  0.6
## looic     -1730.6 68.3
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
summary(model_3)
```

```
## Warning: There were 60 divergent transitions after warmup. Increasing
## adapt_delta above 0.8 may help. See http://mc-stan.org/misc/
## warnings.html#divergent-transitions-after-warmup
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: scaled_charges ~ scaled_age + scaled_bmi + (1 | smoker)
##    Data: train_data (Number of observations: 1004)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~smoker (Number of levels: 2)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.22      0.11     0.10     0.50 1.01      870      508
##
## Population-Level Effects:
##            Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept      0.14      0.17    -0.19     0.48 1.00      894     1025
## scaled_age     0.19      0.01     0.17     0.21 1.00     3275     2730
## scaled_bmi     0.20      0.02     0.16     0.24 1.00     3411     2224
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.10      0.00     0.09     0.10 1.00     3085     2775
##
## 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).
```

```
loo_3 = loo(model_3)
loo_3
```

```
##
## Computed from 4000 by 1004 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo    916.3 34.4
## p_loo         7.2  0.7
## looic     -1832.6 68.8
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
summary(model_4)
```

```
## Warning: There were 101 divergent transitions after warmup. Increasing
## adapt_delta above 0.8 may help. See http://mc-stan.org/misc/
## warnings.html#divergent-transitions-after-warmup
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: scaled_charges ~ scaled_age + scaled_bmi + (1 | smoker) + (1 | region)
##    Data: train_data (Number of observations: 1004)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~region (Number of levels: 2)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.05      0.05     0.00     0.18 1.00     1169     1597
##
## ~smoker (Number of levels: 2)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.22      0.10     0.10     0.49 1.00      920      346
##
## Population-Level Effects:
##            Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept      0.14      0.16    -0.20     0.48 1.00     1377     1526
## scaled_age     0.19      0.01     0.17     0.21 1.00     3478     2548
## scaled_bmi     0.21      0.02     0.17     0.25 1.00     3473     2243
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.10      0.00     0.09     0.10 1.00     3674     2725
##
## 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).
```

```
loo_4 = loo(model_4)
loo_4
```

```
##
## Computed from 4000 by 1004 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo    918.8 34.3
## p_loo         8.0  0.7
## looic     -1837.5 68.6
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
summary(model_5_non_hier)
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: scaled_charges ~ scaled_age + sex + scaled_bmi + scaled_children + region + smoker
##    Data: train_data (Number of observations: 1004)
##   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          -0.05      0.01    -0.07    -0.03 1.00     8031     3261
## scaled_age          0.19      0.01     0.17     0.21 1.00     6447     3084
## sexmale            -0.01      0.01    -0.02     0.01 1.00     8166     3398
## scaled_bmi           0.21      0.02     0.17     0.25 1.00     6876     3485
## scaled_children      0.04      0.01     0.02     0.07 1.00     6496     3088
## regionsouth         -0.02      0.01    -0.03    -0.00 1.00     6027     3367
## smokeryes            0.38      0.01     0.37     0.40 1.00     6521     3313
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.10      0.00     0.09     0.10 1.00     6980     3080
##
## 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).
```

```
loo_5_non_hier = loo(model_5_non_hier)
loo_5_non_hier
```

```
##
## Computed from 4000 by 1004 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo    923.3 34.6
## p_loo        10.2  0.9
## looic     -1846.6 69.3
## ------
## Monte Carlo SE of elpd_loo is 0.0.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

```
summary(model_5)
```

```
## Warning: There were 30 divergent transitions after warmup. Increasing
## adapt_delta above 0.9 may help. See http://mc-stan.org/misc/
## warnings.html#divergent-transitions-after-warmup
```

```
##  Family: gaussian
##   Links: mu = identity; sigma = identity
## Formula: scaled_charges ~ scaled_age + sex + scaled_bmi + scaled_children + (1 | region) + (1 | smoke
##    Data: train_data (Number of observations: 1004)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~region (Number of levels: 2)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.05      0.06     0.00     0.21 1.00     1558     1964
##
## ~smoker (Number of levels: 2)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)     0.22      0.10     0.10     0.48 1.00     1933     1842
##
## Population-Level Effects:
##                 Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept           0.13      0.18    -0.24     0.47 1.00     1199     1424
## scaled_age          0.19      0.01     0.17     0.21 1.00     4470     3083
## sexmale            -0.01      0.01    -0.02     0.01 1.00     4909     2873
## scaled_bmi          0.21      0.02     0.17     0.25 1.00     3603     2337
## scaled_children     0.04      0.01     0.02     0.07 1.00     4204     3063
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma     0.10      0.00     0.09     0.10 1.00     3963     2837
##
## 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).
```

```
loo_5 = loo(model_5)
loo_5
```

```
##
## Computed from 4000 by 1004 log-likelihood matrix
##
##          Estimate   SE
## elpd_loo    923.0 34.6
## p_loo        10.5  0.9
## looic     -1846.0 69.2
## ------
## Monte Carlo SE of elpd_loo is 0.1.
##
## All Pareto k estimates are good (k < 0.5).
## See help('pareto-k-diagnostic') for details.
```

Loo comparison

```
loo_compare(loo_baseline,loo_2_test1,loo_2_test2,loo_2,loo_3,loo_4,loo_5_non_hier,loo_5)
```

```
##                 elpd_diff se_diff
## model_5_non_hier    0.0       0.0
## model_5            -0.3       0.2
## model_4            -4.5       3.6
## model_3            -7.0       4.6
## model_2_test2     -57.8      11.2
## model_2           -58.0      11.2
## model_2_test1    -636.6      33.4
## model_baseline   -650.7      35.7
```

```
p1 <- pp_check(model_baseline) +
  ggtitle("Baseline model")
```

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

```
p2 <- pp_check(model_2) +
  ggtitle("2 effect model ")
```

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

```
p3 <- pp_check(model_3) +
  ggtitle("3 effect model ")
```

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

```
p4 <- pp_check(model_4) +
  ggtitle("4 effect model ")
```

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

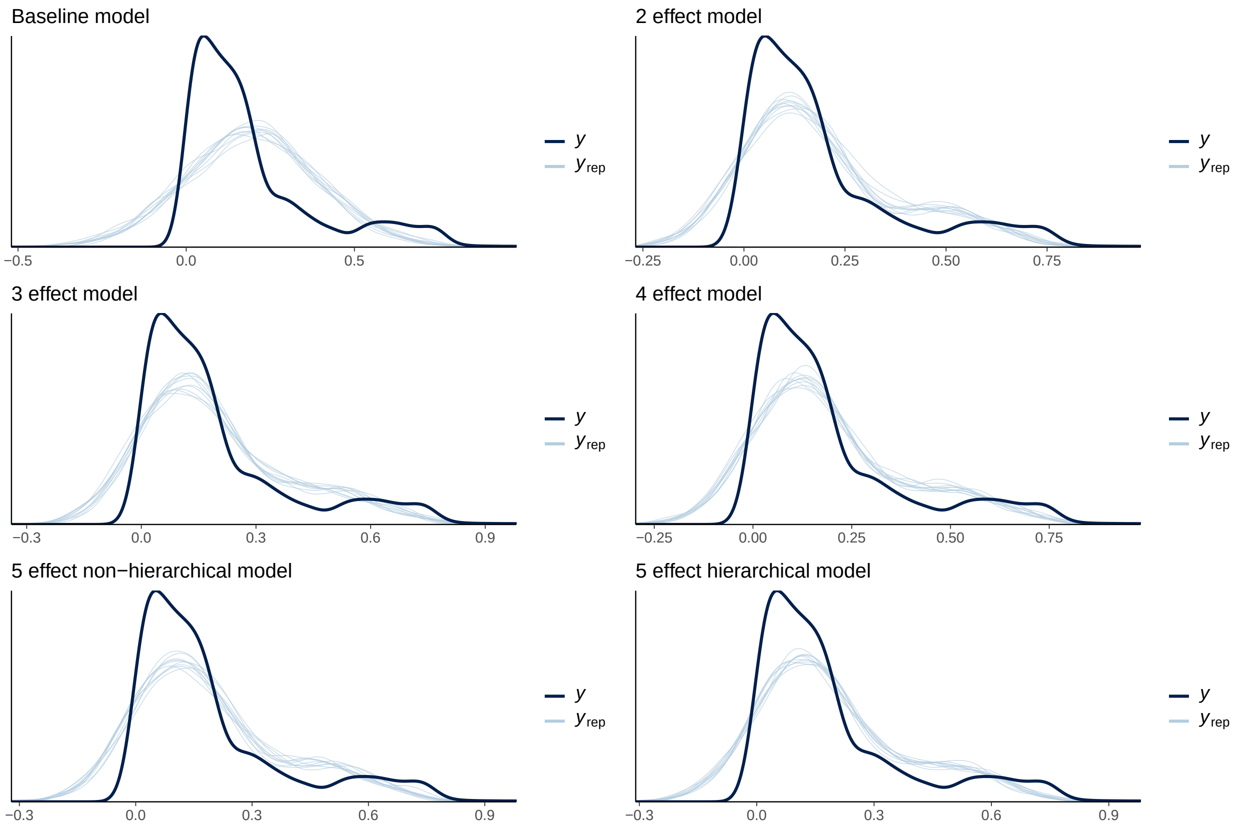
```
p5 <- pp_check(model_5_non_hier) +
  ggtitle("5 effect non-hierarchical model")
```

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

```
p6 <- pp_check(model_5) +
  ggtitle("5 effect hierarchical model ")
```

```
## Using 10 posterior draws for ppc type 'dens_overlay' by default.
```

```
grid.arrange(p1, p2, p3, p4, p5, p6, nrow=3)
```



Posterior prediction with test data

```
pp_baseline = posterior_predict(model_baseline,newdata=test_data)
pp_2 = posterior_predict(model_2,newdata=test_data)
pp_3 = posterior_predict(model_3,newdata=test_data)
pp_4 = posterior_predict(model_4,newdata=test_data)
pp_5_non_hier = posterior_predict(model_5_non_hier,newdata=test_data)
pp_5 = posterior_predict(model_5,newdata=test_data)

# descaling the data
cmax = max(train_data$charges)
cmin = min(train_data$charges)

pp_baseline = descaler(pp_baseline, cmax, cmin)
pp_2 = descaler(pp_2, cmax, cmin)
pp_3 = descaler(pp_3, cmax, cmin)
```

```
pp_4 = descaler(pp_4, cmax, cmin)
pp_5_non_hier = descaler(pp_5_non_hier, cmax, cmin)
pp_5 = descaler(pp_5, cmax, cmin)
```

Calculate prediction errors

```
calculate_rmse <- function(true,predicted){
  rmse = sqrt(mean((true - predicted)^2))
  return(rmse)
}

calculate_r2 <- function(true, predicted){
  rss = sum((predicted - true)^2)
  tss = sum((true - mean(true))^2)
  return(1 - rss/tss)
}

rmse_baseline = calculate_rmse(test_data$charges, pp_baseline)
rmse_2 = calculate_rmse(test_data$charges, pp_2)
rmse_3 = calculate_rmse(test_data$charges, pp_3)
rmse_4 = calculate_rmse(test_data$charges, pp_4)
rmse_5_non_hier = calculate_rmse(test_data$charges, pp_5_non_hier)
rmse_5 = calculate_rmse(test_data$charges, pp_5)

#r2_baseline = calculate_r2(test_data$charges, pp_basic)
#r2_5 = calculate_r2(test_data$charges, pp_complex)

sprintf("Baseline model RMSE: %s", rmse_baseline)
```

```
## [1] "Baseline model RMSE: 16952.4919908481"
```

```
sprintf("2 effect model RMSE: %s", rmse_2)
```

```
## [1] "2 effect model RMSE: 16800.2523887844"
```

```
sprintf("3 effect model RMSE: %s", rmse_3)
```

```
## [1] "3 effect model RMSE: 16829.0142961785"
```

```
sprintf("4 effect model RMSE: %s", rmse_4)
```

```
## [1] "4 effect model RMSE: 16829.4580478745"
```

```
sprintf("5 effect non-hierarchical model RMSE: %s", rmse_5_non_hier)
```

```
## [1] "5 effect non-hierarchical model RMSE: 16845.572685129"
```

```
sprintf("5 effect hierarchical model RMSE: %s", rmse_5)
```

```
## [1] "5 effect hierarchical model RMSE: 16844.5583484642"
```

```
#sprintf("Baseline model R2: %s", r2_baseline)
#sprintf("5 effect model R2: %s", r2_5)
```