BDA - Project brms library test

Arsi Ikäheimonen

Contents

Load packages

```
library(aaltobda)
library(LaplacesDemon)
library(cmdstanr)
library(posterior)
library(loo)
library(tidyr)
library(dplyr)
options(pillar.neg=FALSE)
library(ggplot2)
library(gridExtra)
library(bayesplot)
library(ggdist)
theme_set(bayesplot::theme_default(base_family = "sans"))
library(rprojroot)
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
```

Check for null values

```
colSums(is.na(data))
```

```
##      age      sex      bmi children   smoker   region  charges
##        0        0        0        0        0        0        0
```

Some 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)
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
```

Summary statistics of the data

```
summary(data)
```

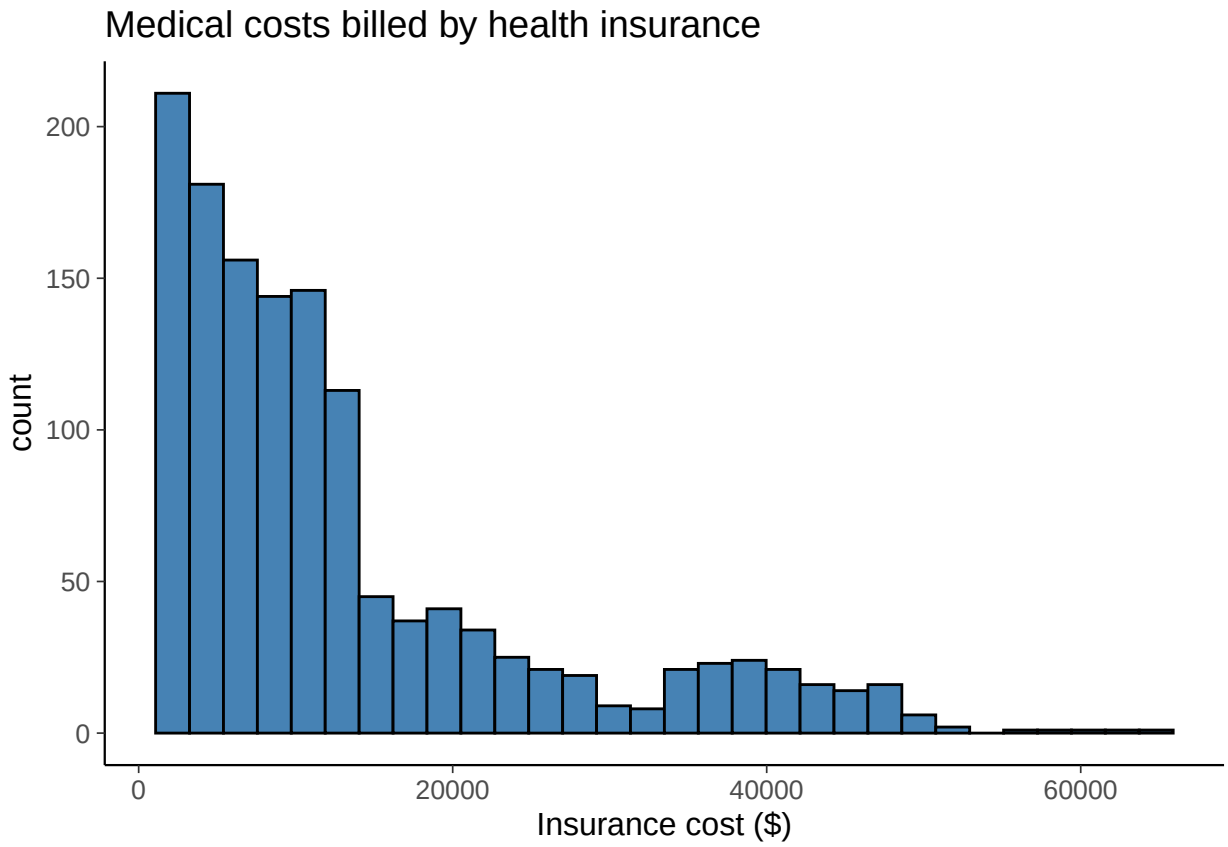
```
##       age            sex           bmi           children     smoker
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064
```

```
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274
##  Median :39.00                 Median :30.40   Median :1.000
##  Mean   :39.21                 Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                 3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                 Max.   :53.13   Max.   :5.000
##        region       charges
##  northeast:324   Min.   : 1122
##  northwest:325   1st Qu.: 4740
##  southeast:364   Median : 9382
##  southwest:325   Mean   :13270
##                  3rd Qu.:16640
##                  Max.   :63770
```

Plot histogram of the insurance costs

```
ggplot() +
  geom_histogram(aes(data$charges), fill = 'steelblue', color = 'black') +
  labs(title = 'Medical costs billed by health insurance', x='Insurance cost ($)')
```

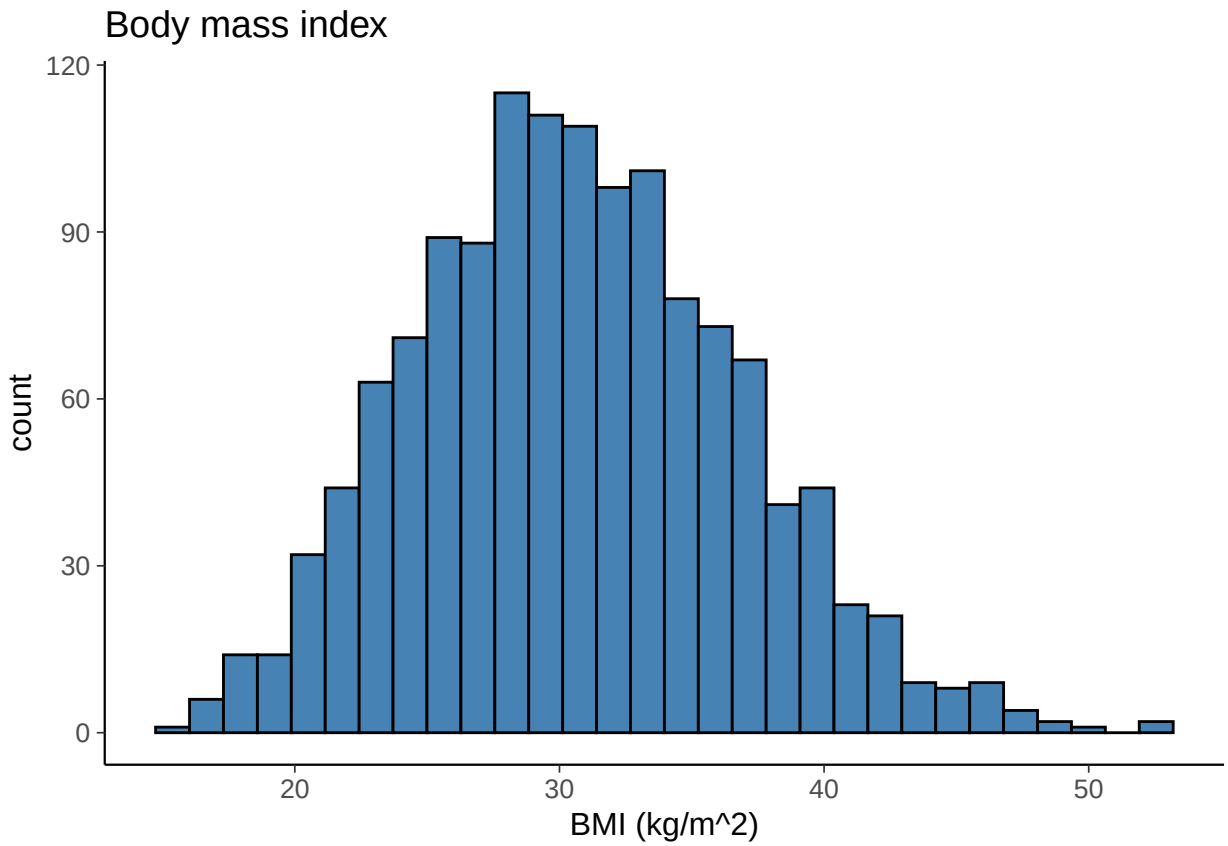
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Plot histogram of the BMI values

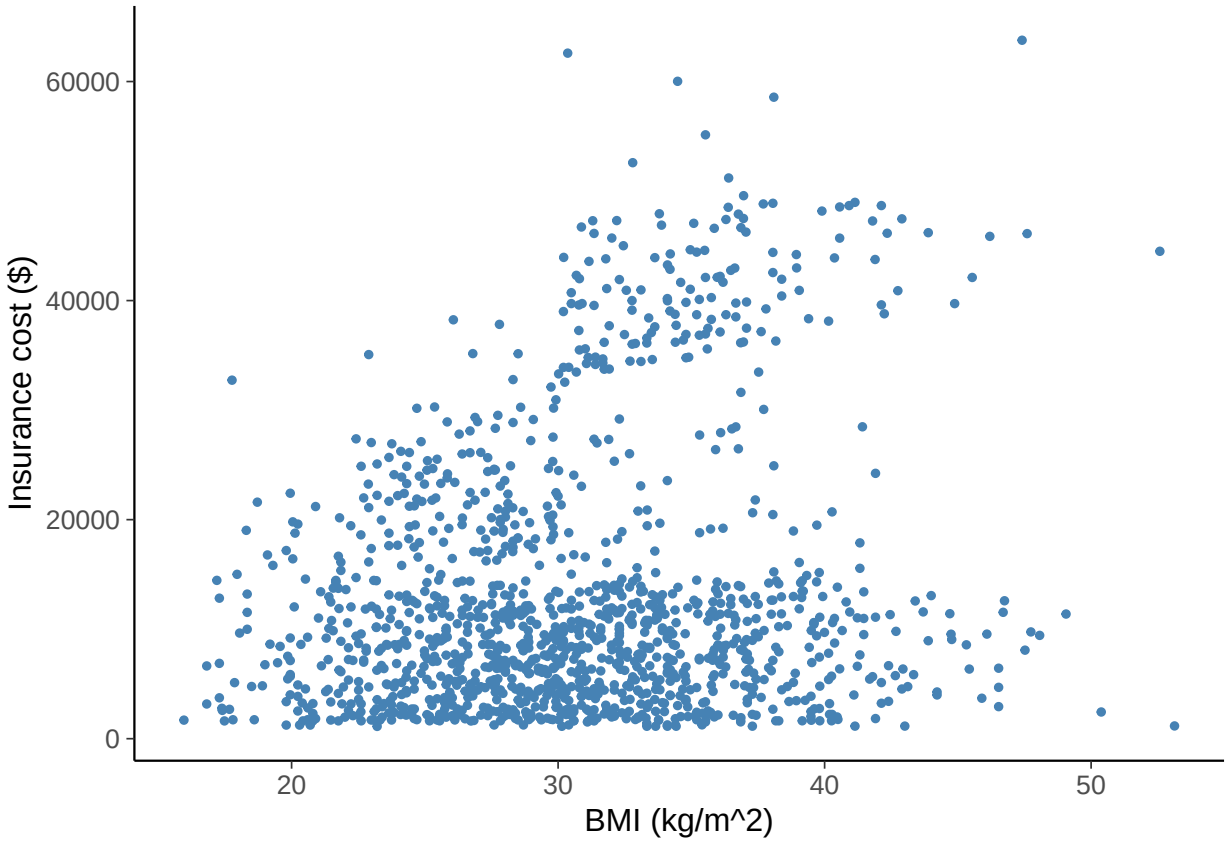
```
ggplot() +
  geom_histogram(aes(data$bmi), fill = 'steelblue', color = 'black') +
  labs(title = 'Body mass index', x='BMI (kg/m^2)')
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



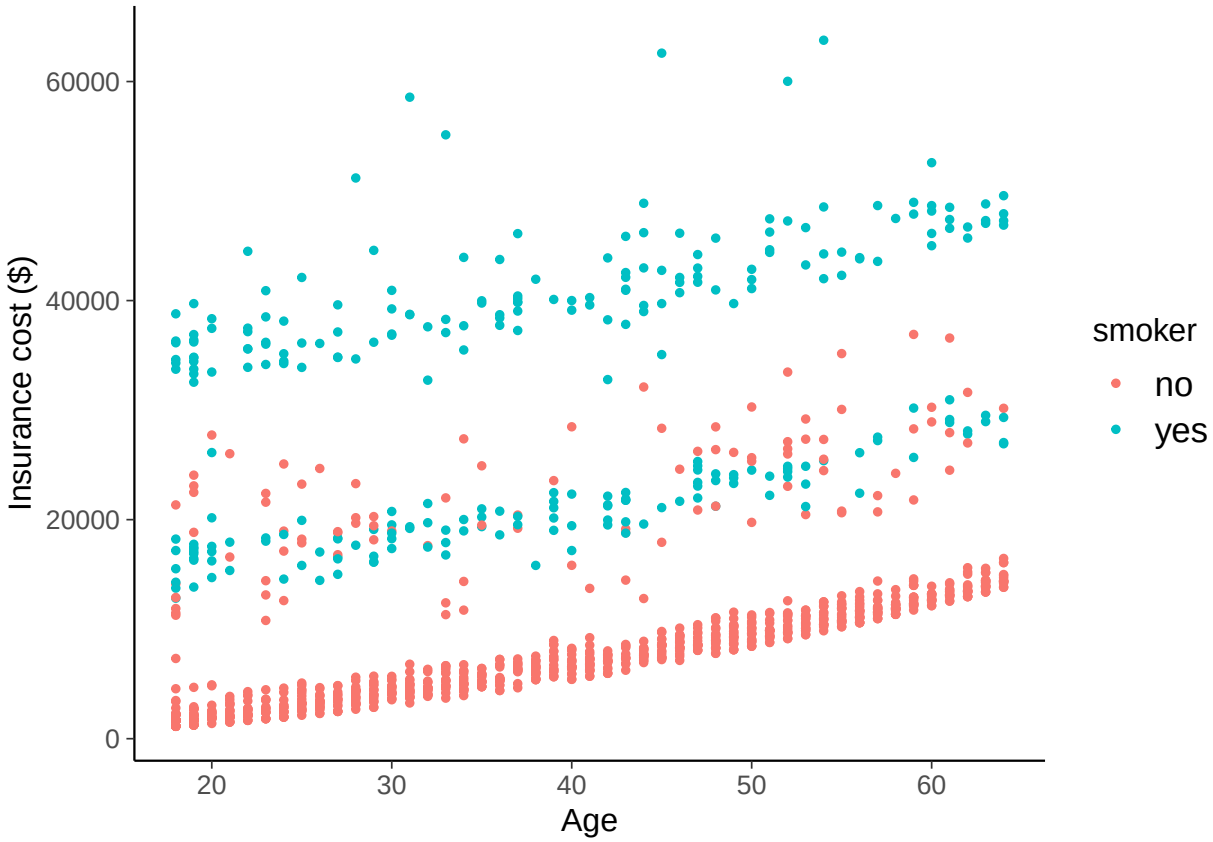
Plot scatter plot of the insurance costs with bmi as x-axis.

```
ggplot(data) +
  geom_point(aes(x=bmi, y=charges), size = 1, color = 'steelblue') +
  labs(y = 'Insurance cost ($)', x= 'BMI (kg/m^2)') +
  guides(linetype = "none")
```



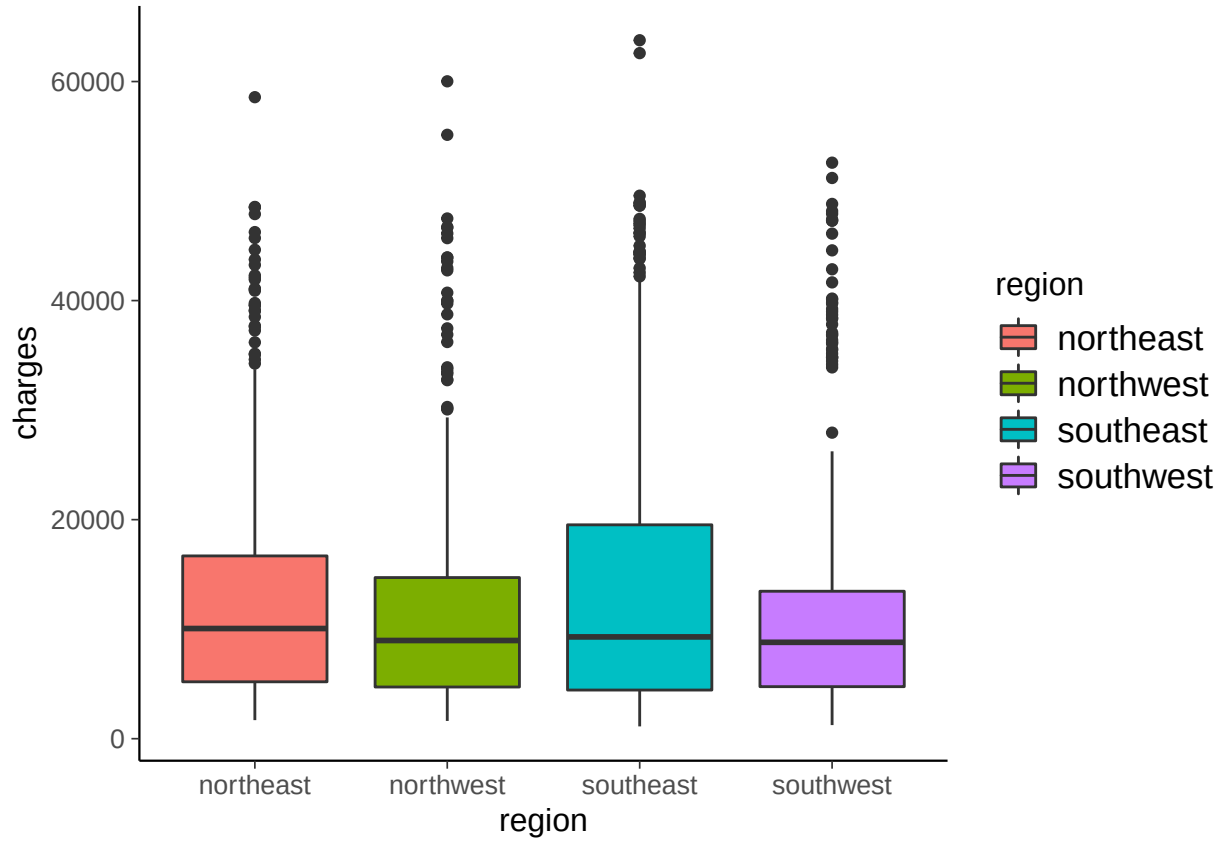
Plot scatter plot of the insurance costs with age as x-axis.

```
ggplot(data,aes(x=age,y=charges,col=smoker)) +
  geom_point(size = 1,) +
  labs(y = 'Insurance cost ($)', x= 'Age') +
  guides(linetype = "none")
```



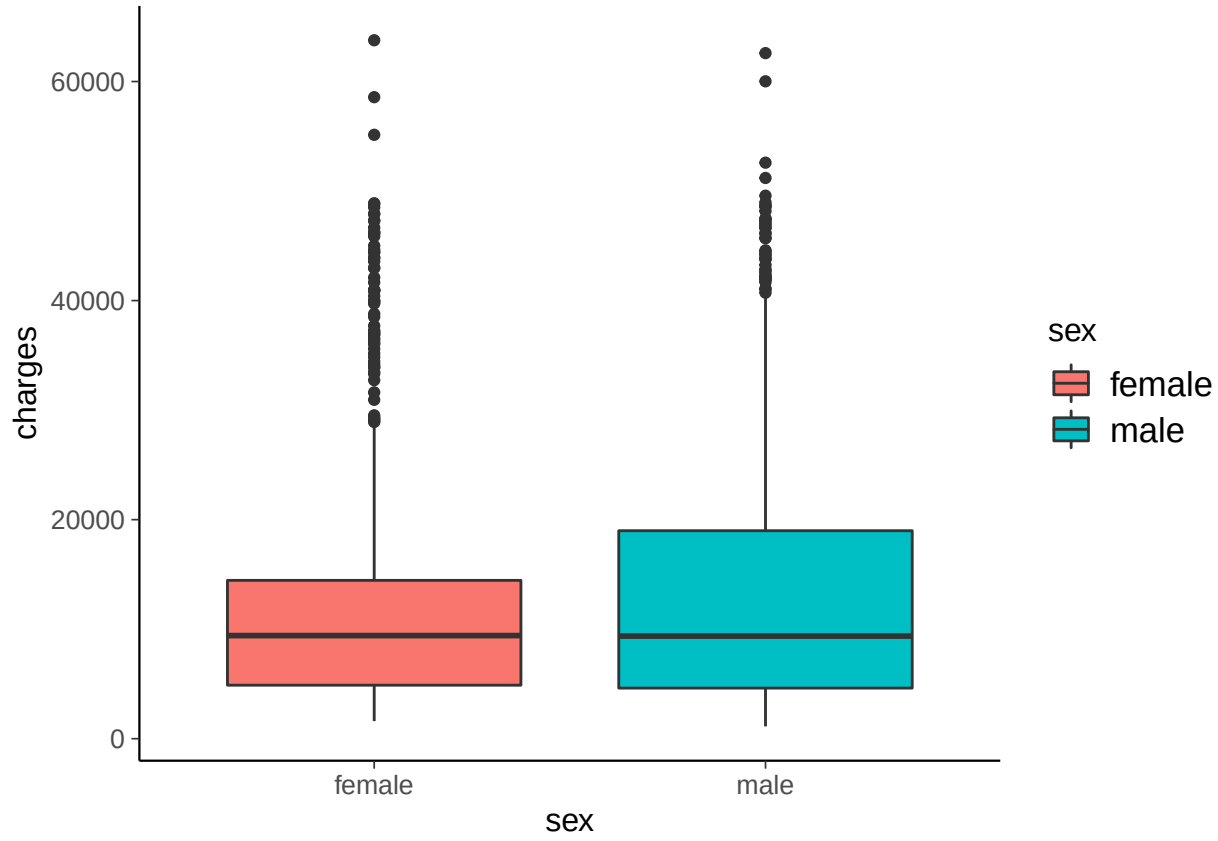
Charges vs region

```
ggplot(data, aes(x=region, y=charges, fill=region)) +
    geom_boxplot()
```



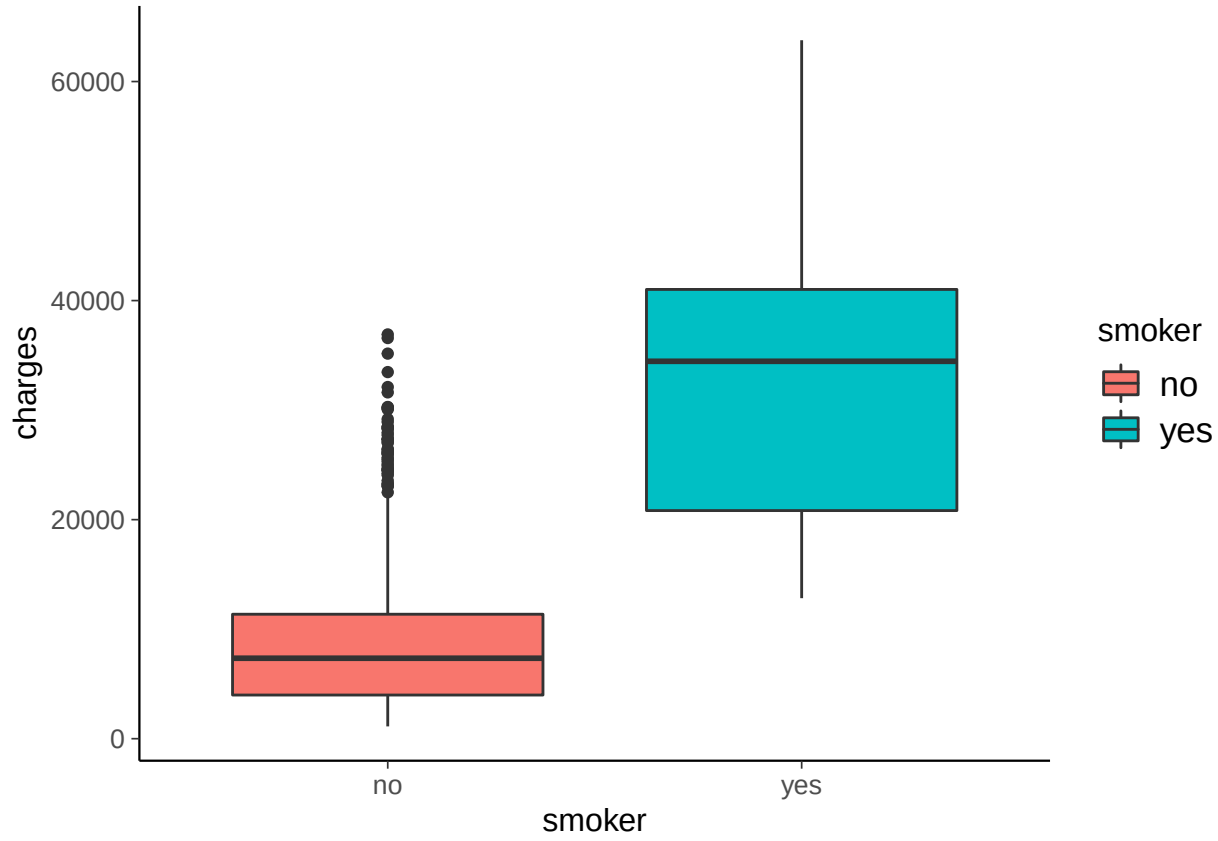
Charges vs sex

```
ggplot(data, aes(x=sex, y=charges, fill=sex)) +
    geom_boxplot()
```



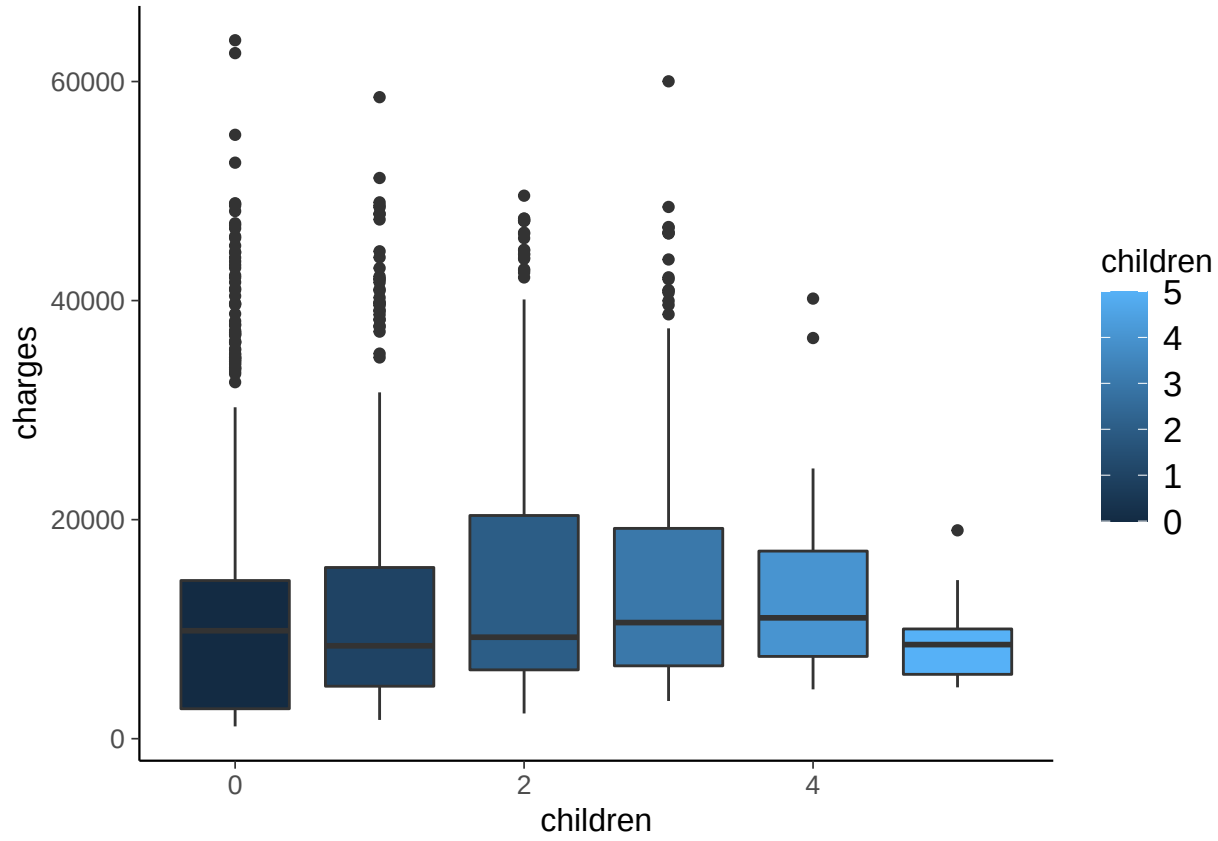
Charges vs smoker

```
ggplot(data, aes(x=smoker, y=charges, fill=smoker)) +
    geom_boxplot()
```



Charges vs children

```
ggplot(data, aes(x=children, y=charges, fill=children, group=children)) +
    geom_boxplot()
```



Basic 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
## -11304.9  -2848.1   -982.1   1393.9  29992.8
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11938.5      987.8 -12.086  < 2e-16 ***
## age                256.9       11.9  21.587  < 2e-16 ***
## sexmale           -131.3      332.9  -0.394 0.693348
## bmi                339.2       28.6  11.860  < 2e-16 ***
## children           475.5      137.8   3.451 0.000577 ***
## smokeryes        23848.5      413.1  57.723  < 2e-16 ***
## regionnorthwest   -353.0      476.3  -0.741 0.458769
## regionsoutheast  -1035.0      478.7  -2.162 0.030782 *
## regionsouthwest   -960.0      477.9  -2.009 0.044765 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.8 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Bauesian mode with brms

```
library(brms)
```

```
## Loading required package: Rcpp
```

```
## Loading 'brms' package (version 2.17.0). Useful instructions
## can be found by typing help('brms'). A more detailed introduction
## to the package is available through vignette('brms_overview').
```

```
##
## Attaching package: 'brms'
```

```
## The following objects are masked from 'package:ggdist':
##
##     dstudent_t, pstudent_t, qstudent_t, rstudent_t
```

```
## The following object is masked from 'package:posterior':
##
##     rhat
```

```
## The following objects are masked from 'package:LaplacesDemon':
##
##     ddirichlet, rdirichlet, WAIC
```

```
## The following object is masked from 'package:stats':
##
##     ar
```

```
pr = prior(normal(0, 10), class = 'b')

bayesian_mixed = brm(
  charges ~ age + sex + bmi + children + (1|region) + (1|smoker),
  data  = data,
  prior = pr,
  cores = 4
)
```

```
## Compiling Stan program...
```

```
## Trying to compile a simple C file
```

```
## Running /usr/lib/R/bin/R CMD SHLIB foo.c
## clang -flto=thin -I"/usr/share/R/include" -DNDEBUG   -I"/usr/local/lib/R/site-library/Rcpp/include/"
## In file included from <built-in>:1:
## In file included from /usr/local/lib/R/site-library/StanHeaders/include/stan/math/prim/mat/fun/Eigen
## In file included from /usr/local/lib/R/site-library/RcppEigen/include/Eigen/Dense:1:
## In file included from /usr/local/lib/R/site-library/RcppEigen/include/Eigen/Core:88:
## /usr/local/lib/R/site-library/RcppEigen/include/Eigen/src/Core/util/Macros.h:628:1: error: unknown ty
## namespace Eigen {
## ^
## /usr/local/lib/R/site-library/RcppEigen/include/Eigen/src/Core/util/Macros.h:628:16: error: expected
## namespace Eigen {
##                ^
##                ;
```

```
## In file included from <built-in>:1:
## In file included from /usr/local/lib/R/site-library/StanHeaders/include/stan/math/prim/mat/fun/Eigen
## In file included from /usr/local/lib/R/site-library/RcppEigen/include/Eigen/Dense:1:
## /usr/local/lib/R/site-library/RcppEigen/include/Eigen/Core:96:10: fatal error: 'complex' file not fou
## #include <complex>
##          ^~~~~~~~
## 3 errors generated.
## make: *** [/usr/lib/R/etc/Makeconf:168: foo.o] Error 1
```

```
## Start sampling
```

```
## Warning: There were 93 divergent transitions after warmup. See
## https://mc-stan.org/misc/warnings.html#divergent-transitions-after-warmup
## to find out why this is a problem and how to eliminate them.
```

```
## Warning: Examine the pairs() plot to diagnose sampling problems
```

Model summary

```
summary(bayesian_mixed, waic=TRUE)
```

```
## Warning: There were 93 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: charges ~ age + sex + bmi + children + (1 | region) + (1 | smoker)
##    Data: data (Number of observations: 1338)
##   Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
##          total post-warmup draws = 4000
##
## Group-Level Effects:
## ~region (Number of levels: 4)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept)   450.43    576.38    10.03  1926.71 1.00     1315     1788
##
## ~smoker (Number of levels: 2)
##               Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sd(Intercept) 13676.06   5425.23  6341.03 27984.04 1.00     1671     1634
##
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept  9721.70   6897.98 -3713.91 23691.93 1.00     1714     2009
## age          99.42      8.43    82.28   115.54 1.00     4133     2972
## sexmale       0.09      9.97   -20.49    19.62 1.00     3746     2918
## bmi          35.09      9.49    16.69    53.45 1.00     3896     2536
## children      2.14      9.97   -17.96    22.29 1.00     3731     2547
##
## Family Specific Parameters:
##       Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## sigma  6794.02    138.12  6522.47  7071.38 1.00     3519     2737
##
## 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).
```

STAN code generation

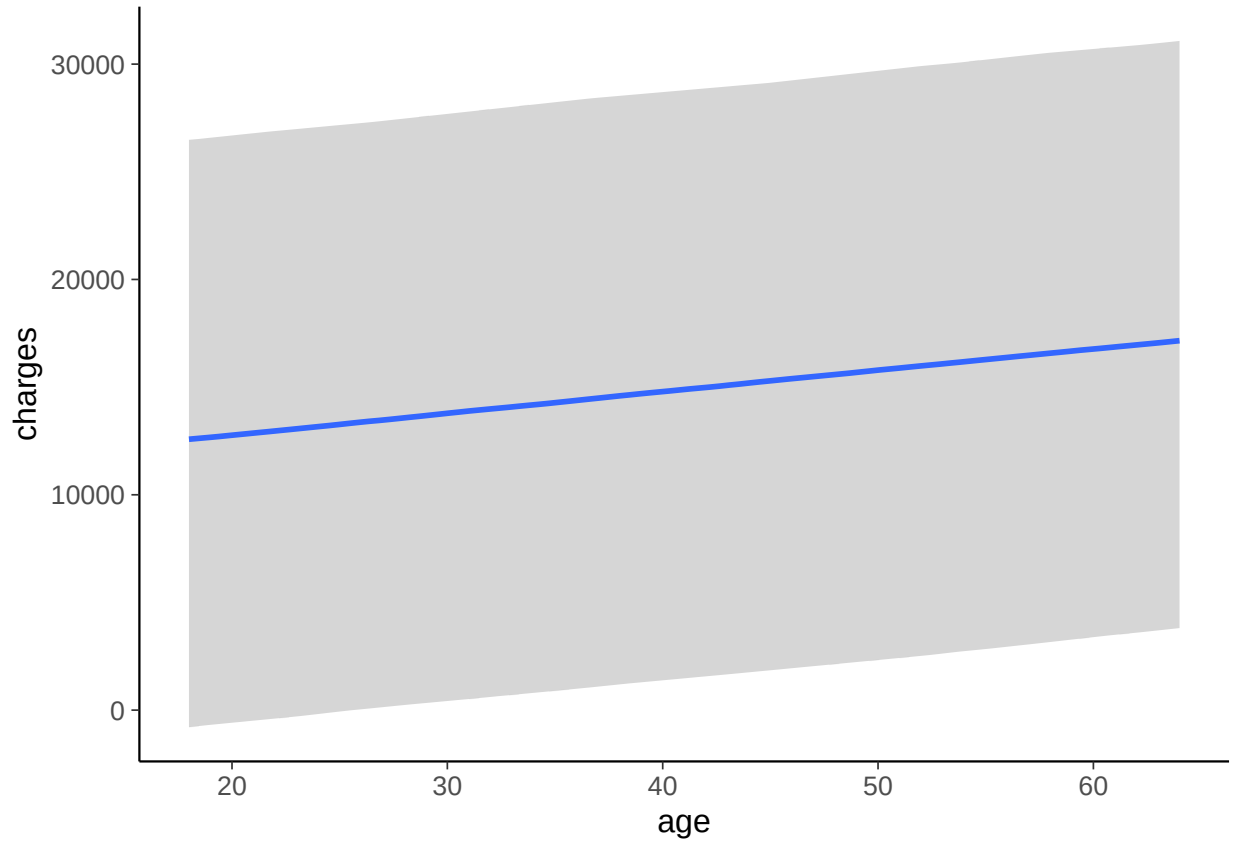
```
make_stancode(charges ~ age + sex + bmi + children + (1|region) + (1|smoker), data=data, family = "gauss
```

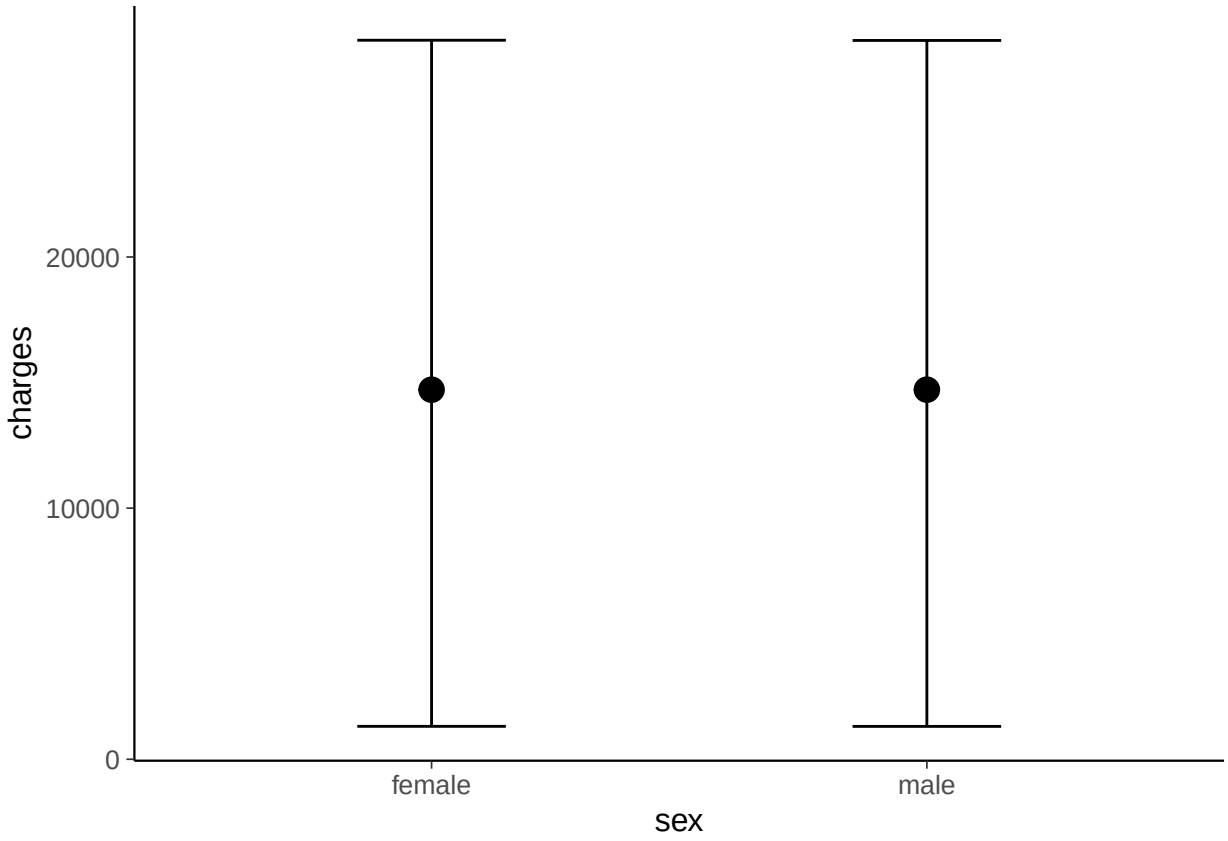
```
## // generated with brms 2.17.0
## functions {
## }
## data {
##   int<lower=1> N;  // total number of observations
##   vector[N] Y;  // response variable
##   int<lower=1> K;  // number of population-level effects
##   matrix[N, K] X;  // population-level design matrix
##   // data for group-level effects of ID 1
##   int<lower=1> N_1;  // number of grouping levels
##   int<lower=1> M_1;  // number of coefficients per level
##   int<lower=1> J_1[N];  // grouping indicator per observation
##   // group-level predictor values
##   vector[N] Z_1_1;
##   // data for group-level effects of ID 2
##   int<lower=1> N_2;  // number of grouping levels
##   int<lower=1> M_2;  // number of coefficients per level
##   int<lower=1> J_2[N];  // grouping indicator per observation
##   // group-level predictor values
##   vector[N] Z_2_1;
##   int prior_only;  // should the likelihood be ignored?
## }
## transformed data {
##   int Kc = K - 1;
##   matrix[N, Kc] Xc;  // centered version of X without an intercept
##   vector[Kc] means_X;  // column means of X before centering
##   for (i in 2:K) {
##     means_X[i - 1] = mean(X[, i]);
##     Xc[, i - 1] = X[, i] - means_X[i - 1];
##   }
## }
## parameters {
##   vector[Kc] b;  // population-level effects
##   real Intercept;  // temporary intercept for centered predictors
##   real<lower=0> sigma;  // dispersion parameter
##   vector<lower=0>[M_1] sd_1;  // group-level standard deviations
##   vector[N_1] z_1[M_1];  // standardized group-level effects
##   vector<lower=0>[M_2] sd_2;  // group-level standard deviations
##   vector[N_2] z_2[M_2];  // standardized group-level effects
## }
## transformed parameters {
##   vector[N_1] r_1_1;  // actual group-level effects
##   vector[N_2] r_2_1;  // actual group-level effects
##   real lprior = 0;  // prior contributions to the log posterior
##   r_1_1 = (sd_1[1] * (z_1[1]));
##   r_2_1 = (sd_2[1] * (z_2[1]));
##   lprior += student_t_lpdf(Intercept | 3, 9382, 7440.8);
##   lprior += student_t_lpdf(sigma | 3, 0, 7440.8)
##     - 1 * student_t_lccdf(0 | 3, 0, 7440.8);
##   lprior += student_t_lpdf(sd_1 | 3, 0, 7440.8)
##     - 1 * student_t_lccdf(0 | 3, 0, 7440.8);
##   lprior += student_t_lpdf(sd_2 | 3, 0, 7440.8)
```

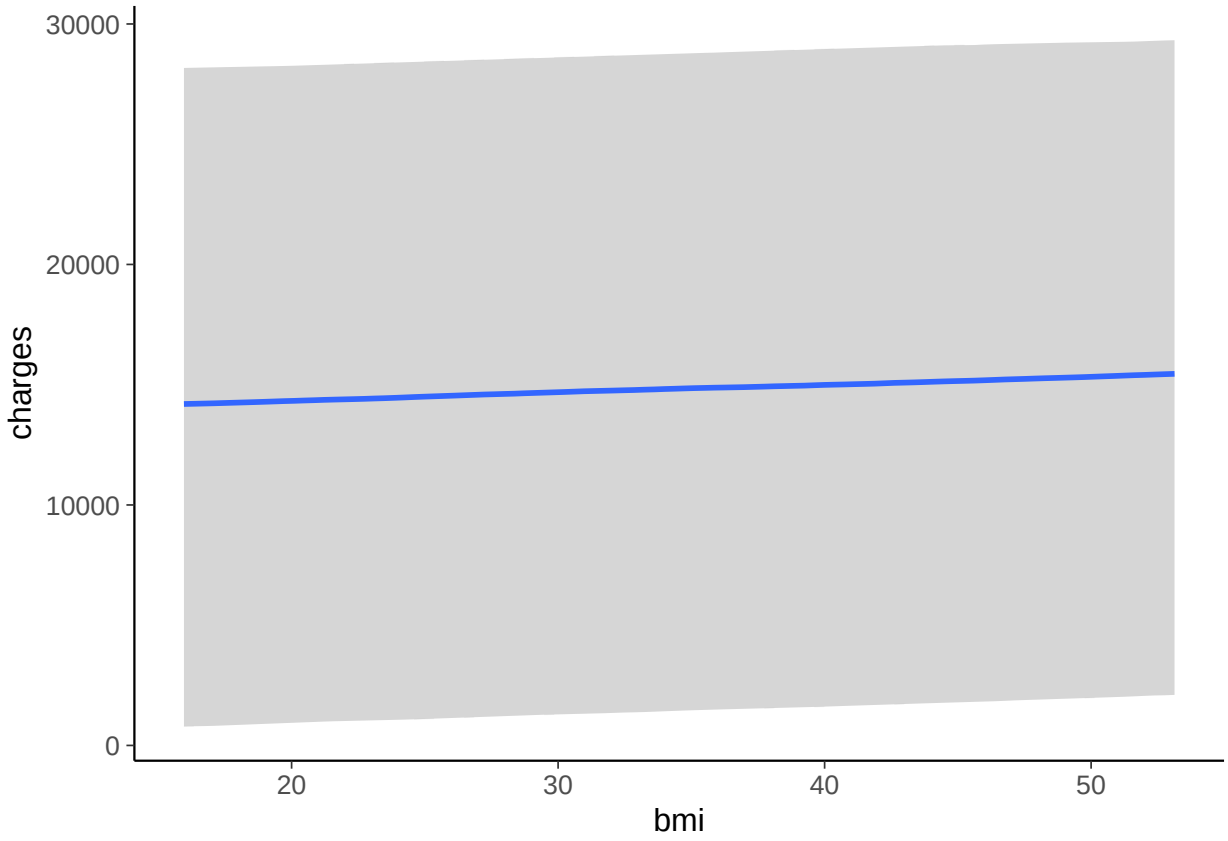
```
##     - 1 * student_t_lccdf(0 | 3, 0, 7440.8);
## }
## model {
##   // likelihood including constants
##   if (!prior_only) {
##     // initialize linear predictor term
##     vector[N] mu = Intercept + rep_vector(0.0, N);
##     for (n in 1:N) {
##       // add more terms to the linear predictor
##       mu[n] += r_1_1[J_1[n]] * Z_1_1[n] + r_2_1[J_2[n]] * Z_2_1[n];
##     }
##     target += normal_id_glm_lpdf(Y | Xc, mu, b, sigma);
##   }
##   // priors including constants
##   target += lprior;
##   target += std_normal_lpdf(z_1[1]);
##   target += std_normal_lpdf(z_2[1]);
## }
## generated quantities {
##   // actual population-level intercept
##   real b_Intercept = Intercept - dot_product(means_X, b);
## }
```

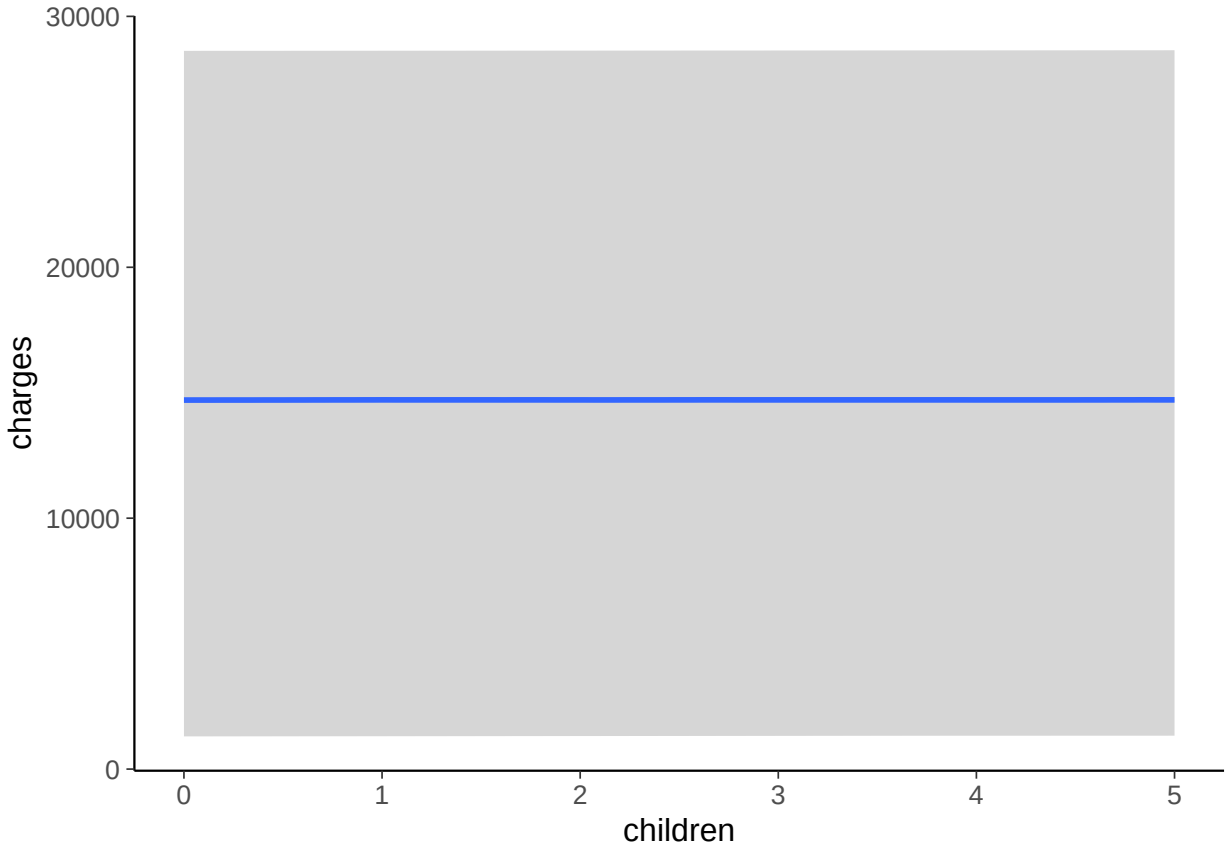
Conditional effects

```
conditional_effects(bayesian_mixed)
```




charges
20000
10000
0
female
male
sex


30000
20000
10000
0
charges
20
30
40
50
bmi


30000
20000
10000
0
charges
0
1
2
3
4
5
children