Adam Work

Adam Kiehl

2022-10-05

Import Data

```
data_raw <- read.csv('ToyotaCorollaData.csv')

data <- data_raw %>%
  mutate_at(c('Metallic', 'Automatic', 'Doors', 'Cylinders', 'Gears',
              'Guarantee', 'BOVAG', 'Fuel_Type', 'Mfg_Month', 'Mfg_Year'),
            as.factor)

head(data)
```

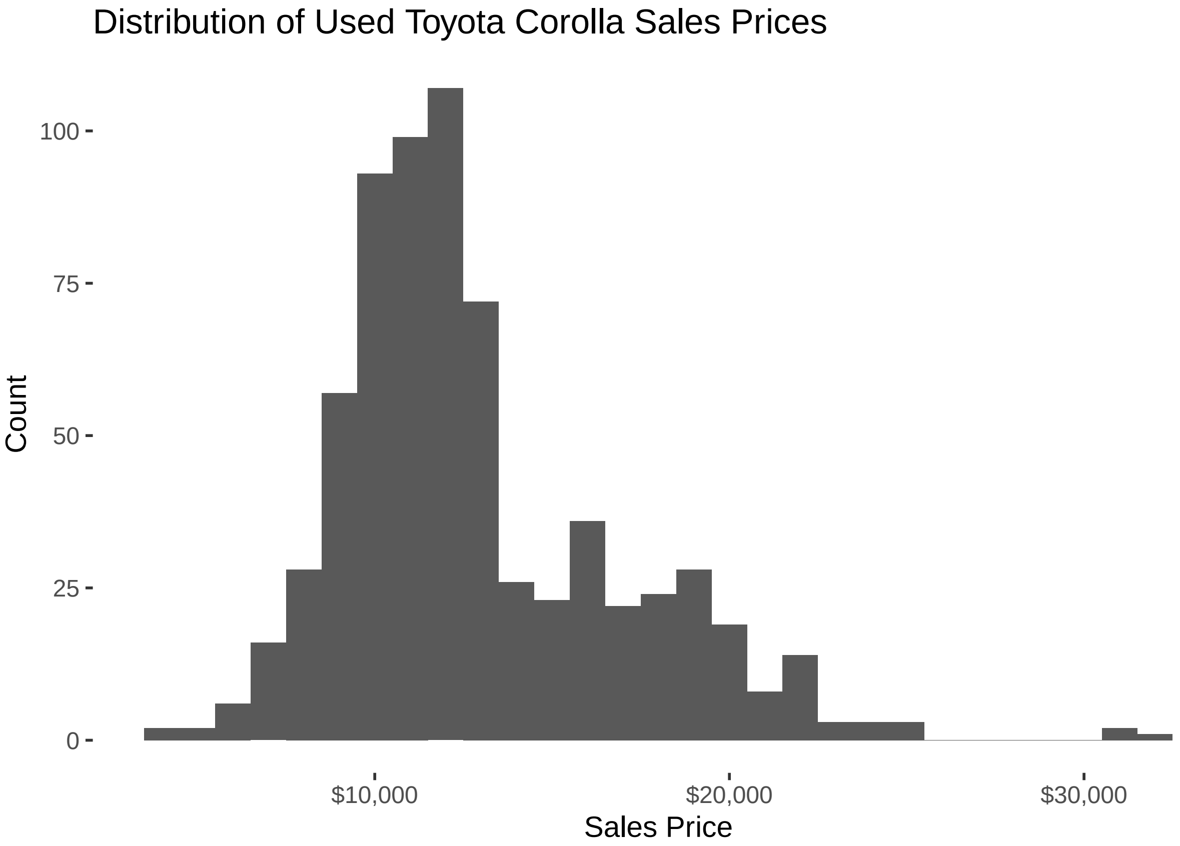
```
##   Price Age Mfg_Month Mfg_Year    KM Fuel_Type HP Metallic Automatic   CC Doors
## 1 13500  23        10     2002 46986    Diesel 90        1         0 2000     3
## 2 13750  23        10     2002 72937    Diesel 90        1         0 2000     3
## 3 13950  24         9     2002 41711    Diesel 90        1         0 2000     3
## 4 14950  26         7     2002 48000    Diesel 90        0         0 2000     3
## 5 13750  30         3     2002 38500    Diesel 90        0         0 2000     3
## 6 12950  32         1     2002 61000    Diesel 90        0         0 2000     3
##   Cylinders Gears QuartTax Weight Guarantee BOVAG Period
## 1         4     5      210   1165         0     1      3
## 2         4     5      210   1165         0     1      3
## 3         4     5      210   1165         1     1      3
## 4         4     5      210   1165         1     1      3
## 5         4     5      210   1170         1     1      3
## 6         4     5      210   1170         0     1      3
```

Scale data and log transform `Price`.

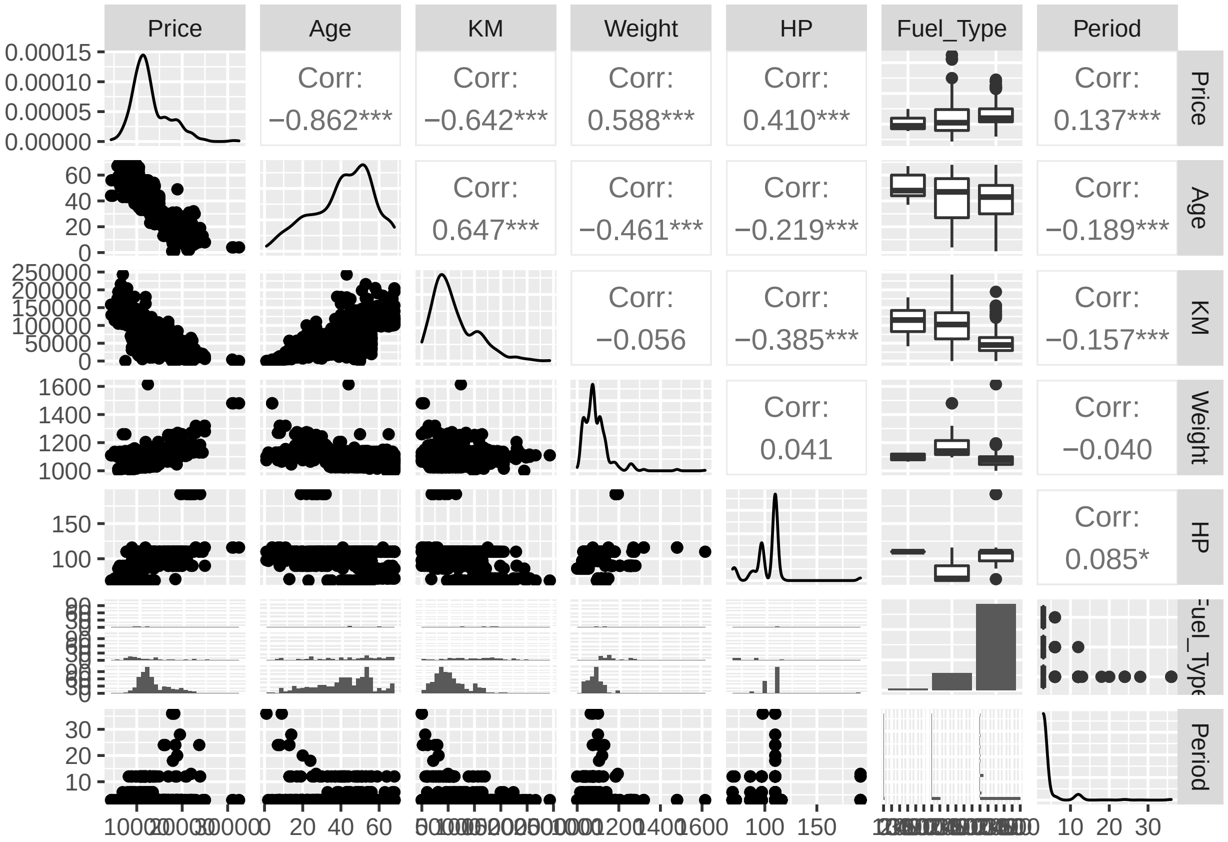
```
data_scale <- cbind(log(data$Price),
                    (data %>%
                       select(-Price) %>%
                       mutate_if(is.numeric, scale))) %>%
  as.data.frame()
names(data_scale) <- names(data)
```

Exploratory Data Analysis

```
ggplot(data,
       mapping = aes(x = Price)) +
  geom_histogram(binwidth = 1000) +
  theme_tufte(base_family = 'sans') +
  scale_x_continuous(labels = label_dollar(prefix = '$')) +
  scale_y_continuous(breaks = seq(0, 100, by = 25)) +
  labs(title = 'Distribution of Used Toyota Corolla Sales Prices',
       x = 'Sales Price',
       y = 'Count')
```



```
data %>%
  select(c(Price, Age, KM, Weight, HP, Fuel_Type, Period)) %>%
  ggpairs(progress = FALSE)
```



Model Fitting

Fit full model, un-transformed, and with default priors.

```
fit1 <- stan_glm(Price ~ .,
                 data = (data %>%
                         select(-Cylinders)),
                 refresh = 0,
                 iter = 5000)

print(fit1,
      digits = 3,
      detail = FALSE)
```

```
##                   Median    MAD_SD
## (Intercept)      -3086.838 22156.510
## Age                -99.754   322.674
## Mfg_Month2         186.789   379.217
## Mfg_Month3         318.657   678.569
## Mfg_Month4          83.981   994.182
## Mfg_Month5         447.616  1304.305
## Mfg_Month6         132.238  1625.509
## Mfg_Month7        -181.544  1944.881
## Mfg_Month8         -15.157  2265.047
## Mfg_Month9        -516.416  2595.839
## Mfg_Month10      -477.748  2921.966
## Mfg_Month11      -542.574  3231.835
## Mfg_Month12      -324.444  3559.140
## Mfg_Year2000     -356.617  3902.437
## Mfg_Year2001     -693.785  7786.776
## Mfg_Year2002     1039.554 11659.111
## Mfg_Year2003     1966.852 15474.961
## Mfg_Year2004     2928.862 19387.906
## KM                 -0.020     0.002
## Fuel_TypeDiesel  1076.683   427.633
## Fuel_TypePetrol  1697.351   427.240
## HP                 50.021     4.763
## Metallic1         181.587   108.975
## Automatic1        426.860   252.793
## CC                  0.075     0.092
## Doors4            -22.223   199.421
## Doors5             91.828   114.820
## Gears6             94.918   334.543
## QuartTax           13.526     2.262
## Weight             11.247     1.689
## Guarantee1        367.683   105.854
## BOVAG1            412.634   188.642
## Period             25.811    14.267
##
## Auxiliary parameter(s):
##       Median   MAD_SD
## sigma 1257.552   35.238
```

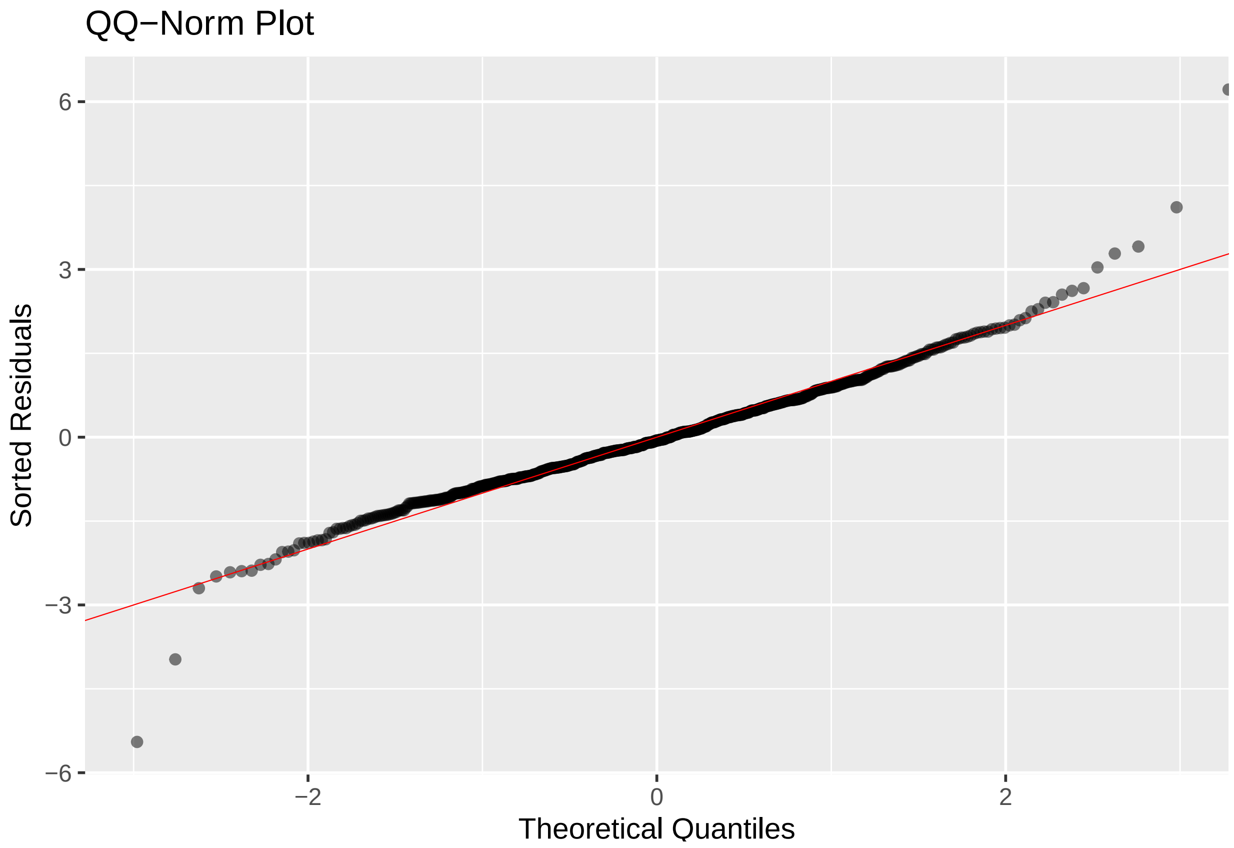
```
print(paste('R^2: ', round(mean(bayes_R2(fit1)), 3)))
```

```
## [1] "R^2:  0.902"
```

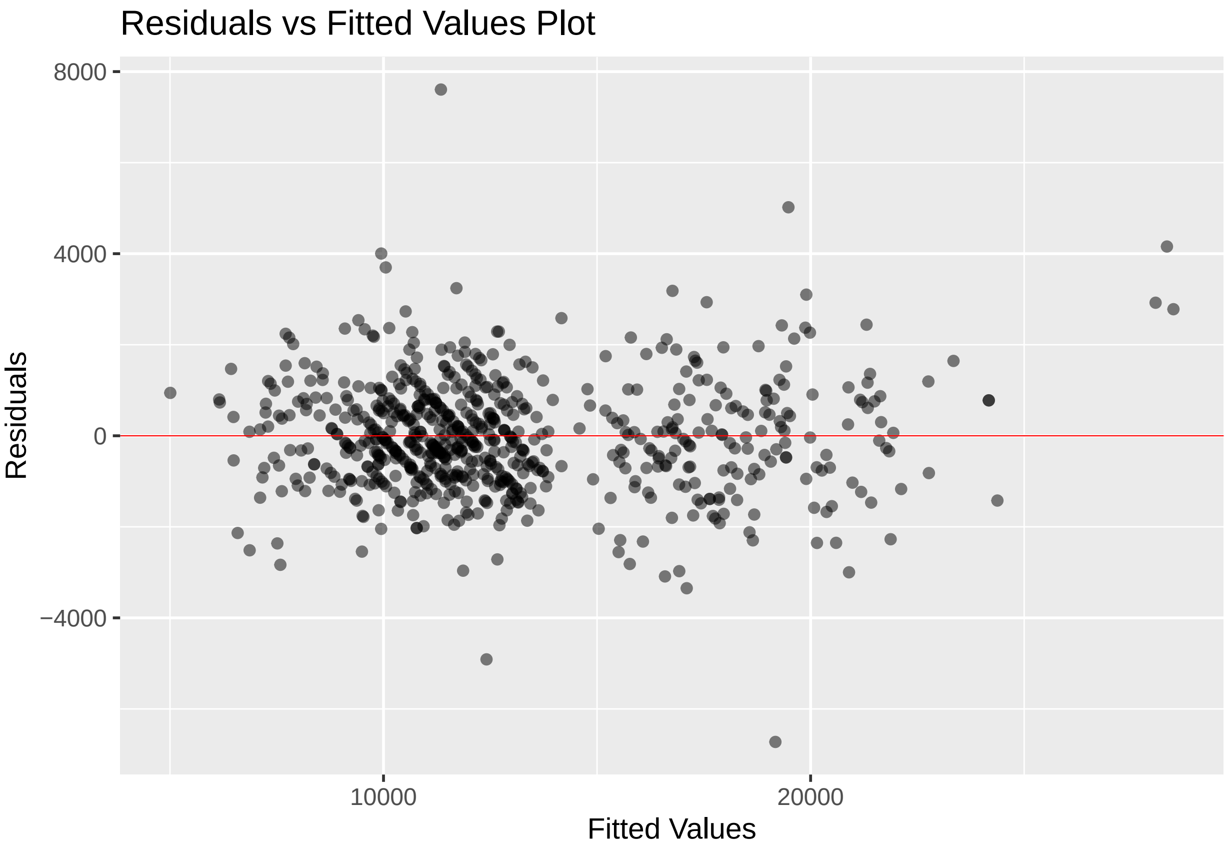
```
print(paste('LOO R^2: ', round(mean(loo_R2(fit1)), 3)))
```

```
## [1] "LOO R^2:  0.89"
```

```
n <- length(fit1$residuals)
quants <- qnorm((1:n / n))

ggplot(mapping = aes(x = quants,
                     y = sort(scale(fit1$residuals)))) +
  geom_point(alpha = .5) +
  geom_abline(intercept = 0,
              slope = 1,
              col = 'red',
              size = .2) +
  labs(title = 'QQ-Norm Plot',
       x = 'Theoretical Quantiles',
       y = 'Sorted Residuals')
```



```
ggplot(mapping = aes(x = fit1$fitted.values,
                     y = fit1$residuals)) +
  geom_point(alpha = .5) +
  geom_hline(yintercept = 0,
             col = 'red',
             size = .2) +
  labs(title = 'Residuals vs Fitted Values Plot',
       x = 'Fitted Values',
       y = 'Residuals')
```



Fit selective model, un-transformed, and with default priors.

```
fit2 <- stan_glm(Price ~ Age + KM + Weight + HP + Fuel_Type + Period,
                 data = data,
                 refresh = 0,
                 iter = 5000)

print(fit2,
      digits = 3,
      detail = FALSE)
```

```
##                   Median    MAD_SD
## (Intercept)     -9796.099  1762.080
## Age              -139.281     5.732
## KM                 -0.017     0.002
## Weight             22.121     1.588
## HP                 44.294     4.669
## Fuel_TypeDiesel   456.197   484.150
## Fuel_TypePetrol   913.159   423.332
## Period             -0.161    15.110
##
## Auxiliary parameter(s):
##       Median   MAD_SD
## sigma 1477.527   39.505
```

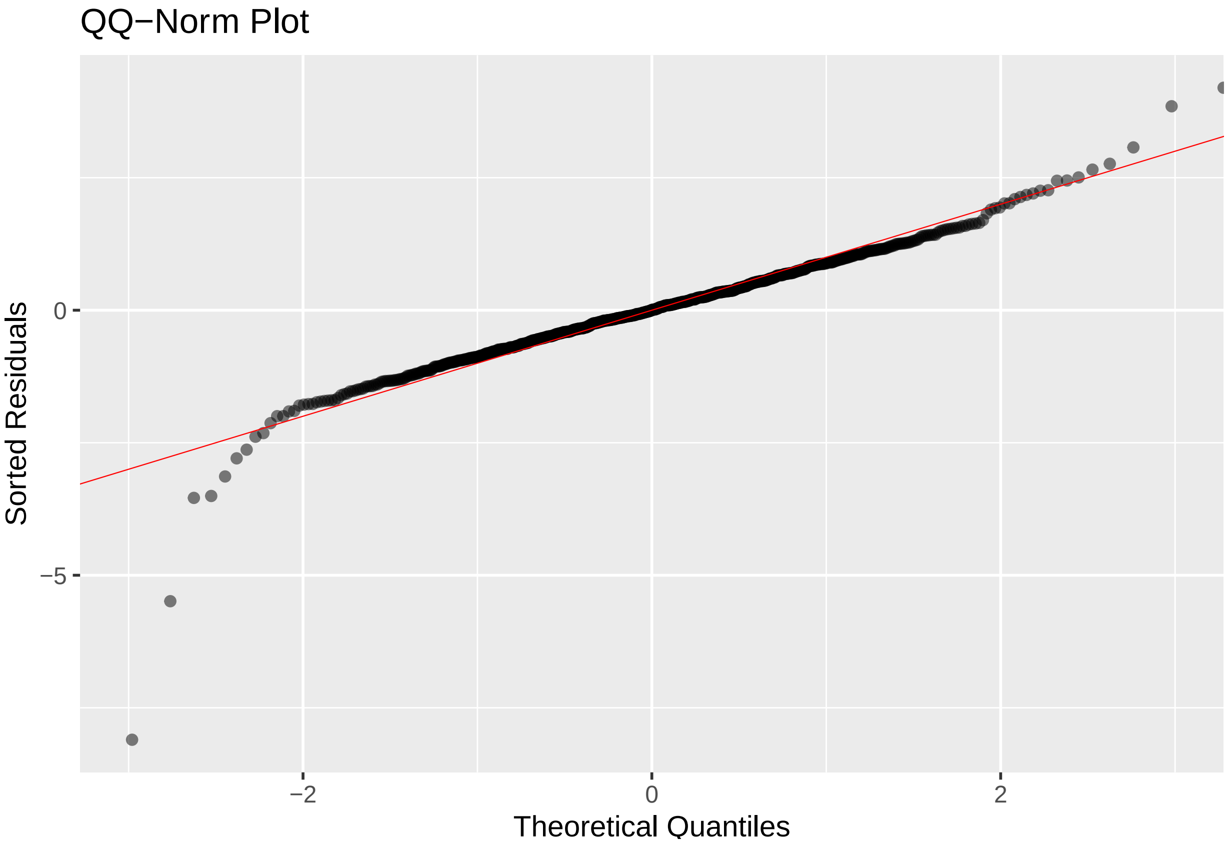
```
print(paste('R^2: ', round(mean(bayes_R2(fit2)), 3)))
```

```
## [1] "R^2:  0.865"
```

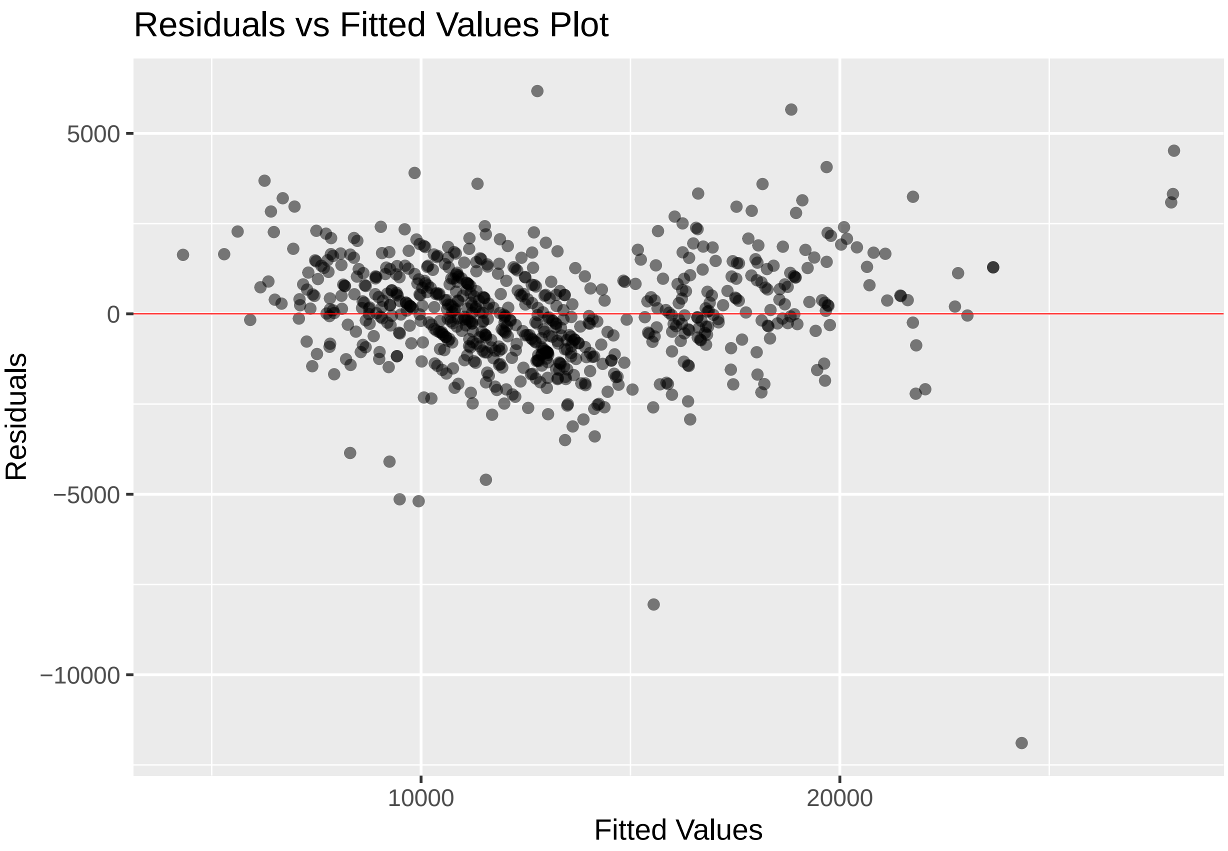
```
print(paste('LOO R^2: ', round(mean(loo_R2(fit2)), 3)))
```

```
## [1] "LOO R^2:  0.855"
```

```
n <- length(fit2$residuals)
quants <- qnorm((1:n / n))

ggplot(mapping = aes(x = quants,
                     y = sort(scale(fit2$residuals)))) +
  geom_point(alpha = .5) +
  geom_abline(intercept = 0,
              slope = 1,
              col = 'red',
              size = .2) +
  labs(title = 'QQ-Norm Plot',
       x = 'Theoretical Quantiles',
       y = 'Sorted Residuals')
```



```
ggplot(mapping = aes(x = fit2$fitted.values,
                     y = fit2$residuals)) +
  geom_point(alpha = .5) +
  geom_hline(yintercept = 0,
             col = 'red',
             size = .2) +
  labs(title = 'Residuals vs Fitted Values Plot',
       x = 'Fitted Values',
       y = 'Residuals')
```



Fit full transformed model with default priors.

```
fit3 <- stan_glm(Price ~ .,
                 data = (data_scale %>%
                           select(-Cylinders)),
                 refresh = 0,
                 iter = 5000)

print(fit3,
      digits = 3,
      detail = FALSE)
```

```
##                   Median MAD_SD
## (Intercept)        9.130  0.658
## Age               -0.097  0.389
## Mfg_Month2         0.024  0.030
## Mfg_Month3         0.037  0.052
## Mfg_Month4         0.016  0.076
## Mfg_Month5       0.040  0.100
## Mfg_Month6       0.024  0.124
## Mfg_Month7       0.002  0.148
## Mfg_Month8       0.021  0.172
## Mfg_Month9      -0.029  0.198
## Mfg_Month10     -0.015  0.220
## Mfg_Month11     -0.016  0.248
## Mfg_Month12      0.010  0.271
## Mfg_Year2000     0.016  0.293
## Mfg_Year2001     0.020  0.590
## Mfg_Year2002     0.158  0.885
## Mfg_Year2003     0.205  1.181
## Mfg_Year2004     0.191  1.482
## KM              -0.078  0.007
## Fuel_TypeDiesel  0.037  0.034
## Fuel_TypePetrol  0.161  0.035
## HP               0.051  0.006
## Metallic1        0.021  0.009
## Automatic1       0.035  0.021
## CC               0.004  0.004
## Doors4          -0.007  0.016
## Doors5           0.016  0.009
## Gears6           0.014  0.028
## QuartTax         0.075  0.009
## Weight           0.032  0.008
## Guarantee1       0.030  0.008
## BOVAG1           0.054  0.015
## Period           0.016  0.005
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.102  0.003
```

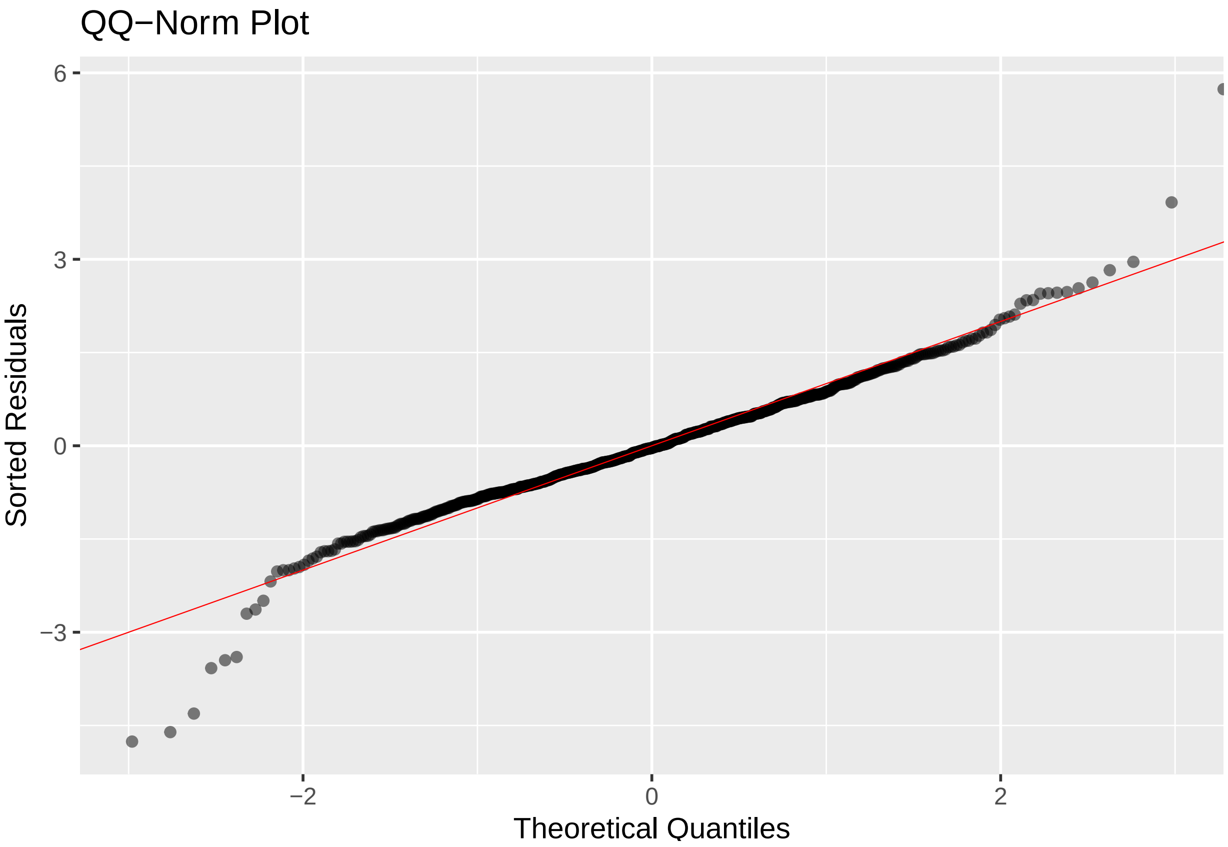
```
print(paste('R^2: ', round(mean(bayes_R2(fit3)), 3)))
```

```
## [1] "R^2:  0.882"
```

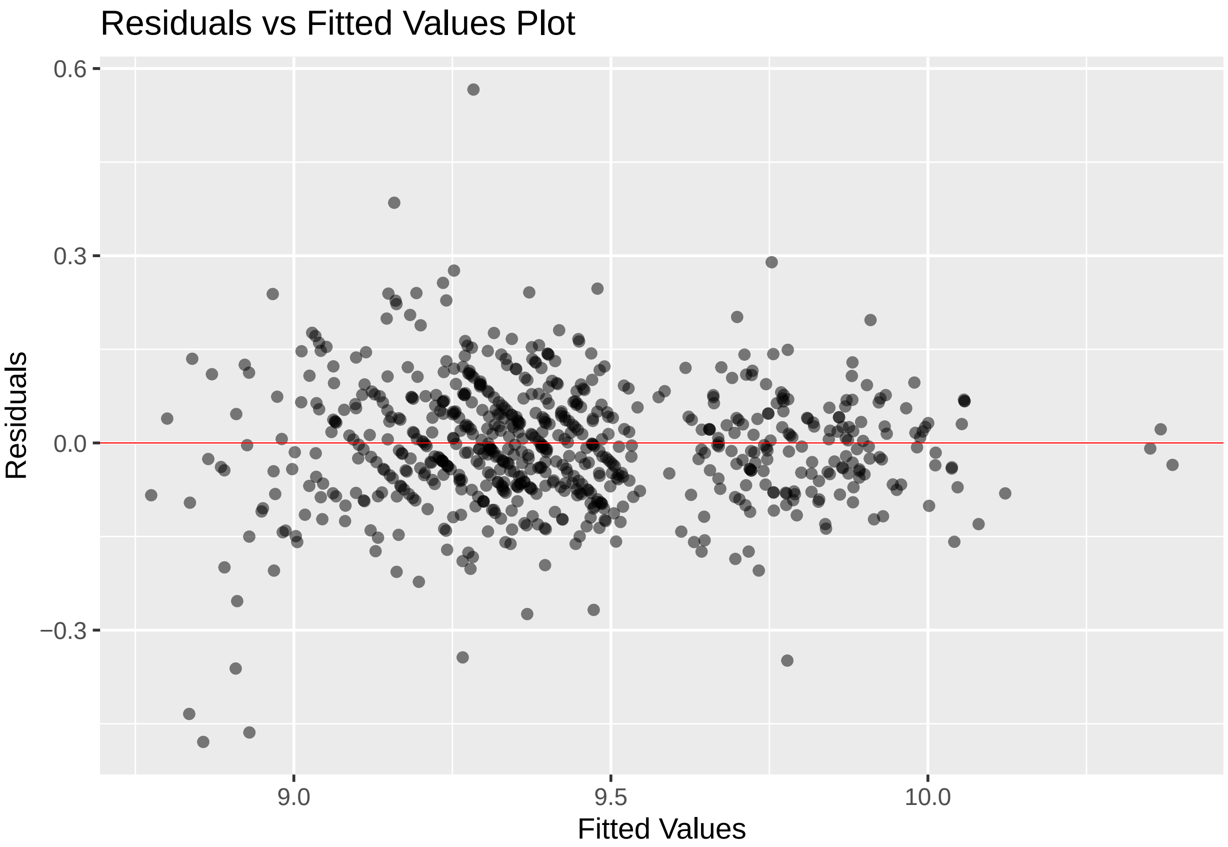
```
print(paste('LOO R^2: ', round(mean(loo_R2(fit3)), 3)))
```

```
## [1] "LOO R^2:  0.87"
```

```
n <- length(fit3$residuals)
quants <- qnorm((1:n / n))

ggplot(mapping = aes(x = quants,
                     y = sort(scale(fit3$residuals)))) +
  geom_point(alpha = .5) +
  geom_abline(intercept = 0,
              slope = 1,
              col = 'red',
              size = .2) +
  labs(title = 'QQ-Norm Plot',
       x = 'Theoretical Quantiles',
       y = 'Sorted Residuals')
```



```
ggplot(mapping = aes(x = fit3$fitted.values,
                     y = fit3$residuals)) +
  geom_point(alpha = .5) +
  geom_hline(yintercept = 0,
             col = 'red',
             size = .2) +
  labs(title = 'Residuals vs Fitted Values Plot',
       x = 'Fitted Values',
       y = 'Residuals')
```



Fit full transformed model with regularized horseshoe prior.

```
p <- ncol(data) - 1
n <- nrow(data)

p0 <- 6

slab_scale <- sqrt(0.3 / p0) * sd(data_scale$Price)
global_scale <- (p0 / (p - p0)) / sqrt(n)

fit4 <- stan_glm(Price ~ .,
                 data = (data_scale %>%
                           select(-Cylinders)),
                 refresh = 0,
                 iter = 5000,
                 prior = hs(global_scale = global_scale,
                            slab_scale = slab_scale))

print(fit4,
      digits = 3,
      detail = FALSE)
```

```
##                   Median MAD_SD
## (Intercept)        9.252  0.036
## Age               -0.147  0.012
## Mfg_Month2         0.015  0.015
## Mfg_Month3         0.027  0.016
## Mfg_Month4       0.003  0.011
## Mfg_Month5       0.024  0.015
## Mfg_Month6       0.005  0.012
## Mfg_Month7      -0.009  0.014
## Mfg_Month8       0.000  0.011
## Mfg_Month9      -0.045  0.021
## Mfg_Month10     -0.038  0.018
## Mfg_Month11     -0.034  0.023
## Mfg_Month12     -0.012  0.019
## Mfg_Year2000    -0.019  0.015
## Mfg_Year2001    -0.053  0.020
## Mfg_Year2002     0.038  0.029
## Mfg_Year2003     0.048  0.034
## Mfg_Year2004     0.001  0.029
## KM              -0.079  0.007
## Fuel_TypeDiesel  0.002  0.019
## Fuel_TypePetrol  0.124  0.030
## HP               0.050  0.006
## Metallic1        0.018  0.009
## Automatic1       0.021  0.022
## CC               0.003  0.004
## Doors4          -0.003  0.011
## Doors5           0.012  0.009
## Gears6           0.004  0.016
## QuartTax         0.070  0.009
## Weight           0.036  0.008
## Guarantee1       0.029  0.009
## BOVAG1           0.051  0.015
## Period           0.014  0.005
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.102  0.003
```

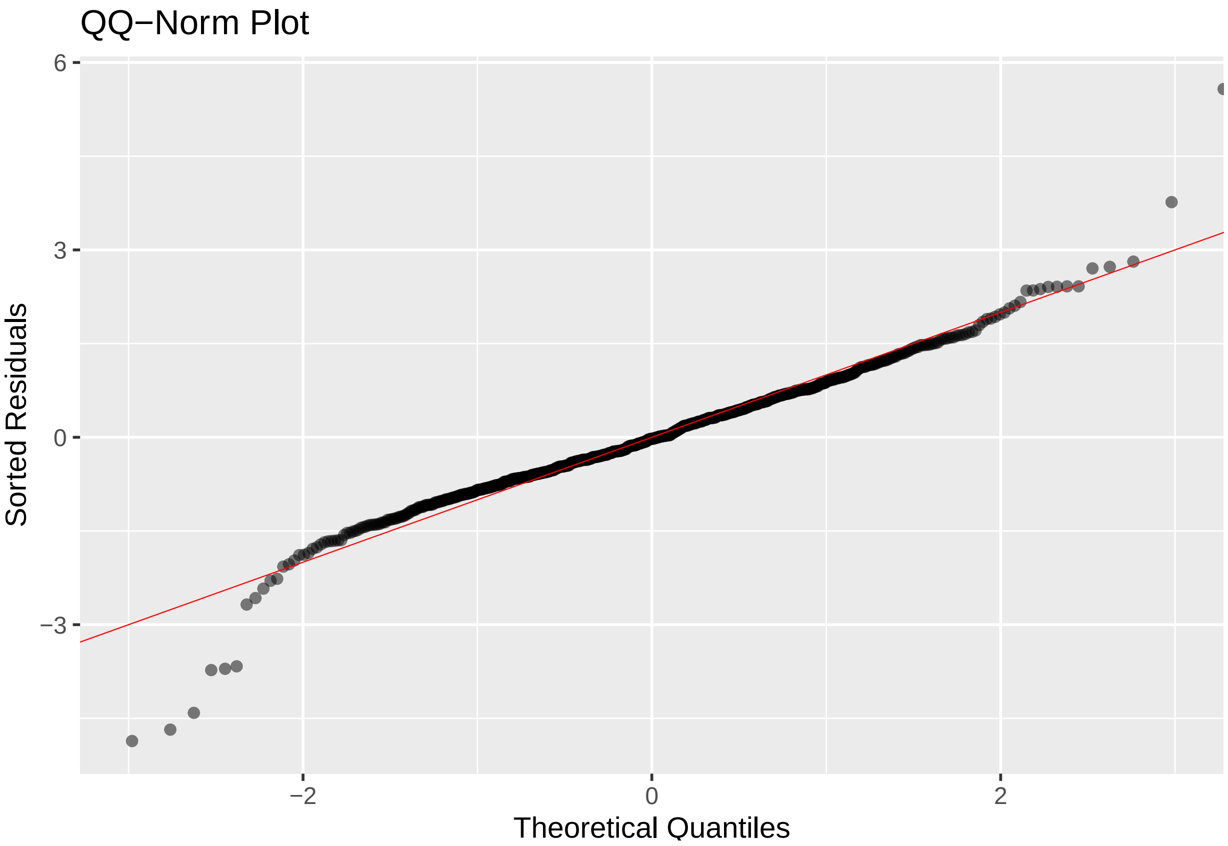
```
print(paste('R^2: ', round(mean(bayes_R2(fit4)), 3)))
```

```
## [1] "R^2:  0.88"
```

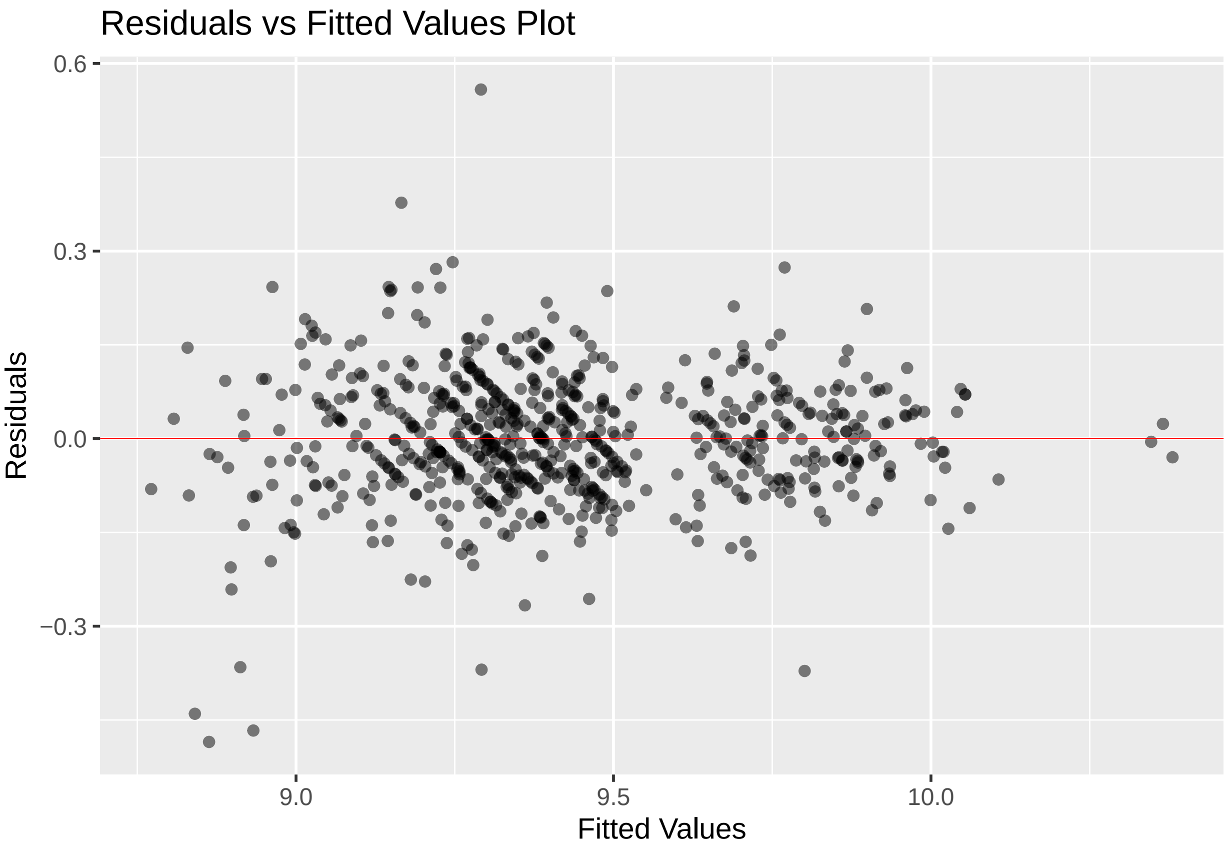
```
print(paste('LOO R^2: ', round(mean(loo_R2(fit4)), 3)))
```

```
## [1] "LOO R^2:  0.87"
```

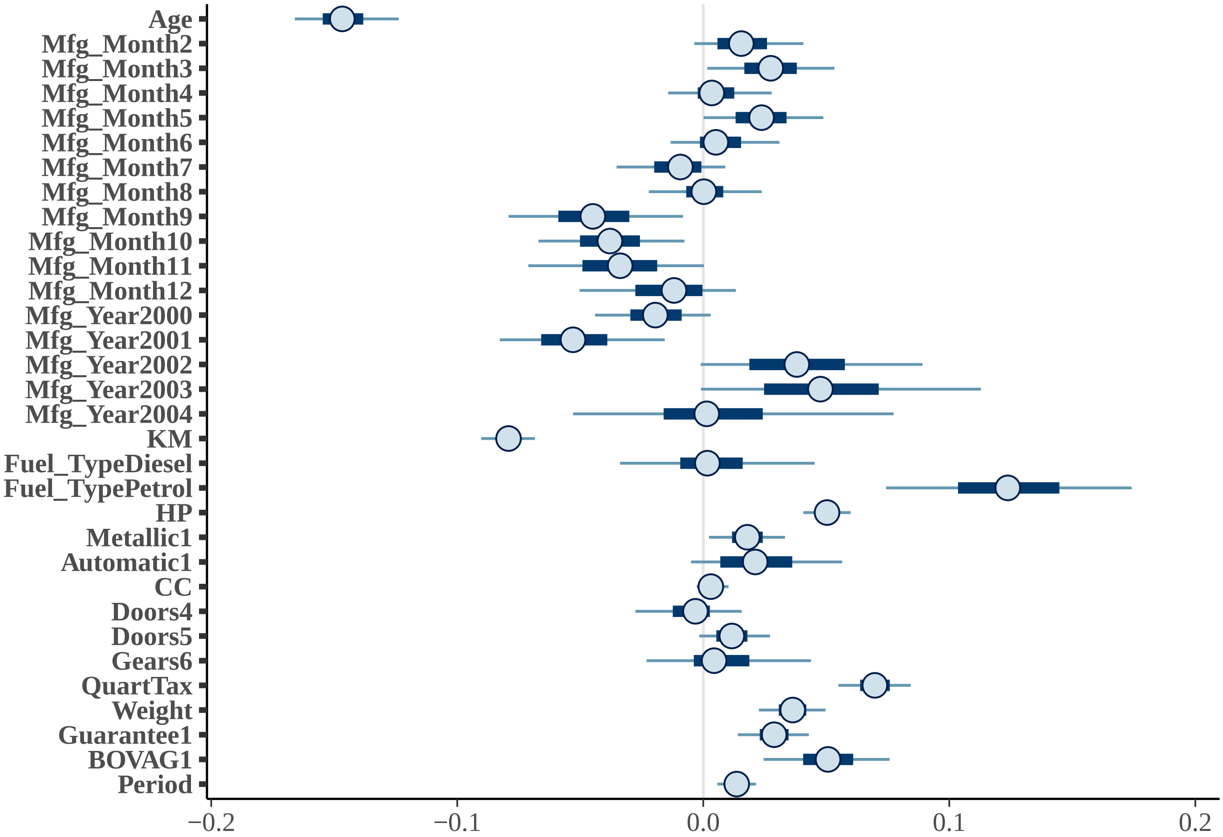
```
n <- length(fit4$residuals)
quants <- qnorm((1:n / n))

ggplot(mapping = aes(x = quants,
                     y = sort(scale(fit4$residuals)))) +
  geom_point(alpha = .5) +
  geom_abline(intercept = 0,
              slope = 1,
              col = 'red',
              size = .2) +
  labs(title = 'QQ-Norm Plot',
       x = 'Theoretical Quantiles',
       y = 'Sorted Residuals')
```



```
ggplot(mapping = aes(x = fit4$fitted.values,
                     y = fit4$residuals)) +
  geom_point(alpha = .5) +
  geom_hline(yintercept = 0,
             col = 'red',
             size = .2) +
  labs(title = 'Residuals vs Fitted Values Plot',
       x = 'Fitted Values',
       y = 'Residuals')
```



```
as.data.frame(fit4) %>%
  select(-c('(Intercept)', 'sigma')) %>%
  mcmc_intervals()
```



Fit horseshoe-selected transformed model with default priors.

```
fit5 <- stan_glm(Price ~ Age + Mfg_Month + Mfg_Year + KM + Fuel_Type + HP +
                   Metallic + QuartTax + Weight + Guarantee + Period,
                 data = data_scale,
                 refresh = 0,
                 iter = 5000)

print(fit5,
      digits = 3,
      detail = FALSE)
```

```
##                   Median MAD_SD
## (Intercept)        9.185  0.654
## Age               -0.112  0.389
## Mfg_Month2         0.024  0.029
## Mfg_Month3         0.039  0.052
## Mfg_Month4         0.017  0.075
## Mfg_Month5         0.038  0.099
## Mfg_Month6         0.023  0.124
## Mfg_Month7         0.003  0.147
## Mfg_Month8         0.021  0.172
## Mfg_Month9        -0.034  0.197
## Mfg_Month10       -0.022  0.221
## Mfg_Month11       -0.023  0.245
## Mfg_Month12        0.002  0.270
## Mfg_Year2000       0.004  0.294
## Mfg_Year2001      0.000  0.588
## Mfg_Year2002      0.121  0.884
## Mfg_Year2003      0.156  1.176
## Mfg_Year2004      0.140  1.471
## KM               -0.076  0.007
## Fuel_TypeDiesel   0.026  0.034
## Fuel_TypePetrol   0.187  0.035
## HP                0.050  0.006
## Metallic1         0.021  0.009
## QuartTax          0.084  0.009
## Weight            0.038  0.008
## Guarantee1        0.035  0.008
## Period            0.011  0.004
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.103  0.003
```

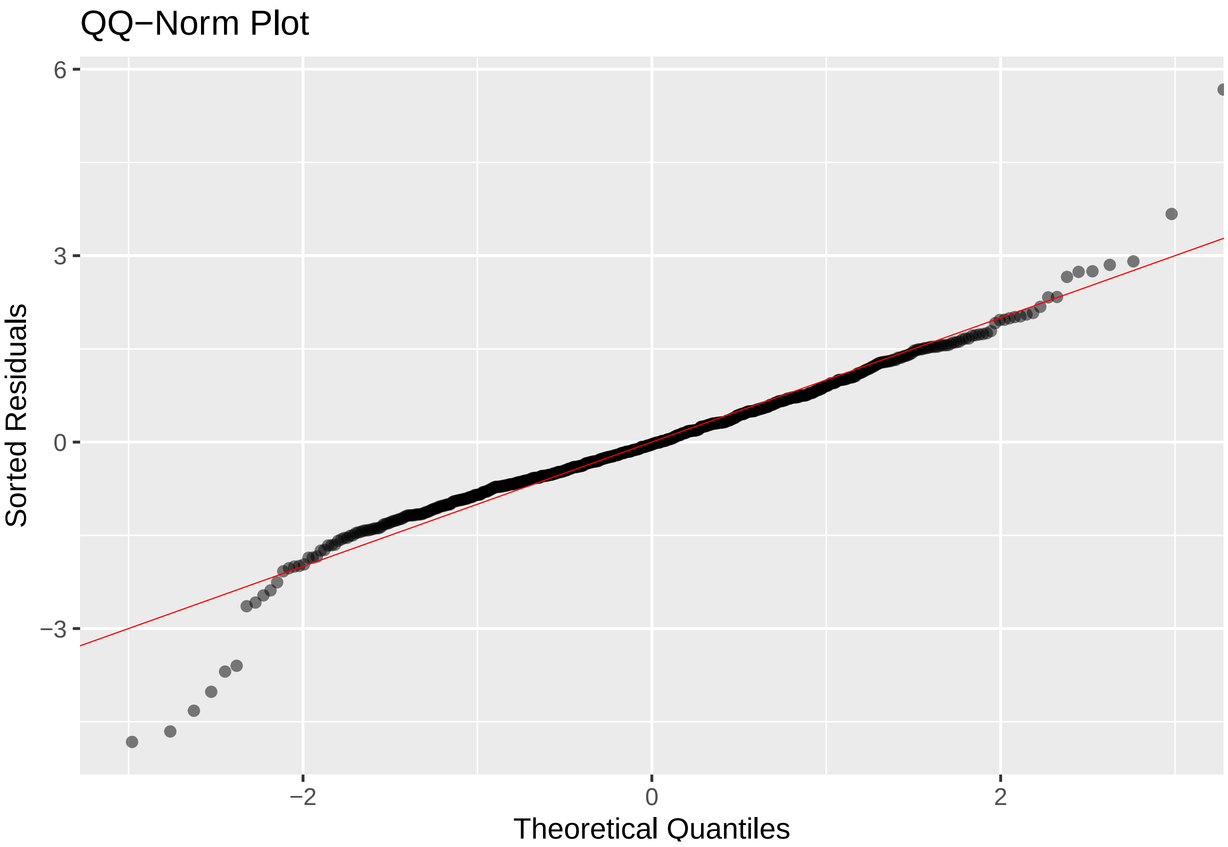
```
print(paste('R^2: ', round(mean(bayes_R2(fit5)), 3)))
```

```
## [1] "R^2:  0.88"
```

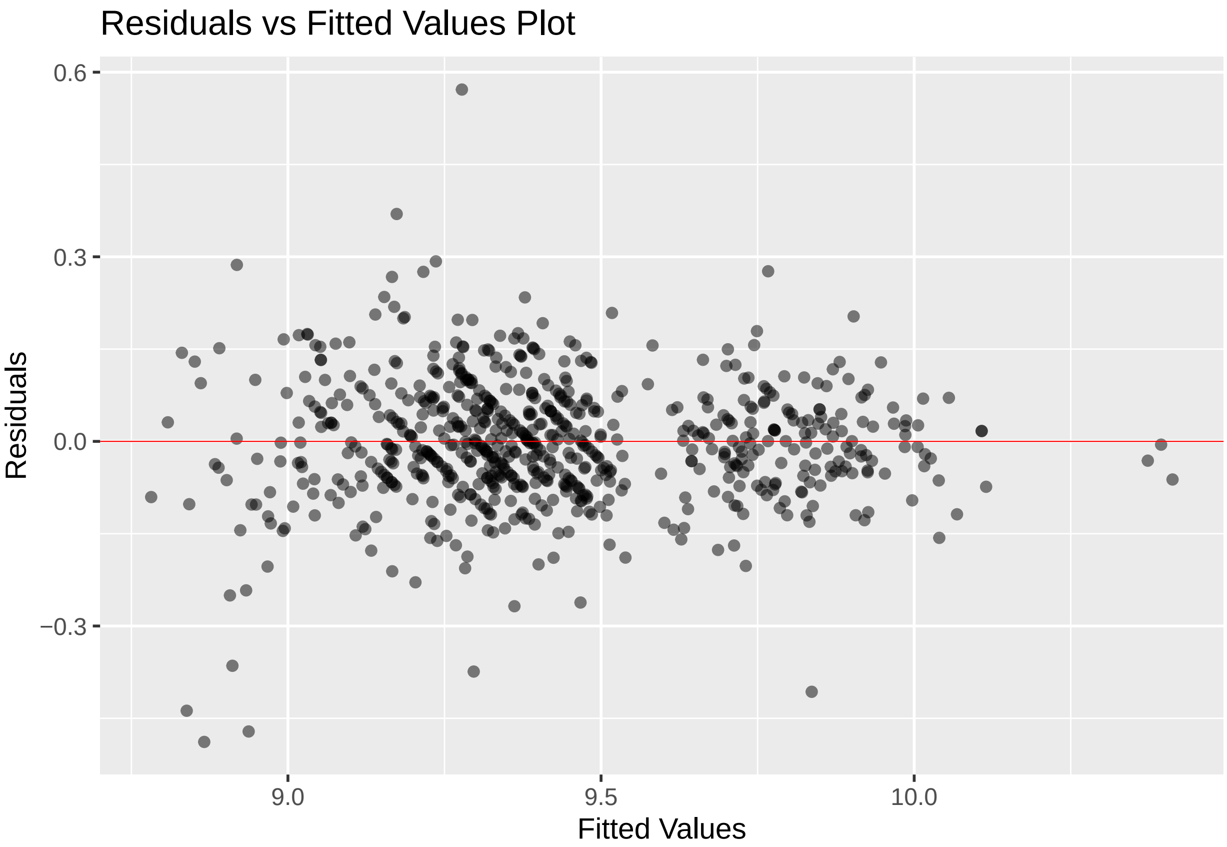
```
print(paste('LOO R^2: ', round(mean(loo_R2(fit5)), 3)))
```

```
## [1] "LOO R^2:  0.869"
```

```
n <- length(fit5$residuals)
quants <- qnorm((1:n / n))

ggplot(mapping = aes(x = quants,
                     y = sort(scale(fit5$residuals)))) +
  geom_point(alpha = .5) +
  geom_abline(intercept = 0,
              slope = 1,
              col = 'red',
              size = .2) +
  labs(title = 'QQ-Norm Plot',
       x = 'Theoretical Quantiles',
       y = 'Sorted Residuals')
```



```
ggplot(mapping = aes(x = fit5$fitted.values,
                     y = fit5$residuals)) +
  geom_point(alpha = .5) +
  geom_hline(yintercept = 0,
             col = 'red',
             size = .2) +
  labs(title = 'Residuals vs Fitted Values Plot',
       x = 'Fitted Values',
       y = 'Residuals')
```



Fit horseshoe-selected (minus months and years) transformed model with default priors.

```
fit6 <- stan_glm(Price ~ Age + KM + Fuel_Type + HP + Metallic + QuartTax +
                   Weight + Guarantee + Period,
                 data = data_scale,
                 refresh = 0,
                 iter = 5000)

print(fit6,
      digits = 3,
      detail = FALSE)
```

```
##                   Median MAD_SD
## (Intercept)        9.218  0.035
## Age               -0.155  0.007
## KM                -0.070  0.007
## Fuel_TypeDiesel    0.017  0.036
## Fuel_TypePetrol    0.217  0.036
## HP                 0.048  0.006
## Metallic1          0.013  0.009
## QuartTax           0.083  0.009
## Weight             0.059  0.008
## Guarantee1         0.031  0.009
## Period             0.013  0.005
##
## Auxiliary parameter(s):
##       Median MAD_SD
## sigma 0.109  0.003
```

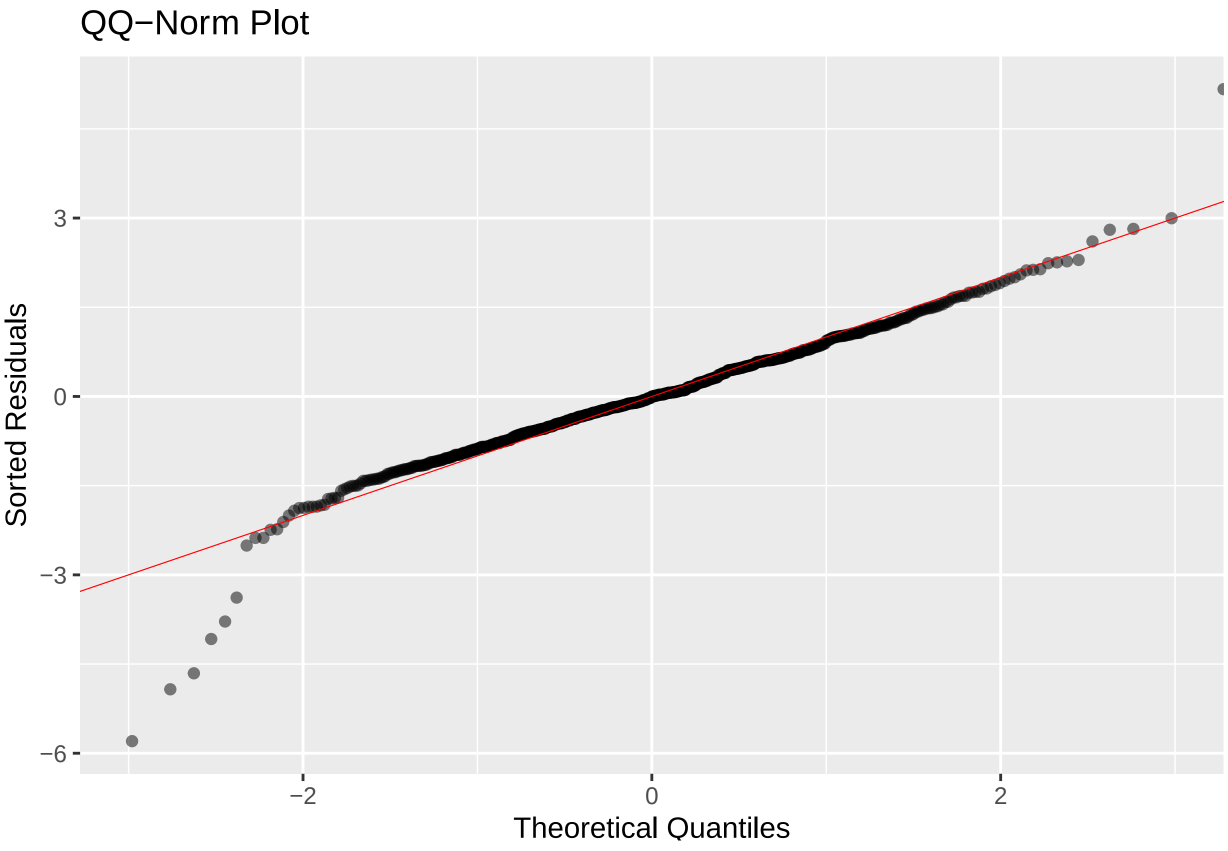
```
print(paste('R^2: ', round(mean(bayes_R2(fit6)), 3)))
```

```
## [1] "R^2:  0.865"
```

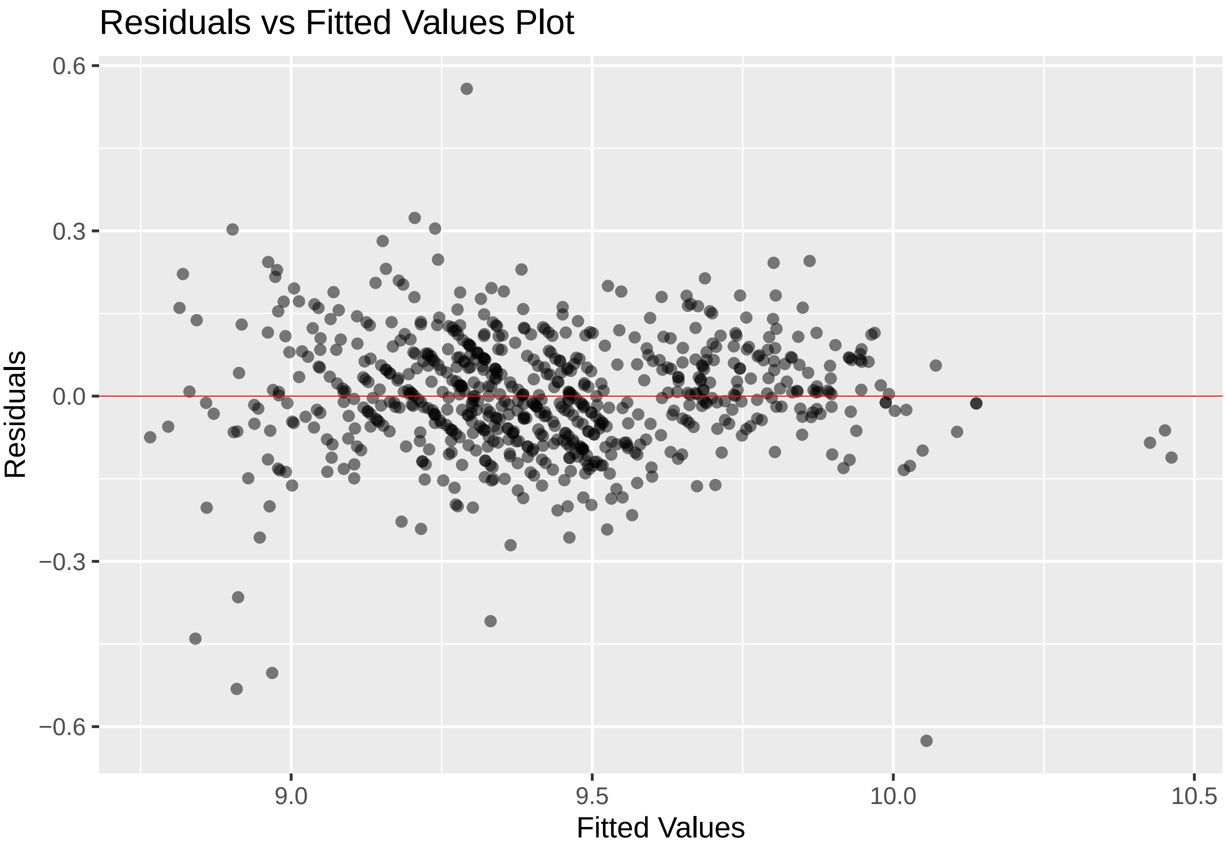
```
print(paste('LOO R^2: ', round(mean(loo_R2(fit6)), 3)))
```

```
## [1] "LOO R^2:  0.855"
```

```
n <- length(fit6$residuals)
quants <- qnorm((1:n / n))

ggplot(mapping = aes(x = quants,
                     y = sort(scale(fit6$residuals)))) +
  geom_point(alpha = .5) +
  geom_abline(intercept = 0,
              slope = 1,
              col = 'red',
              size = .2) +
  labs(title = 'QQ-Norm Plot',
       x = 'Theoretical Quantiles',
       y = 'Sorted Residuals')
```



```r
ggplot(mapping = aes(x = fit6$fitted.values,
                     y = fit6$residuals)) +
  geom_point(alpha = .5) +
  geom_hline(yintercept = 0,
             col = 'red',
             size = .2) +
  labs(title = 'Residuals vs Fitted Values Plot',
       x = 'Fitted Values',
       y = 'Residuals')
```



- Age: N(-100, 50)
- KM N(-.0625, .5)
- Weight N(10, 5)
- HP N(50, 20)
- Fuel Type
- Period

```r
loo1 <- loo(fit1)
loo2 <- loo(fit2)
loo3 <- loo(fit3)
loo4 <- loo(fit4)
loo5 <- loo(fit5)
loo6 <- loo(fit6)

loo_compare(loo1, loo2, loo3, loo4, loo5, loo6)
```

```
##      elpd_diff se_diff
## fit4     0.0       0.0
## fit3    -0.4       3.4
## fit5    -3.9       4.5
## fit6   -37.1      15.2
## fit1 -6544.6      26.2
## fit2 -6645.0      42.7
```