Bayesian Data Analysis project

Housing price prediction

anonymous

1 Introduction

In February 2024, the Pellervo Economic Research (PTT) of Finland forecasts that the housing prices in Espoo will increase by 1.7% because of the influx of people moving to Espoo [1]. Prediction of housing prices helps individuals and businesses make informed decisions about buying, selling, or investing in housing properties. For people planning to buy a house in Espoo, housing price prediction helps with financial planning and estimating the mortgage. For real estate professionals, economists, and policymakers, housing price prediction provides insights into factors that influence housing supply and demand, as well as urban development patterns. For example, housing price predictive models can help identify areas with affordable housing options, address housing inequality, and promote inclusive urban development. In addition, for banks, mortgage lenders, and other financial organizations, predicting housing prices is essential for assessing the risk associated with lending and investment activities.

However, housing price prediction can be challenging. The relationships between housing attributes and prices may not be linear. Factors such as number of rooms and property size can interact in complex ways to influence prices.

In this project, our goal is to predict the Espoo housing prices with linear and non-linear models. Regarding the linear model, we first investigate two variables—the age and the size of the house. For the non-linear model, we also add hierarchy.

The structure of the report is as follows. Section 2 describes the data and the analysis problem. Section 3 describes the models and prior choices. Section 4 presents our analysis with the linear model.

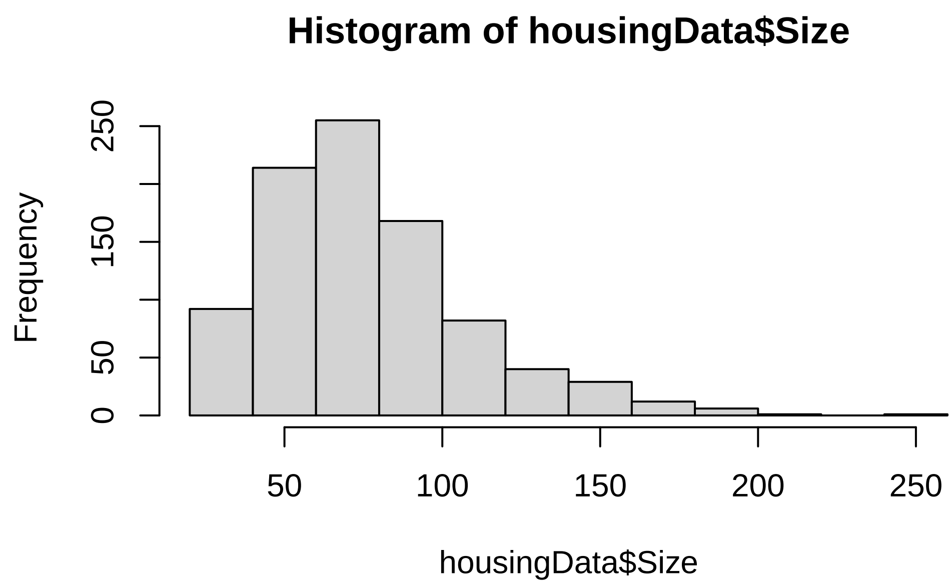
2 Description of the data and the analysis

The housing price dataset is obtained from Asuntojen Hintatiedot [2], which can be translated into Price Information of Housing. This dataset can be viewed and downloaded from here. At the time of writing this report, to our knowledge, there are no existing analyses with this housing dataset.

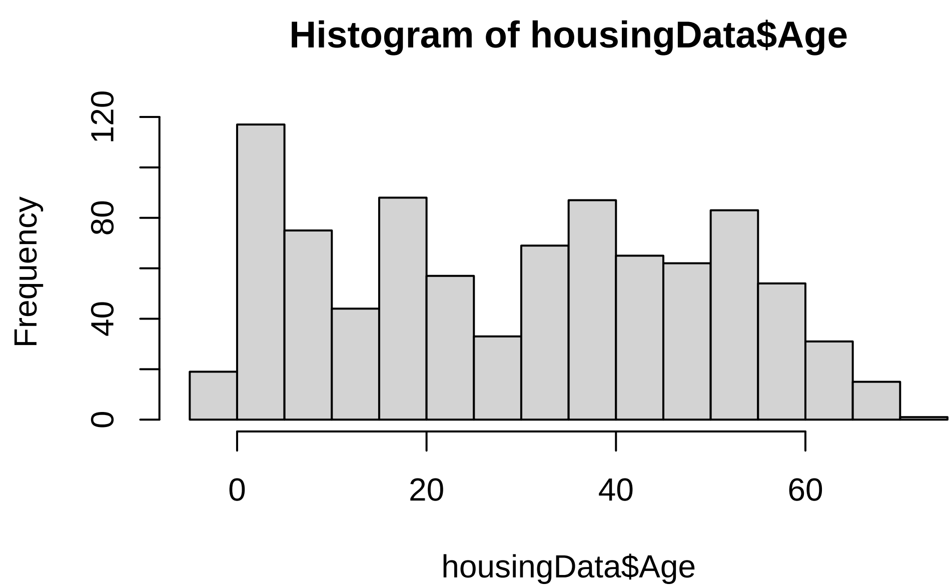
There are 900 observations in the dataset. Each row contains information about a house. In the original dataset, there are 10 variables. However, to investigate the effect , we add a variable called $Age = 2023 - ProductionYear$ to compute the age of the house using the year it was produced.

There are 8 empty cells in column *Condition* and 30 in column *LandOwnership*. However, these two parameters are not used for the Bayesian models in this report.

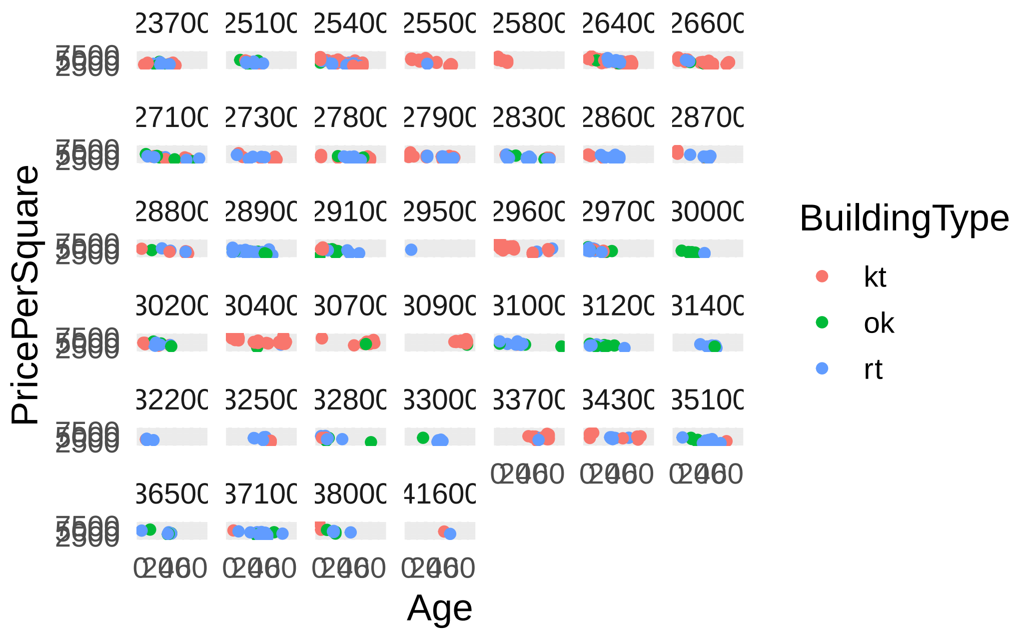
```
hist(housingData$Size)
```




```
hist(housingData$Age)
```




```
# plots for each postal code
ggplot(data = housingData %>% filter(
  Age <= 80)) +
  geom_point(aes(Age, PricePerSquare, color = BuildingType)) + facet_wrap(~IncomeClass)
```



3 Models and prior choices

3.1 Linear model

In the linear model, we choose a normal prior of $(\mu = 5000, 1500)$ for the housing size.

3.2 Non-linear model with hierarchy

4 Analysis with the linear model

This section shows the code of our linear model with Gaussian noise and how the Markov chain Monte Carlo (MCMC) inference was run. We also shows the convergence diagnostic values for the linear model and their interpretation. In addition, we report posterior predictive checks.

4.1 MCMC inference

```
fit1 = brm(Price ~ Size + Age,
           data = housingData,
           family = gaussian(),
           prior = c(
             prior(normal(5000, 1500), class = b, coef = Size),
             prior(normal(-1000, 5000), class = b, coef = Age),
             prior(normal(100000, 30000), class = Intercept)
             ),
           refresh=0,
           show_exceptions = FALSE,
           file="fit1.rds"
           )
```

4.2 Convergence diagnostic

```r
rhat(fit1)
```

```
b_Intercept      b_Size       b_Age       sigma      lprior        lp__ 
  1.0034224   1.0031867   1.0011190   0.9999962   1.0007871   1.0005779 
```

To obtain summaries and convergence diagnostic of the fitted linear model , we call the function `summary()`.

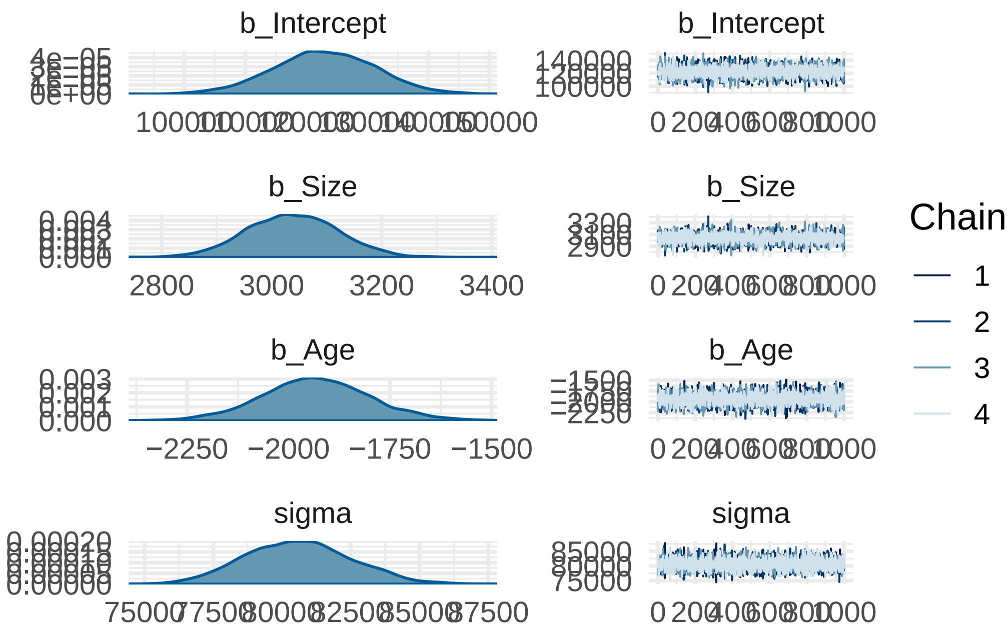
```r
summary(fit1)
```

```
 Family: gaussian 
  Links: mu = identity; sigma = identity 
Formula: Price ~ Size + Age 
   Data: housingData (Number of observations: 900) 
  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 122963.93   8298.85 106276.90 138949.12 1.00     4668     3130
Size        3038.05     83.47   2877.31   3200.37 1.00     4762     2991
Age        -1932.21    133.39  -2196.70  -1665.33 1.00     5194     2985

Family Specific Parameters: 
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma 80600.70   1927.37 77006.40 84418.39 1.00     5302     3513

Draws were sampled using sample(hmc). 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).
```

4.3 Posterior predictive check

By using function `plot()`, we can plot the MCMC chains as well as the posterior distributions for each parameter.

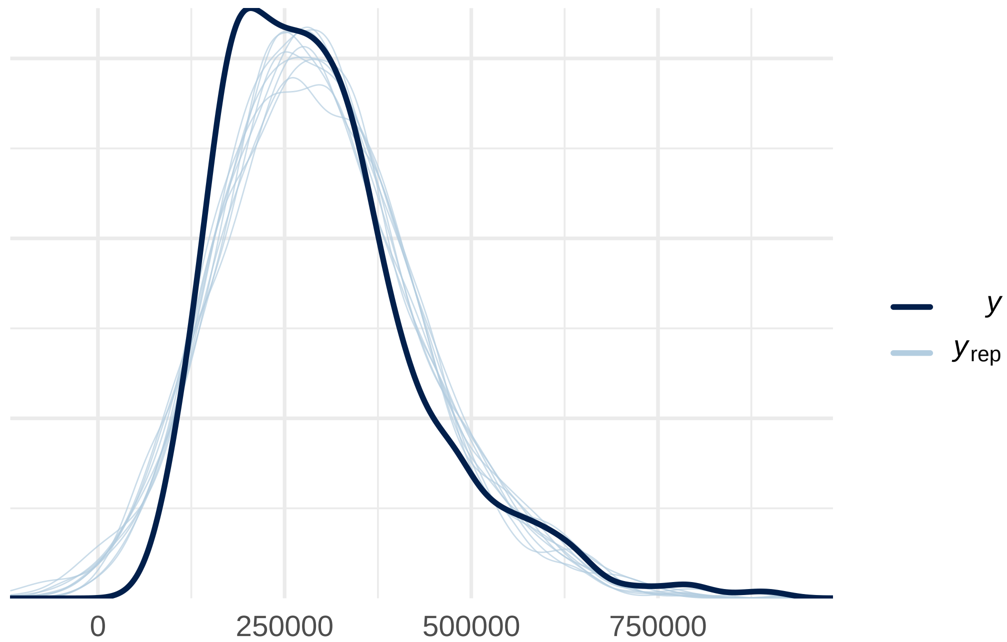
```r
plot(fit1)
```



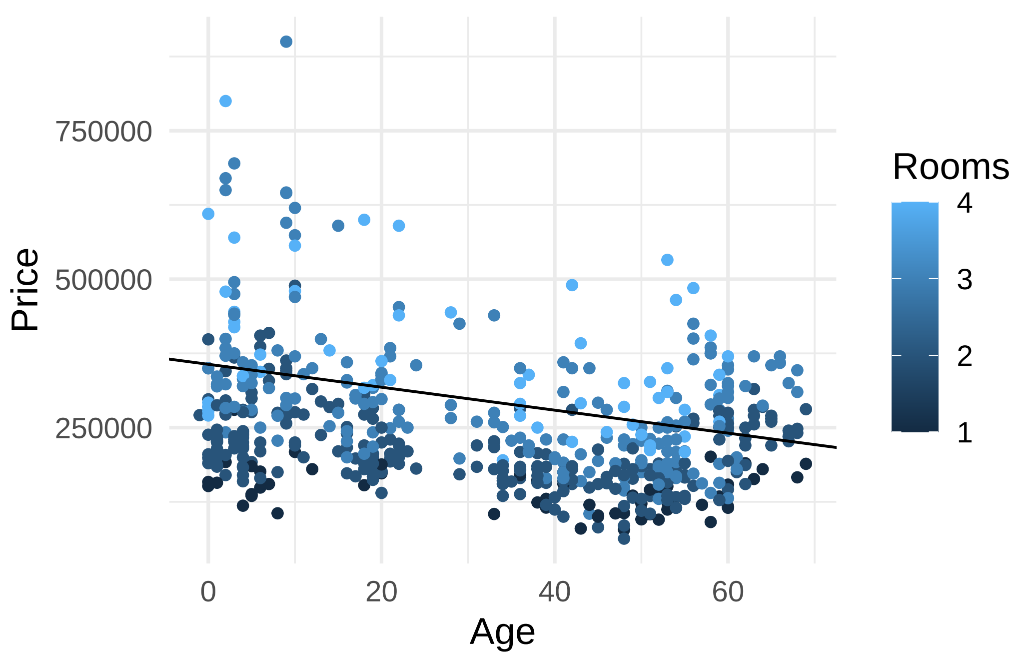

```r
loo(fit1)
```

```
Computed from 4000 by 900 log-likelihood matrix

         Estimate   SE
elpd_loo -11447.9 33.8
p_loo         6.7  1.1
looic     22895.8 67.5
------
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.
```

```r
pp_check(fit1)
```




```
ggplot(data = housingData %>% filter(
  Rooms <= 4,
  Age <= 80,
  BuildingType == 'kt')) +
  geom_point(aes(Age,Price, color = Rooms)) +
  geom_abline(intercept =
              fixef(fit1)[1] +
              fixef(fit1)[2]*mean(housingData$Size),
            slope =
              fixef(fit1)[3])
```



4.4 Sensitivity analysis

Sensitivity analysis is conducted with respect to prior choices (i.e., checking whether the result changes a lot if prior is changed).

5 Non-linear model

The structure of our analysis with the non-linear model is similar to that of the linear model. We first fit the model, then report convergence diagnostics, posterior predictive checks, and

5.1 MCMC inference

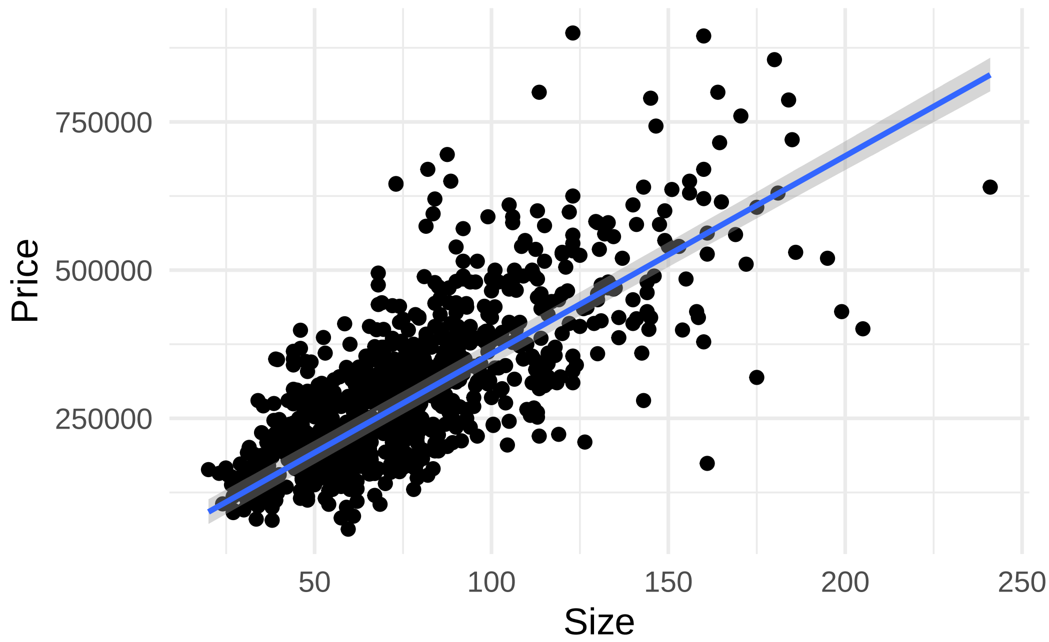
Non-linear models require the bf() around the model specification together with 'nl = TRUE'. The parameters of the model must be specified by b ~ 1 for example, or b ~ 1 + (1|z) if the parameter b varies in groups z.

```
fit2 = brm(bf(Price ~ b1*Size + b2*Age + b3*Age^2 + b4,
                   b1 ~ 1,
                   b2 ~ 1 + (1|PostalCode),
                   b3 ~ 1 + (1|PostalCode),
                   b4 ~ 1 + (1|PostalCode),
                   nl = TRUE
                   ),
           data = housingData,
           family = gaussian(),
           prior = c(
             prior(normal(5000, 1500), nlpar = 'b1'),
             prior(normal(-1000, 5000), nlpar = 'b2'),
             prior(normal(0, 1000), nlpar = 'b3'),
             prior(normal(100000, 30000), nlpar = 'b4')
             ),
           refresh = 0,
           show_exceptions = FALSE,
           file="fit2.rds"
           )
```

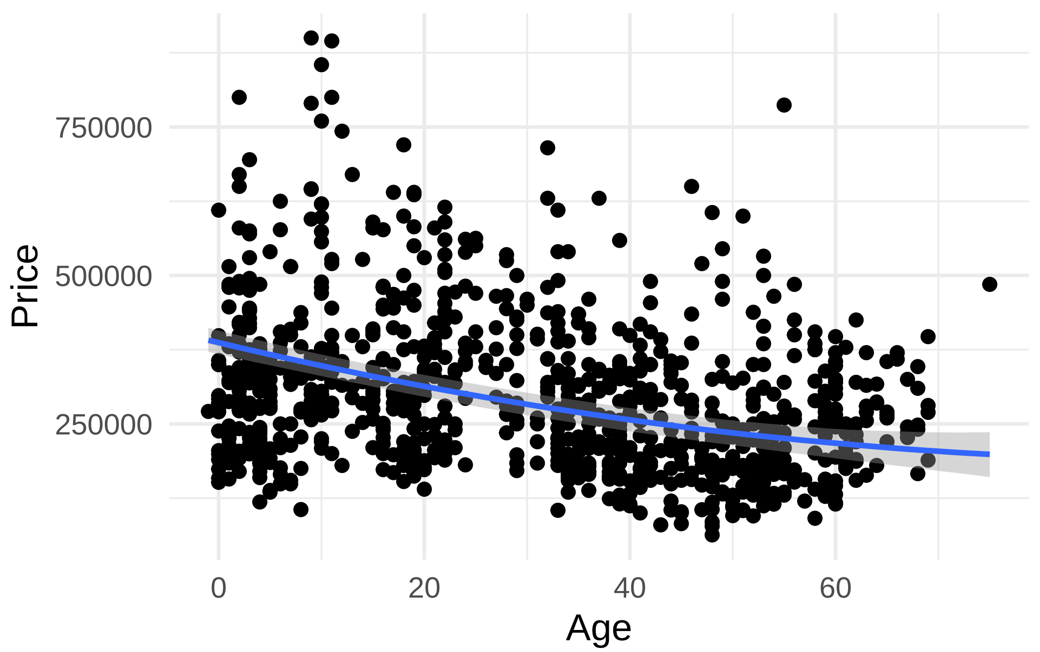
5.2 Convergence diagnostic

To obtain summaries of the fitted model, we again apply function `summary()`.

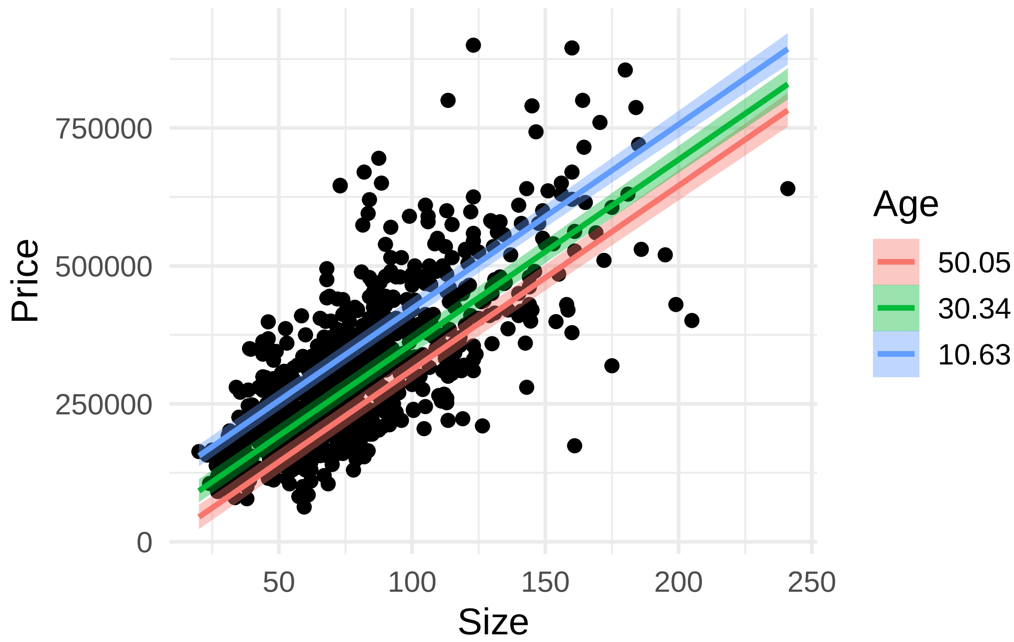
```
plot(conditional_effects(fit2), points = TRUE)
```



Price
750000
500000
250000
50
100
150
200
250
Size



Price
750000
500000
250000
0
20
40
60
Age



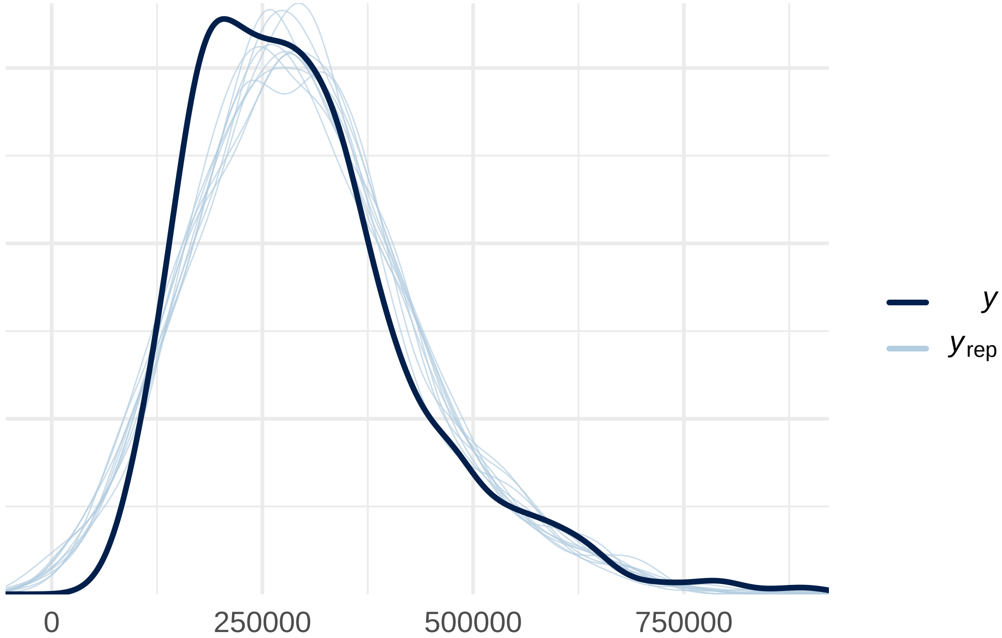

```r
coefs = coef(fit2)
```

5.3 Posterior predictive check

```r
loo(fit2)
```

```
Computed from 4000 by 900 log-likelihood matrix

         Estimate   SE
elpd_loo -11158.7 44.0
p_loo        55.5  6.7
looic     22317.5 88.0
------
Monte Carlo SE of elpd_loo is NA.

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     896   99.6%   436
 (0.5, 0.7]   (ok)         3    0.3%   375
   (0.7, 1]   (bad)        1    0.1%   63
   (1, Inf)   (very bad)   0    0.0%   <NA>
See help('pareto-k-diagnostic') for details.
```

```r
pp_check(fit2)
```



5.4 Sensitivity analysis

6 Model comparison

7 Discussion

8 Conclusion

9 References

[1] Pellervon taloustutkimus PTT, "Alueellinen asuntomarkkinaennuste 2024." Accessed: Mar. 18, 2024. [Online]. Available: https://www.ptt.fi/ennusteet/alueellinen-asuntomarkkinaennuste-2024/

[2] "Asuntojen Hintatiedot." https://asuntojen.hintatiedot.fi/, 2024.