# Bayesian Data Analysis project

## Espoo 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. Many factors such as housing age, size, and average income of housing area can affect house prices. In this project, our goal is to model the effects of Espoo housing size and age on their prices, using linear and non-linear Bayesian multivariate models. Regarding the linear model, we investigate with two variables—the age and the size of the house. For the non-linear model, we also add hierarchy by using average income of the postal area as a grouping variable.

The structure of the report is as follows. Section 2 describes the data and the analysis problem. Section 3 describes the Bayesian models used for analysis. Section 4 presents our analysis with the linear model and Section 5 for the non-linear model. Section 6 shows the results of comparson between our two models. Section 7 discusses issues and potential improvements. Section 8 concludes what was learned from the data analysis. Finally, Section 9 is our self-reflection of what we learned while making the project.

## 2 Description of the data and the analysis

### 2.1 General description

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 conducting the analysis and making this report, to our knowledge, there are no other existing analyses with this housing dataset.

In the original dataset, there are 901 observations and 10 variables. We filter out one house that ages over 100 years from the dataset; therefore, we do the analysis with 900 observations. Each variable contains certain information about a house. We also added two variables to use in our analysis. The first added variable is $Age = 2024 - ProductionYear$, which computes the age of a house based on its production year and is used to investigate the effect of house's age on its price. The second added variable is *IncomeClass*, which rounds the variable *Income*. Below is the first 5 rows of the dataset `HouseData` after adding two variables.

```
  PostalCode Rooms BuildingType Size  Price PricePerSquare ProductionYear
1       2100     4           rt 98.5 397000           4030           1963
2       2110     4           ok 99.0 397000           4010           1954
3       2120     1           kt 28.0 134000           4786           1964
4       2130     3           kt 68.0 245000           3603           1963
5       2130     4           kt 75.5 305000           4040           1964
  Condition LandOwnership Income Age IncomeClass
1         0             1  30435  61       30400
2         0             1  30852  70       30900
3         0             1  30742  60       30700
4         0             1  34342  61       34300
5         0             1  34342  60       34300
```

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

### 2.2 Exploratory data analysis

In this part, we present how we use visualisation to learn more about the dataset `HouseData`. We plot some histograms to explore the range of Espoo housing price, size, and age. As observed in the histograms below, the housing price and size can be assumed to follow normal distribution. Most houses fall in the range of 50 to 100 square meter, with a few outliers of houses over 200 $m^2$. Most houses age from 0 to 10 years. Besides, all house sizes and ages in the dataset are positive values, as they should.

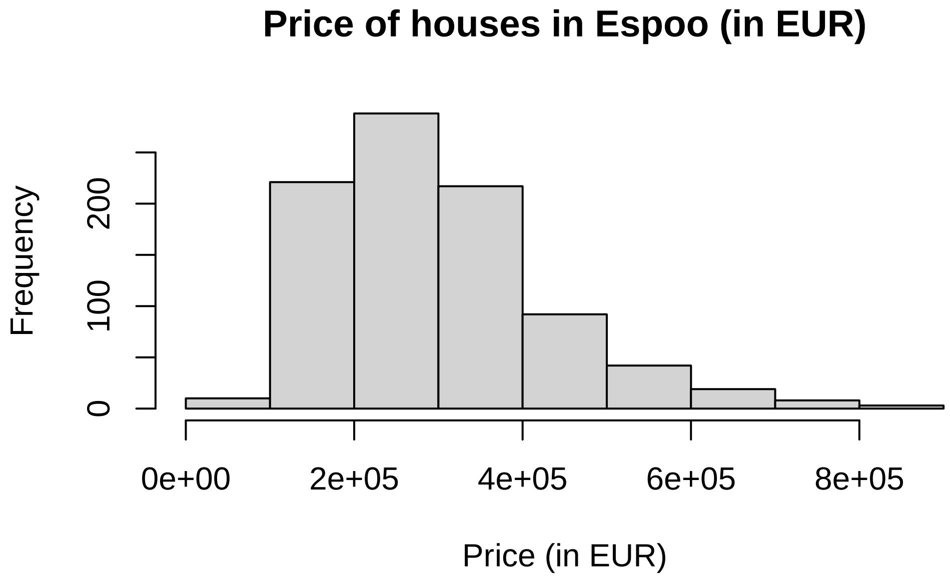


Figure 1: Price of houses in the Espoo housing dataset.

There are three building types in the dataset: apartment, detached, and row house. Since the building type can also affect the price, we group the houses by their building type and explore the overall relationship between housing price and age. The scatter plot in Figure

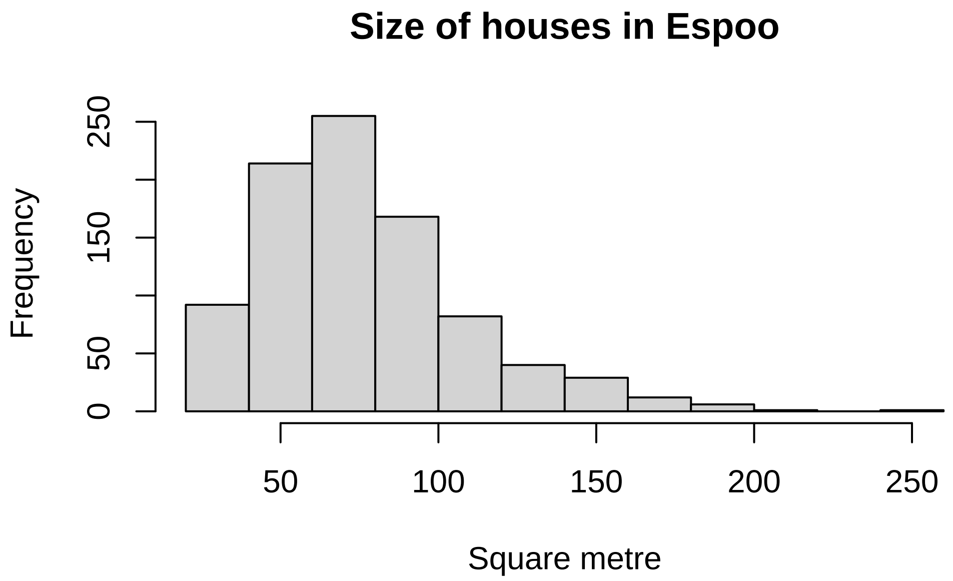


Figure 2: Size range of all houses in the Espoo housing dataset.

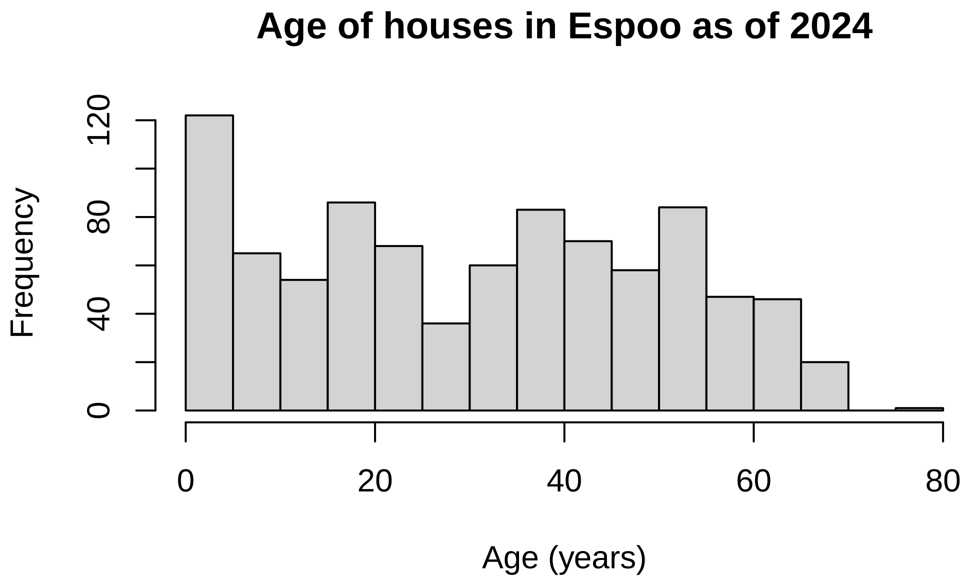


Figure 3: Age range of houses in Espoo as of 2024

suggests that overall, across the ages, apartment is the cheapest type, followed by row house. Detached house is the most expensive type.

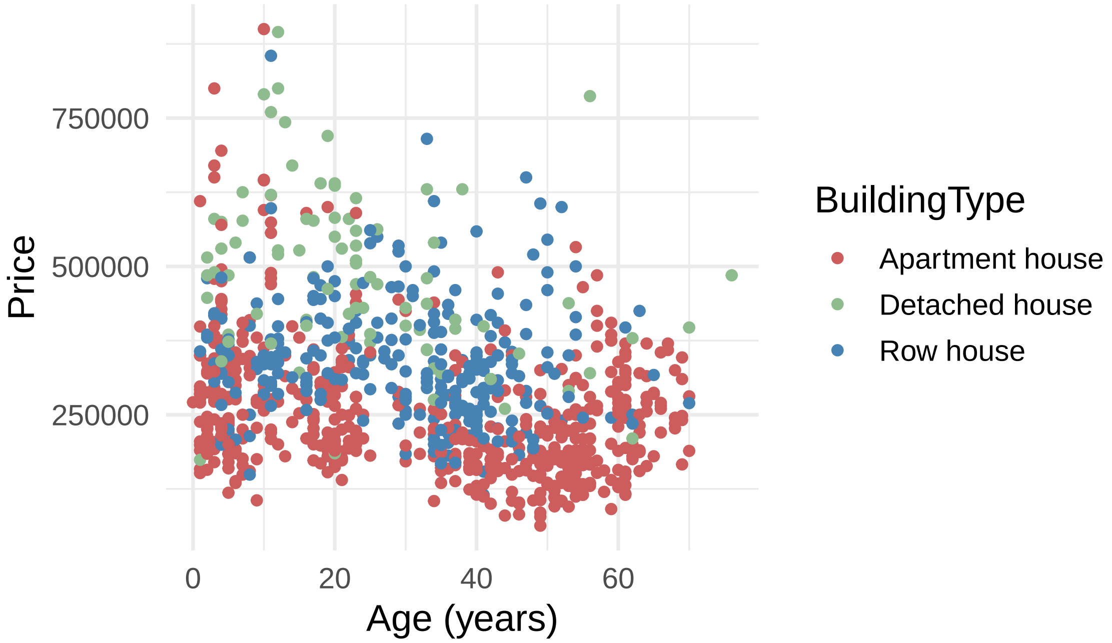


Figure 4: Housing price & age by building type.

Figure below shows the housing size and price, grouped by three building types. We can observe that generally, apartments are in the cheapest and smallest size range, while detached houses are in the most expensive and largest size range.

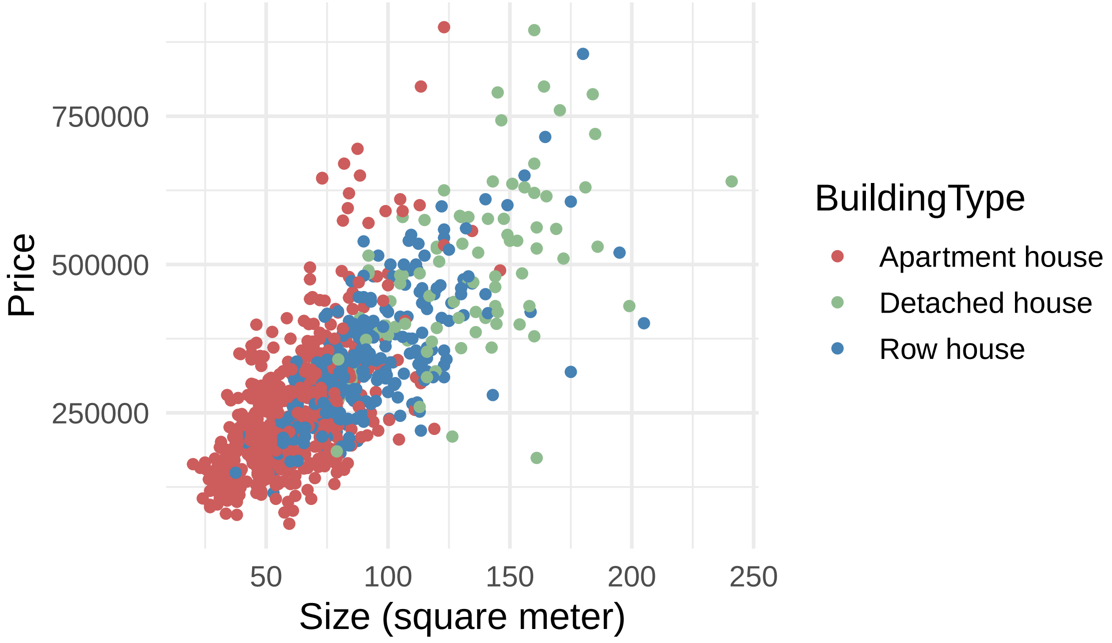


Figure 5: Housing price & size by building type

## 3 Models

From the exploratory data analysis and by intuition, there can be effect of housing size and age on its price. Therefore, we will model the effects of housing size and age in the dataset

`HouseData` on their prices.

### 3.1 Linear model

As a common practice, we start our analysis with a simple, vanilla linear model with Gaussian noise.

### 3.2 Non-linear model with hierarchy

## 4 Analysis with the linear model

Overall, this section shows the choices for priors, the code of our linear model 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 and sensitivity analysis.

### 4.1 Priors

For the priors of the linear model, we will use the weakly informative priors for the housing size and age.

### 4.2 MCMC inference

To fit the model and run MCMC inference, we use **`brms`**—a high-level interface for Stan providing tools to create a wide range of Bayesian models. By default, 4 chains were drawn with 2000 iterations for each chain. The warm-up length for each chain is 1000.

```r
fit1 = brms::brm(Price ~ Size + Age,
                 data = HouseData,
                 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")
                   ),
                 show_exceptions = FALSE,
                 # This causes brms to cache the results
                 file="fit1.rds"
                 )
```

### 4.3 Convergence diagnostic

Below is the summary and convergence diagnostic report for our fitted linear model. The important metric is $Rhat$ ($\hat{R}$) and $ESS$ (effective sample size or $n_{eff}$ ). $\hat{R}$ is computed to monitor the convergence of iterative simulation. For all variables our $\hat{R} = 1$, which indicates possible convergence and means that we can stop the sampling process. In case $\hat{R} > 1.01$, we need to keep sampling to reach convergence.

By using function `check_hmc_diagnostics()`, we can verify that none of 4000 iterations saturated the maximum tree depth of 10.

```
Divergences:

Tree depth:

Energy:


NULL


b_Intercept      b_Size       b_Age       sigma      lprior        lp__
  0.8376779   0.8113703   0.6718708   0.7389249   0.7514043   0.5416204
```

By using function `plot()`, we can plot the MCMC chains and the posterior distributions for each parameter. From the figure, we observe that our MCMC chains have converged and mixed well and to the same posterior.

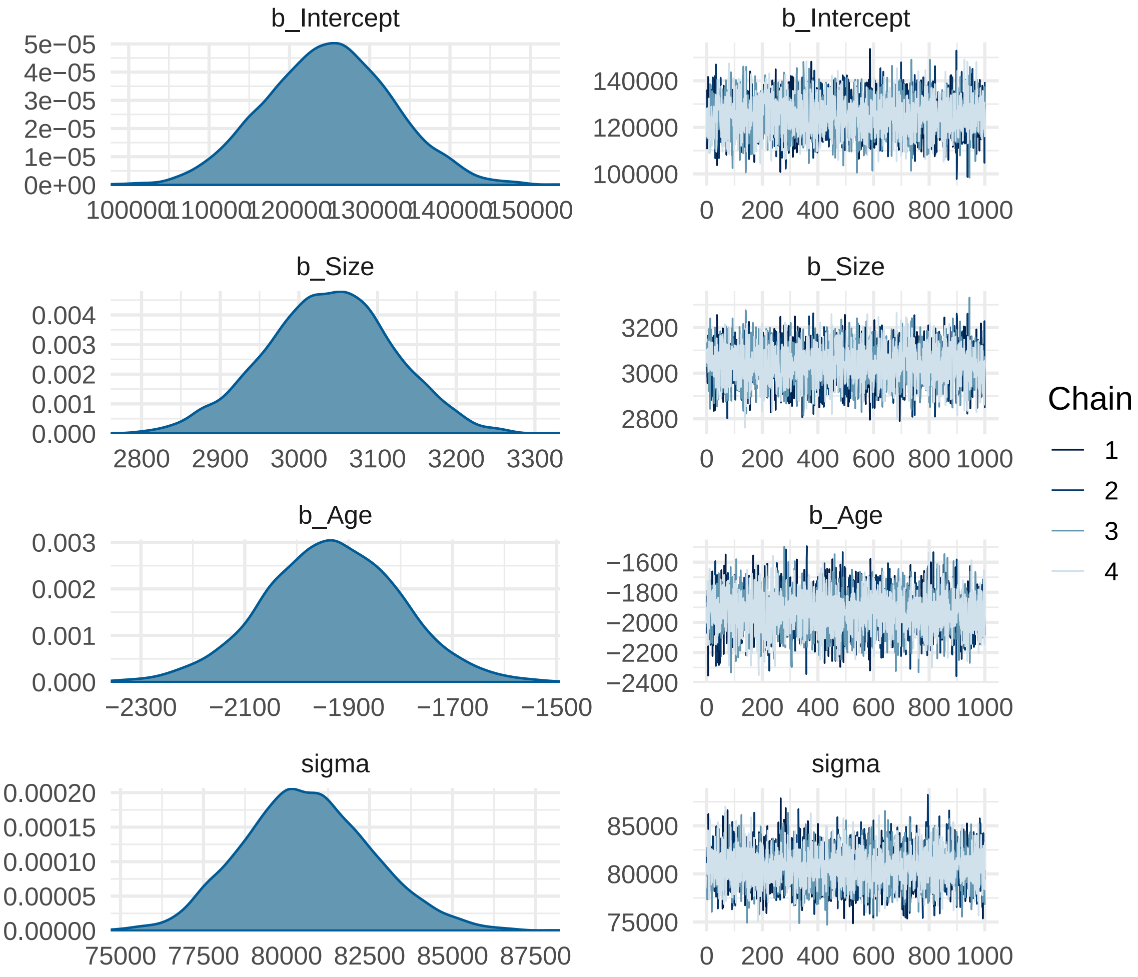


Figure 6: Four MCMC chains and posterior distributions for each parameter.

### 4.4 Posterior predictive check

To investigate and compare model fit, we can apply graphical posterior predictive checks. Let's check the posterior predictions compared to the observed data using the `pp_check` function. In the plot below, the dark blue curve represents the $y$ values, which are the observed data, and the light blue curves represent $y_rep$ values, which are replicated data sets from the posterior predictive distribution. Based on the plot, the posterior prediction roughly encapsulates the main features of the observed data. However, there are negative values $y_rep$ from the posterior predictive distribution.

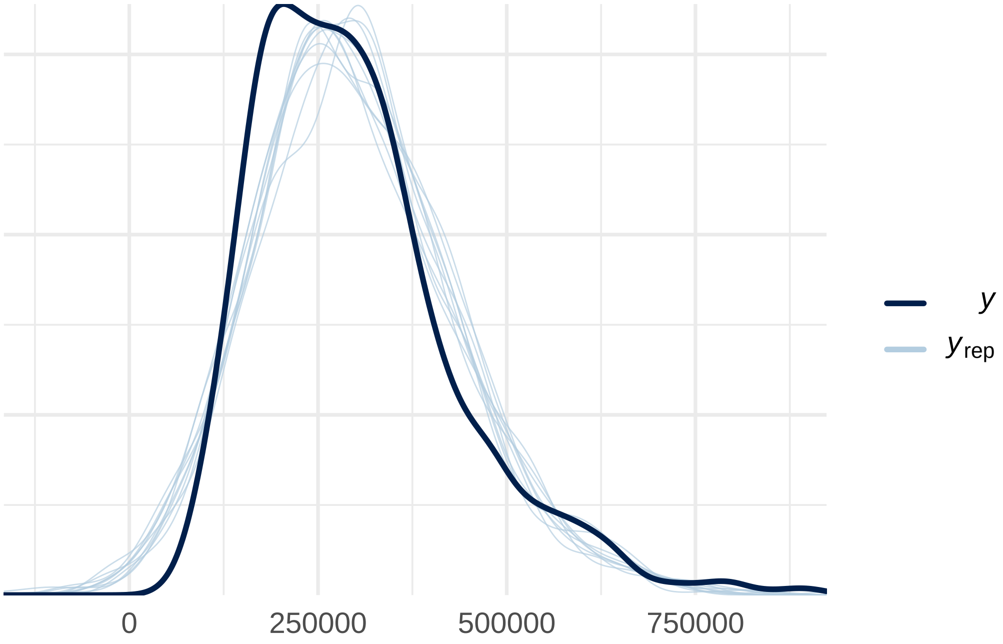


Figure 7: Posterior predictive check

Next, we use the `conditional_effects` method to visualize the model-implied linear relationship between housing size and price as well as housing age and price.

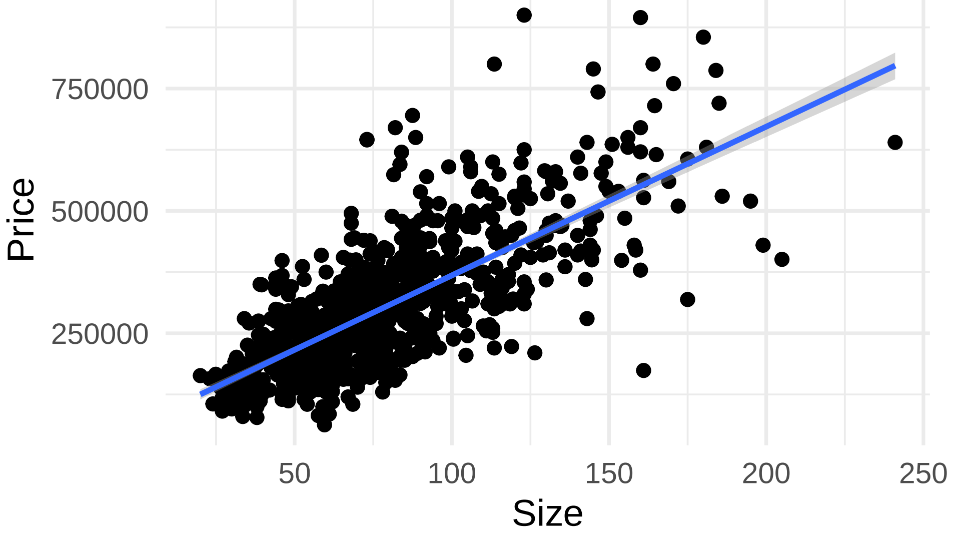


Figure 8: Conditional effects of housing attributes on housing price.

### 4.5 Sensitivity analysis

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

We use external references to pick the new weakly informed prior for the housing size. This technique is more general and doesn't assume there is prior knowledge. In Espoo, the average apartment size is 58 square meters in 2023 [3]. Although average size can be different across

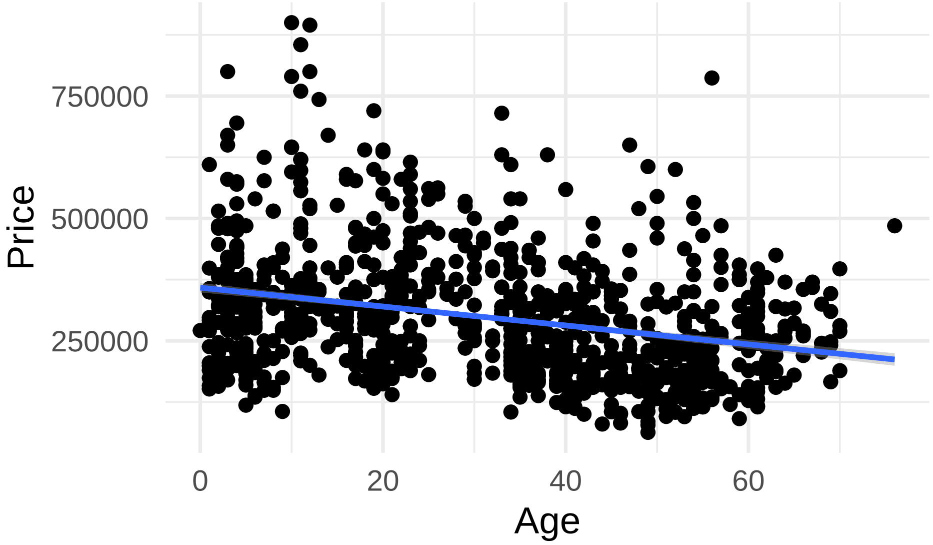


Figure 9: Conditional effects of housing attributes on housing price.

different building types, to keep the sensitivity analysis of the linear model simple, we choose $\mu_0 = 58$ for the new prior of housing size. The housing size range is estimated to be 35-81 square meters.

Assume that 99.7% of all houses in Espoo fall into this range. Under our assumption of a normal distribution, this range will encompass values between $\mu \pm 3\sigma$. Assuming symmetry, the chosen mean $\mu_0$ and either the upper or lower bound of the reference range can be used to solve for $\sigma_0$:

$$Pr(58 - 3\sigma_0 > 35) \approx 0.997 \rightarrow Pr(\sigma_0 < 7.67) \approx 0.997$$

Since we have found that $\sigma_0 < 7.67$, let's choose $\sigma_0 = 7$ for the weakly informative prior of the housing size.

Regarding the prior for the housing age, we slightly change it to see whether the result changes a lot.

```
fit1_1 = brms::brm(Price ~ Size + Age,
                 data = HouseData,
                 family = gaussian(),
                 prior = c(
                   prior(normal(58, 7), class = "b", coef = Size),
                   prior(normal(-1000, 4000), class = "b", coef = Age),
                   prior(normal(100000, 30000), class = "Intercept")
                   ),
                 show_exceptions = FALSE
                 )
```

As shown in the posterior predictive visual check below, the light blue curves, which represent the replicated data sets from the posterior predictive distribution, change dramatically. This behavior suggests that our linear model is quite sensitive to changes in priors.

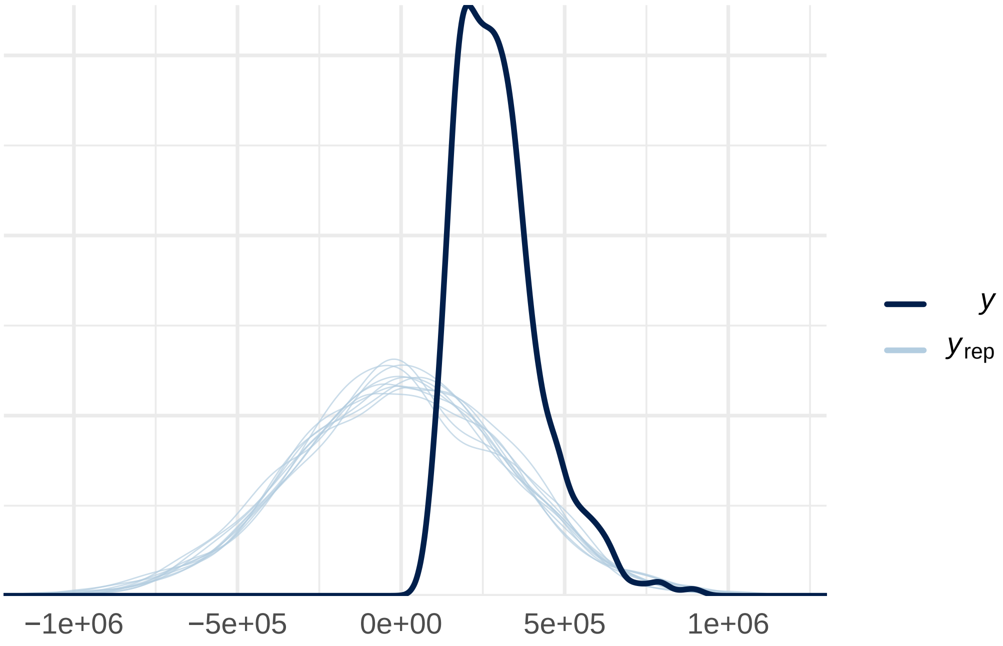


Figure 10: Posterior predictive check for the linear model with new priors

# 5 Non-linear model

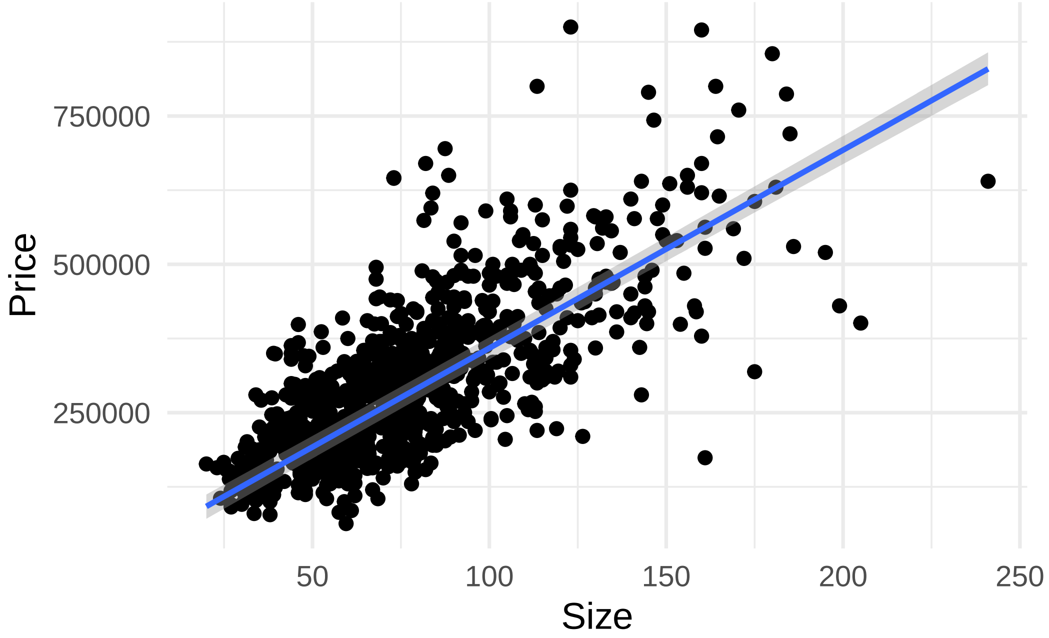
The structure of our analysis with the non-linear model is similar to that of the linear model. We first fit the model, then report convergence diagnostics, posterior predictive checks, and sensitivity analysis.

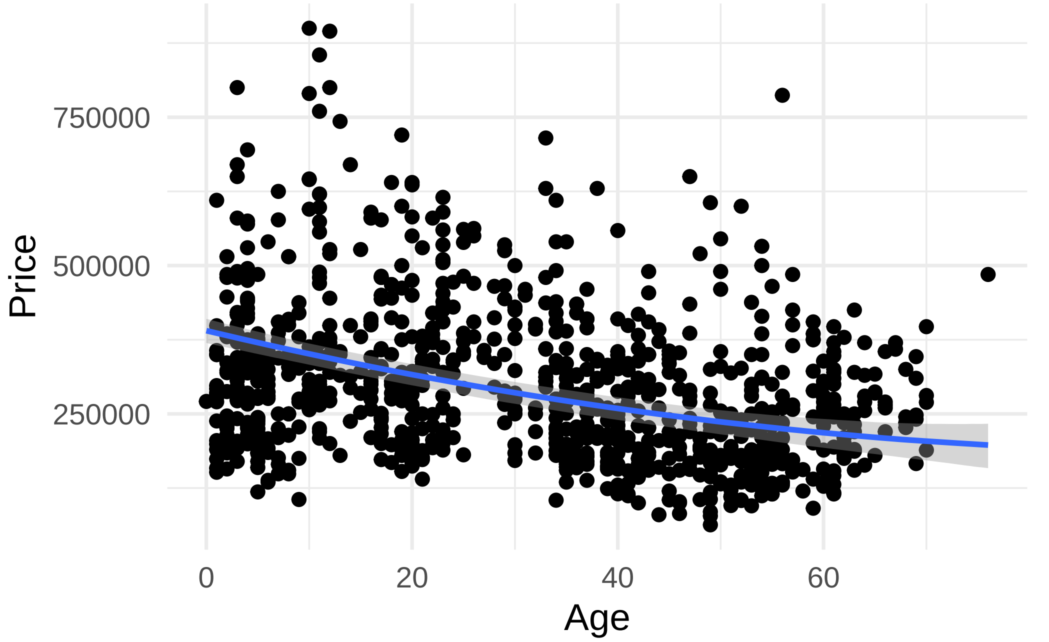
## 5.1 MCMC inference

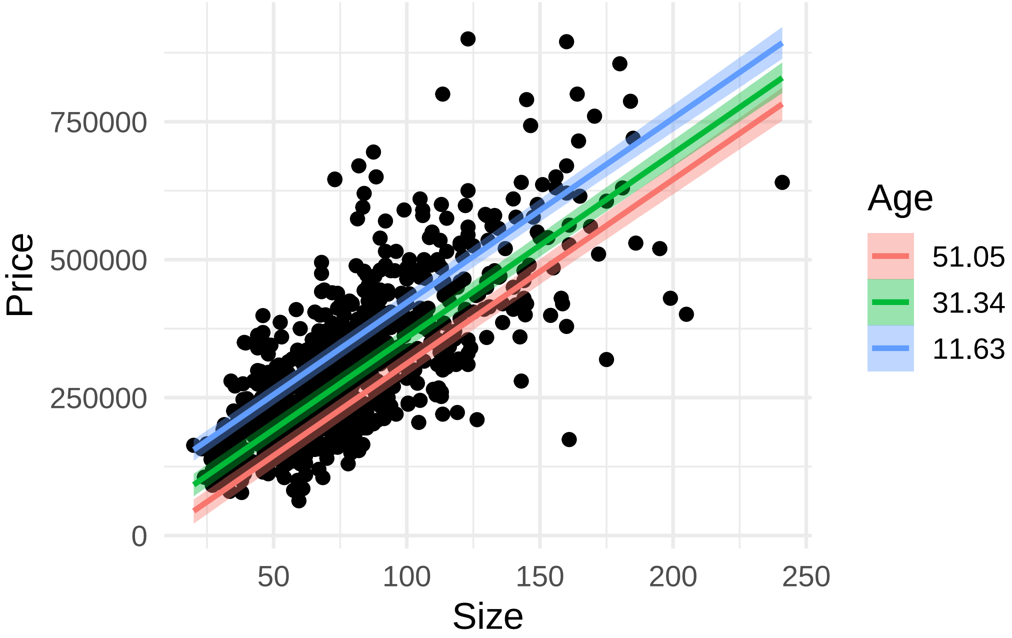
Non-linear models require the bf() around the model specification together with 'nl = TRUE'. The parameters of the model must be specified by b ~ 1 for example, or b ~ 1 + (1|z) if the parameter b varies in groups z.

## 5.2 Convergence diagnostic

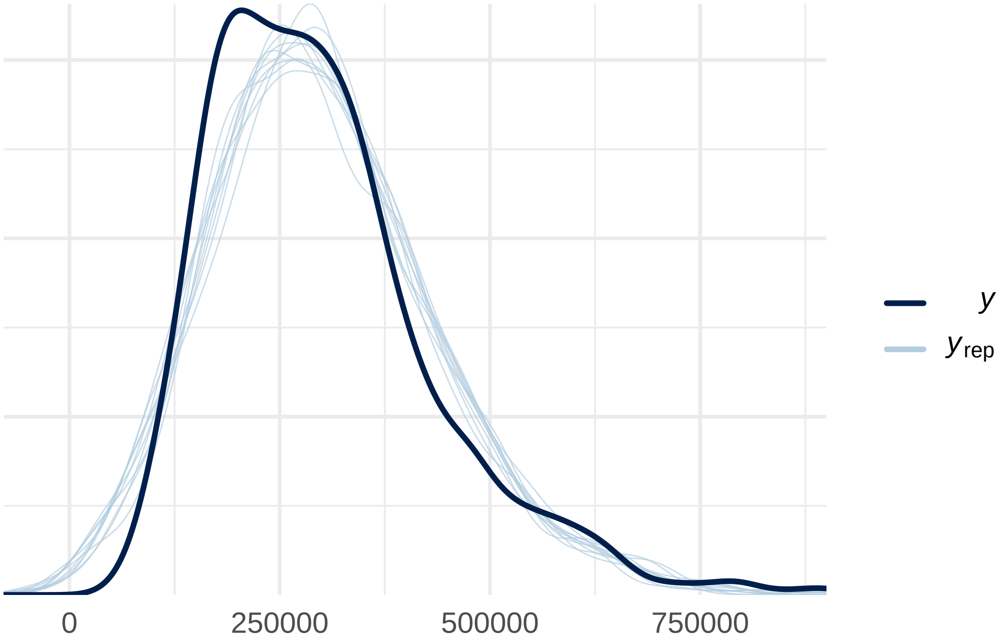
To obtain summaries of the fitted model, we again apply function `summary()` in `brms` package.


Price
750000
500000
250000
50
100
150
200
250
Size


Price
750000
500000
250000
0
20
40
60
Age



### 5.3 Posterior predictive check



### 5.4 Sensitivity analysis

## 6 Model comparison

### 6.1 Model comparison using LOO-CV

We start our model comparison by using leave-one-out cross-validation. For the linear model, the

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

         Estimate   SE
elpd_loo -11447.8 33.7
p_loo         6.7  1.1
looic     22895.6 67.4
------
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.


Computed from 4000 by 900 log-likelihood matrix

         Estimate   SE
elpd_loo -11159.0 43.9
p_loo        56.1  6.9
looic     22318.1 87.9
------
Monte Carlo SE of elpd_loo is NA.

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     897   99.7%   688
 (0.5, 0.7]   (ok)         2    0.2%   154
   (0.7, 1]   (bad)        1    0.1%   43
   (1, Inf)   (very bad)   0    0.0%   <NA>
See help('pareto-k-diagnostic') for details.
```

## 7 Discussion

## 8 Conclusion

## 9 Self-reflection

We learned how to form a Bayesian analysis problem. We revised the linear and hierarchical models covered in this course. We also learned how to structure the report to make it readable and easy to follow.

## 10 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.

[3] “Shift towards small rental apartments in housing construction in Helsinki, Tampere, and the capital region.” https://www.helsinkitimes.fi/finland/finland-news/domestic/23681-shift-towards-small-rental-apartments-in-housing-construction-in-helsinki-tampere-and-the-capital-region.html%0A, 2023.

# 11 Appendix