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 non-hierarchical linear models and hierarchical models. Regarding the linear model, we investigate with two variables—the age and the size of the house. For the hierarchical 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 models used for analysis and prior choices. Section 4 presents our analysis with the non-hierarchical linear model and Section 5 for the hierarchical model. Section 6 shows the results of comparison 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.

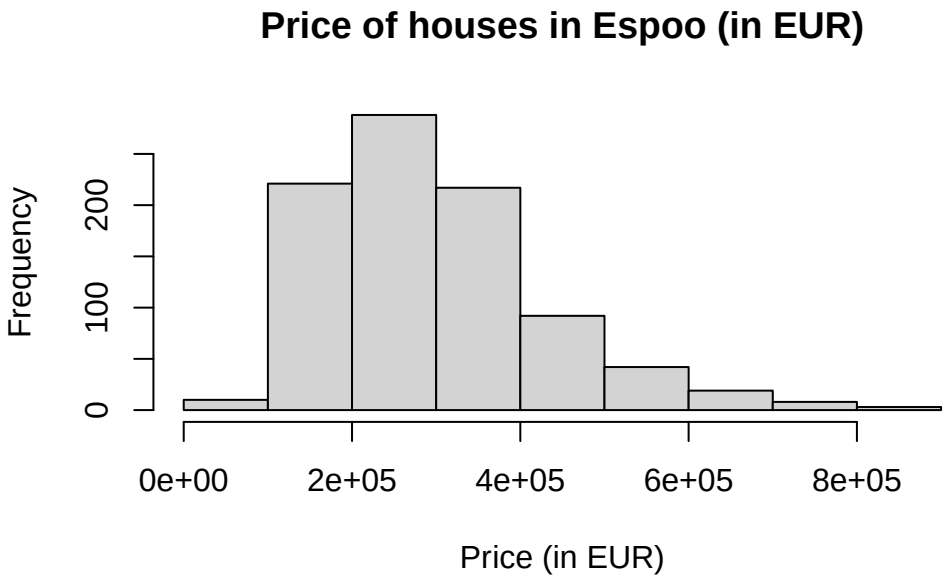

Figure 1: Price of houses in the Espoo housing dataset.

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.

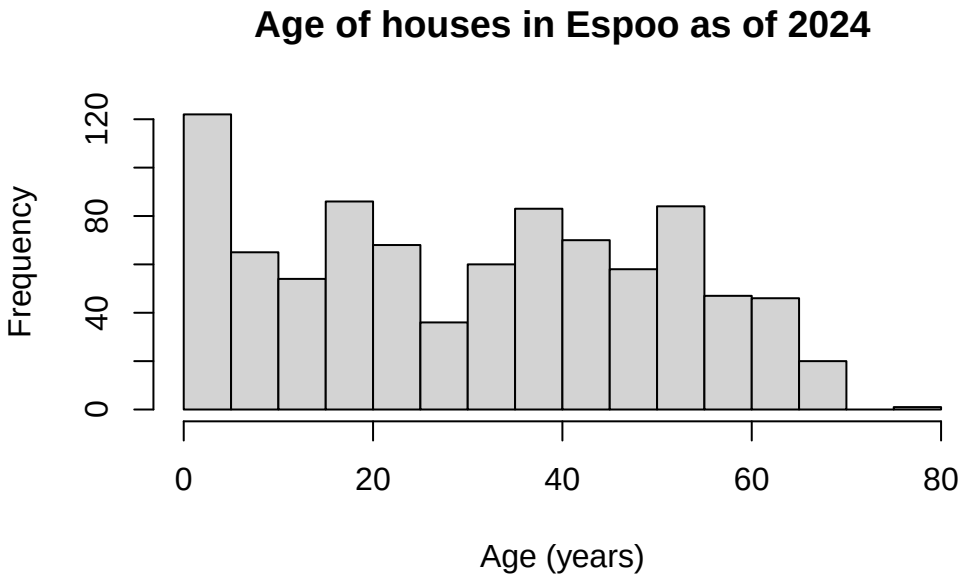

Figure 3: Age range of houses in Espoo as of 2024

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

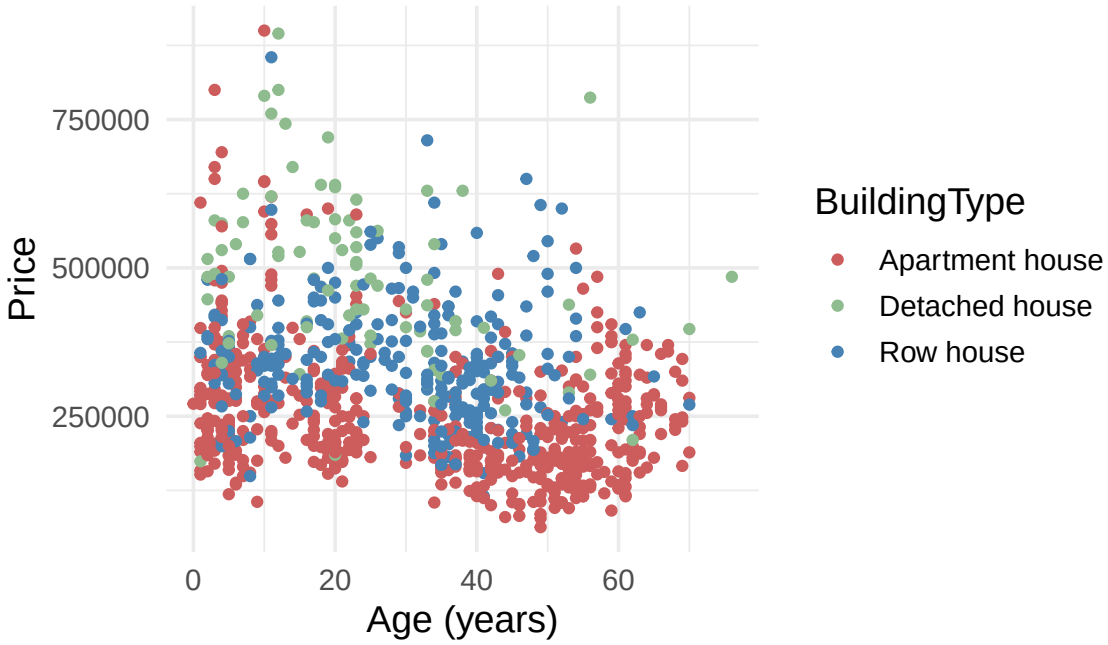

Figure 4: Housing price & age by building type.

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.

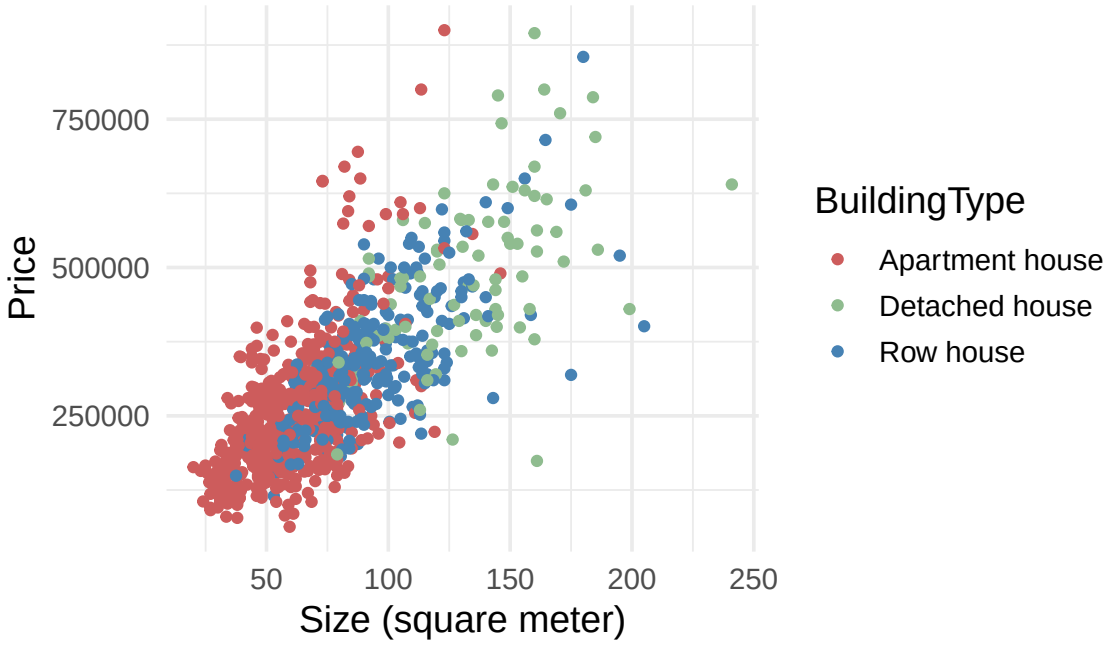

Figure 5: Housing price & size by building type

3 Models and prior choices

From the exploratory data analysis and by intuition, we can see that there can be meaningful 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. As a common practice, we start our analysis with a simple, vanilla linear non-hierarchical model with Gaussian noise.

Next, to take into account the variability of house prices by area, we use the average incomes of the postal areas as the grouping variable for the hierarchical model. We chose to use the income data (instead of the postal codes themselves), because we wanted to aggregate some of the smaller postal areas with other areas and we assume that average income gives a good enough similarity measure.

In both models, we choose normal priors for the parameters with (μ, σ) equal to $(5000, 1500)$ for the price per square meter, $(-1000, 5000)$ for the yearly depreciation of house price and $(1e5, 3e4)$ for the intercept. A reasonable guess for the price per square meter is 5000 euros and we expect that it's not too far off so we set the standard deviation at 1500, meaning values under 0 and over 10000 are very unlikely by our prior knowledge. We estimate the yearly depreciation to be some negative number on the order of thousands of euros per year, so we set its prior mean at -1000 with a fairly large variance by setting the standard deviation at 5000 (because we expect that 50 year old houses, for example, aren't free). We set the intercept's prior mean at 100,000 euros & the standard deviation at 30,000 euros, because we expect that the cheapest new apartments cost about 100,000 euros, but aren't free no matter how small.

4 Analysis with the linear model

Overall, this section shows 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 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.

4.2 Convergence diagnostic

Below is the summary and convergence diagnostic report for our fitted linear model. $\hat{R}$ is computed to monitor the convergence of iterative simulation. For all variables our $\hat{R}$ are under 1.01, 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. All ESS ratios are over 50 percents, which means that the effective sample sizes are sufficient.

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

```
b_Intercept      b_Size       b_Age       sigma      lprior        lp__
   1.000301    1.000093    1.000288    1.000208    1.000941    1.000344
```

```
b_Intercept      b_Size       b_Age       sigma      lprior        lp__
  0.7760635   0.7403750   0.8181959   0.7035505   0.7518872   0.5189594
```

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.

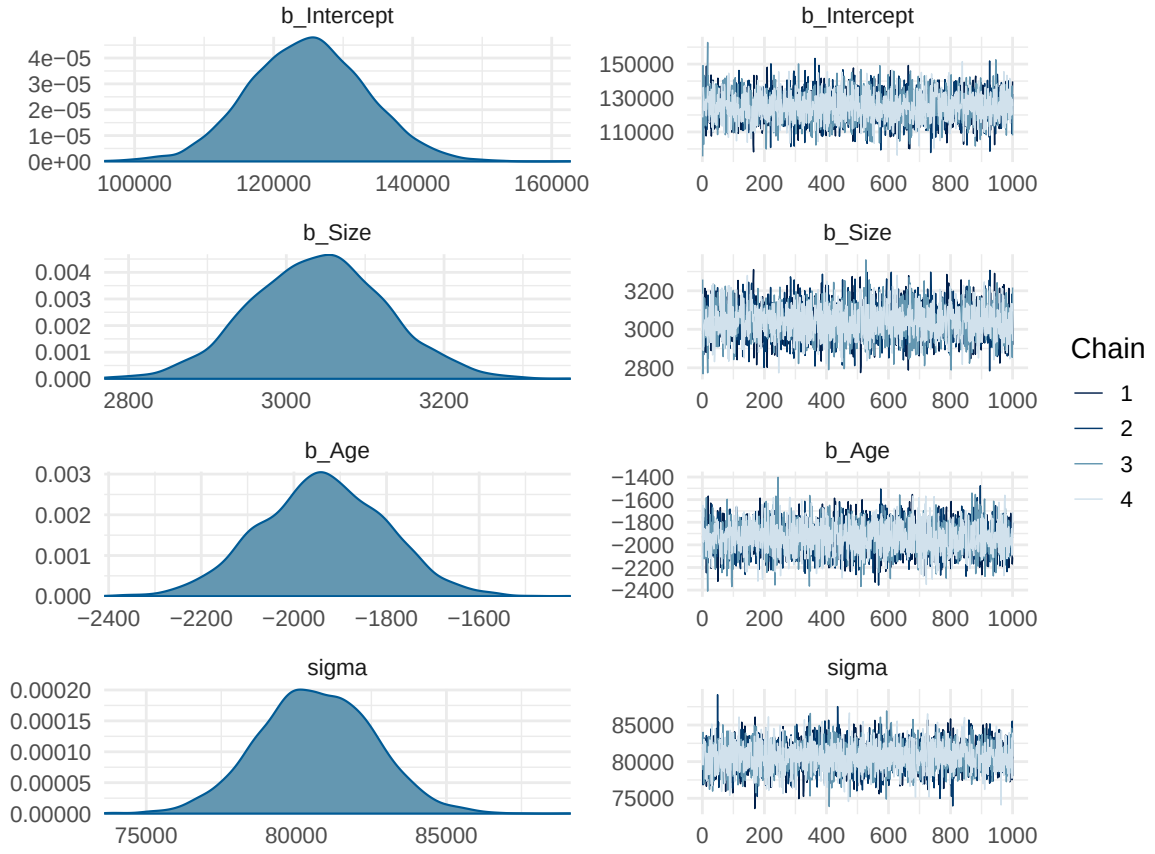

Figure 6: Four MCMC chains and posterior distributions for each parameter.

4.3 Posterior predictive check

To investigate and compare model fit, we can apply graphical posterior predictive checks (Figure 7). 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, which means the positivity of house price is not captured.

Next, in Figure 8, 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.

4.4 Sensitivity analysis

Sensitivity analysis is conducted with respect to prior choices. To keep the sensitivity analysis of the linear model simple and not too time-consuming, we slightly change the priors for all three parameters to see whether the results change a lot.

O A B

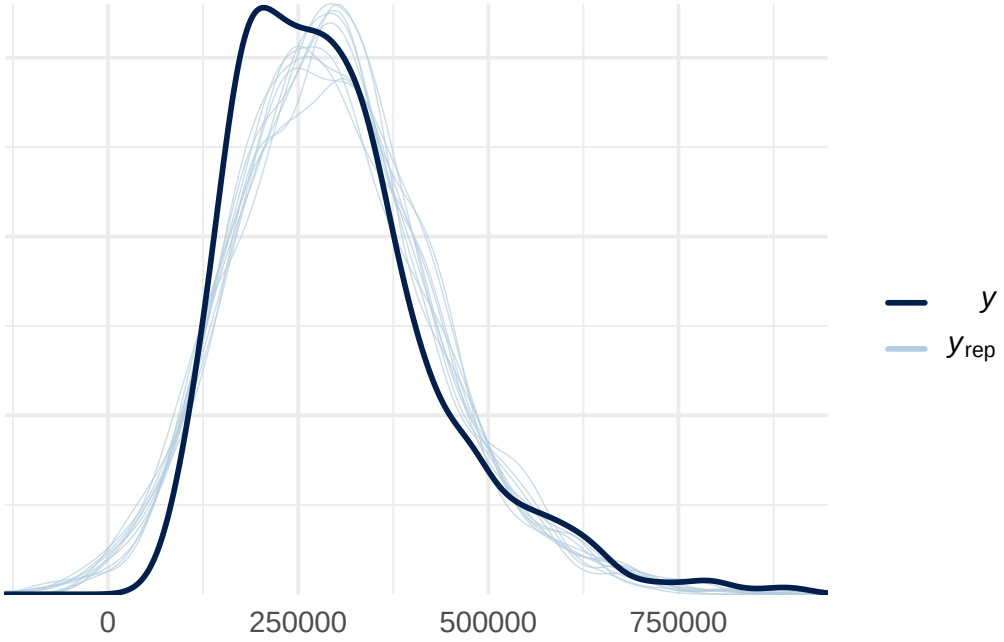

Figure 7: Posterior predictive check for the non-hierarchical model

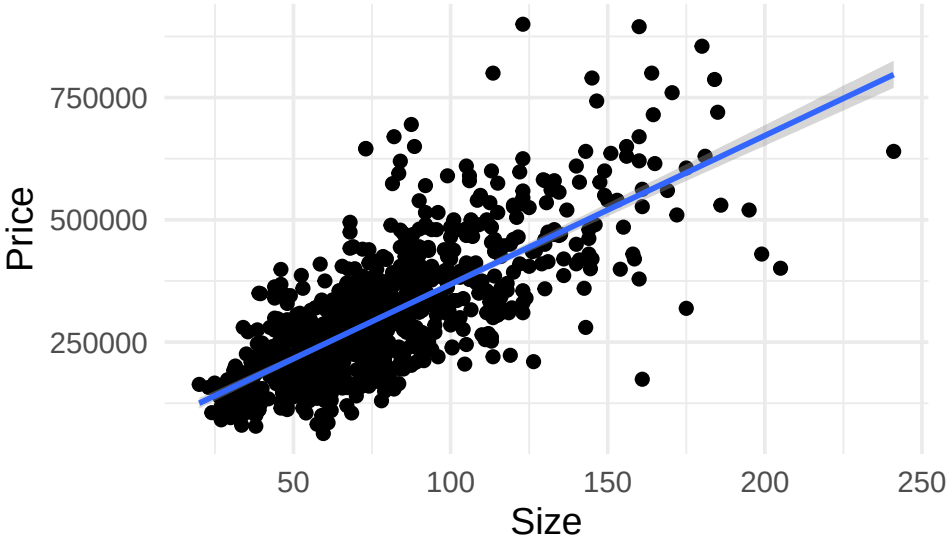

Figure 8: Conditional effects of housing attributes on housing price.

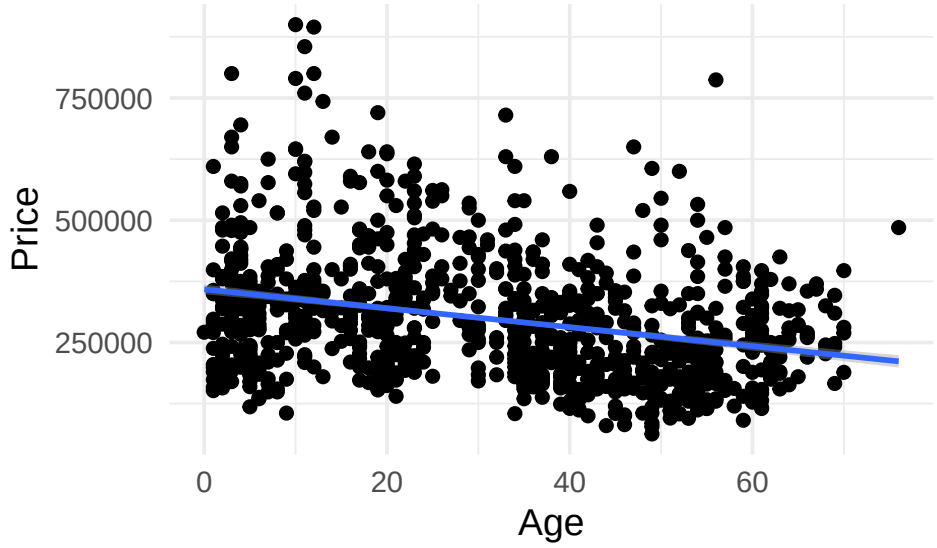

Figure 9: Conditional effects of housing attributes on housing price.

```
Intercept 124886.026 101137.886 114708.412
Size        3040.584   3171.893   2976.306
Age        -1937.329  -1891.169  -1962.101
```

The first column is parameter of our original linear model. The second and third column are for the linear models with modified priors. We see that changing the priors does not change the parameter values much.

As shown in Figure 10 and Figure 11 of the posterior predictive visual check below, the light blue curves, which represent the replicated data sets from the posterior predictive distribution, does not change dramatically. This behavior suggests that our linear model is not sensitive to our changes in priors.

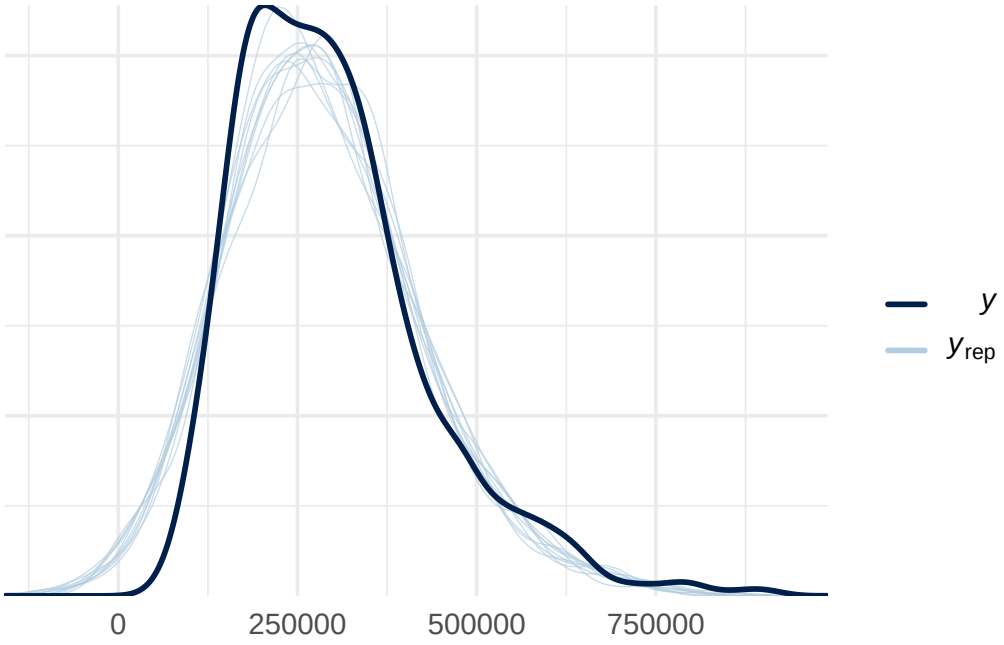

Figure 10: Posterior predictive check for the linear model with new priors

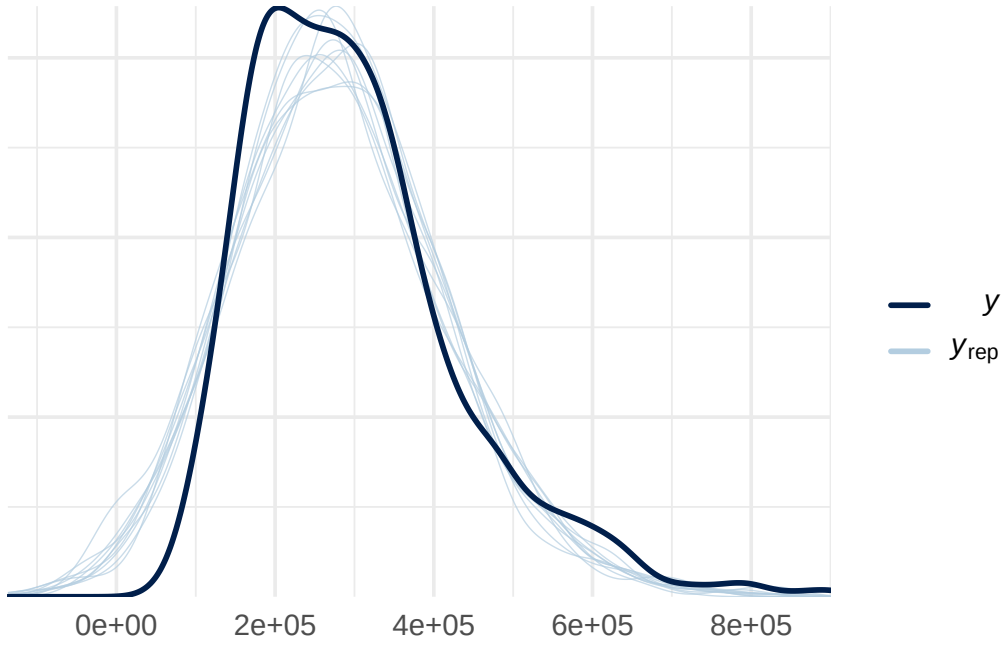

Figure 11: Posterior predictive check for the linear model with new priors

5 Hierarchical 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, and posterior predictive checks. We use the same priors for the hierarchical model as for the non-hierarchical one.

5.1 MCMC inference for the hierarchical model

The MCMC was run by the defaults of the `brm()` function: 4 chains with 2000 iterations and warmup constituting 50 % of the draws. However, the default adaptive step size of 0.8 was changed to +.9, because the runs with the default resulted in divergent transitions.

5.2 Convergence diagnostic

Below is the summary of the fitted hierarchical model. Again, we are checking $\hat{R}$ values and effective sample sizes to diagnose convergence:

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: Price ~ b1 * Size + b2 * Age + b3 * Age^2 + b4
         b1 ~ 1
         b2 ~ 1 + (1 | PostalCode)
         b3 ~ 1 + (1 | PostalCode)
         b4 ~ 1 + (1 | PostalCode)
   Data: HouseData (Number of observations: 900)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Group-Level Effects:
~PostalCode (Number of levels: 41)
```

```
                  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(b2_Intercept)    339.30    217.66    16.53   797.50 1.01      640     1539
sd(b3_Intercept)      5.79      3.62     0.27    13.58 1.00      901     1713
sd(b4_Intercept) 56553.50   7656.46 43494.93 73322.80 1.00     1161     1814

Population-Level Effects:
              Estimate Est.Error  l-95% CI  u-95% CI Rhat Bulk_ESS Tail_ESS
b1_Intercept   3339.13     65.76   3209.13   3466.08 1.00     5509     3358
b2_Intercept  -4092.39    485.33  -5050.45  -3150.51 1.00     3232     3009
b3_Intercept     20.43      7.69      5.05     34.94 1.00     3255     3230
b4_Intercept 133389.01  10798.39 111830.93 154821.17 1.00     1214     1814

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma 56791.73   1410.12 54103.97 59687.70 1.00     5898     2815

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).
```

As shown in the summary above, the $\hat{R}$ value is under 1.01 for every parameter, effective sample sizes around 2000 and there were no divergent transitions. The $\hat{R}$'s imply convergence of the chains and the effective sample sizes seem sufficient.

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

5.3 Posterior predictive check

Visually (Figure 12-16) we can see that the hierarchical model gives a fairly similar result in capturing the distribution of house prices as does the non-hierarchical model, i.e., the distribution matches fairly well, but doesn't capture some areas of the distribution (especially for the houses with below-average prices).

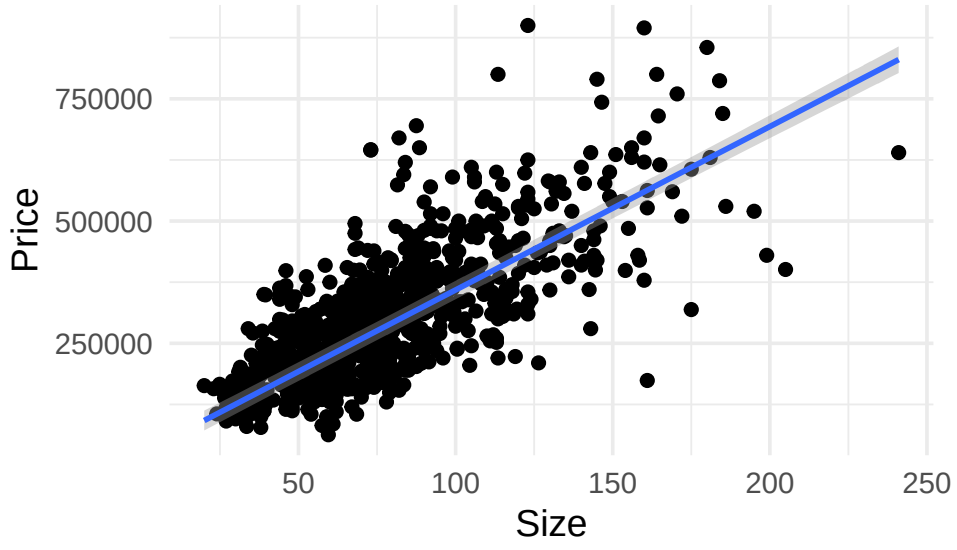

Figure 12: Conditional effects of housing attributes on housing price.

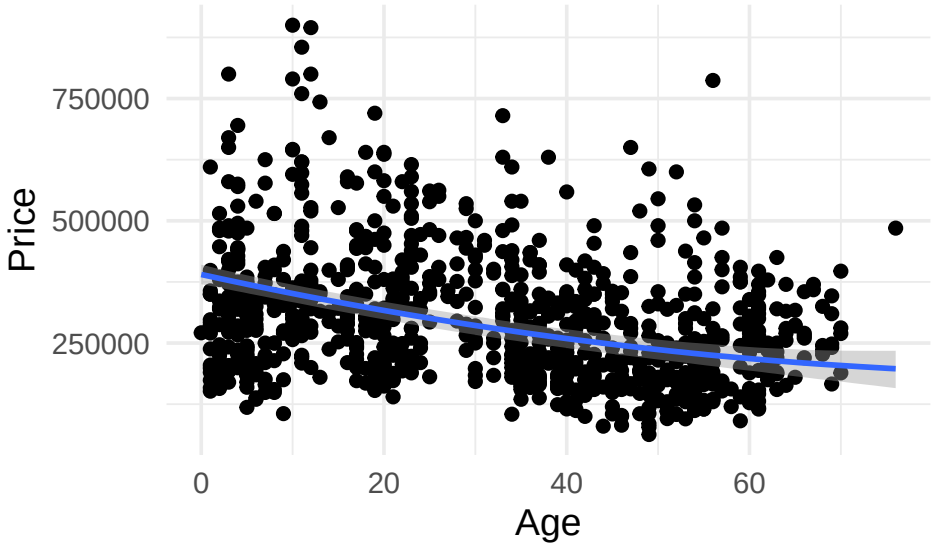

Figure 13: Conditional effects of housing attributes on housing price.

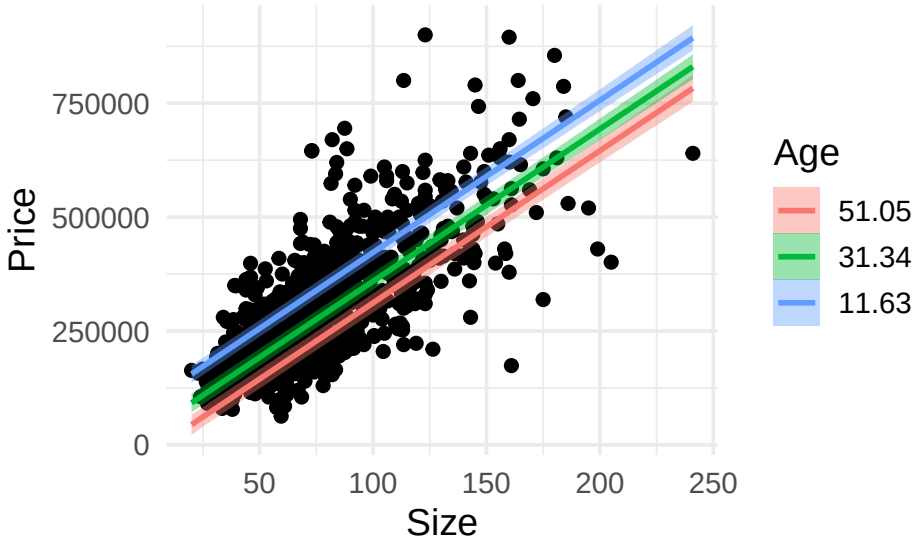

Figure 14: Conditional effects of housing attributes on housing price.

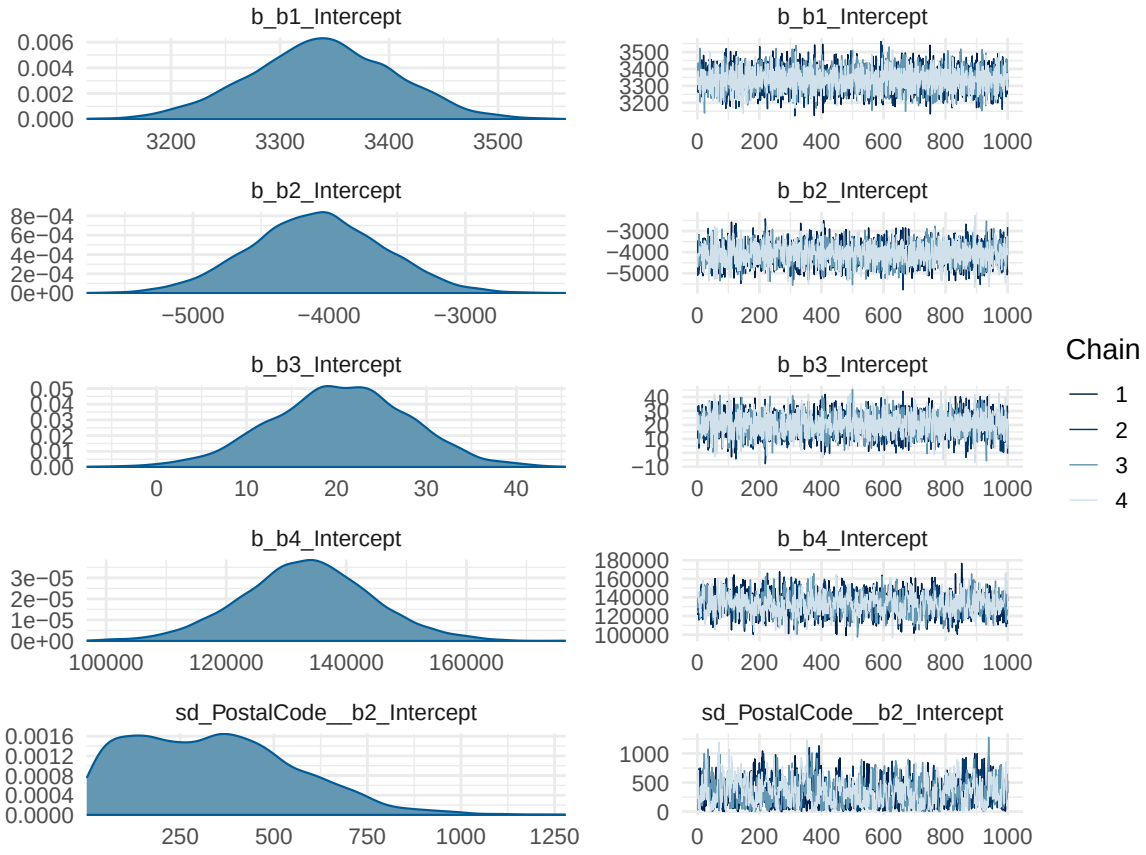

Figure 15: Hierarchical model: Four MCMC chains and posterior distributions for each parameter.

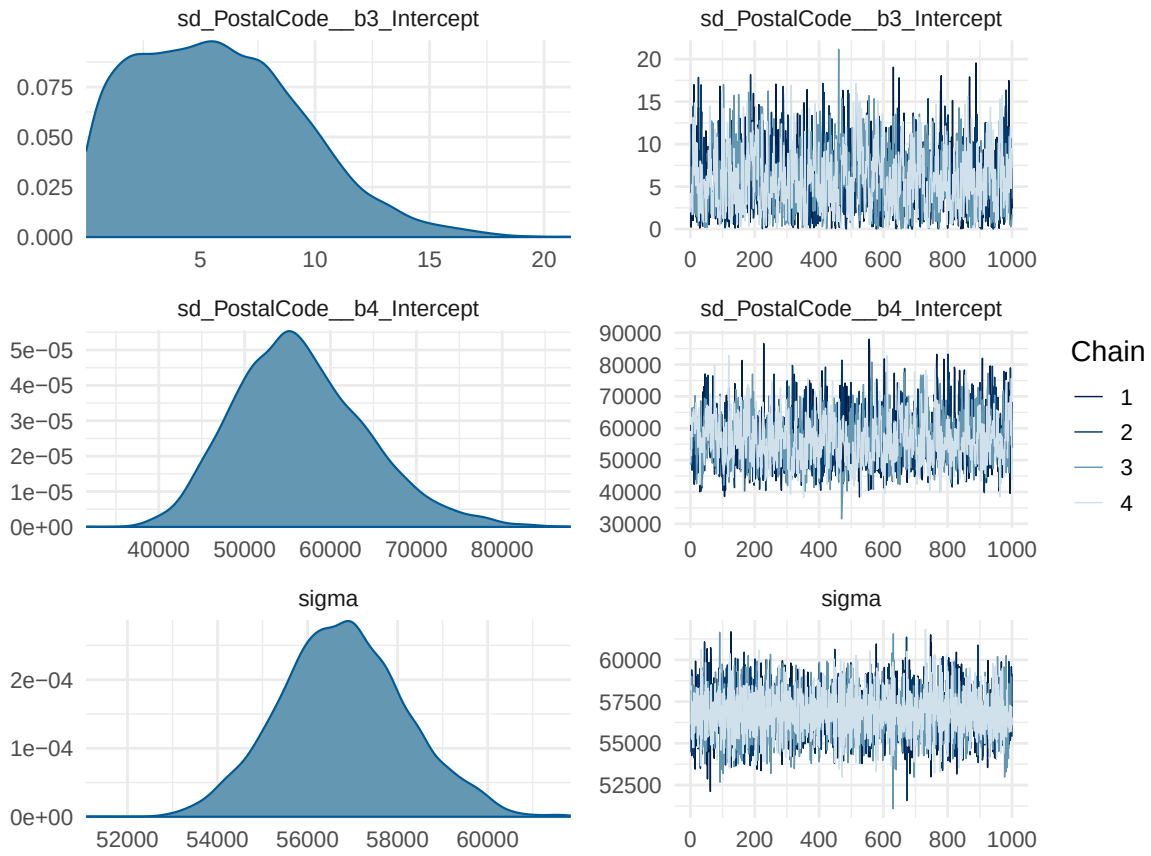

Figure 16: Hierarchical model: Four MCMC chains and posterior distributions for each parameter.

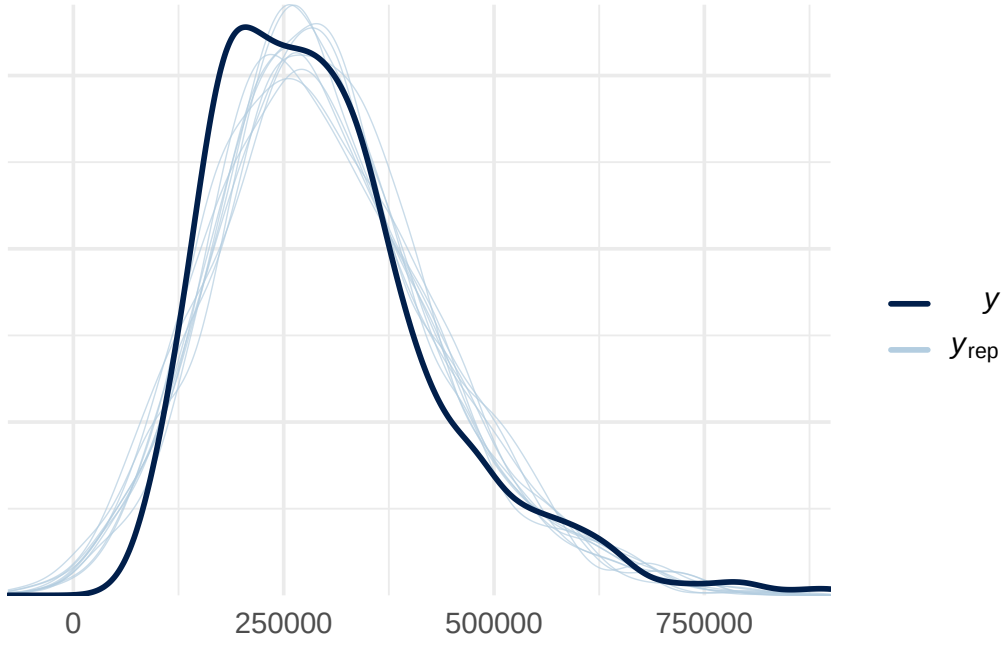

Figure 17: Posterior predictive check for the hierarchical model

5.4 Sensitivity analysis

Sensitivity analysis for the hierarchical model was skipped because the sensitivities of the parameters on the selected priors ought to match those of the linear model. In addition, there are too many parameters to render any visual presentation of the results a bore, and the hierarchical model doesn't seem to be that much better that there would be any utility in completing sensitivity analysis for it.

6 Model comparison

We start our model comparison by using leave-one-out cross-validation (LOO-CV). The better model has higher `elpd_loo` (better predictive model for observed data) and higher `p_loo` values (more complex model).

Below is the leave-one-out cross-validation for the non-hierarchical linear model by using the `loo()` function:

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

         Estimate   SE
elpd_loo -11447.9 33.7
p_loo         6.8  1.2
looic     22895.9 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.
```

Next, we apply the same `loo()` function for the hierarchical model:

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

         Estimate   SE
elpd_loo -11159.1 44.0
p_loo        56.0  6.7
looic     22318.3 87.9
------
Monte Carlo SE of elpd_loo is NA.

Pareto k diagnostic values:
                         Count Pct.    Min. n_eff
(-Inf, 0.5]   (good)     892   99.1%   547
 (0.5, 0.7]   (ok)         7    0.8%   243
   (0.7, 1]   (bad)        1    0.1%   89
   (1, Inf)   (very bad)   0    0.0%   <NA>
See help('pareto-k-diagnostic') for details.
```

`elpd_loo` values of two models are not too significantly different. Comparing the models via LOO-CV, we see that the hierarchical model isn't much better than the non-hierarchical one. `p_loo` values of the hierarchical model are significantly higher which means that it is much more complex than the non-hierarchical model.

```
     elpd_diff se_diff
fit2    0.0       0.0
fit1 -288.8      28.3
```

7 Discussion

The linear model gives a fairly good model of the Espoo house price data, but it doesn't capture some aspects of the distribution, especially the positivity of house price. The shortcomings of the linear model don't seem to be ameliorated by switching to a hierarchical model with the average income per postal area as the grouping variable, i.e., the non-linearity present in the aggregate data isn't explained by heterogeneity between postal areas' average income. As an experiment, a non-linear and non-hierarchical model could be tested to see if it gives better results.

8 Conclusion

Although the hierarchical model is more complex, it doesn't provide much added benefit in comparison to the non-hierarchical one. The linear model in itself is a fairly good model given that it is simple and fast to fit and adjust. However, a non-linear model would be a good experiment for further analysis and might give better results.

9 Self-reflection

We learned how to form a Bayesian analysis problem. We revised the non-hierarchical and hierarchical models covered in this course. We also learned how to manage time, 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.

11 Appendix