Evaluating predictive performance of Bayesian regression models

Max Hansson, Martin Holmqvist, Rashiq Al Tariq[1]

Abstract

This report presents a comprehensive analysis of Bayesian house price modeling, where both traditional regression models and hierarchical models have been implemented. Using a hierarchical model, the geographical differences between municipalities are captured and estimated, to see if it can help explain the complexities of the real estate market, and potentially improve the predictions of properties. The analysis utilizes data from Booli's research, with independent variables such as rent, living area, and municipality. By comparing the pooled regression model with different variations of hierarchical models, the results indicate that the hierarchical models give insight into the complex structure of the housing market and have a better predictive power according to leave one out cross-validation.

1. Introduction

The real estate market is dynamic and complex market which is influenced by many different factors. To be able to predict it effectively is of great interest to multiple groups in society. For example, homeowners, where buying a house or an apartment is often the biggest investment in their lives. Additionally, investors and economists are interested in these predictions to make profitable property investments and make good decisions regarding the housing market. From earlier research, we know that one of the biggest reasons why real estate with rather similar characteristics differ in price is that they are located differently (Kaas et al., 2024). With this argument, house price prediction models that do not account for this would predict worse than models that do. In a traditional regression setting, there is no way of expressing differences across different locations regarding coefficient slopes or intercepts. Hierarchical Bayesian models offer a reliable and suitable framework for handling these complexities. Using random slopes and intercepts, the model can account for variation between different levels of the data. We do, however, also know that macroeconomic variables such as interest rate, inflation, and unemployment rate also impact house prices (Apergis et al., 2003). Still, these variables fluctuate rather slowly, and one could assume these to have limited impact when looking at house sales within a short time span. Hence, the characteristics of a house and the location should in theory explain the majority of the variability of a house price.

In this report, we will be modeling a hierarchical structure between different municipalities using the Bayesian framework, implementing random slopes and random intercepts, which hopefully leads to more flexible predictions of the house prices compared to not taking the different locations into account. A pooled model that does not take the different municipalities into account will be the baseline model for comparison.

2. Data

The data set has been provided by Booli Search Technologies AB. It contains 176655 observations across multiple different municipalities across Sweden. It has 19 variables which include information about the size, location, rent, and other important characteristics regarding the house price. The data is hierarchical as it is measured across different locations across Sweden. These locations likely have significant implications on the cost of the properties, and therefore it makes sense to allow for differences among them. This implies that a Bayesian hierarchical approach would be appropriate for the data. Some of the data is removed from the data set, as they are unimportant for the analysis or because they have too much missing data. The houses in the data set were sold between 2019-2020. We will assume that the macroeconomic environment is rather constant and does not significantly affect the house prices during this time. A possible limitation of the data set is that it is gathered over a limited amount of time meaning it may not be generalizable for different time periods. Table 1 shows a summary of the variables we will use in the model and what they measure.

Due to computational limitations, we will only include 1500 observations from 3 different municipalities, these are "Eskilstuna", "Uppsala", and "Sundsvall". Figure 1 (in the appendix) shows the distribution of sold price and the square root of sold price in these locations. As we can see the distribution of the sold price is rather skewed and we will use the square root transformation since it is pending more towards a normal distribution. Figure 2 (in the appendix) illustrates the distribution in the price of sold houses between different municipalities. We see that there seems to be a noticeable difference in the sold house price between the groups. Because of this, it is of interest to model this difference to improve predictive ability.

Table 1. Variables used in the model where square root of Sold Price is the dependent variable and the parameters they will corresponds to in the models

VARIABLE AND COEFFICIENT	DESCRIPTION
LIVING AREA: β_1	SQUARE METERS OF THE APARTMENT
ROOMS: β_2	NUMBER OF ROOMS IN THE APARTMENT
FLOOR: β_3	THE FLOOR WHERE THE APARTMENT IS LOCATED
CONSTRUCTION YEAR: β_4	YEAR THE BUILDING WAS CONSTRUCTED
RENT: β_5	THE RENTAL PRICE OF THE APARTMENT
OPERATING COST: β_6	OPERATING COST OF THE APARTMENT
MUNICIPALITY j	MUNICIPALITY WHERE THE APARTMENT IS LOCATED
SOLD PRICE: y	PRICE AT WHICH THE APARTMENT WAS SOLD

3. Models and methods

As previously stated, the purpose of the analysis will be to examine the effect of several independent variables on our dependent variable ”sold price”. The observations of the dataset are segmented into different municipalities. Introducing this ”municipality” variable into our models will allow us to account for the hierarchical aspect of housing prices of geographical location.

The specific types of hierarchical Bayesian regression models we will use are random slopes and random intercept models. Random slopes and random intercept models allow for different coefficient estimates for a specified parameter based on ”hierarchical” variables such as municipality. Thus, the main objective is to employ three different regression models to explain and predict housing prices and compare them to establish the best model available.

Initially, a pooled regression model, without random slopes or intercepts, will be fit. This model assumes that all observations, regardless of group, share a common intercept and slope. This first model will serve as the baseline of comparison for the subsequent more complex models. The pooled model is described as

$$y_i = \alpha_0 + \boldsymbol{\beta X} + \epsilon_i \tag{1}$$
$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

The two other models to be fitted are a random intercepts model and a random slopes model. The random intercepts model is described as

$$y_{ij} = \alpha_j + \boldsymbol{\beta X} + \epsilon_{ij} \tag{2}$$
$$\text{where} \quad \alpha_j = \alpha_0 + z_j \times \sigma_\alpha \quad \text{and } z_j \sim \mathcal{N}(0,1)$$
$$\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

The random intercept model captures the variability in the mean of the outcome variable when all predictors are set to 0. The intercept α_j varies by group, with σ_α representing the group-specific deviation from the overall intercept α_0.

For the last model, we instead introduce random slopes to our regression. This aids in capturing the variability in the effect of the predictors on the outcome variable for each hierarchical group, as opposed to assuming equal predictor coefficients across groups.

$$y_{ij} = \alpha_0 + \beta_{1,j} \times x_{i,j,1} + \sum_{k=2}^{6} \beta_k x_{i,j,k} + \epsilon_{i,j} \tag{3}$$
$$\text{where } \beta_{1,j} = \beta_1 + \sigma_\beta \times z_{\beta_j} \text{ and } z_{\beta_j} \sim \mathcal{N}(0,1)$$
$$\sigma_{\beta_{1,j}} \sim \mathcal{N}(0, \theta^2), \quad \epsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$$

This model adds the random slope coefficient β_{1j} to the covariate living area. The specific covariates chosen for all the models are shown in table 1. For the random slope model the variable ”Living area” is given a random slope. This is because the effect of living area on house price presumably differs between different municipalities.

The priors for the model were chosen after examining the distribution of our outcome variable ”sold price”. Since we are now working with the square root of the sold price, the choice of priors will have to reflect that. The priors were chosen by first looking at the scale of the original predictors and assigning reasonable mean and standard deviation values relative to this scale. The final prior values were then taken as the square root of these values. Tables 8, 9 and 10 show the chosen prior distributions for each model (see table 1 to see what each parameter measure).

Each model will be run using 4 chains with 5,000 iterations in each chain. The warm-up length is set to be 2,500. The performance of the models will be evaluated using Leave-one-out cross-validation. The results from this evaluation will serve as an indication of which model has the best predictive capability.

4. Results

4.1. Pooled model

Table 2 and 3 present the summary statistics and LOO estimates respectively for the pooled model.

Table 2. Parameter Estimates for the Pooled Model

PARAMETER	MEAN	SE MEAN	SD	N_EFF	RHAT
ALPHA	972	0.21	17.21	6732.76	1.00
BETA[1]	9.7	0.01	0.49	5683.38	1.00
BETA[2]	121	0.11	9.30	6709.36	1.00
BETA[3]	10.2	0.02	2.34	9025.71	1.00
BETA[4]	-2.00	0.00	0.14	7520.33	1.00
BETA[5]	-0.18	0.00	0.01	7439.68	1.00
BETA[6]	0.29	0.00	0.02	8789.28	1.00
SIGMA_Y	239.7	0.03	2.53	9306.44	1.00

Table 3. LOO-CV Statistics for the Pooled Model

METRIC	ESTIMATE	SE
ELPD_LOO	-31048	54.39
P_LOO	11	1.18
LOOIC	62096	108.78

From the summary statistics, it can be seen that all of the chains for the specified parameters have converged, as all the $\hat{R}$ values are below 1.01. Examining the mean and standard errors, the parameter coefficients are also deemed as significantly separate from 0. From figure 3 we can see that all k-values from the LOO-CV estimation are below 0.7. Indicating that the results from the LOO estimation are reliable. The received $elpd_{loo}$ is -31048.23. All credible intervals for the pooled model coefficients are shown in Appendix.

The estimated effective number of parameters is less than the number of observations, however it is slightly higher than the true number of parameters in the model. This could mean that the given data is not sufficient to accurately predict house prices and that there might be other parameters of interest to consider to improve model fit.

From figure 6 (in the appendix) we can see that the replicated posterior distribution of house prices, for the pooled model, overall looks similar to that of the observed data. When ”sold price” is transformed the replicated distribution deviates from the observed data slightly to the left of the distribution and at the peak. For the original outcome variable, the replicated distribution deviates mostly at the peak. This suggests that the model predicts the given data well but not perfectly.

4.2. Random Intercepts model

Table 4 and 5 present the summary statistics and LOO estimates for the Random Intercepts model. All credible intervals for the random intercept model coefficients are shown in Appendix. From examining the summary statistics, there seems to be clear variability in the mean values for each

Table 4. Parameter Estimates for the Random intercept Model

PARAMETER	MEAN	SE MEAN	SD	N_EFF	RHAT
ALPHA	852	1.96	103.94	2820.85	1.00
BETA[1]	9.38	0.00	0.35	6279.67	1.00
BETA[2]	74.9	0.08	6.87	7438.40	1.00
BETA[3]	10.8	0.02	1.64	11161.97	1.00
BETA[4]	-0.81	0.00	0.10	10736.59	1.00
BETA[5]	-0.11	0.00	0.00	9143.18	1.00
BETA[6]	0.04	0.00	0.01	10106.38	1.00
SIGMA	169.9	0.02	1.80	12395.06	1.00
ALPHA_J[1]	-44	1.95	103.08	2794.17	1.00
ALPHA_J[2]	-187	1.96	103.12	2754.08	1.00
ALPHA_J[3]	261	1.96	103.15	2763.59	1.00

Table 5. LOO-CV Estimates for the Random Intercept model

PARAMETER	ESTIMATE	SE
ELPD_LOO	-29499	57.92
P_LOO	13	0.82
LOOIC	58999	115.83

intercept. This is an indication that the model captures the inherent difference between the baseline values of house prices between municipalities. The chains have converged here as well.

Figure 4 (in the appendix) depicts the estimated k-values, where all k-values fall below the threshold of 0.7. The LOO estimates can then be taken as reliable. The $elpd_{loo}$ is -29499. The same potential issue with a higher effective number of parameters arises here.

In figure 7 (in the appendix) we see the replicated posterior for the random intercept model. The replicated data closely aligns with the observed data, although there is a slight discrepancy around the peak, where the replicated data appears to have a slightly lower density. The model does a reasonably good job of replicating the main distribution. However, there are areas where the prediction could be improved.

4.3. Random Slopes model

Table 6 presents the summary statistics for the random slopes model. The $\hat{R}$ show that all of the chains for the estimation of the parameter estimates have converged. From figure 5 we see that all k-values are below 0.7 suggesting that the LOO estimates are reliable. From table 7 the $elpd_{loo}$ can be seen as -29896. The random slope model has the same issue with a higher effective number of parameter values.

In figure 8 we see the replicated posterior for the random slope model. This is by far the worst replication of the original data. There is a large lack of overlap in the peaks of the distributions.

Table 6. Parameter Estimates for the Random Slope Model

PARAMETER	MEAN	SE MEAN	SD	N_EFF	RHAT
ALPHA	894	0.17	13.50	5978.14	1.00
Z_BETA_J[1]	-0.1	0.01	0.58	3574.97	1.00
Z_BETA_J[2]	-0.6	0.01	0.68	3346.60	1.00
Z_BETA_J[3]	0.9	0.01	0.69	4078.24	1.00
BETA[1]	9.2	0.06	3.50	3103.97	1.00
BETA[2]	64.4	0.06	6.56	10632.54	1.00
BETA[3]	9.7	0.02	1.80	8843.88	1.00
BETA[4]	-1.00	0.00	0.11	6897.41	1.00
BETA[5]	-0.12	0.00	0.00	10445.53	1.00
BETA[6]	0.06	0.00	0.01	9326.04	1.00
SIGMA	185.6	0.02	1.97	11691.92	1.00
SIGMA_BETA	5.5	0.06	3.42	3621.55	1.00
BETA_J[1]	9.02	0.00	0.37	11797.52	1.00
BETA_J[2]	7.12	0.00	0.39	11441.59	1.00
BETA_J[3]	13.11	0.00	0.38	11687.80	1.00

Table 7. LOO-CV Statistics for the Full Random Slopes Model

METRIC	ESTIMATE	SE
ELPD_LOO	-29896	59.8
P_LOO	14	1.03
LOOIC	59793	119.59

4.4. High p-loo estimates

In all models, we can see p-loo $> p$, which could be caused by the model overfitting to the data or is misspecified in some sense. We believe the main reason for this is a potential multicollinearity problem among the predictors. The selected covariates were the ones we found to be most effective, but still, they do overlap somewhat in what they are measuring. For example rooms and living areas will be strongly correlated, potentially increasing the number of efficient parameters. However, since all Pareto k estimates are reasonably good (see figures 3 - 5 in appendix), there is no indication of badly misspecified models. Instead, the high p-loo values may indicate that there are potential improvements to the model, but not that they are completely misspecified. To increase the predictive performance of the models, more experimenting with priors and/or feature engineering may be appropriate.

4.5. Model evaluation

To determine which model has the best predictive capability the $elpd_{loo}$ estimates will be examined and compared with one another. The $elpd_{loo}$ is the Bayesian LOO estimate of the expected log pointwise predictive density and is the sum of pointwise log predictive densities. When evaluating model fit between models, a higher value of the $elpd_{loo}$ is desired. From the tables presented above the random intercepts model has the highest $elpd_{loo}$ value, followed by the random slopes model, and lastly the pooled model. These results imply that the random intercepts model has the best predictive performance, whilst the pooled model has the worst predictive performance. It would be interesting to also let these models give predictions on a completely new test data set and see how they would perform then, as is commonly done when evaluating models in addition to the validation comparison.

5. Conclusion

This paper compares different types of Bayesian hierarchical modeling approaches to analyze and determine the best Bayesian models for predicting house prices. The elpd-loo value suggests that by accounting for hierarchical structures of the data, the model can capture more nuanced relationships that would otherwise go unnoticed in a traditional pooled approach. However, looking at the replicated plots for the test data, the relationship between model complexity and performance is not as definitive. Here we see that the pooled model performs better than the random slopes models. The use of random intercept and random slopes gives a deeper understanding of how the house prices are related to the independent variables and can be useful knowledge when trying to understand the complexities that make up the price of a property. The models generally showed robust results, although some efficient number of parameter estimates were a bit higher than would have been preferred. This indicates that the model is a bit more complex than desired and that there could be higher-quality predictors, or better model structures to be implemented. In general, the analysis improves the understanding of the housing market dynamics and shows the value of more complex models than the pooled regression for this type of data. The largest difficulty met during the analysis was finding optimal priors based on the covariates available. Further experimenting with priors could yield improvements, although the priors used were the ones we found the most efficient. Additionally, some feature engineering or getting more data could also improve the model. We believe some of the large p-loo values could be due to multicollinearity among the predictors, which causes high variance among the predictor coefficients. The major weakness of the models we have created is the poor generalization of the models.

References

Apergis, N. et al. Housing prices and macroeconomic factors: prospects within the european monetary union. *International real estate review*, 6(1):63–74, 2003.

Kaas, L., Kocharkov, G., and Syrichas, N. *Understanding spatial house price dynamics in a housing boom.* CESifo Working Paper, 2024.

A. Appendix

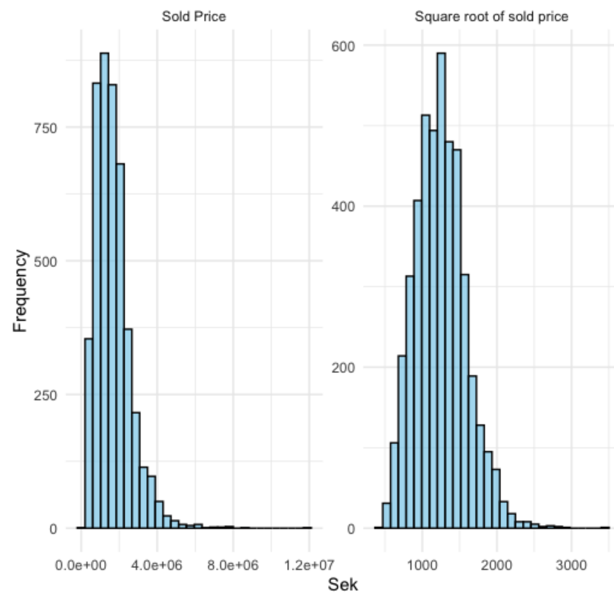

Figure 1. Distribution of sold price and square root of sold price

Table 8. Prior Distributions for the pooled model with adjusted standard deviations

PARAMETER	PRIOR DISTRIBUTION
α	$N(1000, 100)$
β_1	$N(0, 100)$
β_2	$N(100, 100)$
β_3	$N(0, 100)$
β_4	$N(0, 100)$
β_5	$N(0, 100)$
β_6	$N(100, 300)$
σ_y	$N(100, 300)$

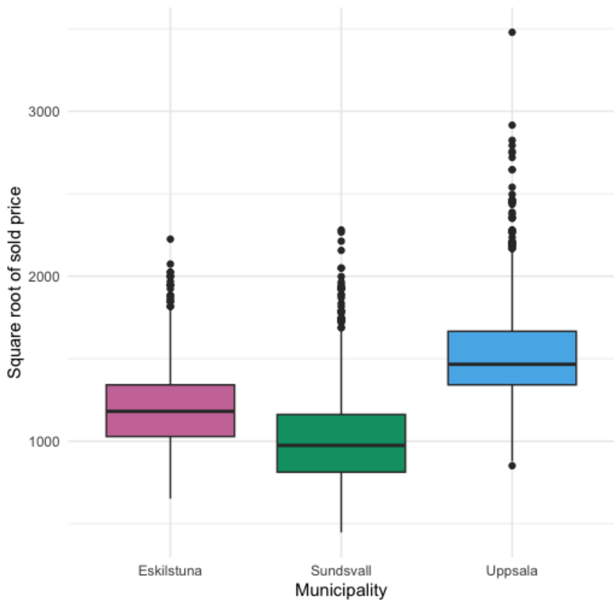

Figure 2. Distribution of square root of sold price in the 3 different municipalities

Table 9. Prior distributions for the random intercepts model

PARAMETER	PRIOR DISTRIBUTION
α	NORMAL$(0, 1000)$
$\beta 1$	NORMAL$(173.2, 100)$
β_2	NORMAL$(173.2, 100)$
β_3	NORMAL$(122.47, 100)$
β_4	NORMAL$(0, 31.6)$
β_5	NORMAL$(0, 31.6)$
β_6	NORMAL$(0, 31.6)$
$z\alpha_j$	NORMAL$(0, 1)$ (STANDARD NORMAL FOR RANDOM INTERCEPTS)
σ^α	NORMAL$(0, 100)$
σ	NORMAL$(0, 316)$

Table 10. Prior distributions for the random slope model

PARAMETER	PRIOR DISTRIBUTION
α	NORMAL$(1000, 100)$
β_1	NORMAL$(0, 15)$
β_2	NORMAL$(0, 15)$
β_3	NORMAL$(0, 15)$
β_4	NORMAL$(0, 10)$
β_5	NORMAL$(0, 10)$
β_6	NORMAL$(0, 10)$
z_{β_j}	NORMAL$(0, 1)$
σ_β	NORMAL$(0, 10)$
σ	NORMAL$(100, 500)$

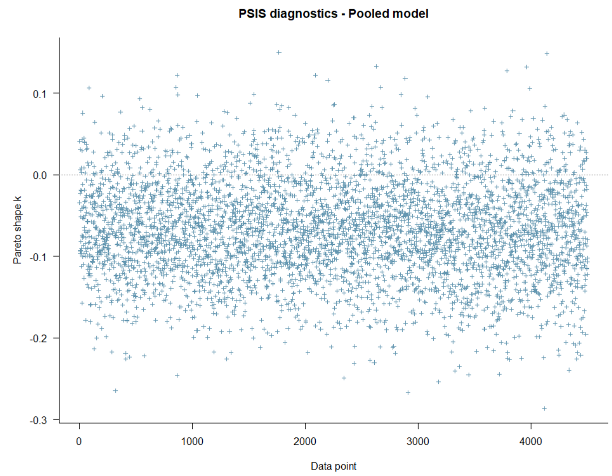

Figure 3. Estimated pareto k-values for the pooled model

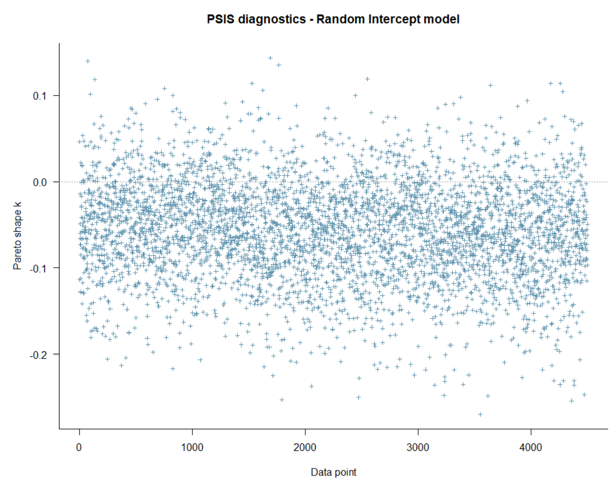

Figure 4. Estimated pareto k-values for the random intercept model

Table 11. 95 percent credible intervals for pooled model parameters

PARAMETER	2.5%	97.5%
α	955	1021
$\beta[1]$	9.5	11.5
$\beta[2]$	86.7	123.5
$\beta[3]$	5.3	14.3
$\beta[4]$	-2.3	-1.7
$\beta[5]$	-0.2	-0.2
$\beta[6]$	0.2	0.3

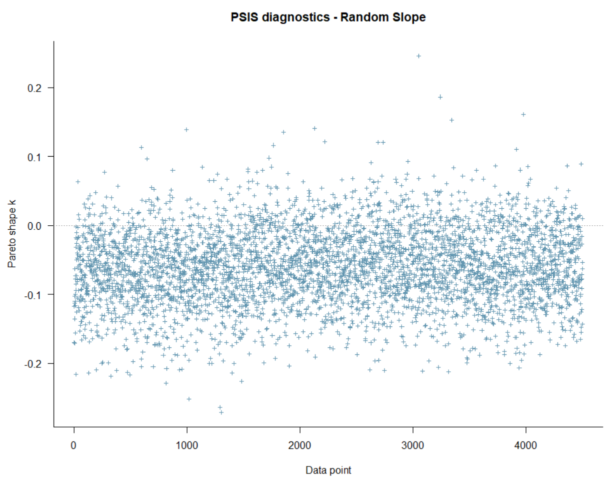

Figure 5. Estimated pareto k-values for the random slopes model

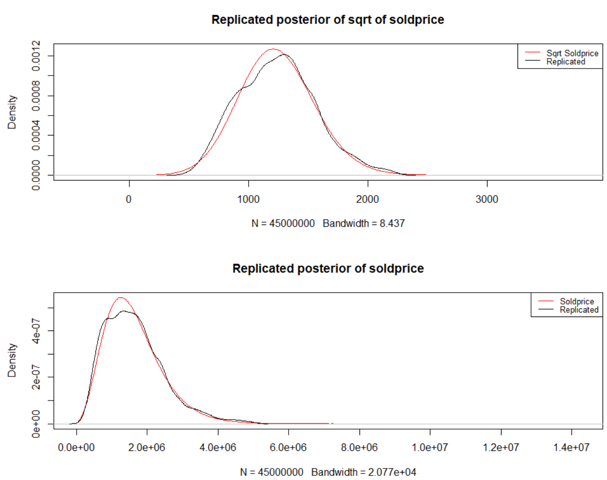

Figure 6. Replicated posterior using the pooled model

Table 12. 95 percent credible intervals for random intercept model parameters

PARAMETER	2.5%	97.5%
α	643	1050
$\beta[1]$	8.7	10.1
$\beta[2]$	61.2	88.4
$\beta[3]$	7.6	14.0
$\beta[4]$	-1.0	-0.6
$\beta[5]$	-0.12	-0.10
$\beta[6]$	0.0	0.1
$\alpha_j[1]$	-1.4	0.8
$\alpha_j[2]$	-2.5	0.1
$\alpha_j[3]$	0.3	3.0

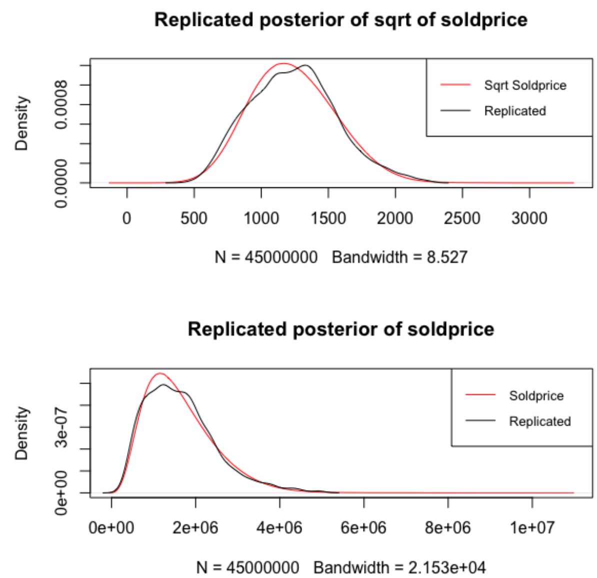

Figure 7. Replicated posterior using the random intercept model

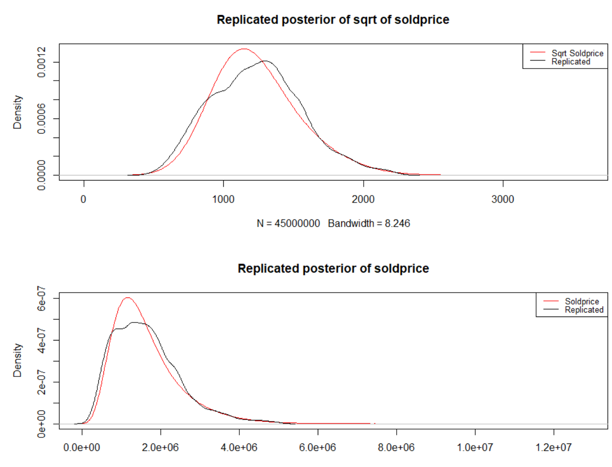

Figure 8. Replicated posterior using the random slope model

```
data {
  int<lower=0> N;
  matrix[N, 5] X;
  vector[N] y;
}
parameters {
  real alpha;
  vector[5] beta;
  real<lower=0> sigma_y;
}
transformed parameters {
  vector[N] y_hat;

  y_hat = alpha + X * beta;
}
model {
  alpha ~ normal(1000, 100);
  beta[1] ~ normal(0, 100);
  beta[2] ~ normal(100, 100);
  beta[3] ~ normal(0, 100);
  beta[4] ~ normal(0, 100);
  beta[5] ~ normal(0, 100);
  sigma_y ~ normal(100, 300);

  y ~ normal(y_hat, sigma_y);
}
generated quantities {
  vector[N] log_lik;
  vector[N] y_rep;

  for (n in 1:N) {
    log_lik[n] = normal_lpdf(y[n] | y_hat[n], sigma_y);
    y_rep[n] = normal_rng(alpha + X[n] * beta, sigma_y);
  }
}
```

Listing 1. Stan code for pooled model

Table 13. 95 percent credible intervals for random slope model parameters

PARAMETER	2.5%	97.5%
α	867	921
$\beta[1]$	1.2	15.7
$\beta[2]$	51.5	76.9
$\beta[3]$	6.3	13.3
$\beta[4]$	-1.2	-0.8
$\beta[5]$	-0.1	-0.1
$\beta[6]$	0.0	0.1
$\beta_j[1]$	8.3	9.7
$\beta_j[2]$	6.4	7.9
$\beta_j[3]$	12.4	13.8

```
data {
  int<lower=0> N;
  int<lower=1> J;
  int<lower=0> P;
  matrix[N, 6] X;
  int<lower=1, upper=J> municipality[N];
  vector[N] y;
}
parameters {
  real alpha;
  vector[J] z_alpha_j;
  vector[P] beta;
  real<lower=0> sigma;
  real<lower=0> sigma_alpha;
  vector<lower=0>[P] sigma_beta;
}
transformed parameters {
  vector[J] alpha_j = sigma_alpha * z_alpha_j;

  vector[N] y_hat;
  for (n in 1:N) {
    y_hat[n] = alpha + alpha_j[municipality[n]] + dot_product(beta, to_vector(X[n]));
  }
}
model {
  alpha ~ normal(0, sqrt(1000000));
  beta[1] ~ normal(sqrt(30000), sqrt(10000));
  beta[2] ~ normal(sqrt(30000), sqrt(10000));
  beta[3] ~ normal(sqrt(15000), sqrt(10000));
  beta[4] ~ normal(0, sqrt(1000));
  beta[5] ~ normal(0, sqrt(1000));
  beta[6] ~ normal(0, sqrt(1000));
  z_alpha_j ~ normal(0, 1);

  sigma_alpha ~ normal(0, sqrt(10000));
  sigma ~ normal(0, sqrt(100000));

  y ~ normal(y_hat, sigma);
}
generated quantities {
  vector[N] log_lik;
  vector[N] y_rep;

  for (n in 1:N) {
    log_lik[n] = normal_lpdf(y[n] | y_hat[n], sigma);
    y_rep[n] = normal_rng(y_hat[n], sigma);
  }
}
```

Listing 2. Stan code for random intercept model

```
data {
  int<lower=0> N;
  int<lower=1> J;
  int<lower=0> P;
  matrix[N, P] X;
  int<lower=1, upper=J> municipality[N];
  vector[N] y;
}

parameters {
  real alpha;
  vector[J] z_beta_j;
  vector[P] beta;
  real<lower=0> sigma;
  real<lower=0> sigma_beta;
}

transformed parameters {
  vector[N] y_hat;
  vector[J] beta_j;


  for (j in 1:J) {
    beta_j[j] = beta[1] + sigma_beta * z_beta_j[j];
  }

  for (n in 1:N) {
    y_hat[n] = alpha + beta_j[municipality[n]] * X[n, 1] + dot_product(beta[2:P],
        to_vector(X[n, 2:P]));
  }
}

model {

  alpha ~ normal(1000, 100);
  beta[1] ~ normal(0, 15);
  beta[2] ~ normal(0, 15);
  beta[3] ~ normal(0, 15);
  beta[4] ~ normal(0, 10);
  beta[5] ~ normal(0, 10);
  beta[6] ~ normal(0, 10);
  sigma_beta ~ normal(0, 10);
  sigma ~ normal(100, 500);
  z_beta_j ~ normal(0,1);

  y ~ normal(y_hat, sigma);
}


generated quantities {
  vector[N] log_lik;
  vector[N] y_rep;



  for (n in 1:N) {
    log_lik[n] = normal_lpdf(y[n] | y_hat[n], sigma);
    y_rep[n] = normal_rng(y_hat[n], sigma);
  }
}
```

Listing 3. Stan code for random slopes model