Implementing a BYM Model in Stan to Fit Boston Housing Price Data

Marcel Gietzmann-Sanders

STAT641 - Bayesian Statistics
University of Alaska Fairbanks

Contents

1 Introduction

The question - what drives property values - is one that new homeowners and city planners alike have been asking for ages. We will be looking back in time at a 1978 survey of housing prices and their possible covariates in Boston, Massachusetts (David and Rubinfeld (1978)). This data was made accessible through the `spData` package in R (Bivand (2022)) and contains median house prices and potential covariates for 506 tracts in the Greater Boston Area.

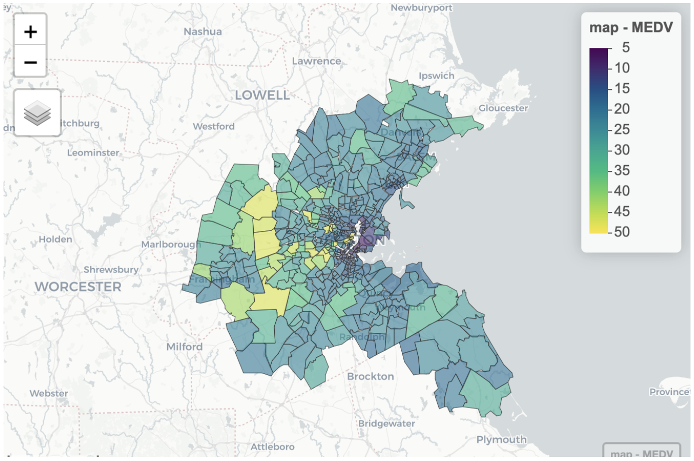

Figure 1: Median Housing Price

The median house value (in $1000USD) by census tract in the Greater Boston Area in 1978.

Fig. 1 shows the median price per tract of land. It should be understood that this data is censored and median values over $50,000 are capped. A full list of covariates can be found in the documentation for the `boston` dataset in `spData`. Our question is simple - which of these covariates are indeed related to housing prices after spatial structure is accounted for.

1.1 Candidate Covariates

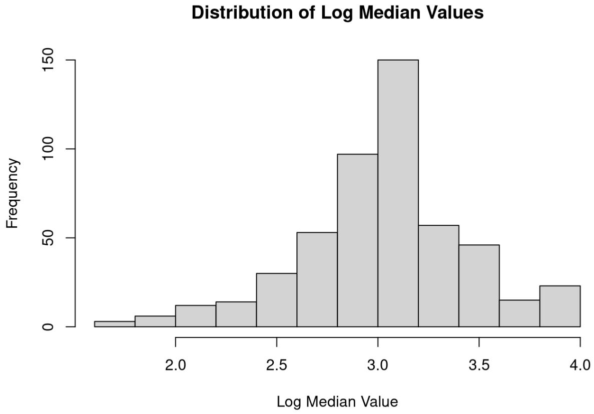

Figure 2: Log Median Housing Price Distribution

As the distribution of values is right skewed (and following in the footsteps of (Moraga (2023))) we took as our target variable the logarithm of the median house value per tract instead of the median value itself. Fig. 2 shows the distribution of our target.

After performing exploratory data analysis in `R` (R Core Team (2024)), four covariates of particular interest were identified - crime rates per capita, average number of rooms per dwelling, weighted distance to employment centers, and the nitric oxide concentrations per town.

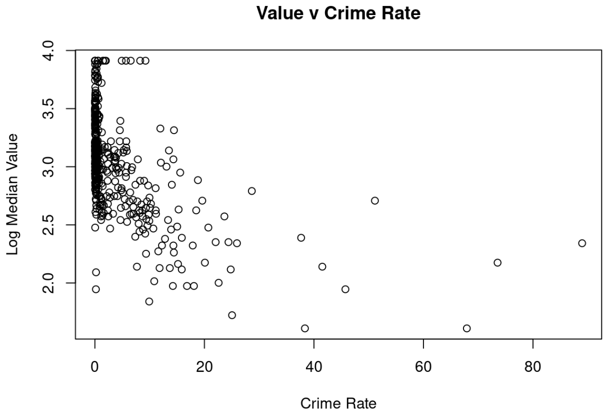

Figure 3: Log Median Price vs Crime Rate

Fig. 3 shows the relationship between per capita crime rate and our target. There appears to be a clear negative relationship between then two and a rather wide spread of values in and around 0.

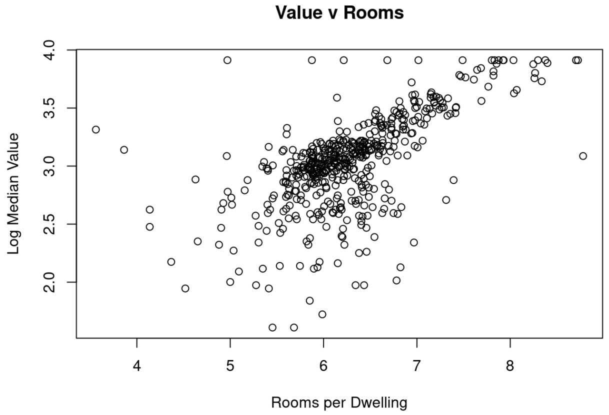

Figure 4: Log Median Price vs Rooms per Dwelling

Fig. 4 shows how our target varies with the average number of rooms per dwelling in each tract. Here we can see a strong positive relationship between the two which makes a great deal of sense.

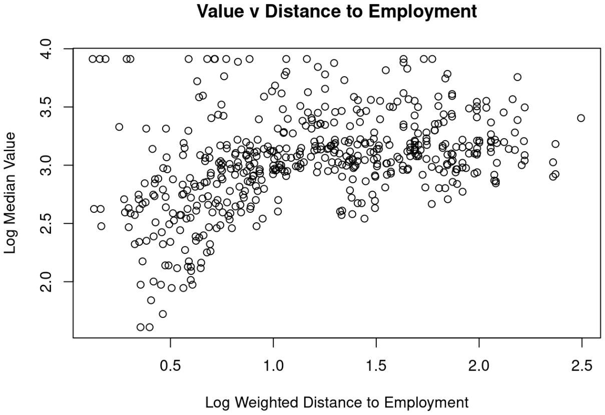

Figure 5: Log Median Price vs Log Weighted Distance to Employment Centers

In Fig. 5 we see what looks to be a similar positive relationship between the logarithm of the average weighted distance to employment centers and our target. However this relationship is not as clear or strong as the one between our target and the number of rooms per dwelling.

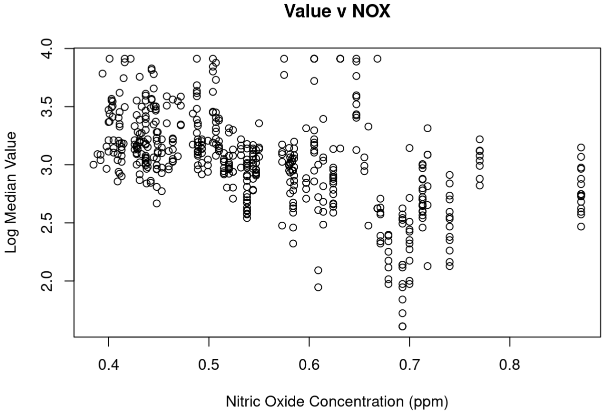

Figure 6: Log Median Price vs NOX

Finally Fig. 6 shows us the relationship between nitric oxide concentrations in parts per million and our target with a somewhat spurious negative relationship.

1.2 Modeling with Spatial Structure

We will be using a variant of the Besag-York-Mollié (BYM) model (Moraga (2023))(Mitzi Morris (2019)). In this model we assume that we have an observations of our target variable Y_i (log median housing value in our case) for each tract i. Furthermore we assume our Y_i can be modeled as a normal distribution:

$$Y_i \sim Normal(\mu_i, \sigma^2)$$

where it is the μ_i that will be a function by our covariates. Specifically:

$$\mu_i = \beta_0 + \vec{\beta}\vec{x_i} + \sigma_r \left(\sqrt{\rho}\phi_i + \sqrt{1-\rho}\theta_i\right)$$

where β_0 is our intercept, $\vec{\beta}$ are the coefficients for our models effects from each covariate, x_i are our covariates corresponding to tract i, ϕ_i and θ_i are spatial and random effects respectively, and σ_r and ρ allow us to control the effect of the random variables as well as the degree to which our model has spatial and/or unstructured noise (if $\rho = 1$ we have only spatial structure whereas if $\rho = 0$ it is totally unstructured).

Of especial interest here is the ϕ_i as they determine our spatial structure.

Each spatial interaction term ϕ_i is modeled as conditional on the other terms:

$$\phi_i | \phi_j \sim N\left(\sum_j w_{ij}\phi_j, \sigma^2\right), i \neq j$$

which defines a conditional autoregressive model (CAR). A key result that Besag proved (Julian (1974)) is that the joint distribution $\boldsymbol{\phi}$ ends up being multivariate normal random variable centered at 0

$$\vec{\phi} \sim N(0, Q^{-1})$$

where $Q = D(I - \alpha A)$. D is a diagonal "neighbors" matrix (each element on the diagonal is the number of neighbors unit i has), A is an adjacency matrix where if i, j are neighbors then the i, j element is 1, and α lets us control spatial dependence. This results in a log probability density of $\vec{\phi}$ which is proportional to (Mitzi Morris (2019)):

$$\frac{n}{2}\log(\det Q) - \frac{1}{2}\vec{\phi}^T Q \vec{\phi}$$

Given $\det Q$ is a constant and MCMC samplers compute the log probability up to a proportionality constant (Mitzi Morris (2019)), the first term drops out of the computation thereby reducing the computational intensity of this evaluation.

In our case, as we will be following the stan implementation from the paper (Mitzi Morris (2019)) we will be setting $\alpha = 1$ and thereby getting an intrinsic conditional autoregressive model (ICAR) in which Q reduces to $D - A$.

With an ICAR model each ϕ_i is distributed with a mean equal to the average of its neighbors (Mitzi Morris (2019)). If we additionally assume that $\vec{\phi}$ is centered at zero with common variance 1, then the joint probability of $\vec{\phi}$ becomes (Mitzi Morris (2019)):

$$p(\vec{\phi}) \propto \exp\left(-\frac{1}{2}\sum_{i\sim j}(\phi_i - \phi_j)^2\right)$$

where $i \sim j$ indicates that i and j are neighbors. This then is the prior for the ϕ_i - an ICAR model centered at 0 with common variance 1.

2 Methods

2.1 Formal Model Definition

As noted in section 1.2, we used a Besag-York-Mollié (BYM) model (Moraga (2023))(Mitzi Morris (2019)) with an ICAR prior. We restate the model here for clarity:

$$Y_i \sim Normal(\mu_i, \sigma^2)$$

$$\mu_i = \beta_0 + \vec{\beta}\vec{x_i} + \sigma_r \left(\sqrt{\rho}\phi_i + \sqrt{1-\rho}\theta_i \right)$$

$$p(\vec{\phi}) \propto \exp\left(-\frac{1}{2} \sum_{i \sim j} (\phi_i - \phi_j)^2 \right)$$

The priors selected for the other components in our model are:

$$\beta_0 \sim Normal(0, 1)$$
$$\beta_i \sim Normal(0, 1)$$
$$\theta_i \sim Normal(0, 1)$$
$$\sigma_r \sim Uniform(0, 1)$$
$$\rho = \frac{e^r}{1+e^r}, r \sim Normal(0, 1)$$

In the initial experimentation with this model we also had the prior:

$$\sigma \sim Normal(0, 1)$$

but consistently found $\sigma \approx 0$. Therefore going forward we simply assume $\sigma = 0.01$ in order to not over complicate the sampling.

2.2 Stan Model Definition

```
functions {
    real icar_normal_lpdf(vector phi, int N, int[] node1, int[] node2) {
        return -0.5 * dot_self(phi[node1] - phi[node2])
            + normal_lpdf(sum(phi) | 0, 0.001 * N);
    }
}
data {
    int<lower=0> N; // number of tracts
    int<lower=0> N_edges; // number of unique edges
    int<lower=1, upper=N> node1[N_edges]; // start of edge
    int<lower=1, upper=N> node2[N_edges]; // end of edge
    int<lower=1> K; // number of covariates
    matrix[N, K] x; // design matrix
    real y[N]; // target
}
parameters {
    real beta0;
    vector[K] betas;

    real logit_rho;

    vector[N] phi;
    vector[N] theta;

    real<lower=0> sigma_r;
}
transformed parameters {
    real<lower=0, upper=1> rho = inv_logit(logit_rho);
    vector[N] convolved_re = sqrt(rho) * phi
                                + sqrt(1 - rho) * theta;
}
model {
    y ~ normal(beta0 + x * betas + convolved_re * sigma_r, 0.01);
    target += icar_normal_lpdf(phi | N, node1, node2);
    beta0 ~ normal(0, 1);
    betas ~ normal(0, 1);
    logit_rho ~ normal(0, 1);
    theta ~ normal(0, 1);
    sigma_r ~ uniform(0, 1);
}
```

Note the term `normal_lpdf(sum(phi)|0,0.001*N)` used to center $\vec{\phi}$ at zero (Mitzi Morris (2019)).

2.3 Fitting the Model

We took advantage of the following `R` packages to transform the data into the appropriate format - `spData` (Bivand (2022)), `sf` (Pebesma and Bivand (2023a) Pebesma (2018)), `spdep` (Bivand and Wong (2018) Roger Bivand (2022) Bivand et al. (2013) Pebesma and Bivand (2023b)).

```r
library(sf)
library(spData)
library(spdep)

map <- st_read(
    system.file("shapes/boston_tracts.shp", package = "spData"),
    quiet = TRUE
)
map$log_median_value <- log(map$MEDV)

# build neighbors arrays
nb <- poly2nb(map)
N = length(map$MEDV)
node1 = c()
node2 = c()
for (i in 1:N) {
    for (j in nb[[i]]) {
        if (j > i) {
            node1 = c(node1, i)
            node2 = c(node2, j)
        }
    }
}
N_edges = length(node1)

# build target and covariates
y = map$log_median_value
x = cbind(map$CRIM, map$RM, log(map$DIS), map$NOX)
K = dim(x)[2]
```

Then, after computing the appropriate features, we fit our model using `rstan` (Stan Development Team (2024)). Each model was fit with 4 chains and 20,000 iterations per chain. 20,000 was selected to ensure each parameter of interest had at least 1,000 effective samples. Two models were fit, one with the crime rates feature and one without (see section 3).

3 Results

3.1 Model 1 - Full Covariates

We began with a model that included all of our covariates of interest.

Parameter	95% CI	Mean	SD	SE	$\hat{R}$	n_{EFF}
β_0	2.127, 2.394	2.260	0.197	0.002	1.000	7584
β_{CRIM}	-0.009, -0.007	-0.008	0.001	≈ 0	1.000	40517
β_{RM}	0.237, 0.255	0.246	0.014	≈ 0	1.000	27304
$\beta_{\log(DIS)}$	-0.166, -0.054	-0.111	0.083	0.002	1.002	1635
β_{NOX}	-1.236, -0.970	-1.102	0.196	0.002	1.000	11994
σ_r	0.342, 0.363	0.352	0.016	≈ 0	1.000	7770
ρ	0.941, 0.966	0.952	0.019	≈ 0	1.001	3800

With our 20,000 iterations per chain we were able to achieve an effective sample size ≥ 1000 for each of the parameters above as well as suitably small $\hat{R}$ and standard error values.

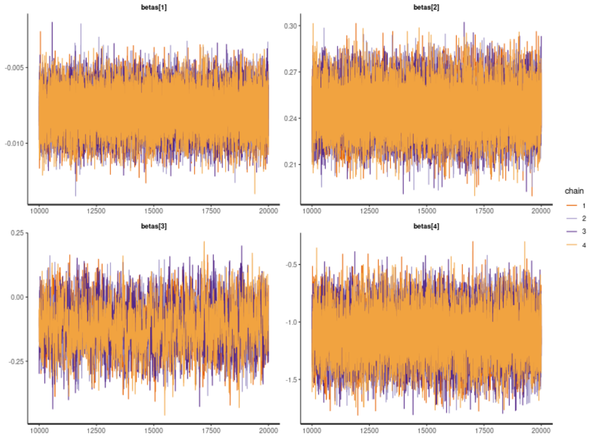

Figure 7: $\vec{\beta}$ Traceplots for 4 Feature Model (The order of features was crime, rooms, distance, and nox.)

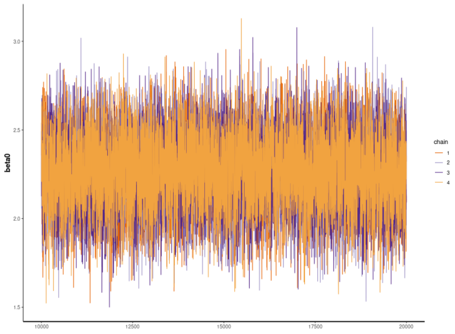

Figure 8: β_0 Traceplot for 4 Feature Model

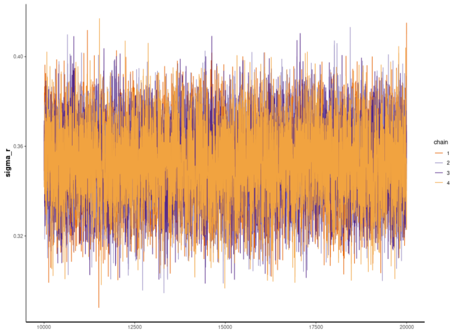

Figure 9: σ_r Traceplot for 4 Feature Model

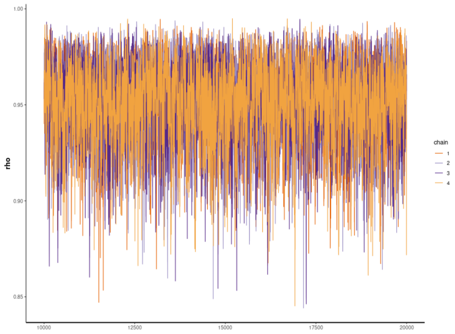

Figure 10: ρ Traceplot for 4 Feature Model

Our traceplots for these features visually do not seem to indicate issues with convergence (7-10).

3.2 Model 2 - No Crime Covariate

However given, the very small value on β_{CRIM} we also trained a model without that feature present.

Parameter	95% CI	Mean	SD	SE	$\hat{R}$	n_{EFF}
β_0	2.040, 2.312	2.176	0.202	0.002	1.000	6716
β_{RM}	0.237, 0.257	0.247	0.015	≈ 0	1.000	20504
$\beta_{\log(DIS)}$	-0.116, 0.002	-0.057	0.088	0.002	1.002	1350
β_{NOX}	-1.267, -0.992	-1.130	0.204	0.002	1.001	8552
σ_r	0.361, 0.382	0.371	0.016	≈ 0	1.001	7265
ρ	0.953, 0.973	0.962	0.016	≈ 0	1.001	2997

As with the model before, we were able to achieve an effective sample size ≥ 1000 for each of the parameters above as well as suitably small $\hat{R}$ and standard error values.

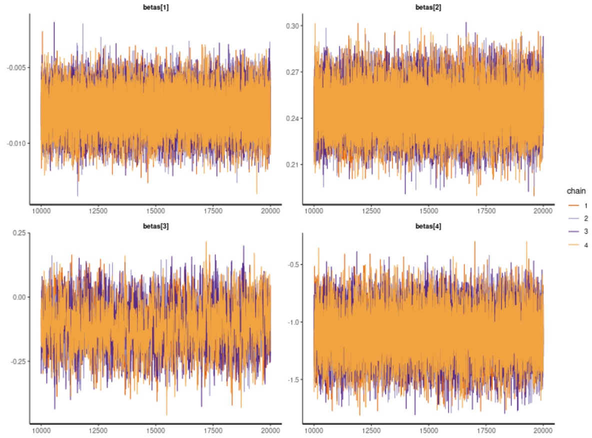

Figure 11: $\vec{\beta}$ Traceplots for 3 Feature Model (The order of features rooms, distance, and nox)

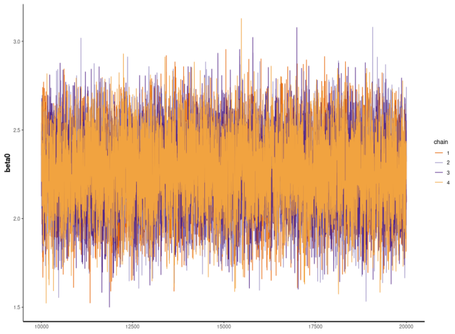

Figure 12: β_0 Traceplot for 3 Feature Model

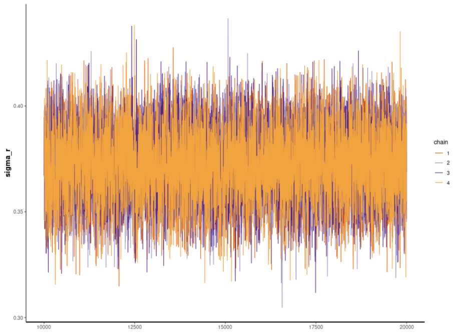

Figure 13: σ_r Traceplot for 3 Feature Model

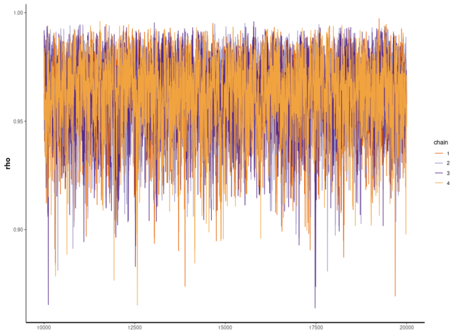

Figure 14: ρ Traceplot for 3 Feature Model

Our traceplots for these features also visually do not seem to indicate issues with convergence (11-14).

4 Discussion

Our results show that, assuming this model, we have a positive relationship between number of rooms per dwelling and our median housing price and a negative relationship between nitric oxide levels and housing prices just as we expected. However the strength of the relationship between nitric oxide levels and housing prices is quite surprising. Also surprising as it is contrary to our initial expectations We is the fact there is, in our model, a negative relationship between the distance to centers of employment and median housing prices per tract ($\beta_{\log DIS}$). Finally it is interesting to note just how inconsequential β_{CRIM} was when we took into account the other features and the spatial correlations among tracts (through the ϕ_i).

Given the relatively small value of $\beta_{\log DIS}$ and the fact that its 95% credible interval crosses 0, we also looked at the bayes factor for the hypothesis that $\beta_{\log DIS} < 0$. The posterior odds were found to be ≈ 2.92 and given the prior was a normal distribution centered at 0 our prior odds are simply 1 giving us a bayes factor of 2.92. Therefore we are just shy of weak evidence for this parameter being non-zero.

Another curious result was the fact that ρ ended up being very nearly one meaning our noise was almost entirely spatially structured.

All in all then, it seems like our two strongest determinants of housing price are the number of rooms per dwelling and the levels of pollution as indicated by nitric oxide levels. Beyond that we also found that there was significant spatial structure in the median housing prices. All of these conclusions seem reasonable given the context.

5 Bibliography

References

Bivand, Roger, J. N.-R. L. (2022). spdata: Datasets for spatial analysis.

Bivand, R. and Wong, D. W. S. (2018). Comparing implementations of global and local indicators of spatial association. *TEST*, 27(3):716–748.

Bivand, R. S., Pebesma, E., and Gómez-Rubio, V. (2013). *Applied spatial data analysis with R, Second edition.* Springer, NY.

David, H. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management.*

Julian, B. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society.*

Mitzi Morris, Katherine Wheeler-Martin, D. S. S. J. M. A. G. C. D. (2019). Bayesian hierarchical spatial models: Implementing the besag york mollié model in stan. *Spat Spatiotemporal Epidemiol.*

Moraga, P. (2023). *Spatial Statistics for Data Science: Theory and Practice with R.* American Fisheries Society.

Pebesma, E. (2018). Simple Features for R: Standardized Support for Spatial Vector Data. *The R Journal*, 10(1):439–446.

Pebesma, E. and Bivand, R. (2023a). *Spatial Data Science: With applications in R.* Chapman and Hall/CRC.

Pebesma, E. and Bivand, R. S. (2023b). *Spatial Data Science With Applications in R.*

R Core Team (2024). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Roger Bivand (2022). R packages for analyzing spatial data: A comparative case study with areal data. *Geographical Analysis*, 54(3):488–518.

Stan Development Team (2024). *RStan: the R interface to Stan.*

6 Density Plots

6.1 Model 1 - Full Covariates

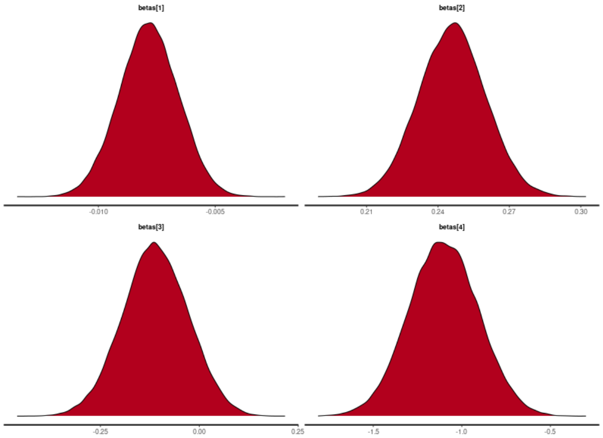

Figure 15: $\vec{\beta}$ Densities for 4 Feature Model

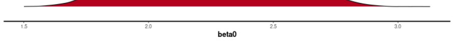

Figure 16: β_0 Density for 4 Feature Model

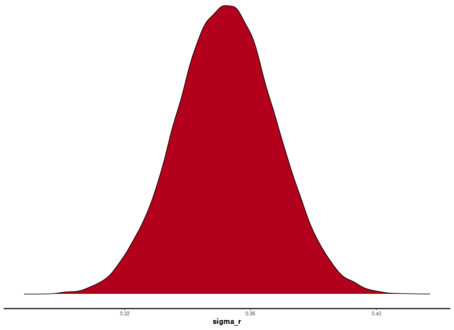

Figure 17: σ_r Density for 4 Feature Model

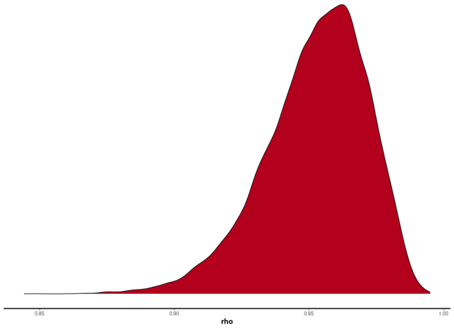

Figure 18: ρ Density for 4 Feature Model

See section 3.1 for summary statistics.

6.2 Model 2 - No Crime Covariate

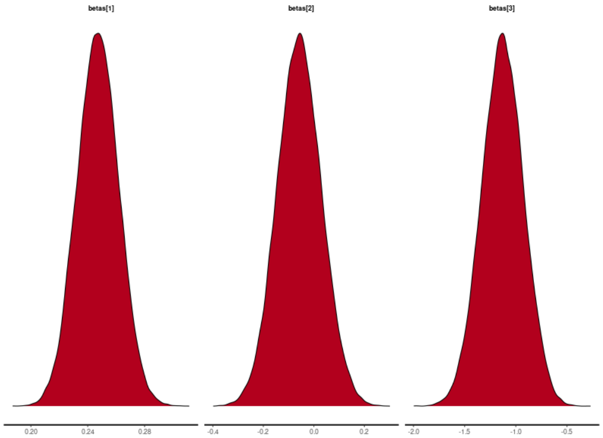

Figure 19: $\vec{\beta}$ Densities for 3 Feature Model

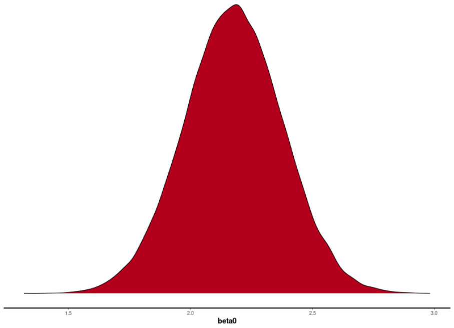

Figure 20: β_0 Density for 3 Feature Model

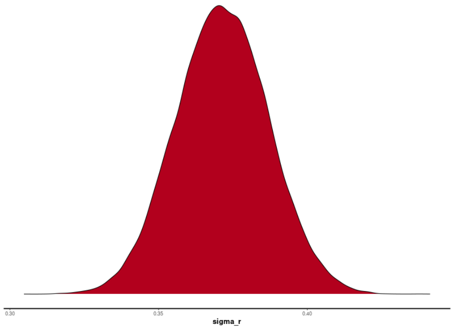

Figure 21: σ_r Density for 3 Feature Model

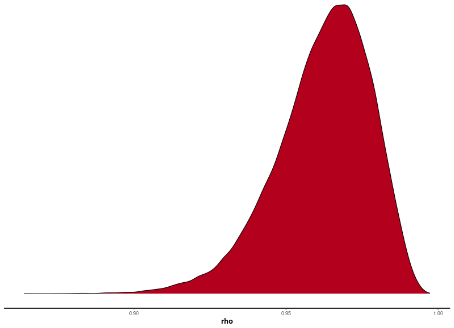

Figure 22: ρ Density for 3 Feature Model

See section 3.2 for summary statistics.