Final Assignment Report – Bayesian Statistics

Annemarie Timmers (6238106) – a.g.j.timmers@uu.nl

Research question

Houses are expensive. Housing prediction models were therefore developed to help people who plan on buying a house in the future plan their finances well. Moreover, they benefit property investors by predicting the trend of prices in a specific location (Alfiyatin et al., 2017). In hedonic pricing models, housing attributes are combined into a function to determine the price of a house (Atkinson & Crocker, 1987). These attributes can generally be divided into two categories: non-geographical and geographical factors (Gao et al., 2019). Non-geographical factors are characteristics of the house itself, such as the number of rooms or the square feet of the living space. Geographical factors are characteristics of the neighborhood in which the house is located, being the pupil teacher ratio or the distance to the city center. Dubin and Sung (1990) further divided the geographical factors into socioeconomic status of the neighborhood residents, quality of the municipal services and racial composition. They found that the socioeconomic and racial composition of the neighborhood was more important than the quality of the municipal services offered.

I am interested in how the relative importance of socioeconomic status in determining the price of a house relates to the relative importance of non-geographical factors. More specifically, I want to know whether the socioeconomic status of the neighborhood residents also trumps the relative importance of non-geographical factors. Given the small size of this project, I will limit the number of non-geographical factors that I examine to one. I will consider the research question: Is the socioeconomic status of the neighborhood residents more important than the number of rooms in a house in determining the price of a house?

Boston Housing dataset

The Boston Housing dataset was used to answer the research question. This dataset was first published by Harrison and Rubinfield (1978). It was collected by the U.S. Census Service in 1970 in the Boston Standard Metropolitan Statistical Area and contains 506 census tracts (geographical areas). The dependent variable was *medv*, the median value of the housing prices in the respective census tracts, hereafter referred to as just housing price. The independent variables were *rm*, the average number of rooms per house in the tracts, and *lstat*, the percentage of the population of the tracts that is of a lower socioeconomic status. For these variables, the descriptive statistics and bivariate associations are presented in Table 1 and 2, respectively.

Table 1

Descriptive statistics of the relevant variables

Variable	Mean	SD	Median	Min.	Max.	N
medv	22.53	9.20	21.20	5.00	50.00	506
rm	6.28	0.70	6.21	3.56	8.78	506
lstat	12.65	7.14	11.36	1.73	37.97	506

Note: medv is in $1000's and lstat is a percentage.

Table 2

Bivariate associations between the relevant variables

Variable	medv	rm	lstat
medv	1.000		
rm	0.695	1.000	
lstat	-0.738	-0.614	1.000

Estimation

In order to get estimates of the effect of the number of rooms and the percentage of people with a lower socioeconomic status on the price of a house, a joint posterior distribution needs to be obtained. This can be derived with the use of a prior distribution for the parameters and the density of the data. The density of the housing price is assumed to be normally distributed with mean μ and standard deviation σ^2, where the mean is a function of the predictors. However, since the model contains multiple parameters, it is difficult to sample from the joint posterior directly. Therefore, the joint posterior can be approximated by means of Markov Chain Monte Carlo (MCMC) algorithm instead. I will use a combination of the Gibbs and Metropolis-Hastings (M-H) sampling algorithms to obtain the (proportional) conditional posterior distributions of the parameters in the model.

Gibbs sampling

Gibbs sampling was used to derive conditional posteriors for the parameters in the model employing conjugate priors where the posterior distribution of the parameters looks like the prior. Therefore, a normal distribution was chosen for the regression parameters and the variance was estimated using an inverse Gamma prior distribution. Moreover, the hyperparameters of the prior were chosen to be vague, hereby influencing the posterior distribution to a minimal extent. I chose to do this, because I have a very limited understanding of the housing market and do not feel confident to quantify the influence of either of the predictors on the housing market in Boston in 1970.

Arbitrary initial values were chosen to start the MCMC algorithm as these should not affect the outcome when the burn-in period is removed. Then, values were sampled iteratively from the conditional posterior distributions given the current values of the other parameters in the model. The number of iterations was set to 10,000 after which the first 1000 draws were discarded as the burn-in period, which allows for the removal of the dependency on the initial values.

As can be seen in Table 2, the predictors were negatively correlated (r = -0.614). Because this strong correlation could cause convergence problems, both variables were centered before they were used in the sampler.

The Metropolis-Hastings algorithm

The M-H algorithm is a form of accept-reject sampling that can be used when the normalizing constant cannot be found or the type of conditional posterior cannot be recognized. In this case, I chose to apply M-H algorithm to estimate the effect of the number of rooms, because of its limited range. Using a non-standardized t-distribution as prior then allows for more uncertainty in the tails of the distribution, which would highlight a more diverse range of houses. To allow for the maximum amount of uncertainty, the degrees of freedom were set to one.

The M-H algorithm derives a function that is proportional to the conditional posterior which is then used to sample from a proposal distribution. First, a value was drawn from a proposal distribution. I chose a normal distribution with the current value of the parameter as its mean. This choice made it a dependent M-H sampler as the proposed value depended on the current value of the parameter. The standard deviation was tuned to 1, providing an acceptance rate between 0.20 and 0.50. This acceptance rate ensured that the steps taken over the posterior distribution by the proposal were not too small and that the autocorrelations was not too high, saving time.

Then, the acceptance probability was computed by dividing the prior density of the proposed value by the prior density of the current value and multiplying that by the current value divided by the proposed value. Lastly, the prosed value was accepted when the acceptance probability was higher than a random draw from the uniform distribution. Otherwise, the current value was retained.

Convergence

Two chains were run and multiple convergence measures were used to check whether the sampled parameters had converted to a stable constant distribution. This was needed as it is not possible to decisively conclude that a parameter converged, only that it has not converged. So, convergence was assumed when these measures did not detect nonconvergence in any of the parameters.

First, trace plots were made that showed the parameter estimates over the iterations without the burn-in period. All trace plots looked like fat, hairy caterpillars, which indicated stability. Because the chains overlapped, this meant that they were not stuck at a local maximum. Additionally, autocorrelation plots were made to investigate the dependence of the sampled values on previously sampled values. These plots also indicated convergence as the autocorrelations quickly went down after a couple of lags, which implied that the consecutive sampled values were not highly correlated with each other. Moreover, density and running mean plots were made for the two chains to investigate their overlap. When density plots overlap and running means end up closely together, this suggests that both chains sampled from the same posterior distribution, which is the desired outcome. Since no abnormalities were found here, these plots combined also indicated that the algorithm had converged for all parameters.

Next to the plots, Gelman-Rubin statistics and Monte Carlo (MC) errors were computed. The Gelman-Rubin static compares the variances within the chains to the variance between the chains and indicates convergence when it is close to one. The MC errors were calculated by dividing the posterior standard deviation by the square root of the number of iterations and should not be larger than five percent. As the results for both convergence measures were in line with what was required, it was assumed that convergence was reached. Subsequently, the chains were combined to extract the point estimates from the sampled data, which are presented in Table 3.

Interpretation of estimates and intervals

The 95% central credible intervals were used to interpret the results and consequently form an answer to the research question. A 95% central credible interval is the range of the posterior distribution that contains 95% of the values. Thus, when there is no zero in there, the predictor probably affects the outcome to some extent. Therefore, I believe that both the number of rooms and the percentage of people with a lower socioeconomic status in the tracts predict the housing price (*EAP* = 5.040 [4.171 – 5.905] and *EAP* = -0.646 [-0.732 – -0.561], respectively). If a house has one room above the average, the house price increases with 5040 dollars, controlling for the percentage of people with a lower socioeconomic status. In contrast, the price of a house decreases with 646 dollars as the percentage of people with a lower socioeconomic status increases with one percent above the average, controlling for the number of rooms. Comparing the two indicators, I believe that the average socioeconomic status of the neighborhood is more important in determining the housing price than the number of rooms a house has, given that its standardized posterior mean is larger in size.

Table 3

Point estimates extracted from the sampled data under Model 1

	EAP	Beta	PSD	Naïve.se	95% CCI
(Intercept)	22.528		0.245	0.002	[22.048 – 23.004]
rm	5.040	0.385	0.443	0.004	[4.171 – 5.905]
lstat	-0.646	-0.501	0.043	0.000	[-0.732 – -0.561]

Note: all predictors were mean-centered.

Bayes Factor

The relative support for a hypothesis can be quantified by means of the Bayes factor. The Bayes factor is the ratio of two marginal likelihoods and can be interpreted as how much more likely one hypothesis is over the other, given the data. A Bayes factor around one would thus mean that both hypotheses are equally likely under the observed data. When multiple hypotheses are considered, the Bayes factor will select the best hypothesis among them.

With reference to my research question, I compared two hypotheses using the *bain* package in R (Gu et al., 2020). The first hypothesis stated that the standardized effects of socioeconomic status and the number of rooms were the same, while socioeconomic status was specified to have a larger standardized effect than the number of rooms in second hypothesis. Considering the different

sign of the relationships a complication in their comparison for magnitude, I changed *lstat* into *hstat*, now indicating the percentage of the population in the tracts that is of a higher socioeconomic status. The resulting Bayes factor was BF_{12} = 0.393. This indicates that there is 0.393 times less evidence in support of the hypothesis that the socioeconomic status of the neighborhood residents is more important in determining the price of a house than the number of rooms a house has, compared to the hypothesis that these indicators are equally important. Similarly, there is 1/0.393 = 2.543 times more support for the hypothesis that both predictor have an equally large effect on the housing price. However, according to Kass and Raftery (1995), Bayes factors below 3.2 are not worth more than a bare mention, so this is no decisive evidence that both predictors are equally important.

Posterior predictive check

A posterior predictive check can be used to compare the model to simulated datasets under the estimated model parameters. By doing this, systematic differences can be detected that could inform whether the chosen model is adequate. Aside from testing the model, assumptions can also be checked using this approach. I will use a posterior predictive check to check an assumption of multiple regression, namely the normality of the residuals. The normal distribution is characterized by a number of assumptions. For one, it is assumed that the mean, median and mode are equal. Therefore, the assumption of the normality of the residuals can be checked by a discrepancy measure comparing the residual mean and mode with each other.

A posterior predictive check involves a number of steps. First, a null hypothesis needs to be formulated. Applied to this project, the residuals were assumed to be normally distributed, so the residual mean was expected to be equal to the residual mode. Second, I sampled from the posterior distribution of the model parameters. Third, datasets were simulated with the same number of observations as the observed data (N = 506) using the model parameters over all iterations, with the burn-in period removed. Then, the residuals were computed for each observed dataset, as well as for each simulated dataset. Afterwards, the mode of the residuals was subtracted from the residual mean, for the observed and simulated datasets, respectively. The posterior predictive p-value could then be computed by calculating the proportion where the absolute simulated mean-mode difference exceeded the absolute observed mean-mode difference.

Following these steps, the posterior predictive p-value was 0.370. This Bayesian p-value is compared to 0.5 as the outcome is a normal distribution where 0.5 would indicate that the absolute observed mean-mode difference is larger than the simulated mean-mode difference in 50% of the posterior samples. As 37% is pretty close to 50%, this means that the observed residuals reflect the residuals I would expect under this model. I therefore believe that the normality assumption was not violated and that residuals are normally distributed.

Model selection with the DIC

Up until now, the relationship between the predictors and the dependent variable has been considered linear. However, Engle, Lilien and Watson (1985) considered a nonlinear relationship between the number of rooms and the housing price. More specifically, they found that the size of the positive relationship decreased as the number of rooms increased. Therefore, a second model was examined in which a quadratic term was added for the number of rooms (Model 2). With the use of Gibbs sampling, a fifth parameter was added to the model for the quadratic term and a normal distribution with the same vague hyperparameters similar of the intercept and the percentage of people with a lower socioeconomic status was chosen as its prior.

The results indicated that the relationship between the number of rooms in a house and the housing price was probably indeed nonlinear (Table 4). Contradictory to the findings of Engle, Lilien and Watson (1985), I found that the size of the positive relationship further increased as the number of rooms increased (*EAP* = 2.989 [2.542– 3.428]). This means that the price of a house goes up faster, the more rooms above the average a house has, controlling for the percentage of people with a lower socioeconomic status. Still, the standardized posterior mean of the percentage of people with

a lower socioeconomic status was larger than the standardized posterior means of the number of rooms. Model 1 and 2 were compared for fit using the deviance information criterium (DIC).

Table 4
Point estimates extracted from the sampled data under Model 2

	EAP	Beta	PSD	Naïve.se	95% CCI
(Intercept)	21.056		0.239	0.002	[20.589 – 21.532]
rm	3.829	0.292	0.407	0.004	[3.046 – 4.620]
rm^2	2.989	0.315	0.226	0.002	[2.542– 3.428]
lstat	-0.706	-0.549	0.039	0.000	[-0.782 – -0.631]

Note: all predictors were mean-centered.

Deviance Information Criterium

The DIC is a measure that can be used to compare non-nested models. Similar to other information criteria, it adds measures of misfit and complexity up and can be interpreted in the same sense that a smaller value equals a better fitting model. While its misfit measure is the same as that of the AIC and BIC, the DIC differentiates itself with its complexity measure of the estimated number of effective parameters. This is calculated by subtracting the likelihood evaluated at the posterior means from the average likelihood over the posterior distribution and multiplying the resulting value by two.

Calculating the DIC for Model 1 and 2 resulted in a DIC of 3174 and 3025, respectively. Model 2 was therefore preferred over Model 1, due to its lower DIC value. The inclusion of a quadratic term for the number of rooms in a house fits the data better than when it is left out.

Bayes Factor

The Bayes factor can quantify the relative support for a model, just as it does for a hypothesis. So, instead of using the DIC to merely decide that Model 2 fits the data better than Model 1, it is possible to determine by how much Model 2 fits the data better than Model 1. Using the package *BayesFactor* in R (Morey et al., 2018), the two models were compared. The Bayes factor exceeded 1000 by far, providing evidence that Model 2 is much more likely than Model 1 and further solidifying that Model 2 is to be preferred.

Comparison of frequentist and Bayesian approaches

This small research project can be used to highlight some of the differences between the frequentist and Bayesian approaches. I want to focus on the similarities and differences before, during and after data analysis.

First, the approaches were comparable in this project in how prior information was considered. In the frequentist approach, prior research is used to inform the research question(s) and hypotheses. In contrast, the influence of prior research reaches further in the Bayesian approach where it is directly incorporated into the analyses. As I did not know how the predictors would have affected the housing price in Boston in 1970, I did not incorporate influential prior information in the algorithm, which is why the point estimates are comparable to the maximum likelihood (ML) estimates. However, when I tried more informative priors for the number of rooms in a small sensitivity analysis (μ_{01} = 2 or 20 and σ^2_{01} = 30 or 50), the results hardly changed. This is likely because the dataset was large, hereby having more weight to affect the results. So, when the number of observations is large, the influence of the prior diminishes and the results of the frequentist and Bayesian approach are comparable.

Regardless of similar results, the data itself is however analyzed in a different manner. Regression analysis in the frequentist approach would involve deriving the ML estimates of the parameters under the given data. These estimates do not change when a regression is run again on the same data and with the same model. In the Bayesian approach, however, the outcome is

variable. This was why it was needed to set a seed before running the algorithm, as running it again would result in slightly different estimates.

Finally, the way in which the results of the data analysis are interpreted differs between the approaches. For example, I did not refer to a p-value while discussing the results or deemed any result significant. Rather, the focus was on the central credible interval and whether or not this interval included a zero. While the confidence interval can also be used to determine one's judgement of a hypothesis, this would entail something different, as the confidence interval is based on the hypothetical concept of repeated sampling. Also, there is more room for my own judgement of the credible interval and whether I really think there is evidence that a predictor affects the dependent variable.

Additionally, the Bayes factor quantifies relative evidence for all of the hypotheses under consideration. In contrast, the frequentist approach only provides an estimate of the likelihood of the observed outcome under the null hypothesis, regardless of the likelihood of the alternative hypothesis. I would thus not have been able to calculate the support for one hypothesis over the other. Returning to the variability of the outcome as discussed in a previous paragraph, I would also not have been able to calculate a DIC under the frequentist approach. This is because I would only have the likelihood evaluated at the posterior means and not the average likelihood over the posterior distribution, considering there would not be multiple samples that differed from each other. Hence, there would be no penalty for the complexity of the model as both of its components would be equal. Still, I could have used another information criterium, like the AIC, and the conclusion would likely have been the same. So, while not all information criteria can be used under both approaches, the conclusions drawn from them are likely similar.

References

Alfiyatin, A. N., Febrita, R. E., Taufiq, H., & Mahmudy, W. F. (2017). Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization Case Study: Malang, East Java, Indonesia. *International Journal of Advanced Computer Science and Applications (IJACSA)*, *8*(10), Article 10. https://doi.org/10.14569/IJACSA.2017.081042

Atkinson, S. E., & Crocker, T. D. (1987). A Bayesian Approach to Assessing the Robustness of Hedonic Property Value Studies. *Journal of Applied Econometrics*, *2*(1), 27–45.

Dubin, R. A., & Sung, C.-H. (1990). Specification of hedonic regressions: Non-nested tests on measures of neighborhood quality. *Journal of Urban Economics*, *27*(1), 97–110. https://doi.org/10.1016/0094-1190(90)90027-K

Engle, R. F., Lilien, D. M., & Watson, M. (1985). A dymimic model of housing price determination. *Journal of Econometrics*, *28*(3), 307–326. https://doi.org/10.1016/0304-4076(85)90003-X

Gao, G., Bao, Z., Cao, J., Qin, A. K., Sellis, T., Fellow, IEEE, & Wu, Z. (2019). Location-Centered House Price Prediction: A Multi-Task Learning Approach. *ArXiv:1901.01774 [Cs, Stat]*. http://arxiv.org/abs/1901.01774

Gu, X., Hoijtink, H., Mulder, J., & van Lissa, C. (2020, March 9). *Bayes Factors for Informative Hypotheses [R package bain version 0.2.4]*. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=bain

Harrison, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, *5*(1), 81–102. https://doi.org/10.1016/0095-0696(78)90006-2

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.2307/2291091

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018, May 19). *Computation of Bayes Factors for Common Designs [R package BayesFactor version 0.9.12-4.2]*. Comprehensive R Archive Network (CRAN). https://CRAN.R-project.org/package=BayesFactor