An Introduction to Bayesian Linear Regression

William Duquette and Soham Changani

December 18, 2021

Abstract

The purpose of this paper is to provide an overview of Bayesian Linear Regression. This paper discusses the rationale behind both Bayesian Inference and Bayesian Linear Regression. Three crucial sections will build on each other so that the final section, Bayesian Linear Regression, can be better understood. First, we discuss Bayes Rule and how the rule is updated to be used for Bayesian Inference. Next, we will walk through a Bayesian Inference example to show how prior knowledge can be implemented and its effect on our statistical analysis. Finally, we outline Bayesian Linear Regression and provide an example where we will compare Frequentist Linear Regression to Bayesian Linear Regression. We hope this paper can introduce Bayesian Linear Regression to those with a basic background in statistics.

Introduction

As can be seen, by the name, Bayesian Linear Regression is a sub-field of Bayesian Statistics. When most think of Bayesian statistics, they tend to think of Thomas Bayes' famous formula:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Bayesian statistics is built off of this extremely powerful formula. A slight update to this formula yields powerful tools for Bayesian inference. Bayesian linear regression, which uses Bayesian inference, is a form of statistical analysis. One might consider performing Bayesian Linear Regression instead of Frequentist Linear Regression if they have beliefs or prior evidence about the distribution for a particular parameter. There are two highly related and significant fundamental differences between the Frequentist and Bayesian approaches to statistics: the inclusion of prior knowledge into the analysis and the Bayesian approach providing a range of possible solutions instead of one estimate for the best value of a parameter. One thing to note is that Bayesian statistics has exploded in the last

30 years due to the increased ability of computing power, allowing for the use of Markov Chain Monte Carlo methods, which will be discussed in detail later (Kruschke, 144).

There is a relatively intense debate between Frequentists and Bayesians as to which approach is the best and whether you should ever include your own prior beliefs, which we will discuss more in later sections. Ultimately, our beliefs fall somewhere in the middle, and we will explain why in future sections. Unlike Frequentist Linear Regression, the goal of Bayesian Linear Regression is not to find the ideal single value for a parameter; instead, it is to create a posterior distribution for the model's parameters. The aforementioned posterior distribution is proportional to the combination of the prior distribution and the probability of your data given your parameters. One of the significant strengths of Bayesian Regression is that it is easier to account for insufficient, inadequate, or poorly distributed data by including the prior.

This paper roughly follows the approach taken by John K. Kruschke in his book *Doing Bayesian Data Analysis.* To start, we discuss Bayes rule in a broad sense. Next, we introduce and explain Bayesian Inference and walk through several examples. Finally, we present Bayesian Linear Regression and the tools needed to accomplish this.

Introduction to Bayesian Inference

As was mentioned in the introduction, Thomas Bayes is best known for the famous theorem named after him: Bayes Theorem. The simple yet powerful formula for Bayes theorem is

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

This, on its own, is an extremely powerful formula. It is typically used for conditional probability purposes. An example of a common question asked in most statistics classes at some point is "What is the probability that person A has cancer, given they smoke?" or something similar. Bayes formula is key to answering this question because to find this probability we need the $P(A), P(B)$, and $P(B|A)$. So, for this example,

$$P(Y_{cancer}|smoker) = \frac{P(smoker|Y_{cancer})P(Y_{cancer})}{P(smoker)}.$$

Bayes rule (Equation 1) can be even more powerful when you think of A as a parameter value, say θ, and B as data, say D. Thinking about Equation 1 this way allows us to build

a formula for Bayesian Inference:

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)} \tag{2}$$

$P(\theta)$ is the prior distribution for θ, meaning that the data (D) has not been taken into account when deciding on this distribution. $P(D|\theta)$ is the probability that we would see the data we saw given θ. $P(D|\theta)$ is a likelihood function, which is also why it is known as the likelihood. The denominator $P(D)$ is also equal to $\Sigma_{\theta^*} P(D|\theta^*)P(\theta^*)$ for discrete data and $\int P(D|\theta^*)P(\theta^*)\, d\theta^*$ for continuous data, where θ^* is a variable that can take all values that are possible for θ. We know that $P(A|B)P(B) = P(B \cap A)$, so $\Sigma_{\theta^*} P(D|\theta^*)P(\theta^*)$ could be rewritten as $\Sigma_{\theta^*} P(D \cap \theta^*)$ and $\int P(D|\theta^*)P(\theta^*)\, d\theta^*$ could be rewritten as $\int P(D \cap \theta^*)\, d\theta^*$. $P(\theta|D)$ is the posterior distribution of θ. In other words, it is the probability of θ given the data. We will now apply this to an example where our likelihood is Bernoulli distributed.

Bayesian Inference: Dice Example

The best way to understand Bayesian Inference is to work through an example thoroughly. Suppose Joe wants to experiment with an unbiased die. The rules for his experiment are simple: if a one is rolled, he considers it a success; if anything else is rolled, he considers it a failure, which means that the probability of an individual outcome is

$$\theta^y (1-\theta)^{1-y} \tag{3}$$

where y can take on the values 0 for failure and 1 for success. The formula shown in Equation 3 is the formula for the Bernoulli Distribution. In this case, θ can be thought of as the probability of rolling a one. Remember that y is a fixed observation, but θ is a variable, meaning different values of θ give different probabilities of seeing success or failure. When looking at Equation 3 in that light, it can be thought of as the likelihood function for one roll of the die. What about many rolls? What if Joe rolls five times? Clearly, the likelihood function we have for one roll will longer suffice. For this example, assume that each roll is independent of every other roll. Let's say that Joe rolls N times, meaning that our data (the number of rolls) is $y =< y_1, y_2, y_3, \ldots y_n >$, where each y is a

roll result (1 or 0). We know θ is the probability of rolling a one, so

$$P(y|\theta) = \prod_i P(y_i|\theta).$$

We now plug in our Bernoulli formula and get

$$P(y|\theta) = \prod_i \theta^{y_i}(1-\theta)^{1-y_i},$$

which we know equals

$$P(y|\theta) = \theta^z(1-\theta)^{N-z} \tag{4}$$

This is the Bernoulli likelihood function for a set of dice rolls, where z is the number of successful rolls and $N - z$ is the number of failures. This likelihood function is our $P(D|\theta)$. A Frequentist would stop here, find the Maximum Likelihood Estimation (MLE), and consider this the best possible value for θ. However, let's say Joe rolled his die 12 times and got 9 ones; would it make sense to take the MLE of that distribution? Simply finding the MLE would tell us that the probability of rolling a one is 0.75, meaning that if I rolled a die 100 times in the future, I should expect 75 ones. Rolling 9 ones out of 12 rolls is possible but extremely unlikely. In fact, the probability of rolling 9 ones given 12 rolls on a fair die is 0.00001263333 ($\binom{12}{9}(\frac{1}{6})^9(1-\frac{1}{6})^{12-9}$). This extreme unlikeliness is exactly why we would want to include our prior knowledge because we know that the true probability of rolling a one is 1/6. We can both include our prior knowledge and use our data to "update" our knowledge by doing this.

There are several differences between Frequentist and Bayesian statistics, but we would argue that the biggest difference is the inclusion of prior knowledge. Including $P(\theta)$ allows you to consider prior knowledge or personal belief. For the dice example, we know that the true probability of rolling a one is $\frac{1}{6} = 0.167$; however, when Joe rolled his die 12 times, he got 9 ones. Again, it is certainly possible but improbable, which is exactly why we would want to include prior knowledge. In this situation, we would want to use the Beta distribution for our prior distribution, which in this case is

$$P(\theta|\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \tag{5}$$

where $B(\alpha,\beta) = \int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\,d\theta$, ensuring the area under the curve is 1. The α and

β in Equation 4 can be thought of as the number of successes and the number of failures, meaning that $n = \alpha + \beta$. For example, if our prior was based on a study where out of 120 rolls, there were 21 ones, our $\alpha = 21$ and $\beta = 99$. It should be noted that any distribution can be used for the prior, but in this case, we will use the Beta distribution because it is an excellent prior for Bernoulli data. Generally, the Beta distribution is the prior for the Bernoulli, Geometric, and Exponential distributions. One other reason to use the Beta distribution is it allows you to take into account the "sample size" of your prior. This means that a study based on 12 dice rolls would have less of an effect than a study based on 120 dice rolls.

Now comes the time to talk about the posterior distribution. The posterior can be thought of as the "middle ground" between the likelihood and the prior. Remember from Equation 2 that $P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$. We know our $P(\theta)$ (Equation 5) and $P(D|\theta)$ (Equation 4), so we will update our Bayesian Inference formula with these distributions:

$$P(\theta|D) = \frac{\theta^z(1-\theta)^{N-z} * \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}}{P(z,N)}$$

$$P(\theta|D) = \frac{\theta^{(z+\alpha)-1}(1-\theta)^{(N-z+\beta)-1}}{B(\alpha,\beta)P(z,N)}$$

$$P(\theta|D) = \frac{\theta^{(z+\alpha)-1}(1-\theta)^{(N-z+\beta)-1}}{B(z+\alpha, N-z+\beta)} \tag{6}$$

where N is the number of dice rolls we roll, z is the number of ones we rolled, α is the number of ones rolled in a previous study for example, and β is the number of rolls that were not a one from the same study.

Looking at this, it would appear as though $P(z, N)$ disappears. There was no complicated math involved with combining $B(\alpha, \beta)P(z, N)$, it was simply by seeing that the α and β on the top and the bottom need to match. As we can see, the equation for our posterior is also a Beta distribution. When both the prior and the posterior share the same distribution, the prior is a conjugate prior.

Now, let's suppose that Joe rolls his die 12 times and sees 9 ones. Let's also suppose that he read some study that had 12 rolls as well, but they only saw 2 ones. The top graph shows our prior. This is our Beta distribution with 2 and 10 plugged in. The next graph shows our likelihood. This is our Bernoulli likelihood function for a set of dice rolls. Finally, you see our posterior, which, as we showed in Equation 6, can be thought of as a

combination between the likelihood and prior.

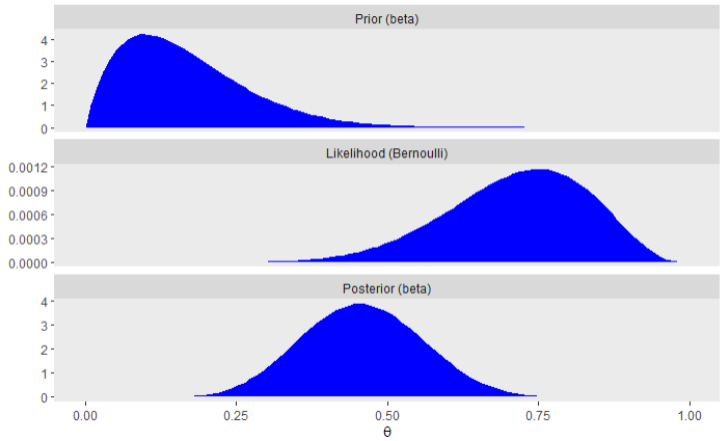


Figure 1: Prior, likelihood and posterior distributions for prior: $\alpha = 2, \beta = 10$

As you can see, the peak of the prior distribution is 0.167, which is what we expect to see. The peak of our likelihood is 0.75, which is also precisely what we would expect to see given our data. However, if you look at the posterior, you can see that its peek is neither 0.167 nor 0.75. The peak of the posterior distribution is the maximum of the posterior, which can be thought of as the best value for θ with both our prior and likelihood taken into account. In this example, the maximum of our posterior is the best value for θ given our likelihood and prior.

Recall that the Beta distribution allows you to take into account the "sample size" that our prior is based on. In the previous example, the study we based our prior on was only 12 rolls. What if our prior was instead based on a study that had 120 rolls. Shown below is what the distributions would look like with a more significant prior:

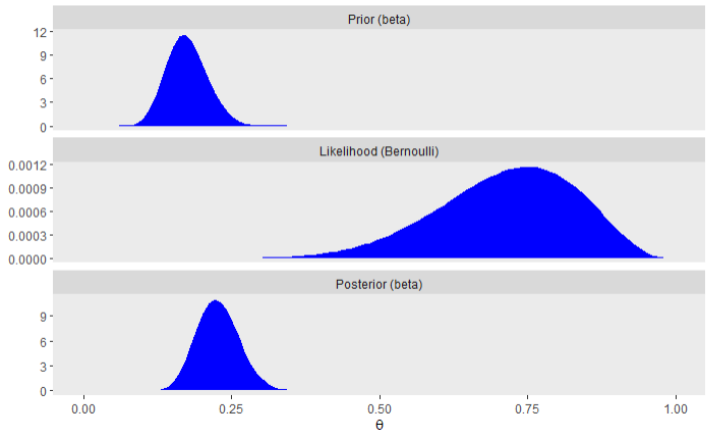


Figure 2: Prior, likelihood and posterior distributions for prior: $\alpha = 21, \beta = 99$

In this case, the prior had a much more significant effect on the posterior than the

likelihood, which can be attributed to our prior being based on a significantly larger sample size than our data. Again, this is not true of all priors. It is a unique property of the Beta distribution. The key takeaway is as our prior gets more significant, the effect of our observed values diminishes. It is worth noting again, however, that the prior distribution does not need to be a Beta distribution. The Beta distribution works well with a Bernoulli likelihood, but any probability distribution that takes values between 0 and 1 would work. The prior distribution is chosen based on our belief of what the distribution for θ should look like, so any probability distribution could work.

Bayesian Inference for Continuous Distributions

Although the dice example is simple and we use arbitrary values for example purposes, the concepts are the same for other distributions. The math can get far more complicated with continuous distributions, which is why simple data like the die roll example is best for showing the concepts. However, again, the concepts and approach are exactly the same regardless of the probability distributions you are working with. In the past, Bayesian Inference was difficult with continuous data, but modern computing and Markov Chain Monte Carlo methods have made Bayesian Inference with continuous distributions far more practical.

Something worth mentioning is how influential and informed the prior is considered when working with normally distributed data. For example, let's say there are two normally distributed variables. Both have a mean of 0, but distribution A has a standard deviation of 6, and distribution B has a standard deviation of 3. Which distribution will have a more significant effect on their respective posteriors? Distribution B will have more of an effect because it has a smaller standard deviation, indicating that the peak of the distribution is higher with distribution B than distribution A. In other words, we are more sure of the variable's value in distribution B than with distribution A. Hence, a prior with a relatively large standard deviation suggests we are less certain about the prior distribution. Given the same likelihood functions, a prior Gaussian distribution with a smaller standard deviation will be more informed and, as a result, will affect the posterior more.

Markov Chain Monte Carlo Methods

This paper will now provide a quick overview of Markov Chain Monte Carlo methods since it is often used to find a posterior distribution for continuous data and linear regression.

Given that you could write an entire paper on both Markov Chains and Monte Carlo simulations, we will briefly explain the method and how it applies to Bayesian linear regression. Monte Carlo simulations are a mathematical technique used to estimate the possible outcomes of an uncertain event. Specifically, Monte Carlo simulations use random repeated sampling to solve deterministic problems. Markov Chains are models describing a sequence of possible events in which the probability of each event depends only on the state attained in the previous event, i.e., a sequence of events where the only factor determining the next event is the present event. When combined, Markov Chain Monte Carlo methods (MCMC Methods) draw random samples from a probability distribution to approximate a parameter's posterior distribution. For example, the Metropolis-Hastings sampling algorithm, one of the simplest MCMC methods, works by first assessing the probability at a sample (sample A) and then taking a "step" towards another sample (sample B). In the case where the probability of the data (likelihood) given the prior is higher at B than A, the chain method then takes a sample and compares it to B. Otherwise, the following sample's probability is compared to the probability at sample A while storing the probability difference between A and B in memory. When this is done multiple times, the posterior probability distribution of the parameter converges with the probability distribution generated by all the samples.

There are several other famous MCMC algorithms, but the most common tend to be the Gibbs Sampler and No U-Turn Sampler (a Hamiltonian Monte Carlo Method). The Gibbs Sampler, a special case of the Metropolis-Hastings algorithm, draws a value from the distributions of each variable, conditional on the current values of the other variables. Our code uses the No-U-Turn Sampler (NUTS), which applies a similar approach but does not use a "random walk" approach over the probabilistic space. Instead, NUTS uses a recursive algorithm to build the posterior distribution that will stop automatically when it starts to double back and retrace its steps. Hence, it is more efficient and reduces costs in comparison to other methods. We use the No-U-Turn Sampler for our project and code for the same reason. An interesting follow-up would be to thoroughly study these algorithms to determine the benefits and downsides of each algorithm and analyze the many facets of each. In general, MCMC methods are used to numerically approximate complex or multi-dimensional integrals. For example, in the linear regression case that follows, we approximate the values for two normally distributed continuous variables, which is when MCMC methods would be useful.

Bayesian Linear Regression

We will now go through an example of Bayesian linear regression. The main idea behind Bayesian linear regression is that we want to create distributions for the intercept, and the coefficients, i.e., $\hat{\beta}$'s based on prior knowledge, while our likelihood distribution is based on our data. For simplicity, we will discuss a simple linear regression model, i.e., one with only one predictor variable. The concepts remain the same for multiple regression, but we will be doing simple linear regression for example purposes.

Ordinary Least Squares vs. Bayesian Linear Regression

In order to explain Bayesian linear regression, we will walk through an example data set and apply the Frequentist approach (Ordinary Least Squares method) and compare it to the Bayesian approach to show how the two methods differ. The data set we are working with, *Zagat*, contains data about 168 Italian restaurants in New York and ratings customers provided for each restaurant's food, ambiance, etc., along with the average price. The question we are interested in answering: *Is the average price of food at restaurants dependent on the average food rating?*

For this question, the restaurant's average price will be our response variable (i.e., dependent), and the food rating will be our explanatory variable (i.e., independent). To better understand the data, we have included a scatter plot.

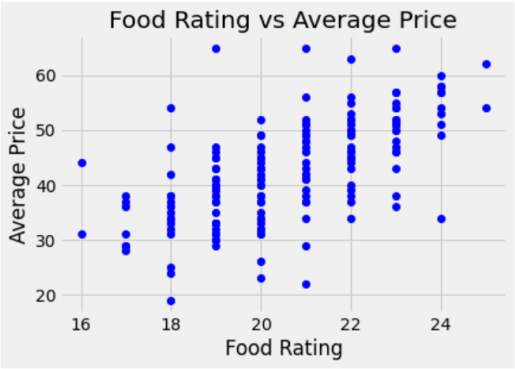


Figure 3: Scatter plot of Food Rating vs. Average Price

Ordinary Least Squares Regression

We will first start with Ordinary Least Squares methods, also known as the Frequentist approach. We will not explain this in-depth, but the output is as follows: the intercept for this model is -17.83, and our slope is 2.93. This means that our linear regression equation

is: $\widehat{\text{Price}} = -17.83 + 2.93 * (\text{Food Rating})$. The graph below shows the OLS fitted line for the scatter plot above.

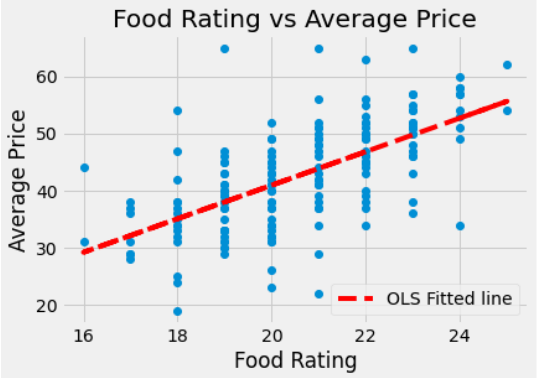


Figure 4: Scatter plot of Food Rating vs. Average Price with OLS Fitted Line

Bayesian Inference: Informed Prior

We will now fit a linear regression model using Bayesian inference to the same data set. First, we will do an example with an informed prior. Let's say we read a study that suggested price is not related to food rating, however, we want to take into account some variability and not fix the prior at 0. Hence, we include a standard deviation of 5. This would mean that the prior distribution of $\hat{\beta}_1$ is $\hat{\beta}_1 \sim N(0, 5)$. Moreover, assume that from some prior knowledge, we believe that the intercept $\hat{\beta}_0$ also has a mean of 0 and a standard deviation of 5, i.e., $\hat{\beta}_0 \sim N(0, 5)$. We then observe some data, specifically the Zagat data set, to generate a posterior distribution for both the parameters.

Similar to the dice example, we combine our prior knowledge ($P(\hat{\beta}_0, \hat{\beta}_1, \sigma)$) and the likelihood ($P(D|\hat{\beta}_0, \hat{\beta}_1, \sigma)$) to find the posterior ($P(\hat{\beta}_0, \hat{\beta}_1, \sigma|D)$). In this case, we use NUTS to find the posterior distribution for $\hat{\beta}_0, \hat{\beta}_1,$ and σ.

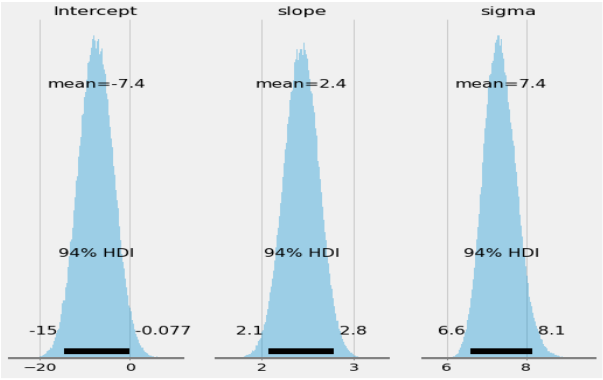
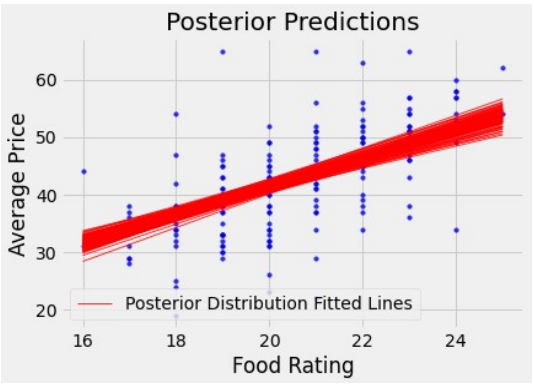


Figure 5: Posterior Distribution for the intercept, slope, and sigma

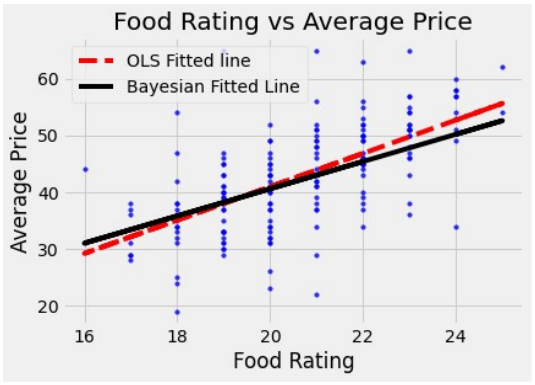
As you can see, the posterior was heavily affected by the prior. For example, the OLS

model's intercept was -17.83 (the MLE of likelihood distribution for the intercept), but the mean for the intercept's posterior is -7.4 (Figure 5). This extremely large shift is caused by our informed prior that has a mean of 0 with a relatively small standard deviation, meaning it will have a significant effect on the posterior.

In Figure 6(a), you will notice many possible linear regression equations. Remember, the posterior is not the one "best" equation but rather a distribution. Typically, people tend to take the mean, median, or mode of the posterior to find the maximum of the posterior (similar to the MLE of the posterior). It can be helpful to see what all the different possible models look like. Using the means from the posterior distributions, the equation of the line will be: $\widehat{\text{Price}} = -7.4 + 2.4(\text{Food Rating})$, which is a little different from the OLS regression since we included prior knowledge. Additionally, it is worth noting that if we get more data about restaurant prices and food ratings in New York in the future, we could use the posterior from this analysis as our prior for our subsequent regression analysis.



(a) Scatter plot of Food Rating vs. Average Price with possible Posterior Distribution Lines



(b) Food Rating vs Average Price with both OLS Fitted Line and "best" Bayesian model (maximum of posterior)

Figure 6

Prediction

What about when we want to predict a single value, say for a food rating of 23? We take every possible posterior value for the intercept and slope distributions and multiply those by 23. This is the same as $\widehat{\text{Price}} = P(\hat{\beta}_0|D) + P(\hat{\beta}_1|D) * 23$. The distribution below shows the posterior distribution given a food rating of 23, which seems to be centered around 48.7. The red dashed vertical line represents the OLS predicted value.

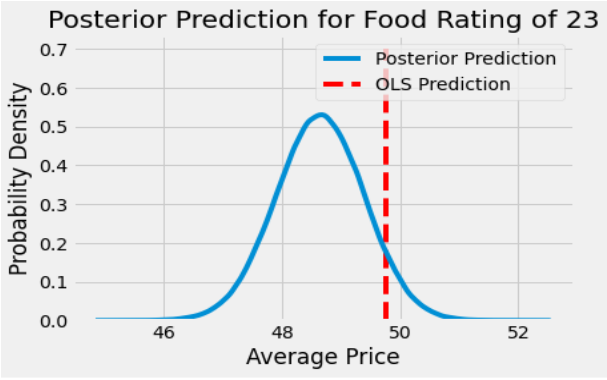
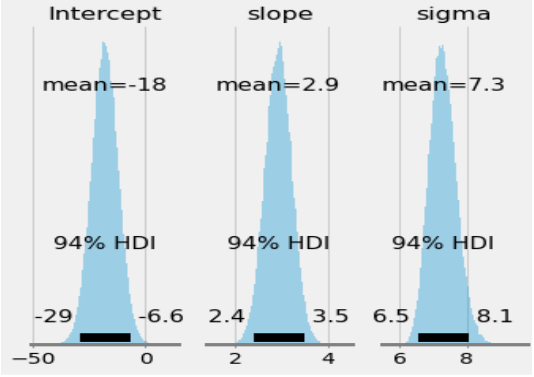


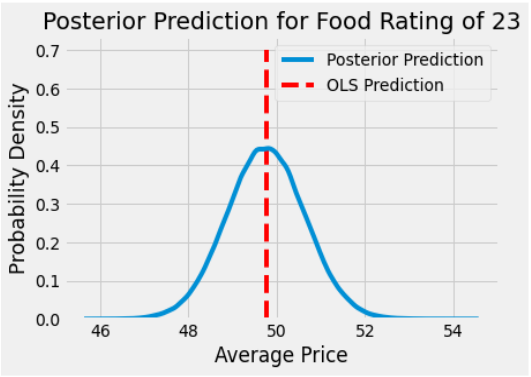
Figure 7: Prediction for one Food Rating value (Food Rating = 23)

Bayesian Inference: Uninformed Prior

We will now walk through an example of Bayesian inference where the prior knowledge is not that significant. An important note here is that uninformed or non-informative priors are a misnomer. Any prior distribution we choose always holds some information, but if the prior distribution has large variability or is a uniform distribution with a large range, it would suggest that we are "less" informed about the true relationship. Thus, the data would have more effect on the posterior distribution. Assume that we have little knowledge of how food ratings and prices are related, but we have some intuition that there is not much relation between the two variables. Similarly, we do not have previous knowledge about the intercept. Then, we would choose a larger standard deviation for both the intercept and the slope. For our uninformed prior example, we use $\hat{\beta}_0 \sim N(0, 100)$ and $\hat{\beta}_1 \sim N(0, 100)$. We will follow the same steps to get a posterior distribution of the intercept and slope using the data and the prior.



(a) Posterior Distribution for the intercept, slope, and sigma



(b) Prediction for one Food Rating value (Food Rating = 23)

Figure 8

Notice that since the standard deviation of the prior in this example is very large, it has almost no effect. The intercept for the OLS model was -17.83, and the mean for the intercept's posterior distribution is -18 (Figure 8(a)). Despite having some effect, it was minimal. This is starkly different from the informed prior example. With the informed prior, the mean for the intercepts posterior is -7.4. The data is the same, and the mean of the prior is exactly the same; the only difference between the two priors in the informed and uniformed example is the standard deviations of each prior. The informed prior has a standard deviation of 5, and the uninformed prior has a standard deviation of 100.

Using the means from this posterior distributions, the equation of the line will be: $\widehat{\text{Price}} = -18 + 2.9(\text{Food Rating})$. Notice that these values are close to the OLS regression values since our prior knowledge did not affect the posterior much. You can see this in Figure 8(b), where the OLS predicted value is almost exactly the same as the maximum of the posterior.

Should Prior Knowledge be Included?

The debate between Frequentist and Bayesian statistics boils down to two fundamental questions: should prior knowledge ever be included, and should the parameter's value have a probability distribution (posterior)? Unfortunately, there is no correct answer to either of those questions, but we will try and briefly go over the first debate.

A Frequentist would argue that you should base your model on the data because that is what took place, and including your prior belief introduces extra bias into your model. But what about when the data we observe cannot possibly be used for prediction due to how improbable it is? Our die example is an excellent illustration of this. We know that the probability of rolling a one on an unbiased die is 1/6. However, we got 9 ones out of 12 rolls. Like we mentioned above, that is extremely unlikely, but unlikely does not mean impossible. Anything is possible, but that certainly does not mean that the true probability of rolling a one on a fair die is 0.75. So, maybe taking the maximum likelihood estimation and moving on is not the best approach. This is an excellent example of when including prior knowledge might be helpful. With that said, it is also crucial to remain open-minded and take steps to prevent any bias because it is unquestionably true that introducing a prior introduces some level of bias. There are certainly benefits and downsides of each approach, and no one method is always correct, which is why it is essential to think about your situation and whether it warrants including a prior.

Conclusion

In conclusion, Bayesian Linear Regression and Bayesian inference are powerful techniques that allow us to generate probability distributions for parameters that we would like to investigate using the data collected and any prior knowledge. The philosophy behind Bayesian inference will likely always be debated. Like anything in statistics, it is important to move cautiously to avoid any bias. But that is not to say that Bayesian Statistics is not worthwhile. In fact, it has made somewhat of a comeback in the statistics world, given its uses in artificial intelligence. Many machine learning algorithms work by updating existing knowledge as new information is gathered; as a result, Bayesian statistics is extremely applicable to machine learning. Similarly, Bayesian statistics is also highly applicable for deep learning algorithms. Many e-commerce, insurance, and healthcare companies are currently implementing Bayesian statistics in their algorithms. Although our paper focuses primarily on Bayesian inference and simple linear regression, the applications of conditional probabilities are endless.

References

Hoffman, Matthew D., and Andrew Gelman. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." J. Mach. Learn. Res. 15.1 (2014): 1593-1623.

Koehrsen, Will. "Introduction to Bayesian Linear Regression." Medium, 20 Apr. 2018, towardsdatascience.com/introduction-to-bayesian-linear-regression-e66e60791ea7.

Kruschke, John. "Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan." (2014).

Shaver, Ben. "A Zero-Math Introduction to Markov Chain Monte Carlo Methods." Medium, 24 Dec. 2017, towardsdatascience.com/a-zero-math-introduction-to-markov-chain-monte-carlo-methods-dcba889e0c50.

GitHub

All code that is used for this paper and a complete linear regression example in Python can be found at: https://github.com/wjduq/NYC-Bayesian-Regression