Likelihood Functions and Bayesian Priors: An Introduction to Alternative Methods for Estimating Regression Coefficients

Eli Mecinas Cruz, Dexter Corley, John Green, Walker Hughes

September 2024

1 Introduction

We often try to predict things based on educated guesses. These educated guesses are based on past experiences, our understanding of how things should work, and sometimes, data. Imagine you are at a carnival, given the chance to win a prize if you can guess how many candies are inside a jar. To win, you will make smart, educated guesses based on the size of the jar, the size of the candies, and maybe even a little math. Even though you can't know the exact number, some guesses will be more likely than others. Maximum Likelihood Estimation works in a similar way: it uses all the available data to narrow down to the most likely answer. It's a tool that we use to make the best possible prediction based on what we know whether we're making simple predictions like in a linear regression or more complex ones like in advanced machine learning models.

When building predictive models, one of the first challenges we encounter is finding the best set of parameters to explain the relationship between our inputs X and outputs y. A powerful way of tackling this problem is through Maximum Likelihood Estimation (MLE). MLE focuses on identifying the parameters that maximize the likelihood of the observed data under the model, effectively providing us with the parameters that make our data most probable. We will begin by using it in the simplest form of regression: linear regression. By applying MLE to Ordinary Least Squares, we will develop an intuition that allows us to understand more nuanced forms of regression, Ridge and Lasso. For these models, a Bayesian framework around the distribution of our regression coefficients will help us derive regularized regression coefficients.

Before moving on, make sure to check out the interactive Streamlit App that accompanies this blog post on GitHub!

2 Maximum Likelihood Estimation

To understand Maximum Likelihood Estimation, we will start by breaking down its key components. The term "Maximum" refers to the highest value or peak of

a function, and in MLE we are looking for the parameter values that maximize what's called a likelihood function, described below. This ensures we are looking for parameters that make the observed data most probable.

Next, "Likelihood" refers to how likely it is that a given set of parameters explains the observed data. In the context of MLE, the likelihood function represents the probability of observing the given data based on the model's parameters. Although, likelihood is not the same as probability; rather it is a function of the parameters and not the data itself. The higher the likelihood, the better the parameters explain the observed data, which is why we want to maximize the likelihood function.

The term "Estimator" refers to a rule or method for calculating an estimate of a parameter based on observed data. In MLE, the estimator is the formula or approach that we use to calculate the best parameter values that will maximize the likelihood function. This process provides an estimate, or approximation, of the true underlying parameters of the model based on the data. Remember, the goal is to estimate something that we don't already know, so MLE will give us our best guess for understanding our data.

We can now introduce some mathematical notation. If we let θ represent the parameters of our model, and X represent the data, the likelihood function is defined as $L(\theta|X)$. We want to find the parameter values $\hat{\theta}$ that maximize this function, therefore finding the parameters that make the observed data X most probable. We can do this by solving the following optimization problem

$$\hat{\theta} = \arg\max_{\theta} L(\theta|X) = \arg\max_{\theta} \prod_{i=1}^{n} f(\theta|X)$$

where $\hat{\theta}$ represents the parameter values that maximize the likelihood function, and $f(\theta|X)$ represents the probability density function. It is very common to work with the log-likelihood instead of the likelihood, since it simplifies many calculations. The benefit of using the log-likelihood is that its maximum occurs at the same parameter values as the likelihood function, but when we take the negative of the log-likelihood, we transform the problem into a minimization problem. Thus, minimizing the negative log-likelihood is equivalent to maximizing the likelihood function. When computing the maximum in MLE, it is often easier to minimize the negative log-likelihood because taking the derivative of the log function simplifies the math compared to the original likelihood function. The logarithmic function allows us to convert exponents into factors and products into sums, further simplifying our complex functions. To compute the maximum in our MLE, it is easy to take the derivative of a log function as opposed to our original likelihood function, which we will see later on. So now we have the following:

$$\hat{\theta} = \arg\max_{\theta} \log L(\theta|X) = \arg\max_{\theta} \sum_{i=1}^{n} log(f(\theta|X))$$

To see Maximum Likelihood Estimators in action, we will apply them to Linear Regression, Ridge Regression, and Lasso Regression. It is important to note

that in regression models, MLE is closely related to minimizing errors. For example, in Linear Regression, the MLE solution coincides with minimizing the sum of squared residuals, which is the basis for Ordinary Least Squares (OLS).

3 Deriving the OLS Coefficients Through Maximum Likelihood Estimation

Consider the linear regression model

$$y = X\beta + \epsilon$$

where y is an nx1 vector of our dependent variable observations, X is an nxp matrrix with columns representing our features, and ϵ is an nx1 vector of iid random errors, with

$$\epsilon \sim N(0, \sigma^2 I)$$

which implies that

$$y \sim N(X\beta, \sigma^2 I)$$

With this in mind, we can formulate our likelihood function since we know the assumed normal distribution of our target variable y and error $\epsilon = y - X\beta$. Assuming a normal distribution, we use MLE to find the coefficients that maximize the likelihood of having seen our data (y, X). Thus we solve the optimization problem

$$\hat{\beta} = \arg\max_{\beta} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(\frac{-(y_i - X_i\beta)^2}{2\sigma^2}\right)$$

and taking logs gives the log-likelihood optimization problem

$$\hat{\beta} = \arg\max_{\beta} l(\beta) = \arg\max_{\beta} \left(-\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - X_i\beta)^2\right)$$

and simplifying notation gives

$$\hat{\beta} = \arg\max_{\beta} \left(-\frac{n}{2} ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta)\right)$$

By considering the first-order necessary conditions for an optimal β we derive the following

$$\frac{\partial l(\beta)}{\partial \beta} = 2X^T(y - X\beta) = 0$$

implying that

$$0 = X^T(y - X\beta) = X^T y - X^T X\beta \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

which agrees with the closed-form solution for Linear Regression coefficients we are familiar with when derived with the Lease Squares principle.

Deriving the OLS coefficients through Maximum Likelihood Estimation is useful to build our intuition around the distributions of the various data and variables at play. Indeed, our assumption of normally distributed errors motivated our use of a Normal probability density function for y in our likelihood function, which led to deriving the same closed-form solution for $\hat{\beta}$ that we are familiar with. Importantly, we did not need to make any assumption about the distribution of β in order to do so as we treat β as fixed in this approach. But what if we had reason to believe that β followed a specific distribution in our estimation? What is we wanted to incorporate prior beliefs about β and treat it as a random variable rather than a fixed parameter?

4 Incorporating Prior Beliefs About β Through a Bayesian Framework

A fundamental shift in how we think about the β coefficients happens when we assume they are random variables. This is a deviation from the frequentist approach used in Maximum Likelihood Estimation, which assumes that β is fixed. Assuming that β is in fact a random variable is helpful in several ways, however. For example, perhaps a domain expert for the problem at hand has reason to believe there are likely values for β based on previous research. We can incorporate this information through a prior distribution that we assume β follows and as we incorporate new data into our model, we can update our beliefs about β through Bayes' Theorem to form an updated 'posterior' distribution.

Assume that $\beta \sim f(\beta)$ for some pdf f. While the MLE computes

$$\hat{\beta} = \arg\max_{\beta} f(X|\beta)$$

we may wish to estimate what is called the Bayesian Maximum a Posteriori (MAP) estimator

$$\hat{\beta} = \arg\max_{\beta} P(\beta|X)$$

which by Bayes theorem is equivalent to

$$\hat{\beta} \propto \arg\max_{\beta} P(X|\beta)P(\beta) = \arg\max_{\beta} f(X|\beta)f(\beta)$$

and now incorporates our prior belief about the distribution of β in our optimization function.

It should be noted that in the two derivations covered hereafter, the regression coefficients we obtain will differ slightly from the OLS coefficients. This means that the Ridge and Lasso solutions will not be the Best Linear Unbiased Estimators for β. In fact, these forms of regression actually introduce bias into our models intentionally! This is done at the expense of *variance*, and highlights the bias-variance *trade-off* we hear about in machine learning. In fact, there are

many real-world situations where a savvy regression practitioner may be willing to accept some extra bias in order to mitigate unwanted variance. While a full discussion of these trade-offs is outside the scope of this article, this Bayesian framework will soon prove useful for us, and we use it to derive the Ridge and Lasso Regression coefficients with clever choices of prior distributions for β.

5 Ridge Regression

Ridge Regression is a form of regularized regression, which means we include a penalization term in our objective function that penalizes large coefficients. This is often helpful in mitigating the effects of multicollinearity and also helps prevent overfitting. We can derive the Ridge Regression coefficients by minimising the Sum of Squared Errors as in OLS, but assuming a Gaussian prior for β centered at zero:

$$\beta \sim N(0, \frac{\sigma^2}{\lambda} I)$$

This expresses the belief that large coefficients of β should be penalized towards zero, but not exactly zero. The strength of this belief is controlled by a user-defined parameter λ, which acts as a regularization parameter. Building on our previous derivation, in Maximum A Posteriori (MAP) Estimation, we combine the likelihood (from the data) and the prior (our belief about the parameters) using Bayes' theorem:

$$P(\beta|Y, X) \propto P(Y|\beta, X) \cdot P(\beta)$$

Where:

- $P(\beta|Y, X)$ is the posterior distribution (what we want to maximize).
- $P(Y|\beta, X)$ is the likelihood (what we maximize in MLE).
- $P(\beta)$ is the prior distribution (our belief about β).

We can take logs as we did in MLE to obtain

$$\log P(\beta|Y, X) = \log P(Y|\beta, X) + \log P(\beta)$$

The log-likelihood $\log P(Y|\beta, X)$ is:

$$\log P(Y|\beta, X) = -\frac{1}{2\sigma^2} \|Y - X\beta\|^2$$

The log-prior $\log P(\beta)$ is:

$$\log P(\beta) = -\frac{\lambda}{2\sigma^2} \|\beta\|^2$$

Combining these, the log-posterior becomes:

$$\log P(\beta|Y, X) = -\frac{1}{2\sigma^2} \|Y - X\beta\|^2 - \frac{\lambda}{2\sigma^2} \|\beta\|^2$$

To maximize the posterior distribution, we minimize the negative log-posterior. This gives us the following objective function for Ridge regression:

$$\arg\min_{\beta} Q(\beta) = \arg\min_{\beta} \|Y - X\beta\|^2 + \lambda\|\beta\|^2$$

This is the key equation for Ridge regression, where:

- The first term $\|Y - X\beta\|^2$ is the sum of squared residuals, representing the data fit (MLE).
- The second term $\lambda\|\beta\|^2$ is the penalty term, representing the influence of the prior belief about β (that it should be small).

This transition from MLE to MAP estimation explains how Ridge regression arises as a regularized version of linear regression, where we shrink the coefficients by introducing a Gaussian prior. Instead of just maximizing the likelihood (MLE), we incorporate prior beliefs and use MAP estimation. The multiple linear regression optimization problem

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|^2$$

becomes

$$\hat{\beta} = \arg\min_{\beta} \left(\|Y - X\beta\|^2 + \lambda\|\beta\|^2\right)$$

Where:

- Y is the $n \times 1$ vector of observed values (dependent variable).
- X is the $n \times (p+1)$ matrix of input features (independent variables), where n is the number of observations and p is the number of predictors. The first column of X is typically all ones, corresponding to the intercept β_0.
- β is the $(p + 1) \times 1$ vector of coefficients (parameters) that we want to estimate, including the intercept β_0.
- $\|\cdot\|^2$ is the squared Euclidean norm.

The second term $\lambda\|\beta\|^2$ is the regularization term that penalizes large values of β. This λ term acts as a trade-off between minimizing the residual sum of squares and controlling the magnitude of the coefficients. The Ridge regression optimization problem combines the least squares loss with the regularization penalty:

$$\min_{\beta} \left[(Y - X\beta)^T(Y - X\beta) + \lambda\beta^T\beta\right]$$

Expanding this:

$$\min_{\beta} \left[Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta + \lambda\beta^T\beta\right]$$

We have to take the derivative with respect to β and set the derivative to zero. We can differentiate each term individually:

1. $\frac{\partial}{\partial \beta} Y^T Y$: This is a constant with respect to β, so its derivative is zero.
2. $\frac{\partial}{\partial \beta}(-2\beta^T X^T Y)$: The derivative of this term with respect to β is: $-2X^T Y$
3. $\frac{\partial}{\partial \beta}\beta^T X^T X\beta$: This is a quadratic form, and the derivative with respect to β is: $2X^T X\beta$
4. $\frac{\partial}{\partial \beta}\lambda\beta^T\beta$: This is the regularization term, and the derivative with respect to β is: $2\lambda\beta$

Combining the results and setting the derivative to zero:

$$-2X^T Y + 2X^T X\beta + 2\lambda\beta = 0$$

Solving for β by simplifying:

$$X^T Y + X^T X\beta + \lambda\beta = 0$$

$$X^T X\beta + \lambda\beta = X^T Y$$

$$(X^T X + \lambda I)\beta = X^T Y$$

Now, solving for the estimated coefficients, we have:

$$\hat{\beta} = (X^T X + \lambda I)^{-1} X^T Y$$

The I is the identity matrix of size $(p+1) \times (p+1)$, and λI is the regularization term. We now have better estimates for β that we can use in our models.

Ridge Regression provides a powerful extension to Linear Regression, particularly when dealing with issues such as multicollinearity or high-dimensional datasets. By incorporating a regularization term into the model, it balances model complexity and predictive accuracy. The adaptation of MLE to MAP estimation incorporates prior information about β, allowing for a more robust estimation process. This creates a more stable model that can generalize to new data, making the model more robust.

6 Lasso Regression

Lasso Regression is another form of regularized regression similar to Ridge Regression, but that finds a *sparse* vector of coefficients for β. Where Ridge penalized large coefficients through its regularization term to make them smaller in magnitude than the OLS coefficients, Lasso Regression will actually set some of these coefficients to 0, essentially performing feature selection for us in addition to mitigating the effects of overfitting. This may be useful when we have a large featureset but are not sure which are the most relevant features to include in our regression model; Lasso can help us perform this feature selection by setting some of their coefficients to zero.

To begin, we assume a Laplace prior for β. That is

$$\beta \sim \frac{\lambda}{2} exp(-\lambda|\beta|)$$

Following the same process as we did for Ridge Regression and assuming the same dimensions for y, X, and β as found before, we note that

$$\hat{\beta} = \arg\max_{\beta} l(\beta) = \arg\max_{\beta} \log P(\beta|Y, X)$$

$$= \arg\max_{\beta} \log P(Y|\beta, X) + \log P(\beta)$$

$$= \arg\max_{\beta} -\frac{1}{2\sigma^2}\|Y - X\beta\|^2 + log(\frac{\lambda}{2} exp(-\lambda|\beta|))$$

or equivalently

$$= \arg\min_{\beta} \frac{1}{2\sigma^2}\|Y - X\beta\|^2 + \lambda||\beta||_1 - log(\frac{\lambda}{2})$$

Since constant terms and scaling factors will not affect our optimal β for this objective function, we arrive at the Lasso Regression objective

$$= \arg\min_{\beta} \|Y - X\beta\|^2 + \lambda||\beta||_1$$

which is typically computed numerically with gradient descent due to the 1-norm making differentiation intractable. However, by assuming a Laplace prior for βs in this problem, we successfully derived the Lasso Regression objective function, showing that our Bayesian approach can also lead to the popular regularization technique in regression problems when assuming a Laplace prior distribution for β.

7 Conclusion

In this article we explored how Maximum Likelihood Estimation for the OLS Regression coefficients β developed our intuition around the distributions of the variables at play in regression problems. This motivated our exploration of explicitly assuming a prior distribution for β by treating it as a random variable, a key difference between Bayesian statistics and Frequentist statistics like MLE. With clever choices for these prior distributions, we were able to derive the popular Ridge and Lasso Regression solutions. We also learned that while Ridge and Lasso introduce bias into our regressions, they can help reduce unwanted variance too.

You may be asking yourself if there are other priors we could assume for β that lead to additional interesting regression problems, like Elastic Net Regression. We will leave that as an exercise for the reader =]

Make sure to check out the interactive Streamlit App that accompanies this blog post on GitHub!