Crash Course on Linear Regression: A Comparison Between Classical and Bayesian Statistics

**Santiago Lozano Sandino[[1]](#footnote-1)**

April 30, 2018

**Abstract**

This document discusses linear regression from two different perspectives: classical (also known as frequentist) and Bayesian statistics. I provide the theoretical considerations of both approaches, commenting on the similarities and differences. Also I study the application of each of these approaches when estimating the model's parameters and use three different methods to achieve the parameter estimation: ordinary least squares, maximum likelihood, and posterior distribution sampling. The corresponding code written in R Programming Language and Python is available on my GitHub webpage or may be requested via e-mail. To fully harness the content of this document it's necessary to have some basic knowledge of linear algebra, programming and probability theory.

Keywords: Linear Regression, Classical Statistics, Bayesian Statistics, Multivariate Normal Distribution, Ordinary Least Squares, Maximum Likelihood, Posterior Distribution Simulation.

# Introduction

Linear regression is an approach used to model the relationship between a dependent variable, , and a set of explanatory (*known* or *fixed*) variables, . In other words, linear models are used to model the conditional mean of given , meaning that is not random, and can be explained as a linear combination of plus a disturbance. It's important that contains all relevant explanatory variables. Consider the following equation:

where is the random vector of dependent variables, is the matrix containing the explanatory variables, is the vector of parameters and is the random vector of error, i.e. the disturbance.

Linear regression widely used in multiple fields such as economics, engineering, finance, among others. Its main advantages include that it's very easy to estimate and its results are easy to interpret.

Theoretically, the model is considered to be a good model if it satisfies the assumptions of the Gauss-Markov theorem. There are many versions of the Gauss-Markov theorem assumptions (see Rencher & Schaalje, 2008, p. 143; Stapleton, 1995, p. 87), but essentially they demand the following requirements to be fulfilled (along with the conditions stated in the previous paragraph):

* The expected value of is for each , i.e. .
* There is no correlation between and for all , i.e. .
* Homoscedasticity, i.e. for all and .
* There is no perfect multicollinearity, i.e. the columns of matrix are linearly independent from each other.

Usually there are ways to diagnose and correct the infringement of these assumptions, but that's not an objective of this document. In other words, we will assume that the Gauss-Markov theorem assumptions are satisfied. With a classical approach the Gauss-Markov theorem assumptions are required in order to obtain the *best linear unbiased estimator* (BLUE), and some of them are even needed to estimate the parameters. On the other hand, Bayesian approach does not *need* these assumptions to obtain the *best estimator*.

Classical statistics differ from Bayesian statistics mainly because the former considers the parameters to be fixed, while the latter considers them to be random. This has major implications: from a classical point of view parameters are fixed, while the Bayesian perspective argues that parameters are not fixed, but rather random. I will expand on these ideas in the following sections.

The data used in the estimation was simulated using R and Python random number generator, so the parameters' real values are known. I use three different sample sizes to study how these affect the estimations: samples of size 30, 3,000 and 300,000. The data simulated corresponds to a simple linear regression, with the intention to display the results in a 2-dimensional scatterplot. However, the function bayes\_lm can be used in a multiple linear regression. Data was simulated as follows:

The document is organised as follows. First, I briefly comment about the theory concerning the classical approach concerning the linear model and I detail the procedure to estimate the parameters with the classical approach, using ordinary least squares and maximum likelihood, and present the results obtained. Secondly, I explain the Bayesian approach, giving a brief explanation of the reasons behind its implementation, comparing it with the classical approach and giving an example of a Bayesian linear model. Then I detail one method commonly used in Bayesian statistics: posterior distribution sampling. Finally, I give my final comments and discuss both approaches.

# Classical Approach

The classical approach relies on the Gauss-Markov theorem assumptions. It considers that parameters are fixed and unknown values, and that the data used to estimate them was randomly sampled. Following this, the model can be expressed as:

where the length of vector is , the dimension of is , with and , and the properties of : and , being the identity matrix of dimension . Considering the latter, by properties of the expected value and variance of random vector it's evident that and . The model previously described has parameters, i.e. and .

Note that until now I haven't assumed any distribution on or . That is because with these conditions alone one can obtain the BLUE of , using ordinary least squares. However, when estimating the parameters with maximum likelihood a distribution on must be assumed. Most typically the distribution of the disturbances is assumed to be a multivariate centred normal, i.e. , often as a result of the central limit theorem. There are other options, such as the multivariate , but this document will assume that the vector of disturbances are normally distributed. The distributional assumption is also required when performing parametrical hypothesis testing, such as significance tests.

Here I present the procedures used to estimate the parameters using ordinary least squares and maximum likelihood.

## Ordinary Least Squares

The idea behind ordinary least squares is that we want to minimise the errors . Since the errors are unknown, we use the residuals the Euclidean distance between the real observations and their estimations. To do this we consider:

The squared Euclidean distance of the residuals, or the squared sum of residuals, can be written as:

The previous expression is continuous and strictly convex respect to , so optimization by differentiating the expression allows us to obtain a minimum. The result is that the minimum is:

Here comes evident the necessity of matrix being full rank; if it's not then there would be no inverse of . The estimation of suffices to make an estimation . Also, under the assumption of independence and identically distributed[[2]](#footnote-2) samples, a prediction can be made using a matrix of dimension as .

The mean squared residuals are often used as an estimator of :

However this alone doesn't suffice to perform hypothesis testing, to check for instance if all or any elements of vector differ from 0, unless we appeal to the central limit theorem. If we appeal to the central limit theorem, then we can produce an hypothesis about real value of a particular , , via hypothesis testing and construct a confidence interval by acknowledging that:

where is the hypothesised value of . In a two-tailed hypothesis testing the absolute value of lies beyond (i.e., the percentile of the distribution with degrees of freedom), then the null hypothesis that the real value of is can be rejected. The value of is fixed beforehand and is usually 0.1, 0.05 or 0.01.

Using the same distribution used in hypothesis testing we can construct a confidence interval. Confidence intervals are interpreted as a range which will contain the real value of the parameter 95% of the time (when ), in repeated measurements. The boundaries of this range are considered random, whereas the true value of the parameter is fixed.

We can calculate the confidence interval of the estimation , for :

where is a vector of explanatory variables associated with observation (i.e. a row of matrix ). Likewise, we can obtain the prediction interval of a new observation :

where , and vector a row of matrix .

## Maximum Likelihood

Maximum likelihood estimation assumes a distribution on , along with the rest of the Gauss-Markov theorem assumptions, so that:

where[[3]](#footnote-3),

The purpose of this estimation is to find the values that maximise the probability density , or in other words find its mode. A close inspection of the probability density function shows as that maximising it with respect to gives us the same result as minimising the squared sum of residuals.

On the other hand, the maximum likelihood estimator of is:

but since it can easily be determined that the estimator is not unbiased. The corrected (unbiased) estimator is the mean squared residuals.

The comparison between the two methods of estimation makes it evident how similar they are. With maximum likelihood the constriction of hypothesis tests, and confidence and prediction intervals are pretty straightforward and identical to those explained previously.

# Bayesian Approach

This approach also considers a linear relationship between the response variable and the explanatory variables:

However, the main difference between Bayesian and classical approaches is that the former doesn't consider the parameters to be fixed. In other words, parameters are treated as random, and so they have an associated random distribution determined by a set of parameters (the hyperparameters).

The Bayesian approach uses the Bayes' theorem in order to obtain information about the parameters' distribution. Consider:

where is density function of the distribution of the data given a set of parameters , is the distribution of those parameters, and is the posterior distribution of the parameters. The latter is the distribution of the parameters given the data and with it we can answer the question *what does the data tell us about the parameters' distribution*; the posterior distribution is a interpreted as the probability distribution of the unknown random parameters conditional on the evidence .

It's up to the investigators to choose the distribution of the parameters and its hyperparameters. The selection of the parameter's distributions and the values of the hyperparameters are set according to the investigators *belief*. This means that there's no *correct* way to set the model and it's subjective to the investigator's belief, which contrasts with the classical approach. While there is no fixed recipe, there are some considerations that allow choosing distribution for the parameters that's more convenient and makes model mathematically tractable. In this document we will consider only one configuration, which uses a normal-scaled inverse chi-squared distribution for the parameters and .

Consider the following model describing the data distribution and the parameter distribution:

Choosing the normal-scaled inverse chi-squared distribution for the parameters and is quite convenient as the posterior distribution is also a normal-scaled inverse chi-squared, meaning that this distribution is *conjugate*. The values , , and are the hyperparameters. They are chosen accordingly to the investigator's belief or rather in a way that's ambiguous enough not to bias the model, also known as a weakly informative prior. By a simple mathematical manipulation it's easy to find that the conditional posterior distribution of the parameters:

Furthermore, the conditional posterior distribution of can be simplified by marginalising with respect to , and thus it can be proved that

where . The proof goes as follows:

(using the Pythagorean Theorem we get)

Note that if and the elements of the diagonal of tend to , then the posterior conditional distribution is identical to the distribution of under the classical approach. Additional to this, if hyperparameters of the prior distribution of are and , then the priori would be *non-informative prior*.

On the other hand, the predictive posterior distribution of , given a matrix of new observations (of dimension ), can be obtained by solving this:

which results in a multivariate *t* distribution:

Using the posterior distributions of both the parameters and the new observations we can construct credible intervals. Bayesian credible intervals are similar to frequentist confidence intervals but their interpretation differs, because the Bayesian approach considers the boundaries of the interval to be fixed and the parameter to be random. The values of the boundaries correspond to the percentile of the posterior distribution.

## Simulation from the Posterior Distributions

There are many ways to approach this problem. One is to perform a Gibbs sampling, sampling from and , but it requires initial values to initialise the chain. Another method is to use the properties of the posterior distributions to obtain the mean, the variance, and the inverse cumulative functions to identify the credible intervals.

The function I programmed to perform the simulation exploits vectorization in Python and R, trying to make it as efficient as possible. The steps that describe simulation process are the following:

1. Sample values from . This way the simulation doesn't require initial values. The values simulated are stored in a vector .

2. Sample values from a distribution and then transform the vector into a matrix of dimension . Sampling from a standard normal distribution is less expensive than sampling form a multivariate normal distribution with a non-diagonal variance-covariance matrix. To obtain the sample of it's necessary to perform the following operation using both and :

where is the vector corresponding to the row of and matrix is the spectral decomposition of matrix , such that .

3. Estimate the simulated value of the mean .

4. Sample times from , and re-scale and relocate the sample accordingly.

1. MSc in statistics from Pontificia Universidad Católica de Chile and bachelor's degree in economics from Universidad de los Andes (Colombia). Contact: lozsandino@gmail.com. GitHub: github.com/lozsandino. [↑](#footnote-ref-1)
2. Independence implies no correlation, while no correlation doesn't necessarily imply independence. However, under normal distribution it does. [↑](#footnote-ref-2)
3. The likelihood is the probability density function of all observations, i.e. the joint distribution, which in this case is equivalent to the multivariate normal distribution. [↑](#footnote-ref-3)