Regression :
Bayesian vs. Frequentist

11.23.2021

Hannah Do

0. Introduction

Bayesian **Frequentist**

Statistical **inference** methods that evaluate evidence for a hypothesis.

Inference from data can take many forms, but primary inferential aims will often be point estimation, to provide a “best guess” of an unknown parameter, and interval estimation, to produce ranges for unknown parameters that are supported by the data.

○. Introduction

Heads or Tails?

Probability of a Coin Flip

○. Introduction

Bayesian:

“Based on my knowledge and logic, it’s 50/50”
“It could be different for you though”

Frequentist:

“I don’t know yet.. give me that coin. Let me try”

1. Definition

Bayesian

The Bayesian approach begins by specifying a **prior** distribution over parameters that must be estimated.

The prior reflects the information known to the researcher without reference to the dataset on which the model is estimated.

In the Bayesian approach, the parameter θ is considered as a random variable with a certain probability distribution, referred to as the prior distribution, whose role is that of representing the experimenter's belief before observing the data.

The resulting distribution is referred to as the posterior distribution and summarizes the information in both the prior distribution and in the data.

$$p\left(\theta|y\right) = \frac{P(y|\theta)\ p(\theta)}{p(y)}$$

Random variable : A variable whose value is unknown or a function that assigns values to each of an experiment's outcomes.

https://www.sciencedirect.com/topics/engineering/bayesian-approach

Frequentist

The frequentist approach makes predictions on the underlying truths of the experiment using only data from the **current** experiment.

In the Frequentist approach, the parameter θ is considered an unknown, but a fixed quantity, only the information coming from the sampling data is relevant for inference. There is no random variable.

For example, MLE - Maximum Likelihood Estimates are used to compute the probability density as a frequentist approach.

The probability density of observing a single data point x, generated from a Gaussian distribution :

$$P(x;\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

https://www.sciencedirect.com/topics/mathematics/frequentist-approach

2. Applications in Regression

Based on the size of the dataset…

For small samples, the Bayesian approach with thoughtfully specified priors is often the only way to go because of the difficulty in obtaining well-calibrated frequentist intervals.

For medium to large samples, unless there is strong prior information that one wishes to incorporate, a robust frequentist approach using sandwich estimation (or quasi-likelihood if one has faith in the variance model) is very appealing since consistency is guaranteed under relatively mild conditions.

For highly complex models (e.g., with many random effects), a Bayesian approach is often the most convenient way to formulate the model, and computation under the Bayesian approach is the most straightforward.

2. Applications in Regression - BAYESIAN

Bayes Theorem :

$$p(A \mid B) = \frac{p(B \mid A)\, p(A)}{p(B)}$$

In practice, calculating the exact posterior distribution is computationally intractable for continuous values and so we turn to sampling methods such as Markov Chain Monte Carlo (MCMC) to draw samples from the dataset in order to approximate the posterior.

Monte Carlo refers to the general technique of drawing random samples, and Markov Chain means the next sample drawn is based only on the previous sample value.

The concept is that as we draw more samples, the approximation of the posterior will eventually converge on the true posterior distribution for the model parameters.

2. Applications in Regression - BAYESIAN

PyMC3 :

Well known library for probabilistic programming and Bayesian Inference

GLM :

Generalized Linear Models that helps construct Bayesian Linear Model in Python

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

β is the coefficient matrix (model parameters), X is the data matrix, and σ is the standard deviation. If we want to make a prediction for a new data point, we can find a **normal distribution** of estimated outputs by multiplying the model parameters by our data point to find the mean and using the standard deviation from the model parameters.

1. Build a formula relating the features to the target and decide on a prior distribution for the data likelihood

2. Sample from the parameter posterior distribution using MCMC

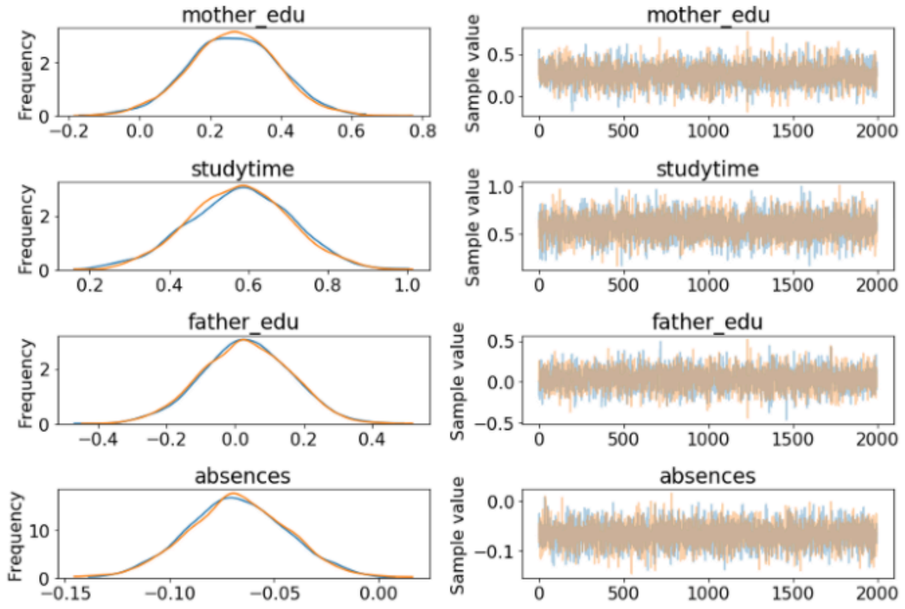
```
import pymc3 as pm

# Context for the model
with pm.Model() as normal_model:

    # The prior for the data likelihood is a Normal Distribution
    family = pm.glm.families.Normal()

    # Creating the model requires a formula and data (and optionally a family)
    pm.GLM.from_formula(formula, data = X_train, family = family)

    # Perform Markov Chain Monte Carlo sampling letting PyMC3 choose the algorithm
    normal_trace = pm.sample(draws=2000, chains = 2, tune = 500, njobs=-1)
```

https://towardsdatascience.com/bayesian-linear-regression-in-python-using-machine-learning-to-predict-student-grades-part-2-b72059a8ac7e

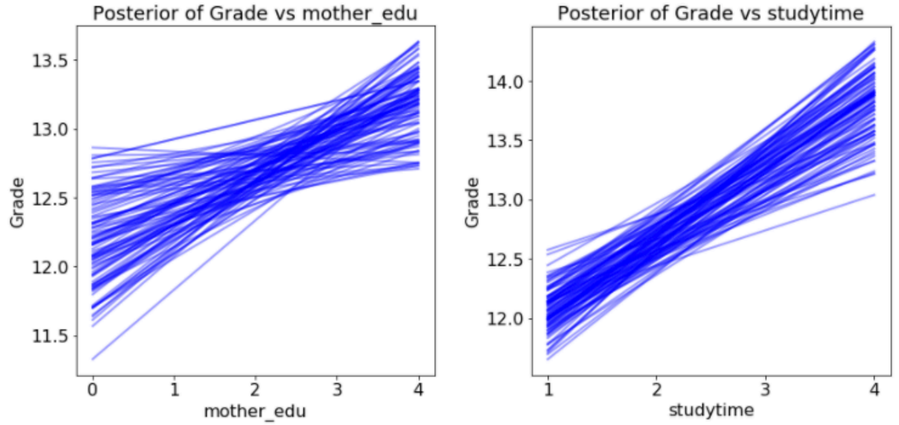
2. Applications in Regression - BAYESIAN

Regression Example:

Predicting a student's test score with multiple parameters



Model parameters are not point estimates but distributions. The mean of each distribution can be taken as the most likely estimate, but we also use the entire range of values to show we are uncertain about the true values.

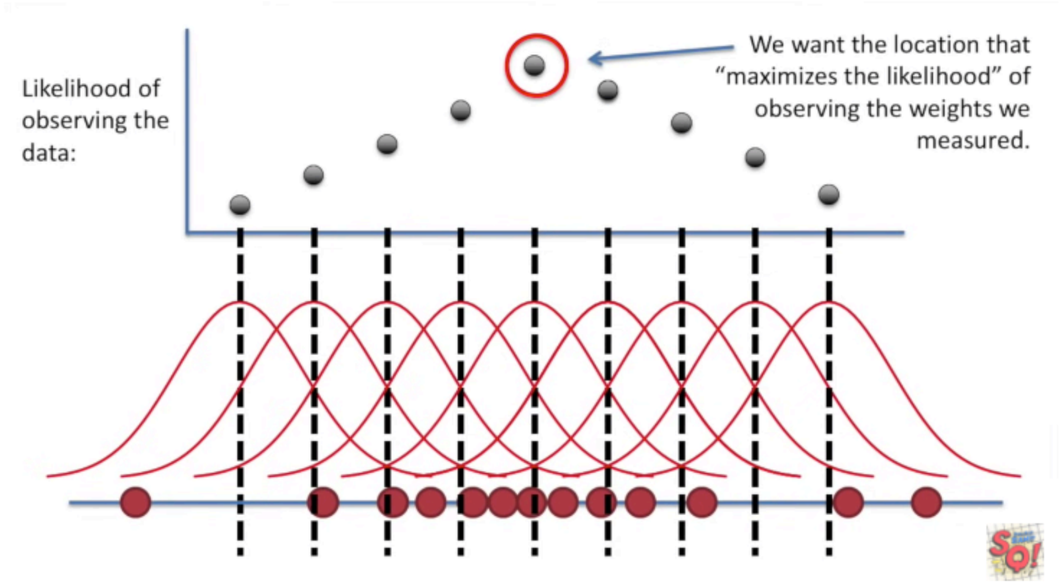


https://towardsdatascience.com/bayesian-linear-regression-in-python-using-machine-learning-to-predict-student-grades-part-2-b72059a8ac7e

2. Applications in Regression - FREQUENTIST

Under the frequentist approach, parameters and hypotheses are viewed as unknown but fixed (nonrandom) quantities, and consequently there is no possibility of making probability statements about these unknowns.

Likelihood - MLE Quasi-likelihood Sandwich Estimation Bootstrap Methods



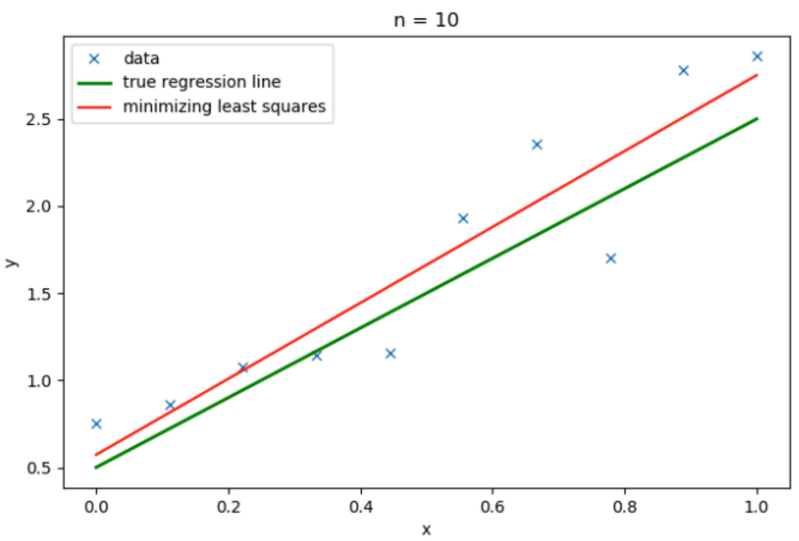
$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i \mid \boldsymbol{\theta}).$$

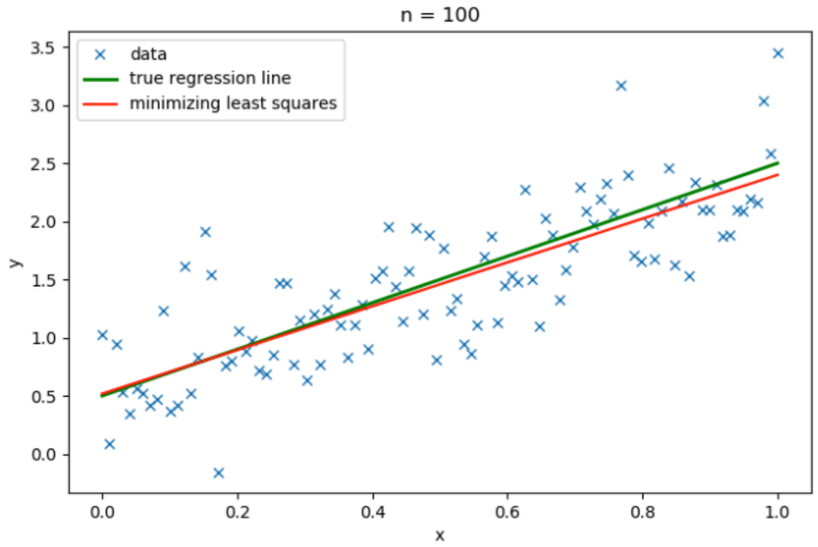
MLE, Maximum Likelihood Estimation:

Method of estimating the parameters of an assumed probability distribution, given some observed data. This is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable.

2. Applications in Regression - FREQUENTIST

Least squares method on different # of samples





OLS can be also identified as a frequentist approach, assuming we do not create any prior or posterior for the distribution.

We can see the regression line obtained minimizing least squares that provides a single regression line specific for the given data. If we do not have sufficient data, then we do not have a way of determining the uncertainty of the model parameters or the confidence attached to each parameter.

3. Conclusion

$$\hat{Y} = bX + a$$

Bayesian views parameter as a *random variable*, which means the value is NOT a single value, while Frequentists view parameters as a limited quantity.

From a predictive point of view, there is no significant difference between both approaches.

1. References

1. Probability concepts explained: Maximum likelihood estimation
   https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1

2. Bayesian and Frequentist Regression Methods (Springer) — Jon Wakefield

3. Bayesian Linear Regression in Python: Using Machine Learning to Predict Student Grades Part 2
   https://towardsdatascience.com/bayesian-linear-regression-in-python-using-machine-learning-to-predict-student-grades-part-2-b72059a8ac7e