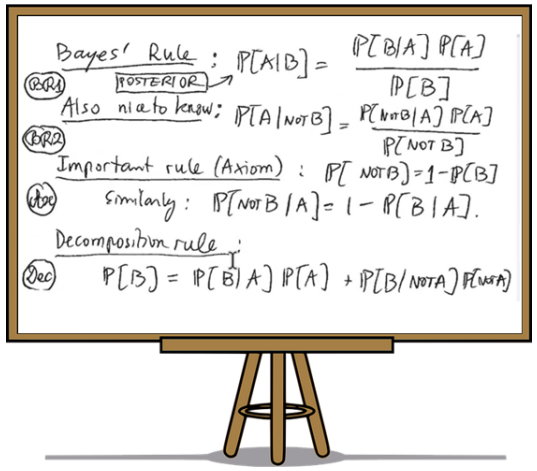
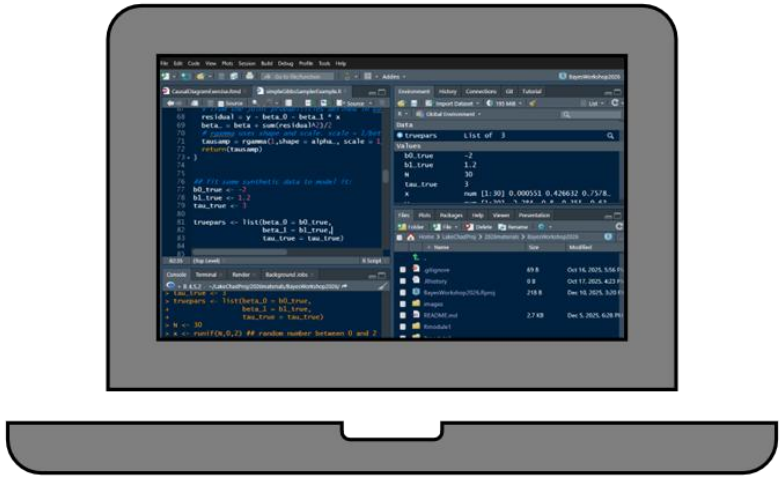
Bridging Theory and Practice Part 1: Introducing Bayesian inference with simple linear regression

Katherine Muller



Theory

Bayes Rule!

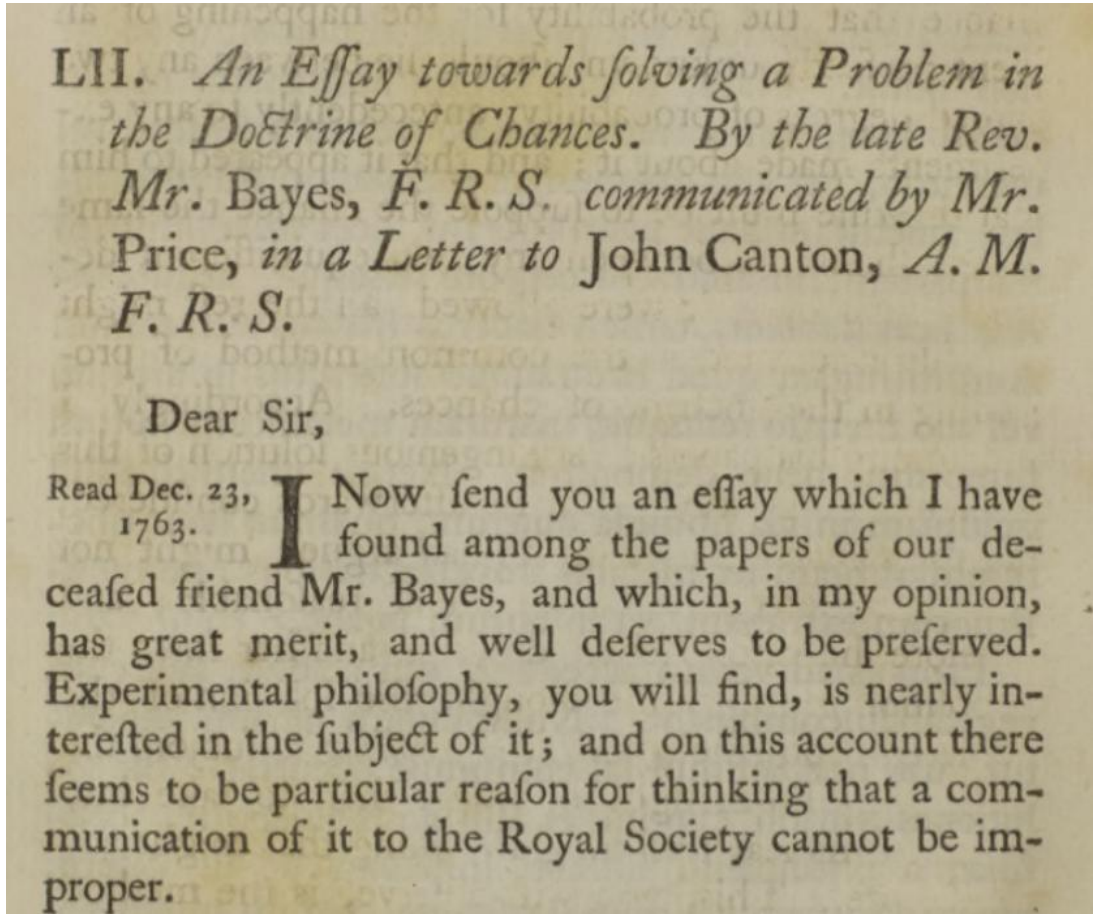


Computation

Outline

I. Origins of Bayesian inference
II. Simple linear regression example
III. Comparison with frequentist inference

Bayes' Theorem was first published in 1763



LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr.* Bayes, *F. R. S. communicated by Mr.* Price, *in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763.

I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

"Bayesian inference" came later

Bayesian Analysis (2006) 1, Number 1, pp. 1–40

When Did Bayesian Inference Become "Bayesian"?

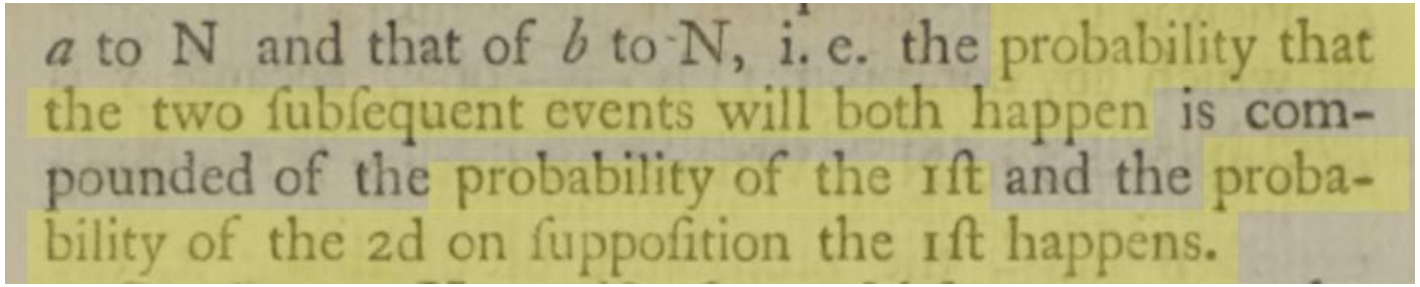
Stephen E. Fienberg*

Abstract. While Bayes' theorem has a 250-year history, and the method of inverse probability that flowed from it dominated statistical thinking into the twentieth century, the adjective "Bayesian" was not part of the statistical lexicon until relatively recently. This paper provides an overview of key Bayesian developments, beginning with Bayes' posthumously published 1763 paper and continuing up through approximately 1970, including the period of time when "Bayesian" emerged as the label of choice for those who advocated Bayesian methods.

Keywords: Bayes' Theorem; Classical statistical methods; Frequentist methods; Inverse probability; Neo-Bayesian revival; Stigler's Law of Eponymy; Subjective probability.

https://projecteuclid.org/journals/bayesian-analysis/volume-1/issue-1/When-did-Bayesian-inference-become-Bayesian/10.1214/06-BA101.pdf

What we call Bayes' Rule is elementary probability



a to N and that of *b* to N, i. e. the probability that the two ſubſequent events will both happen is compounded of the probability of the 1ſt and the probability of the 2d on ſuppoſition the 1ſt happens.

$$P(A,B) = P(A)\,P(B|A)$$

By the same logic:

$$P(A,B) = P(B)\,P(A|B)$$

Therefore:

Rearrange:

$$P(B)\,P(A|B) = P(A)\,P(B|A)$$

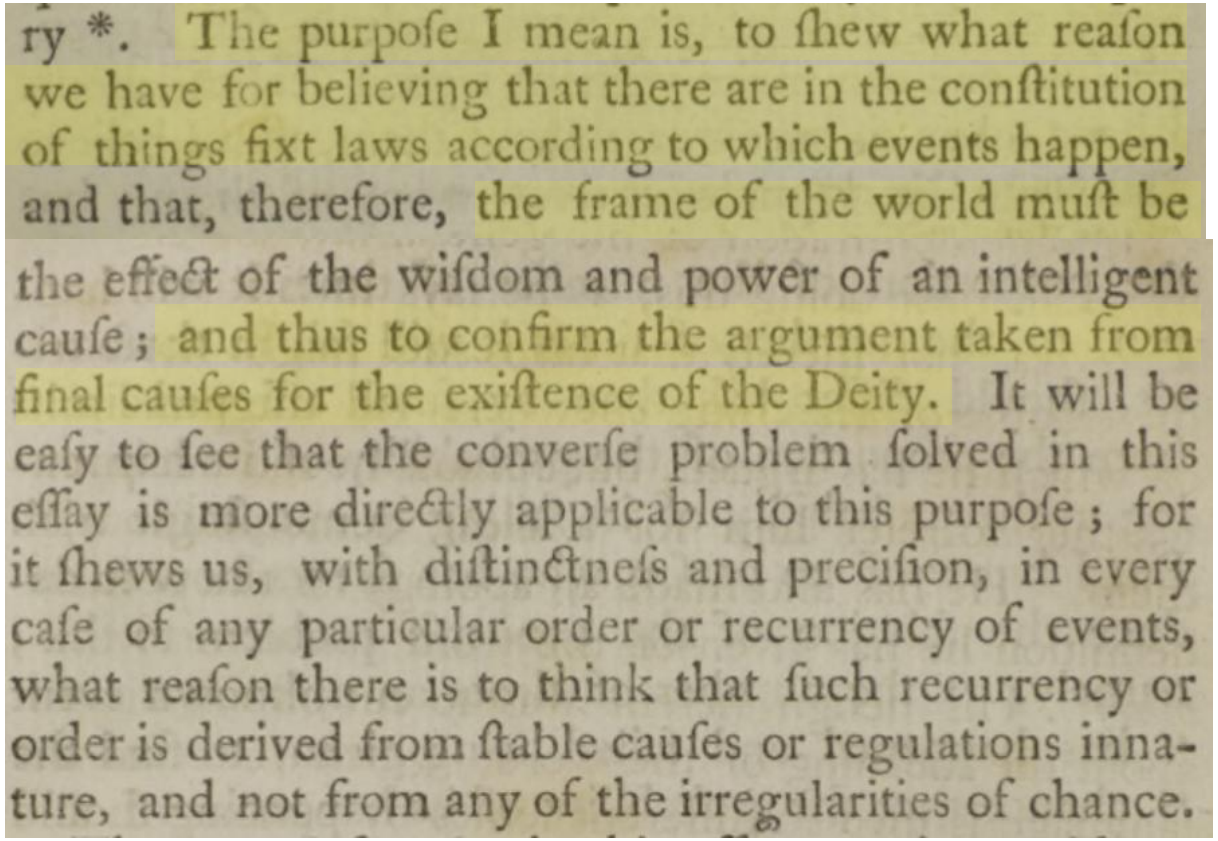
Bayes' Rule

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

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

Predict **unknown/unobservable** parameters based on **known/observable** data and prior understanding about parameters.

Fun fact: Bayes' friend Richard Price thought Bayes' theories confirmed the existence of God.



ry *. The purpoſe I mean is, to ſhew what reaſon we have for believing that there are in the conſtitution of things fixt laws according to which events happen, and that, therefore, the frame of the world muſt be the effect of the wiſdom and power of an intelligent cauſe; and thus to confirm the argument taken from final cauſes for the exiſtence of the Deity. It will be eaſy to ſee that the converſe problem ſolved in this eſſay is more directly applicable to this purpoſe; for it ſhews us, with diſtinctneſs and preciſion, in every caſe of any particular order or recurrency of events, what reaſon there is to think that ſuch recurrency or order is derived from ſtable cauſes or regulations inna-ture, and not from any of the irregularities of chance.

your very humble ſervant,
Newington-Green, Nov. 10, 1763.
Richard Price.

Ccc 2 SEC-

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

Predict **unknown/unobservable** parameters based on **known/observable** data and prior understanding about parameters.

Agroecology/sustainability science often involves estimating parameters from data

unknown/unobservable parameters		**known/observable** data
Disease resistance		Occurrence or severity of disease
Drought resilience		Yields in dry and wet years
Effect of management on soil health		Management history and soil health indicators

Estimating unknown parameters from data is not uniquely “Bayesian”

Part 2: Simple linear regression example

- Set up a model to explore theory and computation concepts
- Callback to causal inference and data story

Simple linear regression

Rainfall
mm/year

Rainfed rice yield
kg/ha

Yield = Function(Rainfall, Random)

Data story

- Rainfed rice is grown **on one farm** under optimal temperature, soil and management.
- Rice is planted and harvested **once per year every year** (no carryover).
- Rice yield is limited by **water availability through rainfall**.
- Rice yield varies randomly from year to year for reasons unrelated to rainfall

Simple linear regression

- Two continuous random variables, **X**, and **Y**
- A change in **X** is associated with a change in **Y**.

$$y = \underset{\text{intercept}}{\beta_0} + \underset{\text{slope}}{\beta_1 x} + \underset{\text{error}}{\varepsilon}$$

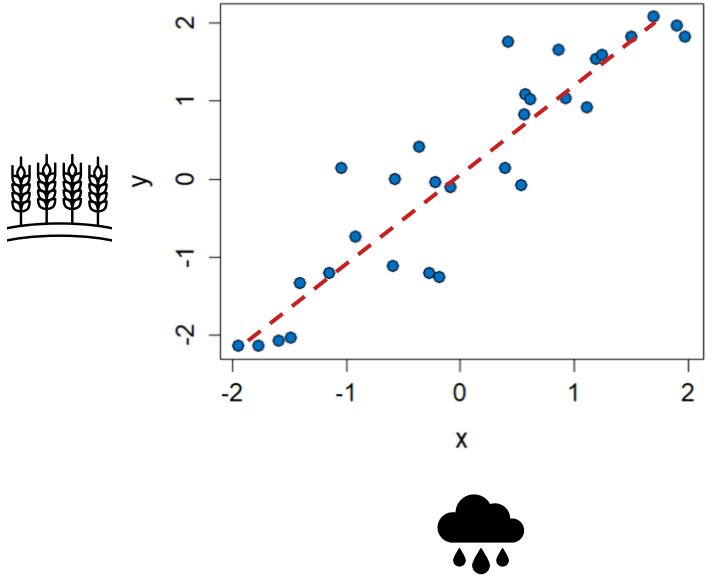
Error has Gaussian noise $\varepsilon \sim \mathcal{N}(0, 1/\tau)$

Yield = Function(Rainfall, Random)

Data processing: *standardize variables*

$$y = \frac{\boldsymbol{yield} - \text{mean}(\boldsymbol{yield})}{\text{sd}\,(\boldsymbol{yield})}$$

$$x = \frac{\boldsymbol{rainfall} - \text{mean}(\boldsymbol{rainfall})}{\text{sd}\,(\boldsymbol{rainfall})}$$



Precision (τ) is the reciprocal of variance

Data: *x* and *y*

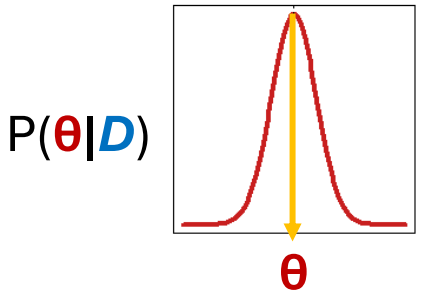
Parameters (*θ*): β_0, β_1, τ

Part 3: Frequentist vs. Bayesian inference

Overview

Frequentist inference

- Focus is on parameter **values**



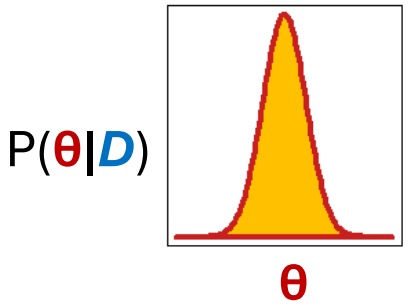
- Assumes **no prior understanding** about parameters

- **Uncertainty** about parameters must be estimated indirectly, using resampling or asymptotic theory.

With weak priors, Bayesian and frequentist methods give the same parameter estimates.

Bayesian inference

- Focus is on the parameter **distributions**



- **Incorporates prior understanding** (or lack thereof).

- **Uncertainty** about parameters is baked into the model

Frequentist Inference

Maximum Likelihood Estimation

Find **values** for each **parameter** that maximize the likelihood of the **data**

Likelihood function: $P(D|\theta)$

Probability of the **data** given **parameters**

Easily solvable with math! (for simple models)

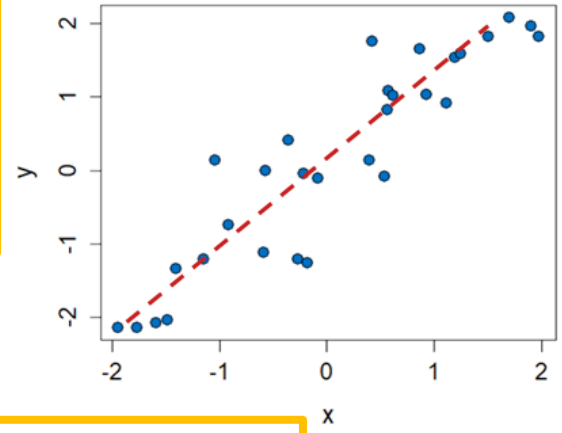
Simple linear regression:

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, 1/\tau)$$

$$p(y, x|\, \beta_0, \beta_1, \tau) = \prod_{i=1}^{n} \mathcal{N}(\underbrace{\beta_0 + \beta_1 x_i}_{\text{line}}, \underbrace{1/\tau}_{\text{scatter}})$$

Find the values of β_0, β_1, τ that maximize this function with the observed **x** and **y**.

Equivalent for β_0 and β_1



Ordinary Least Squares:
Find the values of β_0, β_1 that draw the best fit line, minimizing the distance between line and points—i.e., sum of squared residuals

Bayesian Inference

Update our prior understanding about the **distribution** of **parameters** based on the **data.**

Likelihood of the data given parameters

Prior understanding about parameters

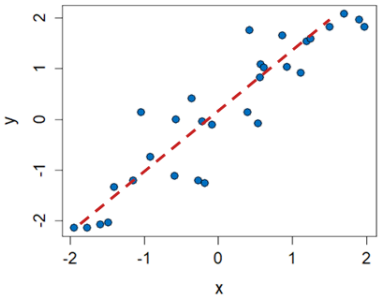
$$P(\boldsymbol{\theta}|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{D})}$$

Updated understanding about parameters (posterior)

Support over plausible values of the data: Normalizing constant so probability sums to 1

Simple linear regression:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, 1/\tau)$$



$$p(\beta_0, \beta_1, \tau \mid \boldsymbol{x}, \boldsymbol{y}) = \frac{p(\boldsymbol{x}, \boldsymbol{y}|\beta_0, \beta_1, \tau)p(\beta_0, \beta_1, \tau)}{p(\boldsymbol{x}, \boldsymbol{y})}$$

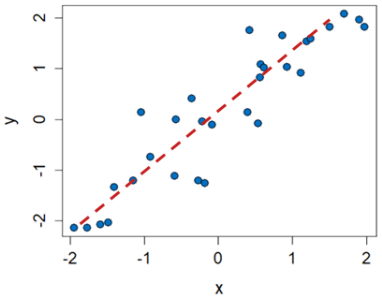
Find **most likely distributions** of β_0, β_1, τ based on observed **x** and **y**.

Bayesian Inference

Update our prior understanding about the **distribution** of **parameters** based on the **data.**

Simple linear regression:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, 1/\tau)$$



Specify prior distributions for parameters

$$P(\boldsymbol{\theta}) \quad \begin{matrix} \beta_0 \sim \\ \beta_1 \sim \\ \tau \sim \end{matrix} \quad \boxed{f(\boldsymbol{\theta}^*)}$$

intercept

slope

precision

hyperparameters

$$p(\beta_0, \beta_1, \tau \mid \boldsymbol{x}, \boldsymbol{y}) = \frac{p(\boldsymbol{x}, \boldsymbol{y} \mid \beta_0, \beta_1, \tau) p(\beta_0, \beta_1, \tau)}{p(\boldsymbol{x}, \boldsymbol{y})}$$

Find **most likely distributions** of β_0, β_1, τ based on observed **x** and **y**.

Choose priors based on:

- Analytical convenience
- How they represent your knowledge (or lack of knowledge) about your system

Bayesian Inference

Update our prior understanding about the **distribution** of **parameters** based on the **data.**

Simple linear regression:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, 1/\tau)$$

Specify prior distributions for parameters

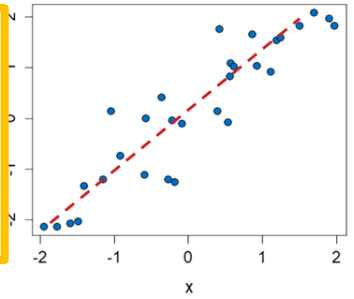
$$P(\boldsymbol{\theta})$$

$$\beta_0 \sim \mathcal{N}(\mu_0, (\mathrm{v}_0\tau)^{-1}) \text{ intercept}$$

$$\beta_1 \sim \mathcal{N}(\mu_1, (\mathrm{v}_1\tau)^{-1}) \text{ slope}$$

$$\tau \sim \mathrm{Gamma}(\mathrm{a}, \mathrm{b}) \text{ precision}$$

Prior uncertainty about the regression coefficients must be proportional to prediction uncertainty for the data



$$p(\beta_0, \beta_1, \tau \mid \boldsymbol{x}, \boldsymbol{y}) = \frac{p(\boldsymbol{x}, \boldsymbol{y} \mid \beta_0, \beta_1, \tau)\, p(\beta_0, \beta_1, \tau)}{p(\boldsymbol{x}, \boldsymbol{y})}$$

Find **most likely distributions** of β_0, β_1, τ based on observed **x** and **y**.

Choose priors based on:

- Analytical convenience

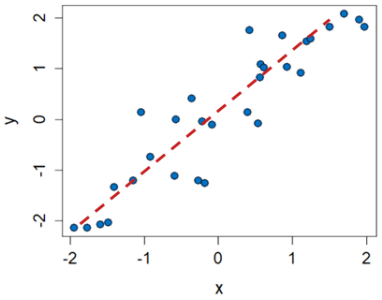
With very specific priors, the posterior can be solved analytically (without computer sampling)

Bayesian Inference

Update our prior understanding about the **distribution** of **parameters** based on the **data.**

Simple linear regression:

$$\boldsymbol{y} = \beta_0 + \beta_1 \boldsymbol{x} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(0, 1/\tau)$$



Specify prior distributions for parameters

$$P(\boldsymbol{\theta}) \quad \begin{aligned} \beta_0 &\sim \mathcal{N}(\mu_0, {\tau_0}^{-1}) \\ \beta_1 &\sim \mathcal{N}(\mu_1, {\tau_1}^{-1}) \\ \tau &\sim \text{Gamma}(a, b) \end{aligned}$$

intercept

slope

precision

Independent hyperparameters

$$p(\beta_0, \beta_1, \tau \mid \boldsymbol{x}, \boldsymbol{y}) = \frac{p(\boldsymbol{x}, \boldsymbol{y} \mid \beta_0, \beta_1, \tau) p(\beta_0, \beta_1, \tau)}{p(\boldsymbol{x}, \boldsymbol{y})}$$

Find **most likely distributions** of β_0, β_1, τ based on observed **x** and **y**.

The posterior can no longer be solved without computer sampling

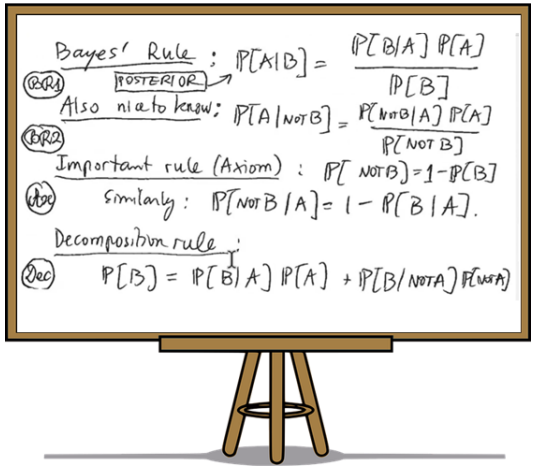
Recap

- Statistical inference (Bayesian and otherwise) involves estimating unknown or unobservable parameters from observable data (e.g., disease resistance from disease occurrence).
- Frequentist inference differs from Bayesian in that it:
  1. Focuses on parameter values instead of distributions
  2. Estimates parameters solely from the data (no priors)
  3. Is less well-suited to dealing with uncertainty about parameters
  4. Is typically easier and faster to compute.
- Frequentist and Bayesian methods produce similar estimates for parameter values when prior expectations are weak.
- Without computer sampling, Bayesian inference can only be performed on simple models with very restricted priors.

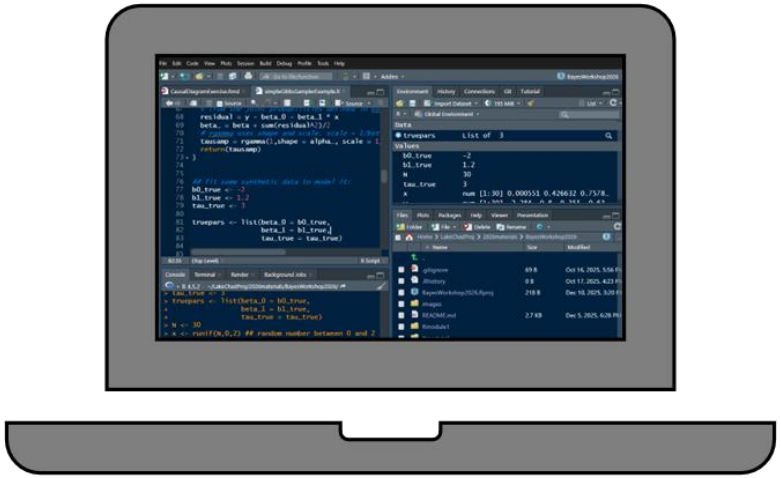
Frequentist vs. Bayesian inference: Computation

Aspect	Frequentist MLE	Bayesian
Goal	Find the best parameters that fit the data	Estimate the full distribution of parameters
Computation	One optimization run	Many repeated calculations (sampling or approximation)
Speed / Effort	Usually fast, moderate computation	Can be slow, especially for complex models
Determinism	Gives the same result each time	Results can vary due to randomness in sampling
Ease	Easy to do with standard tools	Harder; often requires specialized software

Next: Why computation makes Bayesian inference tractable



Theory



Computation