Bayesian Linear Regression

Vasilis Gkolemis

ATHENA RC — HUA

June 2025

A Synthetic Dataset

- We will use a synthetic dataset for demonstration.
- The dataset is generated using a simple linear function with added Gaussian noise.

$$y = wx + \beta + \epsilon$$

where:

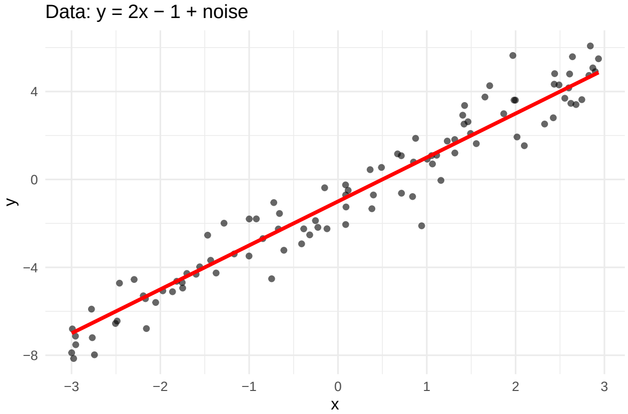
$w = 2$, $\beta = -1$ and $\epsilon \sim \mathcal{N}(0, \sigma^2), \sigma^2 = 1$.

$x \sim \mathcal{U}([-3, 3])$ is uniformly distributed.

y is the target variable.

A Synthetic Dataset

- The dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ consists of $N = 100$ samples.



Program

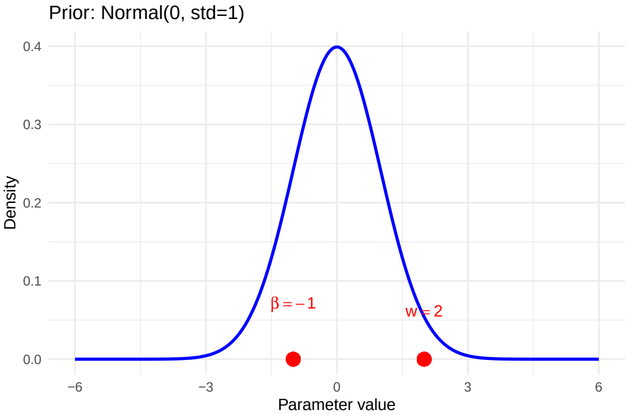
1. Prior
2. Posterior
3. Conjugate Priors

Prior

- In Bayesian linear regression, we assume a prior distribution over the model parameters.
- The prior reflects our beliefs about the parameters before observing any data.
- A common choice is a Gaussian prior or a uniform prior.
- The prior is denoted as $p(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (w, b)$ are the parameters of the linear model.

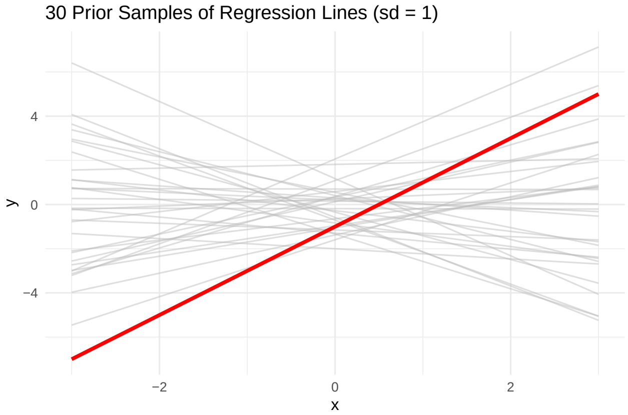
Prior on the parameters

Prior for slope w **and intercept** b**:** $w, b \sim \mathcal{N}(0, 1)$



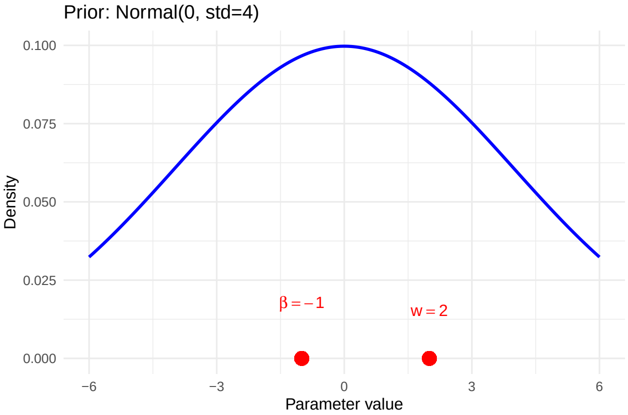
Prior on the parameters

Prior for slope w and intercept b: $w, b \sim \mathcal{N}(0, 1)$



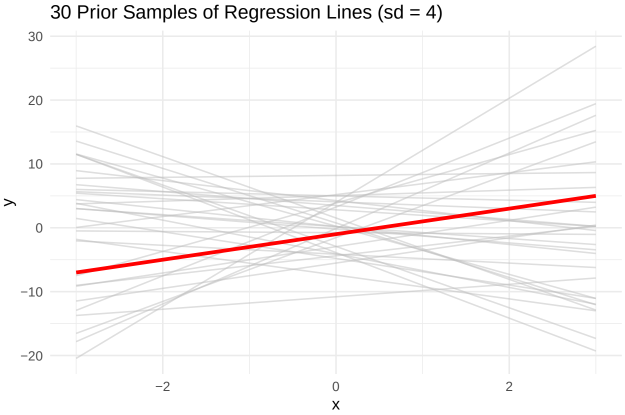
Prior with Larger Variance (Less Informative)

Increasing the prior variance expresses less certainty about w, b:
$w, b \sim \mathcal{N}(0, 4)$



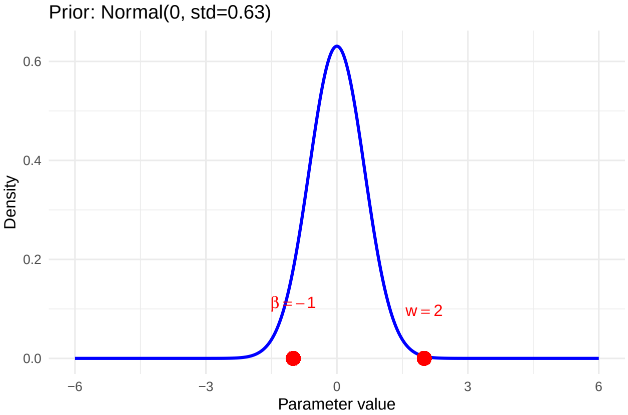
Prior with Larger Variance (Less Informative)

Increasing the prior variance expresses less certainty about w, b:
$w, b \sim \mathcal{N}(0, 4)$



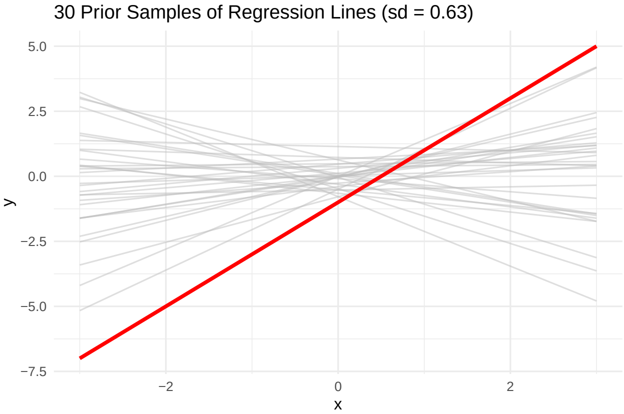
Incorrect Prior that Excludes True Parameters

A poorly chosen prior far from the truth, with low variance:
$w, b \sim \mathcal{N}(0, 0.4)$



Incorrect Prior that Excludes True Parameters

A poorly chosen prior far from the truth, with low variance:
$w, b \sim \mathcal{N}(0, 0.4)$



Program

1. Prior
2. Posterior
3. Conjugate Priors

Posterior

- The posterior distribution combines the prior and the likelihood of the observed data.
- It is computed using Bayes' theorem:

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \prod_{i=1}^{N} p(y^i|x^i, \boldsymbol{\theta})$$

- The posterior reflects our updated beliefs about the parameters after observing the data.

Posterior

We can compute the posterior in analytic form for linear regression with Gaussian noise:

- The posterior distribution is also Gaussian:

$$p(\boldsymbol{\theta}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- The posterior mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ can be computed as:

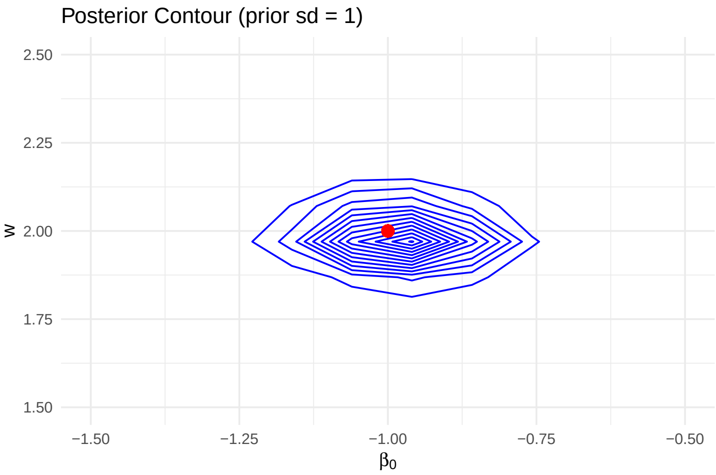
$$\boldsymbol{\mu} = (\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$$\boldsymbol{\Sigma} = (\boldsymbol{X}^T\boldsymbol{X} + \sigma^2\boldsymbol{I})^{-1}$$

- Skip the maths for now!

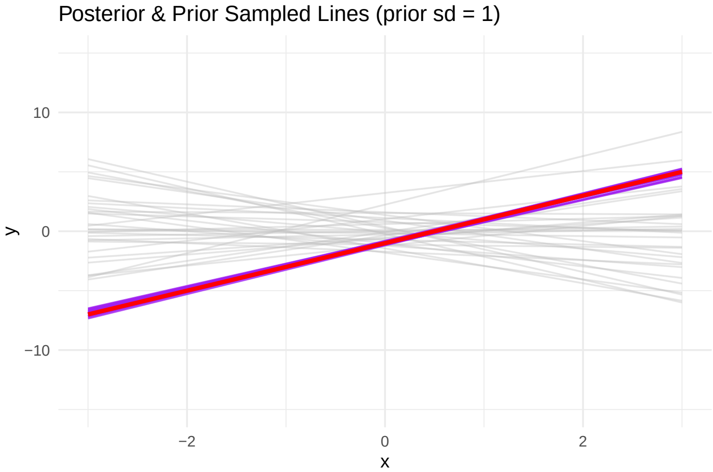
Posterior

Posterior when prior was $\mathcal{N}(0, \sigma^2 = 1)$: $w, b|\mathcal{D} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$



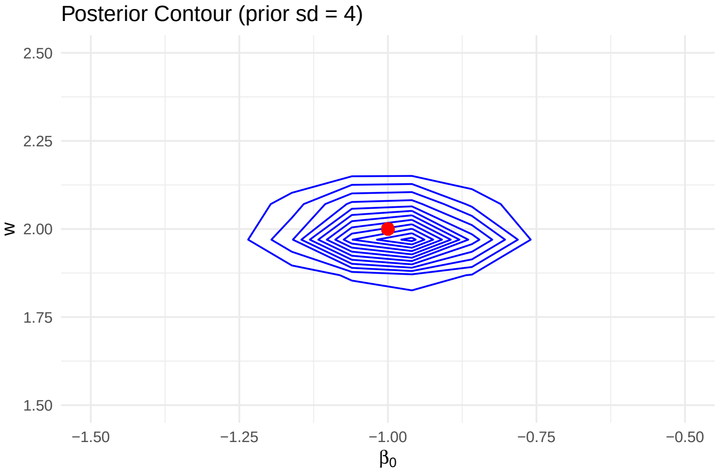
Posterior

Posterior when prior was $\mathcal{N}(0, \sigma^2 = 1)$:



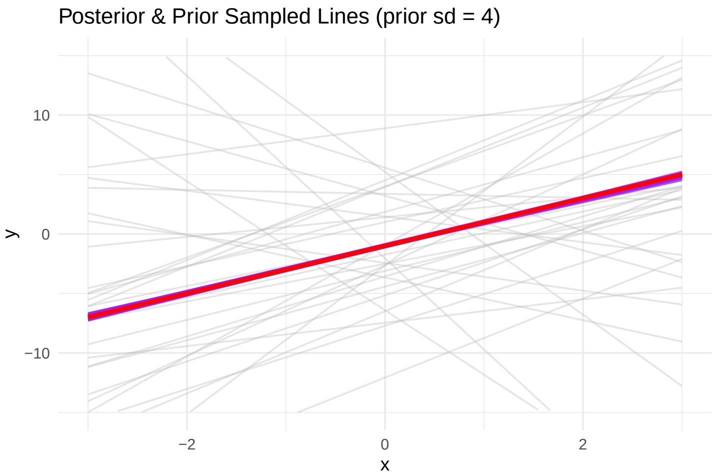
Posterior

Posterior when prior was $\mathcal{N}(0, \sigma^2 = 4)$:

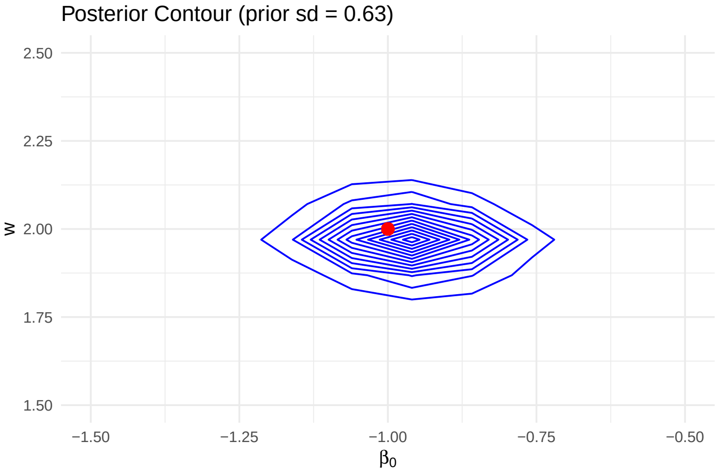


Posterior

Posterior when prior was $\mathcal{N}(0, \sigma^2 = 4)$:

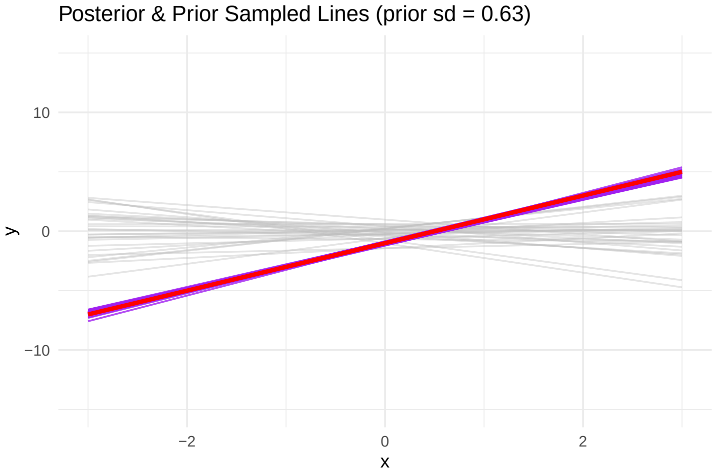


Posterior when prior was $\mathcal{N}(0, \sigma^2 = 0.4)$:



Posterior

Posterior when prior was $\mathcal{N}(0, \sigma^2 = 0.4)$:



Summary

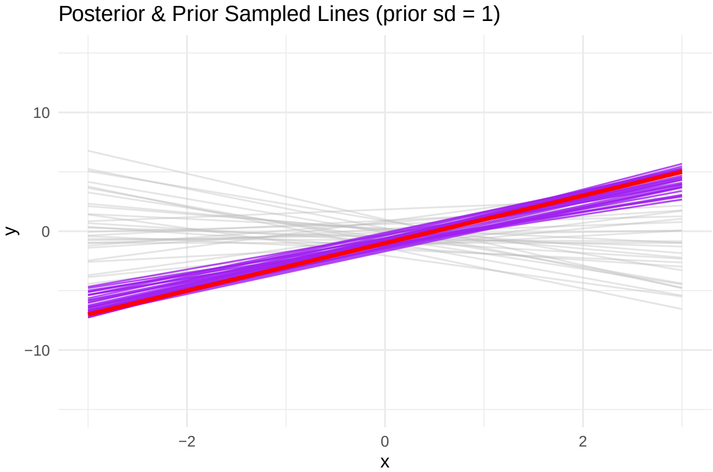
- Surprisingly (?), the posterior is accurate even when the prior does not match the true parameters.
- The posterior is influenced by the prior, but the data (likelihood) has a strong effect.
- However, do not be that overconfident!
- Our example is simple, and the prior is Gaussian which has infinite support.
- If the prior was a uniform distribution, i.e., $\mathcal{U}([-1, 1])$, with the true parameters outside this range, Bayesian linear regression would fail to learn the true parameters.

Summary

- Surprisingly (?), the posterior is accurate even when the prior does not match the true parameters.
- The posterior is influenced by the prior, but the data (likelihood) has a strong effect.
- However, do not be that overconfident!
- Our example is simple, and the prior is Gaussian which has infinite support.
- Even with fewer data points, $N = 3$, the posterior becomes worse if the prior is not informative enough.

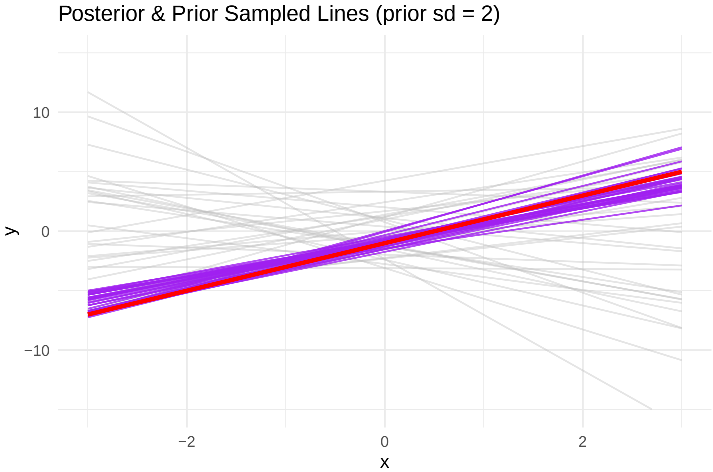
Posterior with $N = 5$

Posterior when prior was $\mathcal{N}(0, \sigma^2 = 1)$ and $N = 5$



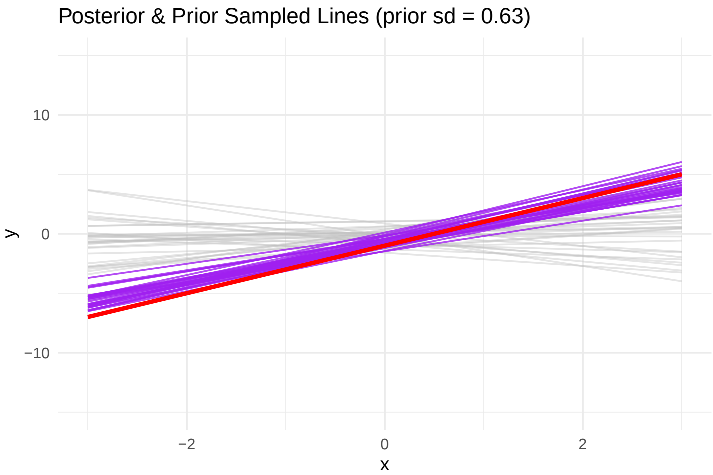
Posterior with $N = 5$

Posterior when prior was $\mathcal{N}(0, \sigma^2 = 4)$ and $N = 5$



Posterior with $N = 5$

Posterior when prior was $\mathcal{N}(0, \sigma^2 = 0.4)$ and $N = 5$



Program

Conjugate Priors: Motivation

- In Bayesian inference, we update our belief (the prior) after seeing data (via the likelihood) to get the posterior:

$$p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})$$

- In general, computing this posterior is hard—often requires numerical methods (e.g., MCMC, variational inference).
- But in some special cases, we get a **closed-form** posterior.
- These special cases arise when the **prior is conjugate to the likelihood**

Definition: Conjugate Prior

Definition

A prior is said to be **conjugate** to the likelihood if the posterior is in the same family as the prior.

- Example: Gaussian likelihood $+$ Gaussian prior $\Rightarrow$ Gaussian posterior
- This allows efficient inference—no need for numerical approximations.
- Conjugate priors are available for many common likelihoods:
  - Binomial likelihood $\rightarrow$ Beta prior
  - Poisson likelihood $\rightarrow$ Gamma prior
  - Gaussian likelihood $\rightarrow$ Gaussian prior (as we'll see!)

Conjugate Prior for Linear Regression

- Suppose a Bayesian linear regression model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$$

- Prior: $\boldsymbol{\theta} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$
- Likelihood: $p(\boldsymbol{y}|\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2 \boldsymbol{I})$
- Posterior is also Gaussian:

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with:

$$\boldsymbol{\mu} = (\boldsymbol{X}^T \boldsymbol{X} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{X}^T \boldsymbol{y}, \quad \boldsymbol{\Sigma} = (\boldsymbol{X}^T \boldsymbol{X} + \sigma^2 \boldsymbol{I})^{-1}$$

- Proofs can be found in many Bayesian textbooks, e.g., "Bayesian Data Analysis" by Gelman et al.
- Intuition: The product of Gaussian distributions is another Gaussian.

Why Conjugacy Matters

- **Fast, exact inference:** No sampling or approximation needed.
- **Analytic tractability:** Makes teaching, derivation, and understanding easier.
- **Limitations:**
  - Only available for limited combinations of priors and likelihoods.
  - Sometimes the conjugate prior may not reflect real prior beliefs well.
- In most real-world cases: we rely on approximate inference.
- But conjugacy gives insight into the Bayesian machinery in clean, solvable cases.

Summary

- What we have learned:
  - Bayesian linear regression allows us to incorporate prior beliefs about parameters.
  - The posterior distribution combines prior and likelihood, updating our beliefs after observing data.
  - Conjugate priors provide a powerful framework for efficient inference in Bayesian models.
- What comes next:
  - The world is not linear.
  - Bayesian linear regression is a starting point, but real-world data often requires more complex models.
  - In the next two lectures, we will explore how to perform Bayesian inference in more complex models:
    - Non-linear regression
    - Bayesian neural networks