3.4 Prediction: Posterior Predictive Distributions

From parameter inference to forecasting the next data points

4BIC

Inference vs prediction: what changes?

Inference (what we've done so far)

You observe data $Y_1, \ldots, Y_N$ and build a model with parameter θ.

$$\text{Prior } \pi(\theta) \quad \longrightarrow \quad \text{Posterior } \pi(\theta \mid y).$$

Inference vs prediction: what changes?

Inference (what we've done so far)

You observe data $Y_1, \ldots, Y_N$ and build a model with parameter θ.

$$\text{Prior } \pi(\theta) \quad \longrightarrow \quad \text{Posterior } \pi(\theta \mid y).$$

Prediction (the extra question)

Now ask:

"What are the next data points likely to be?"

Formally, introduce a future observation Z and want the distribution

$$\pi(z \mid y).$$

Inference vs prediction: what changes?

Inference (what we've done so far)

You observe data $Y_1, \ldots, Y_N$ and build a model with parameter θ.

$$\text{Prior } \pi(\theta) \quad \longrightarrow \quad \text{Posterior } \pi(\theta \mid y).$$

Prediction (the extra question)

Now ask:

"What are the next data points likely to be?"

Formally, introduce a future observation Z and want the distribution

$$\pi(z \mid y).$$

Key message

Inference learns *parameters*. Prediction learns *future outcomes*, and must propagate parameter uncertainty into the forecast.

Definition 3.3: Posterior predictive distribution

Set-up

- Observe data y under a model parameterised by θ.
- Choose a prior $\pi(\theta)$ and compute posterior $\pi(\theta \mid y)$.
- Interested in a future observation Z generated by the same model.

Definition 3.3: Posterior predictive distribution

Set-up

- Observe data y under a model parameterised by θ.
- Choose a prior $\pi(\theta)$ and compute posterior $\pi(\theta \mid y)$.
- Interested in a future observation Z generated by the same model.

Posterior predictive (mixture form)

$$\pi(z \mid y) = \int \pi(z \mid \theta)\, \pi(\theta \mid y)\, d\theta.$$

Definition 3.3: Posterior predictive distribution

Set-up

- Observe data y under a model parameterised by θ.
- Choose a prior $\pi(\theta)$ and compute posterior $\pi(\theta \mid y)$.
- Interested in a future observation Z generated by the same model.

Posterior predictive (mixture form)

$$\pi(z \mid y) = \int \pi(z \mid \theta)\, \pi(\theta \mid y)\, d\theta.$$

One-sentence interpretation

It is the model's prediction $\pi(z \mid \theta)$ *averaged over* the posterior uncertainty in θ.

The historical intuition: Bayes and the beanbag

(Story / intuition)

Think of Bayes repeatedly throwing a beanbag over his shoulder and recording where it lands.

- Past observations: y (where it landed previously).
- Unknown "parameter": θ (the underlying tendency / bias).
- Next throw: Z (where will the next beanbag land?).

The historical intuition: Bayes and the beanbag

(Story / intuition)

Think of Bayes repeatedly throwing a beanbag over his shoulder and recording where it lands.

- Past observations: y (where it landed previously).
- Unknown "parameter": θ (the underlying tendency / bias).
- Next throw: Z (where will the next beanbag land?).

What does Bayes really want?

Not only "what is θ?", but also:

"Given what I've seen, where will the *next* one land?"

That is exactly $\pi(z \mid y)$.

Thomas Bayes: the "beanbag" (billiard table) experiment

The story (why this appears in Bayes' original essay)

Imagine a flat table.

- A ball/beanbag is tossed (or rolled) onto the table and lands at some position.
- Bayes repeats this many times and records where the toss lands.
- He wants to use past tosses to predict the next one.

Thomas Bayes: the "beanbag" (billiard table) experiment

The story (why this appears in Bayes' original essay)

Imagine a flat table.

- A ball/beanbag is tossed (or rolled) onto the table and lands at some position.
- Bayes repeats this many times and records where the toss lands.
- He wants to use past tosses to predict the next one.

Modern translation

- Past data: $Y_1, \ldots, Y_n$ (previous landing outcomes)
- Unknown "bias" parameter: θ (how likely the toss is to land in a region)
- Future outcome: Z (next landing outcome)

What the beanbag story teaches

Two uncertainties

- **Parameter uncertainty:** we do not know θ exactly.
- **Outcome randomness:** even if θ were known, the next toss is still random.

What the beanbag story teaches

Two uncertainties

- **Parameter uncertainty:** we do not know θ exactly.
- **Outcome randomness:** even if θ were known, the next toss is still random.

Posterior predictive combines them

$$\pi(z \mid \text{data}) = \int \pi(z \mid \theta)\, \pi(\theta \mid \text{data})\, d\theta.$$

It averages the next-outcome model over all plausible θ values.

What the beanbag story teaches

Two uncertainties

- **Parameter uncertainty:** we do not know θ exactly.
- **Outcome randomness:** even if θ were known, the next toss is still random.

Posterior predictive combines them

$$\pi(z \mid \text{data}) = \int \pi(z \mid \theta)\, \pi(\theta \mid \text{data})\, d\theta.$$

It averages the next-outcome model over all plausible θ values.

Why it feels reasonable

If the data strongly pin down θ, the predictive becomes sharp. If the data are weak, the predictive stays diffuse (more cautious).

What do the two factors mean?

$\pi(z \mid \theta)$

Sampling model for new data.

Fix a candidate parameter value θ.
Ask: "If θ were true, how likely is each z?"

What do the two factors mean?

$\pi(z \mid \theta)$

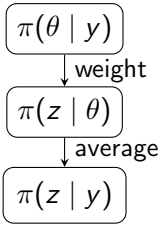
Sampling model for new data.

Fix a candidate parameter value θ.
Ask: "If θ were true, how likely is each z?"

$\pi(\theta \mid y)$

Posterior over parameters.

After seeing y, which θ values are plausible, and how plausible?



What do the two factors mean?

$\pi(z \mid \theta)$

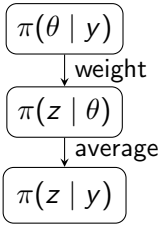
Sampling model for new data.

Fix a candidate parameter value θ.
Ask: "If θ were true, how likely is each z?"

$\pi(\theta \mid y)$

Posterior over parameters.

After seeing y, which θ values are plausible, and how plausible?



Mixture picture

For each θ you get a different predictive curve $\pi(z \mid \theta)$. The posterior predictive is the weighted average of all those curves.

A "mental algorithm" for prediction

Two-stage sampling view (very important intuition)

To draw a predictive sample $Z^\star \sim \pi(z \mid y)$:

1. Sample $\theta^\star \sim \pi(\theta \mid y)$ (pick a plausible world)
2. Sample $Z^\star \sim \pi(z \mid \theta^\star)$ (simulate the next outcome)

A "mental algorithm" for prediction

Two-stage sampling view (very important intuition)

To draw a predictive sample $Z^\star \sim \pi(z \mid y)$:

1. Sample $\theta^\star \sim \pi(\theta \mid y)$ (pick a plausible world)
2. Sample $Z^\star \sim \pi(z \mid \theta^\star)$ (simulate the next outcome)

Why this is helpful

Even if the integral is messy, simulation often remains simple.

Why "constants don't matter" breaks here

Earlier: posterior up to proportionality

For parameter inference we often write

$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta)$$

and say "ignore the normalising constant".

Why "constants don't matter" breaks here

Earlier: posterior up to proportionality

For parameter inference we often write

$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta)$$

and say "ignore the normalising constant".

But now we want actual probabilities for z

The posterior predictive $\pi(z \mid y)$ is a **proper** density/pmf in z. If you want to compute numbers like $\Pr(Z \leq 8 \mid y)$, you need the **correct normalisation**.

Why "constants don't matter" breaks here

Earlier: posterior up to proportionality

For parameter inference we often write

$$\pi(\theta \mid y) \propto \pi(y \mid \theta)\pi(\theta)$$

and say "ignore the normalising constant".

But now we want actual probabilities for z

The posterior predictive $\pi(z \mid y)$ is a **proper** density/pmf in z. If you want to compute numbers like $\Pr(Z \leq 8 \mid y)$, you need the **correct normalisation**.

Practical consequence

You must keep track of constants (or use a known conjugacy result / a computer). That is why the algebra can feel "painful" in predictive calculations.

Example 3.6: late coursework submissions (set-up)

Problem

Last year: $y = 3$ late submissions out of $n_0 = 30$ students.
This year: $n = 30$ students. Predict how many will be late.

Example 3.6: late coursework submissions (set-up)

Problem

Last year: $y = 3$ late submissions out of $n_0 = 30$ students.
This year: $n = 30$ students. Predict how many will be late.

Model

Let θ be the probability a student submits late.

$$Y \mid \theta \sim \text{Bin}(n_0, \theta), \qquad Z \mid \theta \sim \text{Bin}(n, \theta).$$

We assume the same θ carries over from last year to this year.

Example 3.6: late coursework submissions (set-up)

Problem

Last year: $y = 3$ late submissions out of $n_0 = 30$ students.
This year: $n = 30$ students. Predict how many will be late.

Model

Let θ be the probability a student submits late.

$$Y \mid \theta \sim \mathrm{Bin}(n_0, \theta), \qquad Z \mid \theta \sim \mathrm{Bin}(n, \theta).$$

We assume the same θ carries over from last year to this year.

Prior (Uniform / "non-informative" on $[0, 1]$)

$$\theta \sim \mathrm{Unif}[0, 1] = \mathrm{Beta}(1, 1).$$

Step 1: posterior for θ (conjugacy)

Beta–Binomial conjugacy

If

$$Y \mid \theta \sim \mathrm{Bin}(n_0, \theta), \qquad \theta \sim \mathrm{Beta}(\alpha, \beta),$$

then

$$\theta \mid y \sim \mathrm{Beta}(\alpha + y,\ \beta + n_0 - y).$$

Step 1: posterior for θ (conjugacy)

Beta–Binomial conjugacy

If

$$Y \mid \theta \sim \mathrm{Bin}(n_0, \theta), \qquad \theta \sim \mathrm{Beta}(\alpha, \beta),$$

then

$$\theta \mid y \sim \mathrm{Beta}(\alpha + y,\ \beta + n_0 - y).$$

Here ($n_0 = 30$, $y = 3$, $\alpha = \beta = 1$)

$$\theta \mid y \sim \mathrm{Beta}(1 + 3,\ 1 + 30 - 3) = \mathrm{Beta}(4, 28).$$

Step 1: posterior for θ (conjugacy)

Beta–Binomial conjugacy

If

$$Y \mid \theta \sim \mathrm{Bin}(n_0, \theta), \qquad \theta \sim \mathrm{Beta}(\alpha, \beta),$$

then

$$\theta \mid y \sim \mathrm{Beta}(\alpha + y,\ \beta + n_0 - y).$$

Here ($n_0 = 30$, $y = 3$, $\alpha = \beta = 1$)

$$\theta \mid y \sim \mathrm{Beta}(1 + 3,\ 1 + 30 - 3) = \mathrm{Beta}(4, 28).$$

Posterior mean (for intuition)

$$\mathbb{E}[\theta \mid y] = \frac{4}{4 + 28} = \frac{1}{8} = 0.125.$$

Step 2: set up the posterior predictive integral

We want

$$\pi(z \mid y) = \int_0^1 \pi(z \mid \theta)\, \pi(\theta \mid y)\, d\theta.$$

Step 2: set up the posterior predictive integral

We want

$$\pi(z \mid y) = \int_0^1 \pi(z \mid \theta)\, \pi(\theta \mid y)\, d\theta.$$

Write each factor

$$\pi(z \mid \theta) = \binom{30}{z} \theta^z (1-\theta)^{30-z},$$

$$\pi(\theta \mid y) = \frac{1}{\mathrm{B}(4, 28)} \theta^{4-1}(1-\theta)^{28-1} = \frac{1}{\mathrm{B}(4, 28)} \theta^3 (1-\theta)^{27}.$$

Step 2: set up the posterior predictive integral

We want

$$\pi(z \mid y) = \int_0^1 \pi(z \mid \theta)\, \pi(\theta \mid y)\, d\theta.$$

Write each factor

$$\pi(z \mid \theta) = \binom{30}{z} \theta^z (1-\theta)^{30-z},$$

$$\pi(\theta \mid y) = \frac{1}{\mathrm{B}(4, 28)} \theta^{4-1}(1-\theta)^{28-1} = \frac{1}{\mathrm{B}(4, 28)} \theta^3 (1-\theta)^{27}.$$

Combine like terms

$$\pi(z \mid y) = \binom{30}{z} \frac{1}{\mathrm{B}(4, 28)} \int_0^1 \theta^{z+3} (1-\theta)^{57-z}\, d\theta.$$

The "cheat step": recognise a Beta integral

Recall the Beta function

$$\mathrm{B}(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}\, dt.$$

The "cheat step": recognise a Beta integral

Recall the Beta function

$$\mathrm{B}(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}\, dt.$$

Match exponents

$$\theta^{z+3} = \theta^{(z+4)-1}, \qquad (1-\theta)^{57-z} = (1-\theta)^{(58-z)-1}.$$

So

$$\int_0^1 \theta^{z+3}(1-\theta)^{57-z}\, d\theta = \mathrm{B}(z+4, 58-z).$$

The "cheat step": recognise a Beta integral

Recall the Beta function

$$\mathrm{B}(a, b) = \int_0^1 t^{a-1}(1-t)^{b-1}\, dt.$$

Match exponents

$$\theta^{z+3} = \theta^{(z+4)-1}, \qquad (1-\theta)^{57-z} = (1-\theta)^{(58-z)-1}.$$

So

$$\int_0^1 \theta^{z+3}(1-\theta)^{57-z}\, d\theta = \mathrm{B}(z+4, 58-z).$$

Posterior predictive pmf (Beta–Binomial)

$$\boxed{\Pr(Z = z \mid y) = \binom{30}{z} \frac{\mathrm{B}(z+4, 58-z)}{\mathrm{B}(4, 28)}, \qquad z = 0, 1, \ldots, 30.}$$

Predictive mean (fast route, no summing)

Law of iterated expectation

$$\mathbb{E}[Z \mid y] = \mathbb{E}_{\theta|y}\big[\mathbb{E}[Z \mid \theta]\big].$$

For $Z \mid \theta \sim \mathrm{Bin}(30, \theta)$, $\mathbb{E}[Z \mid \theta] = 30\theta$, hence

$$\mathbb{E}[Z \mid y] = 30\,\mathbb{E}[\theta \mid y] = 30 \cdot \frac{4}{4+28} = 30 \cdot \frac{1}{8} = 3.75.$$

Predictive mean (fast route, no summing)

Law of iterated expectation

$$\mathbb{E}[Z \mid y] = \mathbb{E}_{\theta|y}\big[\mathbb{E}[Z \mid \theta]\big].$$

For $Z \mid \theta \sim \mathrm{Bin}(30, \theta)$, $\mathbb{E}[Z \mid \theta] = 30\theta$, hence

$$\mathbb{E}[Z \mid y] = 30\,\mathbb{E}[\theta \mid y] = 30 \cdot \frac{4}{4+28} = 30 \cdot \frac{1}{8} = 3.75.$$

Interpretation

Given last year's data, we expect about **3.75** students to hand in late this year.

Computing the pmf in R (Beta via Gamma)

Beta function identity

$$\mathrm{B}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

Computing the pmf in R (Beta via Gamma)

Beta function identity

$$\mathrm{B}(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}.$$

```
pmf:  choose(n,z) * B(z+alpha, n-z+beta) / B(alpha,beta) num <- gamma(z + alpha) *
gamma(n - z + beta) * gamma(alpha + beta) den <- gamma(alpha) * gamma(beta) *
gamma(n + alpha + beta)
out <- choose(n, z) * (num / den) return(out)
```

Check it is a valid pmf + plot

```
sum(ppd)  should be 1 (up to numerical error)
plot(z, ppd, xlab = "z", ylab = "Posterior predictive mass", main = "Posterior
predictive (Beta-Binomial)")
```

Check it is a valid pmf + plot

```
sum(ppd)  should be 1 (up to numerical error)
plot(z, ppd, xlab = "z", ylab = "Posterior predictive mass", main = "Posterior
predictive (Beta-Binomial)")
```

Sanity checks you should always do

- Non-negativity: $\Pr(Z = z \mid y) \geq 0$ for all z.
- Sums to 1: $\sum_{z=0}^{n} \Pr(Z = z \mid y) = 1$.
- Shape makes sense: mode near the expected value.

Predictive CDF, intervals, and a probability statement

```
CDF cdf <- cumsum(ppd) cbind(z, cdf)
Probability up to 8 late submissions sum(ppd[z <= 8])
```

Predictive CDF, intervals, and a probability statement

```
CDF cdf <- cumsum(ppd) cbind(z, cdf)
Probability up to 8 late submissions sum(ppd[z <= 8])
```

Interpreting outputs

- $\mathbb{E}[Z \mid y] = 3.75$ means: **expected number** of late submissions is 3.75.
- $\Pr(Z \leq 8 \mid y) \approx 0.954$ means: **about a 95.4% chance** of up to 8 late submissions.

Bayesian Linear Model: Matrix Form

- We assume the covariates $\{x_i\}_{i=1}^n$ are **fixed** and the noise variance σ^2 is **known**.
- A compact representation of the linear model is

$$Y = X\beta + \varepsilon,$$

where

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \qquad X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \qquad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

$$\varepsilon \sim \mathcal{N}\left(0,\ \sigma^2 I_{n\times n}\right).$$

Illustration: Stock Market Prediction (Linear Regression)

Setup (supervised learning view)

We want to predict the **next-day price** from the **previous 10 days**.

- Let $p = 10$ (use 10 historical days as features).
- For each training example (window) i:

$$x_i = \begin{pmatrix} s_{i,1} \\ s_{i,2} \\ \vdots \\ s_{i,10} \end{pmatrix} \in \mathbb{R}^{10}, \qquad y_i = s_{i,11} \in \mathbb{R},$$

where $s_{i,t}$ denotes the observed stock price on day t within window i (e.g. closing price).

What is a "window" (10-day lookback)?

Idea

A **window** is one training example made by taking a **contiguous chunk** of the time series:

$$\underbrace{s_t,\ s_{t+1},\ \ldots,\ s_{t+9}}_{\text{inputs (10 days)}} \longrightarrow \underbrace{s_{t+10}}_{\text{target (next day)}} .$$

We then **slide** this window forward one day to create many examples.

day	t	$t+1$	$t+2$	$t+3$	$t+4$	$t+5$	$t+6$	$t+7$	$t+8$	$t+9$	$t+10$
window i			features	$x_i = (s_t, \ldots, s_{t+9})$						label	$y_i = s_{t+10}$

- Window i starts at some time index t; then $x_i \in \mathbb{R}^{10}$ and $y_i \in \mathbb{R}$.
- Next example uses the next start time: $x_{i+1} = (s_{t+1}, \ldots, s_{t+10})$, $y_{i+1} = s_{t+11}$.

.

Matrix form

Stack the n training windows:

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,10} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,10} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n,1} & s_{n,2} & \cdots & s_{n,10} \end{pmatrix}.$$

Interpretation of β (linear predictor)

$$y_i = x_i^\top \beta + \varepsilon_i \quad \Rightarrow \quad \widehat{y}_i = x_i^\top \beta = \beta_1 s_{i,1} + \cdots + \beta_{10} s_{i,10}.$$

Here $\beta \in \mathbb{R}^{10}$ assigns a **weight** to each lag (day) in the 10-day history.

Example 3.7: Bayesian linear model (setup)

For $i = 1, \ldots, n$:

$$Y_i = x_i^\top \beta + \varepsilon_i, \qquad \varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2),$$

where $x_i \in \mathbb{R}^p$ are fixed covariates and $\beta \in \mathbb{R}^p$ are unknown coefficients.

Example 3.7: Bayesian linear model (setup)

For $i = 1, \ldots, n$:

$$Y_i = x_i^\top \beta + \varepsilon_i, \qquad \varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2),$$

where $x_i \in \mathbb{R}^p$ are fixed covariates and $\beta \in \mathbb{R}^p$ are unknown coefficients.

Matrix form

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_{n \times n}).$$

Example 3.7: Bayesian linear model (setup)

For $i = 1, \ldots, n$:

$$Y_i = x_i^\top \beta + \varepsilon_i, \qquad \varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2),$$

where $x_i \in \mathbb{R}^p$ are fixed covariates and $\beta \in \mathbb{R}^p$ are unknown coefficients.

Matrix form

$$Y = X\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_{n\times n}).$$

Two tasks

1. Posterior distribution: $\pi(\beta \mid Y)$
2. Posterior predictive for a new covariate x':

$$Y' = x'^\top \beta + \varepsilon', \quad \varepsilon' \sim \mathcal{N}(0, \sigma^2)$$

Find $\pi(Y' \mid Y, x')$.

Prior and likelihood

Likelihood

Because $Y \mid \beta \sim \mathcal{N}(X\beta, \sigma^2 I)$,

$$\pi(Y \mid \beta) \propto \exp\left(-\frac{1}{2\sigma^2}\|Y - X\beta\|_2^2\right).$$

Prior and likelihood

Likelihood

Because $Y \mid \beta \sim \mathcal{N}(X\beta, \sigma^2 I)$,

$$\pi(Y \mid \beta) \propto \exp\left(-\frac{1}{2\sigma^2}\|Y - X\beta\|_2^2\right).$$

Prior (Gaussian)

Assume

$$\beta \sim \mathcal{N}(0, c^2 I_{p\times p}).$$

So

$$\pi(\beta) \propto \exp\left(-\frac{1}{2c^2}\|\beta\|_2^2\right).$$

.

Why this is convenient

Normal prior + Normal likelihood $\Rightarrow$ Normal posterior (conjugacy).

Posterior distribution (Gaussian update / Bayes rule for Gaussians)

Posterior is multivariate Normal

$$\boxed{\beta \mid Y \sim \mathcal{N}(\mu, \Sigma)}$$

with

$$\boxed{\mu = (X^\top X + \tfrac{\sigma^2}{c^2} I)^{-1} X^\top Y}, \qquad \boxed{\Sigma = \left(\tfrac{1}{\sigma^2} X^\top X + \tfrac{1}{c^2} I\right)^{-1}}.$$

Posterior distribution (Gaussian update / Bayes rule for Gaussians)

Posterior is multivariate Normal

$$\boxed{\beta \mid Y \sim \mathcal{N}(\mu, \Sigma)}$$

with

$$\boxed{\mu = (X^\top X + \tfrac{\sigma^2}{c^2} I)^{-1} X^\top Y}, \qquad \boxed{\Sigma = \left(\tfrac{1}{\sigma^2} X^\top X + \tfrac{1}{c^2} I\right)^{-1}}.$$

Interpretation

- $X^\top X$ measures how informative the design is.
- σ^2 controls observation noise (less noise $\Rightarrow$ sharper posterior).
- c^2 controls prior strength (smaller c^2 $\Rightarrow$ more shrinkage toward 0).

Connection to Ridge Regression

OLS estimator (frequentist)

If $X^\top X$ is invertible (typically $n > p$),

$$\hat{\beta}_{\text{OLS}} = (X^\top X)^{-1} X^\top Y = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2.$$

Connection to Ridge Regression

OLS estimator (frequentist)

If $X^\top X$ is invertible (typically $n > p$),

$$\hat{\beta}_{\text{OLS}} = (X^\top X)^{-1} X^\top Y = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2.$$

Ridge estimator

For $\lambda > 0$,

$$\hat{\beta}_\lambda^{\text{ridge}} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\} = (X^\top X + \lambda I)^{-1} X^\top Y.$$

Connection to Ridge Regression

OLS estimator (frequentist)

If $X^\top X$ is invertible (typically $n > p$),

$$\hat{\beta}_{\mathrm{OLS}} = (X^\top X)^{-1} X^\top Y = \arg\min_{\beta \in \mathbb{R}^p} \|Y - X\beta\|_2^2.$$

Ridge estimator

For $\lambda > 0$,

$$\hat{\beta}_\lambda^{\mathrm{ridge}} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \|Y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\} = (X^\top X + \lambda I)^{-1} X^\top Y.$$

Bayesian link (Gaussian prior)

Assume the likelihood $y \mid \beta \sim \mathcal{N}(X\beta, \sigma^2 I)$ and a prior $\beta \sim \mathcal{N}(0,\, c^2 I)$. Then the posterior mean equals the ridge solution:

$$\mathbb{E}[\beta \mid y] = (X^\top X + \lambda I)^{-1} X^\top Y, \qquad \lambda = \frac{\sigma^2}{c^2}.$$

MAP viewpoint

Posterior mode (MAP)

$$\beta_{\text{MAP}} = \arg\max_{\beta} \ \pi(\beta \mid Y) = \arg\max_{\beta} \ \pi(Y \mid \beta)\pi(\beta).$$

MAP viewpoint

Posterior mode (MAP)

$$\beta_{\text{MAP}} = \arg\max_{\beta} \ \pi(\beta \mid Y) = \arg\max_{\beta} \ \pi(Y \mid \beta)\pi(\beta).$$

Take logs and flip sign

$$\begin{aligned}\beta_{\text{MAP}} &= \arg\max_{\beta} \{\log \pi(Y \mid \beta) + \log \pi(\beta)\} \\ &= \arg\min_{\beta} \left\{ \frac{1}{\sigma^2} \|Y - X\beta\|_2^2 + \frac{1}{c^2} \|\beta\|_2^2 \right\}.\end{aligned}$$

Multiply by σ^2:

$$\beta_{\text{MAP}} = \arg\min_{\beta} \left\{ \|Y - X\beta\|_2^2 + \frac{\sigma^2}{c^2} \|\beta\|_2^2 \right\}.$$

MAP viewpoint

Posterior mode (MAP)

$$\beta_{\mathrm{MAP}} = \arg\max_{\beta} \pi(\beta \mid Y) = \arg\max_{\beta} \pi(Y \mid \beta)\pi(\beta).$$

MAP viewpoint

Posterior mode (MAP)

$$\beta_{\mathrm{MAP}} = \arg\max_{\beta} \pi(\beta \mid Y) = \arg\max_{\beta} \pi(Y \mid \beta)\pi(\beta).$$

For a multivariate Normal posterior, the mean and mode coincide, so $\mu = \beta_{\mathrm{MAP}}$.

Posterior predictive for a new input x'

New observation model

$$Y' = x'^{\top}\beta + \varepsilon', \qquad \varepsilon' \sim \mathcal{N}(0, \sigma^2), \quad \varepsilon' \perp \beta.$$

Posterior predictive for a new input x'

New observation model

$$Y' = x'^{\top}\beta + \varepsilon', \qquad \varepsilon' \sim \mathcal{N}(0, \sigma^2), \quad \varepsilon' \perp \beta.$$

Posterior predictive definition

$$\pi(y' \mid y, x') = \int \pi(y' \mid \beta, x')\, \pi(\beta \mid y)\, d\beta.$$

Posterior predictive for a new input x'

New observation model

$$Y' = x'^{\top}\beta + \varepsilon', \qquad \varepsilon' \sim \mathcal{N}(0, \sigma^2), \quad \varepsilon' \perp \beta.$$

Posterior predictive definition

$$\pi(y' \mid y, x') = \int \pi(y' \mid \beta, x')\, \pi(\beta \mid y)\, d\beta.$$

Key simplification

$\pi(y' \mid \beta, x')$ is Normal and $\pi(\beta \mid y)$ is Normal, so the posterior predictive is also Normal.

Posterior predictive via "linear combination of Gaussians"

Write

$$Y' = x'^{\top}\beta + \varepsilon'.$$

Given Y, we have $\beta \mid Y \sim \mathcal{N}(\mu, \Sigma)$ and $\varepsilon' \sim \mathcal{N}(0, \sigma^2)$ independent.

Posterior predictive via "linear combination of Gaussians"

Write

$$Y' = x'^{\top}\beta + \varepsilon'.$$

Given Y, we have $\beta \mid Y \sim \mathcal{N}(\mu, \Sigma)$ and $\varepsilon' \sim \mathcal{N}(0, \sigma^2)$ independent.

Mean

$$\mathbb{E}[Y' \mid Y, x'] = \mathbb{E}[x'^{\top}\beta \mid Y] + \mathbb{E}[\varepsilon'] = x'^{\top}\mu.$$

Posterior predictive via "linear combination of Gaussians"

Write

$$Y' = x'^{\top}\beta + \varepsilon'.$$

Given Y, we have $\beta \mid Y \sim \mathcal{N}(\mu, \Sigma)$ and $\varepsilon' \sim \mathcal{N}(0, \sigma^2)$ independent.

Mean

$$\mathbb{E}[Y' \mid Y, x'] = \mathbb{E}[x'^{\top}\beta \mid Y] + \mathbb{E}[\varepsilon'] = x'^{\top}\mu.$$

Variance

$$\mathrm{Var}(Y' \mid Y, x') = \mathrm{Var}(x'^{\top}\beta \mid Y) + \mathrm{Var}(\varepsilon') = x'^{\top}\Sigma x' + \sigma^2.$$

Posterior predictive via "linear combination of Gaussians"

Write

$$Y' = x'^{\top}\beta + \varepsilon'.$$

Given Y, we have $\beta \mid Y \sim \mathcal{N}(\mu, \Sigma)$ and $\varepsilon' \sim \mathcal{N}(0, \sigma^2)$ independent.

Mean

$$\mathbb{E}[Y' \mid Y, x'] = \mathbb{E}[x'^{\top}\beta \mid Y] + \mathbb{E}[\varepsilon'] = x'^{\top}\mu.$$

Variance

$$\mathrm{Var}(Y' \mid Y, x') = \mathrm{Var}(x'^{\top}\beta \mid Y) + \mathrm{Var}(\varepsilon') = x'^{\top}\Sigma x' + \sigma^2.$$

Posterior predictive distribution

$$\boxed{Y' \mid Y, x' \sim \mathcal{N}\Big(x'^{\top}\mu,\ x'^{\top}\Sigma x' + \sigma^2\Big).}$$

Interpretation: what contributes to predictive uncertainty?

Two sources

$$\mathrm{Var}(Y' \mid y, x') = \underbrace{x'^{\top}\Sigma x'}_{\text{parameter uncertainty}} + \underbrace{\sigma^2}_{\text{irreducible noise}} .$$

Interpretation: what contributes to predictive uncertainty?

Two sources

$$\mathrm{Var}(Y' \mid y, x') = \underbrace{x'^{\top}\Sigma x'}_{\text{parameter uncertainty}} + \underbrace{\sigma^2}_{\text{irreducible noise}} .$$

Intuition

- If you have tons of data, Σ shrinks and $x'^{\top}\Sigma x'$ gets small.
- Even with infinite data, you cannot beat the observation noise σ^2.

Three common computation routes

1) Closed form (conjugacy)

- Beta–Binomial $\Rightarrow$ Beta–Binomial predictive
- Normal–Normal $\Rightarrow$ Normal predictive
- Gaussian linear model (with known σ^2) $\Rightarrow$ Normal predictive

Three common computation routes

1) Closed form (conjugacy)

- Beta–Binomial $\Rightarrow$ Beta–Binomial predictive
- Normal–Normal $\Rightarrow$ Normal predictive
- Gaussian linear model (with known σ^2) $\Rightarrow$ Normal predictive

2) Monte Carlo (works very generally)

Approximate the integral in (3.1) by:

$$\pi(z \mid y) \approx \frac{1}{M} \sum_{m=1}^{M} \pi(z \mid \theta^{(m)}), \quad \theta^{(m)} \sim \pi(\theta \mid y).$$

Or simulate $Z^{(m)}$ by two-stage sampling.

Three common computation routes

1) Closed form (conjugacy)

- Beta–Binomial $\Rightarrow$ Beta–Binomial predictive
- Normal–Normal $\Rightarrow$ Normal predictive
- Gaussian linear model (with known σ^2) $\Rightarrow$ Normal predictive

2) Monte Carlo (works very generally)

Approximate the integral in (3.1) by:

$$\pi(z \mid y) \approx \frac{1}{M} \sum_{m=1}^{M} \pi(z \mid \theta^{(m)}), \quad \theta^{(m)} \sim \pi(\theta \mid y).$$

Or simulate $Z^{(m)}$ by two-stage sampling.

3) Approximate posteriors

.

3) Approximate posteriors

Laplace approximation, variational Bayes, etc. give an approximate $\pi(\theta \mid y)$, then plug into (3.1).

Posterior predictive sampling algorithm

1. Draw $\theta^{(1)}, \ldots, \theta^{(M)} \sim \pi(\theta \mid y)$.
2. For each draw, simulate $Z^{(m)} \sim \pi(z \mid \theta^{(m)})$.
3. Approximate:

$$\mathbb{E}[Z \mid y] \approx \frac{1}{M} \sum_{m=1}^{M} Z^{(m)}, \qquad \Pr(Z \leq k \mid y) \approx \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{Z^{(m)} \leq k\}.$$

Monte Carlo

Posterior predictive sampling algorithm

1. Draw $\theta^{(1)}, \ldots, \theta^{(M)} \sim \pi(\theta \mid y)$.
2. For each draw, simulate $Z^{(m)} \sim \pi(z \mid \theta^{(m)})$.
3. Approximate:

$$\mathbb{E}[Z \mid y] \approx \frac{1}{M} \sum_{m=1}^{M} Z^{(m)}, \qquad \Pr(Z \leq k \mid y) \approx \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\{Z^{(m)} \leq k\}.$$

Why this is conceptually perfect

It literally follows the Bayesian generative story:

$$\theta \mid y \;\rightarrow\; Z \mid \theta.$$

Summary: what to remember for prediction

Posterior predictive = mixture over posterior

$$\pi(z \mid y) = \int \pi(z \mid \theta)\, \pi(\theta \mid y)\, d\theta.$$

Summary: what to remember for prediction

Posterior predictive = mixture over posterior

$$\pi(z \mid y) = \int \pi(z \mid \theta)\, \pi(\theta \mid y)\, d\theta.$$

Interpretation you can always write in words

- $\pi(z \mid \theta)$: how the model generates new data if θ were known.
- $\pi(\theta \mid y)$: uncertainty about θ after seeing data.
- integral: average predictions over all plausible θ values.

Summary: what to remember for prediction

Posterior predictive = mixture over posterior

$$\pi(z \mid y) = \int \pi(z \mid \theta)\, \pi(\theta \mid y)\, d\theta.$$

Interpretation you can always write in words

- $\pi(z \mid \theta)$: how the model generates new data if θ were known.
- $\pi(\theta \mid y)$: uncertainty about θ after seeing data.
- integral: average predictions over all plausible θ values.

Example 3.6

Beta prior + Binomial likelihood $\Rightarrow$ Beta posterior, and Binomial + Beta $\Rightarrow$ Beta–Binomial predictive.

Summary: what you must remember

Posterior predictive (core formula)

$$\pi(z \mid y) = \int \pi(z \mid \theta)\, \pi(\theta \mid y)\, d\theta.$$

Summary: what you must remember

Posterior predictive (core formula)

$$\pi(z \mid y) = \int \pi(z \mid \theta)\, \pi(\theta \mid y)\, d\theta.$$

Common special cases

- Beta prior + Binomial likelihood $\Rightarrow$ Beta posterior, Beta–Binomial posterior predictive.
- Gaussian linear model with Gaussian prior $\Rightarrow$ Gaussian posterior and Gaussian posterior predictive.

Summary: what you must remember

Posterior predictive (core formula)

$$\pi(z \mid y) = \int \pi(z \mid \theta)\, \pi(\theta \mid y)\, d\theta.$$

Common special cases

- Beta prior + Binomial likelihood $\Rightarrow$ Beta posterior, Beta–Binomial posterior predictive.
- Gaussian linear model with Gaussian prior $\Rightarrow$ Gaussian posterior and Gaussian posterior predictive.

Interpretation

"The posterior predictive averages the sampling model for new data over the posterior uncertainty in the parameters."

Non-informative priors: the idea and the snag

Definition 3.4 (informal)

A prior is "non-informative" if it places equal weight across plausible values of θ. On a bounded interval, this suggests:

$$\theta \sim \text{Unif}[a, b].$$

Non-informative priors: the idea and the snag

Definition 3.4 (informal)

A prior is "non-informative" if it places equal weight across plausible values of θ. On a bounded interval, this suggests:

$$\theta \sim \text{Unif}[a, b].$$

But on $\mathbb{R}$ this becomes tricky

There is no proper uniform distribution on the entire real line. So "flat" priors are approximations (e.g. very large-variance Normal).

Non-informative priors: the idea and the snag

Definition 3.4 (informal)

A prior is "non-informative" if it places equal weight across plausible values of θ. On a bounded interval, this suggests:

$$\theta \sim \text{Unif}[a, b].$$

But on $\mathbb{R}$ this becomes tricky

There is no proper uniform distribution on the entire real line. So "flat" priors are approximations (e.g. very large-variance Normal).

Practical takeaway

Uniform(0,1) is genuinely simple for probabilities. For unbounded parameters, "non-informative" needs more care.

Unintended consequence: non-informative is not invariant

Example: reparameterisation

Let $\theta \sim \text{Unif}(0, 1)$ with density $\pi(\theta) = 1$. Define $\phi = \theta^2$.

Unintended consequence: non-informative is not invariant

Example: reparameterisation

Let $\theta \sim \text{Unif}(0,1)$ with density $\pi(\theta) = 1$. Define $\phi = \theta^2$.

Change of variables

$\theta = \sqrt{\phi}$ and

$$\left|\frac{d\theta}{d\phi}\right| = \frac{1}{2\sqrt{\phi}}.$$

So

$$\pi_\phi(\phi) = \pi_\theta(\sqrt{\phi})\left|\frac{d\theta}{d\phi}\right| = 1 \cdot \frac{1}{2\sqrt{\phi}}, \qquad 0 < \phi < 1.$$

Unintended consequence: non-informative is not invariant

Example: reparameterisation

Let $\theta \sim \mathrm{Unif}(0,1)$ with density $\pi(\theta) = 1$. Define $\phi = \theta^2$.

Change of variables

$\theta = \sqrt{\phi}$ and

$$\left|\frac{d\theta}{d\phi}\right| = \frac{1}{2\sqrt{\phi}}.$$

So

$$\pi_\phi(\phi) = \pi_\theta(\sqrt{\phi})\left|\frac{d\theta}{d\phi}\right| = 1 \cdot \frac{1}{2\sqrt{\phi}}, \qquad 0 < \phi < 1.$$

Meaning

Flat in θ is **not** flat in ϕ. A "non-informative" prior depends on the parameterisation.

Picture intuition for the reparameterisation issue

Why does $\phi = \theta^2$ bias toward small values?

For $\theta \in (0, 1)$, squaring makes numbers smaller:

$$0.3^2 = 0.09, \quad 0.8^2 = 0.64.$$

So lots of θ values get "squeezed" near $\phi = 0$.

Picture intuition for the reparameterisation issue

Why does $\phi = \theta^2$ bias toward small values?

For $\theta \in (0, 1)$, squaring makes numbers smaller:

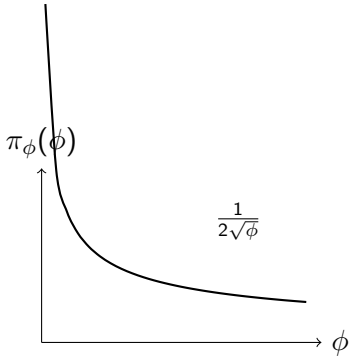
$$0.3^2 = 0.09, \quad 0.8^2 = 0.64.$$

So lots of θ values get "squeezed" near $\phi = 0$.

Density behaviour

$$\pi_\phi(\phi) = \frac{1}{2\sqrt{\phi}}$$

blows up near $\phi = 0$, meaning ϕ is much more likely to be small.



Picture intuition for the reparameterisation issue

Why does $\phi = \theta^2$ bias toward small values?

For $\theta \in (0, 1)$, squaring makes numbers smaller:

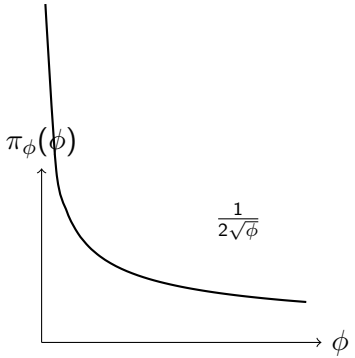
$$0.3^2 = 0.09, \quad 0.8^2 = 0.64.$$

So lots of θ values get "squeezed" near $\phi = 0$.

Density behaviour

$$\pi_\phi(\phi) = \frac{1}{2\sqrt{\phi}}$$

blows up near $\phi = 0$, meaning ϕ is much more likely to be small.



Preview