SSY316, Hand-in 2

Group: *27*
Student: *Jahanvi B Dinesh*
Student: *Sumukh S Moudghalya*

Exercise 1

Part 1

The given linear regression model is:

$$y = w_1x_1 + w_2x_2 + \epsilon, \quad \epsilon \sim N(0, 5)$$

We need to find $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ using the maximum likelihood approach.

The observed data is:

$$X = \begin{bmatrix} 3 & -1 \\ 4 & 2 \\ 2 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}.$$

Using the maximum likelihood approach for a Gaussian noise model, the weights w can be obtained by minimizing the squared error:

$$\hat{w} = (X^TX)^{-1}X^Ty$$

Results:

$$\hat{w}_{\text{MLE}} = \begin{bmatrix} 0.52 \\ -0.44 \end{bmatrix}$$

Part 2

Using the prior distribution:

$$p(w) = N\left(w \mid \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.2 & 0 \\ 0 & 0.2 \end{bmatrix}\right),$$

and the likelihood:

$$p(y|w) = N(y \mid Xw, 5I),$$

The posterior distribution is:

$$p(w|y) = N(w \mid m_N, S_N),$$

where: $S_N^{-1} = S_0^{-1} + \frac{1}{\sigma^2}X^TX$; $m_N = S_N\left(S_0^{-1}m_0 + \frac{1}{\sigma^2}X^Ty\right)$.

Results:

$$m_N = \begin{bmatrix} 0.2246 \\ -0.0185 \end{bmatrix}, \quad S_N = \begin{bmatrix} 0.0954 & -0.0215 \\ -0.0215 & 0.1662 \end{bmatrix}$$

Part 3

- MLE ignores prior information and directly estimates the weights to best fit the data, yielding $\hat{w}_{\text{MLE}} = \begin{bmatrix} 0.52 \\ -0.44 \end{bmatrix}$.
- Bayesian Approach incorporates the prior, resulting in $m_N = \begin{bmatrix} 0.2246 \\ -0.0185 \end{bmatrix}$, which is more conservative (closer to the prior mean $m_0 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$).
- The Bayesian approach also provides uncertainty information via the covariance matrix S_N.

Exercise 2

Part 1

Probabilistic Linear Regression Model,

The merit-value (denoted as y) is modeled as a linear function of the spare-time activities:

$$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + \epsilon,$$

where:

- x_1, x_2, x_3, x_4 represent the normalized scores for reading books, playing computer games, participating in sports, and hanging out with friends.
- w_0 is the intercept term.
- $\epsilon \sim N(0, \sigma^2)$ is Gaussian noise accounting for unobserved factors.

The prior beliefs are:

- 1. Reading books (x_1): Expected to explain up to 20 points.
- 2. Other activities (x_2, x_3, x_4): Expected to explain up to 10 points each.
- 3. Unexplained variance: Expected to account for at least 20 points.

This leads to the probabilistic model:

$$p(y|w, \sigma^2) = N(y; Xw, \sigma^2 I),$$

where:

- $X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,3} & x_{1,4} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N,1} & x_{N,2} & x_{N,3} & x_{N,4} \end{bmatrix}$ (design matrix).
- $w = \begin{bmatrix} w_0 & w_1 & w_2 & w_3 & w_4 \end{bmatrix}^T$.

Prior Distribution for w: Assume independent Gaussian priors:

$$p(w) = N(w \mid 0, S_0), \quad S_0 = \text{diag}([\infty, 20^2, 10^2, 10^2, 10^2]).$$

Likelihood Function:

$$p(y|w, \sigma^2) = N(y \mid Xw, \sigma^2 I).$$

Posterior Distribution for w: Combining the prior and likelihood:

$$p(w|y) = N(w \mid m_N, S_N),$$

where:

$$S_N^{-1} = S_0^{-1} + \frac{1}{\sigma^2} X^T X, \quad m_N = S_N \left(S_0^{-1} m_0 + \frac{1}{\sigma^2} X^T y \right).$$

Part 2

To include gender as a new predictor, we introduce x_5 where:

$$x_5 = \begin{cases} -1, & \text{if male,} \\ 1, & \text{if female.} \end{cases}$$

The updated model becomes:

$$y = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_5 + \epsilon.$$

The prior for w_5 should reflect the assumption that gender is expected to explain no more than 10 points:

$$w_5 \sim N(0, 10^2).$$

The new prior covariance matrix S_0 is:

$$S_0 = \text{diag}([\infty, 20^2, 10^2, 10^2, 10^2, 10^2]).$$

The likelihood, prior, and posterior are updated accordingly to incorporate the new design matrix X with x_5.

Exercise 3

Part 1

The likelihood is defined as:

$$p(y|w, \beta) = \prod_{n=1}^{N} N(y_n; w^T x_n, \beta^{-1})$$

First, we represent the product of individual Gaussian distributions

For a single Gaussian term $N(y_n; w^T x_n, \beta^{-1})$, the probability density is:

$$N(y_n; w^T x_n, \beta^{-1}) = \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left(-\frac{\beta}{2} (y_n - w^T x_n)^2 \right)$$

The product over all $n = 1, \dots, N$ becomes:

$$p(y|w, \beta) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\beta^{-1}}} \exp\left(-\frac{\beta}{2} (y_n - w^T x_n)^2 \right)$$

Next, aggregate into matrix-vector form

Let:

- $y = [y_1, y_2, \ldots, y_N]^T$ (column vector of outputs),
- $X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{bmatrix}$ (design matrix, $N \times d$ size),
- $\epsilon = y - Xw$ (error vector).

The quadratic term aggregates as:

$$\sum_{n=1}^{N} (y_n - w^T x_n)^2 = \|y - Xw\|^2 = (y - Xw)^T (y - Xw)$$

The likelihood becomes:

$$p(y|w, \beta) = \left(\frac{\beta}{2\pi}\right)^{N/2} \exp\left(-\frac{\beta}{2}(y - Xw)^T (y - Xw)\right)$$

Finally, identify as a multivariate Gaussian

Recognizing this as the probability density of a multivariate Gaussian distribution:

$$p(y|w, \beta) = N(y; Xw, \beta^{-1} I_N)$$

where I_N is the $N \times N$ identity matrix.

Part 2

We need to verify the posterior distribution $p(w|y) = N(w; m_N, S_N)$:

The prior is given as $p(w) = N(w; m_0, S_0)$, and the likelihood is $p(y|w, \beta) = N(y; Xw, \beta^{-1} I_N)$.

The posterior using Bayes' theorem

$$p(w|y) \propto p(y|w, \beta) p(w)$$

Substitute the Gaussian forms:

- Likelihood: $p(y|w, \beta) \propto \exp\left(-\frac{\beta}{2}(y - Xw)^T (y - Xw)\right)$,
- Prior: $p(w) \propto \exp\left(-\frac{1}{2}(w - m_0)^T S_0^{-1} (w - m_0)\right)$.

The posterior $p(w|y)$ combines these terms:

$$p(w|y) \propto \exp\left(-\frac{\beta}{2}(y - Xw)^T (y - Xw) - \frac{1}{2}(w - m_0)^T S_0^{-1} (w - m_0)\right)$$

To combine the quadratic terms,

Expand the terms:

1. Likelihood quadratic:

$$-\frac{\beta}{2}(y - Xw)^T(y - Xw) = -\frac{\beta}{2}(y^T y - 2w^T X^T y + w^T X^T X w)$$

2. Prior quadratic:

$$-\frac{1}{2}(w - m_0)^T S_0^{-1}(w - m_0) = -\frac{1}{2}(w^T S_0^{-1} w - 2w^T S_0^{-1} m_0 + m_0^T S_0^{-1} m_0)$$

Combine:

$$p(w|y) \propto \exp\left(-\frac{1}{2}w^T(\beta X^T X + S_0^{-1})w + w^T(\beta X^T y + S_0^{-1} m_0) - \text{const}\right)$$

Complete the square for w:

$$w^T(\beta X^T y + S_0^{-1} m_0) - \frac{1}{2}w^T(\beta X^T X + S_0^{-1})w$$

leads to:

$$m_N = S_N(\beta X^T y + S_0^{-1} m_0), \quad S_N^{-1} = \beta X^T X + S_0^{-1}$$

Thus, the posterior is:

$$p(w|y) = N(w; m_N, S_N)$$

where:

$$m_N = S_N(\beta X^T y + S_0^{-1} m_0), \quad S_N^{-1} = S_0^{-1} + \beta X^T X$$

Exercise 4

The posterior is:

$$p(w, \beta|y) \propto p(y|w, \beta)\, p(w, \beta)$$

Substitute the likelihood and the prior:

$$p(w, \beta|y) \propto N(y; Xw, \beta^{-1} I_N)\, N(w; m_0, \beta^{-1} S_0)\, \text{Gam}(\beta; a_0, b_0)$$

The likelihood is a multivariate Gaussian:

$$p(y|w, \beta) \propto \beta^{N/2} \exp\left(-\frac{\beta}{2}\|y - Xw\|^2\right)$$

The Gaussian prior for w is:

$$p(w|\beta) = N(w; m_0, \beta^{-1} S_0) \propto \beta^{d/2} \exp\left(-\frac{\beta}{2}(w - m_0)^T S_0^{-1}(w - m_0)\right)$$

The Gamma prior for β is:

$$p(\beta) = \mathrm{Gam}(\beta; a_0, b_0) \propto \beta^{a_0-1} e^{-b_0\beta}$$

The posterior becomes:

$$p(w, \beta|y) \propto \beta^{(N+d)/2+a_0-1} \exp\left(-\beta\left[\frac{1}{2}\|y - Xw\|^2 + \frac{1}{2}(w - m_0)^T S_0^{-1}(w - m_0) + b_0\right]\right)$$

The terms inside the exponential can be written as:

$$\|y - Xw\|^2 + (w - m_0)^T S_0^{-1}(w - m_0) = w^T(X^T X + S_0^{-1})w - 2w^T(X^T y + S_0^{-1} m_0) + \text{const}$$

Define:

$$\begin{aligned} S_N^{-1} &= S_0^{-1} + X^T X \\ m_N &= S_N(S_0^{-1} m_0 + X^T y) \end{aligned}$$

Thus:

$$\|y - Xw\|^2 + (w - m_0)^T S_0^{-1}(w - m_0) = w^T S_N^{-1} w - 2w^T S_N^{-1} m_N + \text{const}$$

Completing the square, the posterior for $w|\beta, y$ is:

$$p(w|\beta, y) = N(w; m_N, \beta^{-1} S_N)$$

Substitute $w = m_N$ into the quadratic terms. The effective quadratic term for β is:

$$\beta\left[b_0 + \frac{1}{2}\left(m_0^T S_0^{-1} m_0 - m_N^T S_N^{-1} m_N + \sum_{n=1}^{N} y_n^2\right)\right]$$

Let:

$$b_N = b_0 + \frac{1}{2}\left(m_0^T S_0^{-1} m_0 - m_N^T S_N^{-1} m_N + \sum_{n=1}^{N} y_n^2\right)$$

and:

$$a_N = a_0 + \frac{N}{2}$$

Thus, the posterior for β is:

$$p(\beta|y) = \mathrm{Gam}(\beta; a_N, b_N)$$

The joint posterior is:

$$p(w, \beta|y) = N(w; m_N, \beta^{-1} S_N)\,\mathrm{Gam}(\beta; a_N, b_N)$$

where:

$$m_N = S_N(S_0^{-1} m_0 + X^T y), \quad S_N^{-1} = S_0^{-1} + X^T X$$

$$a_N = a_0 + \frac{N}{2}, \quad b_N = b_0 + \frac{1}{2}\left(m_0^T S_0^{-1} m_0 - m_N^T S_N^{-1} m_N + \sum_{n=1}^{N} y_n^2\right)$$

Lab Report

Obtaining Beta using different methods

model parameter	scikit-learn	from calculation	statsmodel
$beta_0$	0.88516511	0.88516511	0.88516511
$beta_1$	-2.23262046	-2.23262046	-2.23262046
$beta_2$	0.55870444	0.55870444	0.55870444

It is observed that the coefficients match within the tolerance level.

Obtain Beta for 3rd degree polynomial

model parameter	from calculation	statsmodel
$beta_0$	-0.010460	-0.0105
$beta_1$	10.124663	10.1247
$beta_2$	-31.139397	-31.1394
$beta_3$	21.132068	21.1321
$beta_0 - lower95$	-0.627324	-0.627
$beta_1 - lower95$	4.641755	4.642
$beta_2 - lower95$	-44.076122	-44.076
$beta_3 - lower95$	12.638011	12.638
$beta_0 - upper95$	0.606404	0.606
$beta_1 - upper95$	15.607572	15.608
$beta_2 - upper95$	-18.202672	-18.203
$beta_3 - upper95$	29.626125	29.626

It can be observed that the values from the calculation and from statsmodel are almost the same.

Posterior Parameters

Based on calculation, we get -

- mu_n = (9.96183048e-03 ; 9.64253662e+00 ; 2.96565687e+01 ; 2.01257788e+01)
- $omega_n$ = (21. 10.5 7.18421053 5.52631579 ; 10.5 7.18421053 5.52631579 4.53341595 ; 7.18421053 5.52631579 4.53341595 3.87301356 ; 5.52631579 4.53341595 3.87301356 3.40260608)
- a_n: 11.0

- b_n: 2.4290992050004583

Model evidence of different polynomials

polynomial	log model evidence
2	-25.122117003562504
3	-18.212515913837233
4	-18.935759028543394
5	-19.246343841247004
6	-20.427624972944546

It can be observed that the third degree polynomial results in the maximum log-evidence. This is in line with the visual representation about fitting the sinus curve better than the others.

Calculation of posterior parameter for the given dataset

Of the 8 models,

- Most optimal model : Model which only has an attorney
- mu_n: (2.58935851 ; 3.96220259)
- $Omega_n$: (0.00157607 -0.00157607; -0.00157607 0.00308548)
- a_n: 648.1
- b_n : 34572.23820661567
- lmodevid: -4428.352053578861

Of the eight models implemented, the one with the largest log-model evidence was the one having an attorney. So, one might say that the only factor that affects economic loss is whether the claimant has an attorney or not.