STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES
MASTER OF SCIENCE IN DATA SCIENCE & ANALYTICS
CAT 2- **MARKING GUIDE**
DSA 8505: Bayesian Statistics

DATE: 21st Mar 2025

Instruction

(a) Answer All Question

(b) Scan and submit your answer sheet through the Google Classroom by 23h59, 14th Feb 2025.

1. Suppose we are modeling the probability of a patient recovering from a disease ($Y_i = 1$) based on the number of days they adhered to a prescribed treatment (X_{1i}) and their age in years (X_{2i}). We assume a logistic regression model:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

Given the following dataset:

Days on Treatment (X_{1i})	**Age** (X_{2i})	**Recovered** (Y_i)
1	25	0
2	30	0
3	35	1
4	28	1
5	40	1
2	45	0
6	50	1
3	33	1
4	27	1
5	29	1

We assume the following priors:

$$\beta_0 \sim \mathcal{N}(0, 10)$$
$$\beta_1 \sim \mathcal{N}(0, 10)$$
$$\beta_2 \sim \mathcal{N}(0, 10)$$

The likelihood function is given by:

$$P(Y|\beta) = \prod_{i=1}^{n} p_i^{Y_i}(1 - p_i)^{1-Y_i}$$

Using the MCMC method in Python, we estimated the posterior distributions of β_0, β_1, and β_2 as follows:

- β_0 has a mean of -8 with a 94% HDI of (-18, 2).
- β_1 has a mean of 3.8 with a 94% HDI of (0.5, 8.0).
- β_2 has a mean of -0.2 with a 94% HDI of (-1.5, 1.0).

Compute the probability of recovery if a patient follows the treatment for 3 days and is 30 years old, and for a patient who follows the treatment for 5 days and is 35 years old. Interpret the results.

Solution

Step 1: Logistic Regression Model

The logistic regression equation is given by:

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \tag{1}$$

From the estimated posterior means, we have:

$$\beta_0 = -8$$
$$\beta_1 = 3.8$$
$$\beta_2 = -0.2$$

The probability of recovery is computed as:

$$p_i = \frac{e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}}}{1 + e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}}} \tag{2}$$

Step 2: Compute Probability for Given Cases

Case 1: Patient follows treatment for 3 days and is 30 years old

Substituting $X_1 = 3$ and $X_2 = 30$:

$$\begin{aligned}\log\left(\frac{p_1}{1 - p_1}\right) &= -8 + (3.8 \times 3) + (-0.2 \times 30)\\ &= -8 + 11.4 - 6\\ &= -2.6\end{aligned}$$

$$
\begin{aligned}
p_1 &= \frac{e^{-2.6}}{1+e^{-2.6}} \\
&= \frac{0.074}{1+0.074} \\
&= \frac{0.074}{1.074} \approx 0.069
\end{aligned}
$$

Thus, the probability of recovery is **6.9%**.

Case 2: Patient follows treatment for 5 days and is 35 years old

Substituting $X_1 = 5$ and $X_2 = 35$:

$$
\begin{aligned}
\log\left(\frac{p_2}{1-p_2}\right) &= -8 + (3.8 \times 5) + (-0.2 \times 35) \\
&= -8 + 19 - 7 \\
&= 4
\end{aligned}
$$

$$
\begin{aligned}
p_2 &= \frac{e^{4}}{1+e^{4}} \\
&= \frac{54.6}{1+54.6} \\
&= \frac{54.6}{55.6} \approx 0.982
\end{aligned}
$$

Thus, the probability of recovery is **98.2%**.

Step 3: Interpretation of Results

- A patient who follows the treatment for **3 days** and is ****30 years old**** has only a **6.9% chance** of recovery. This suggests that short adherence to the treatment is insufficient for a high probability of recovery.
- A patient who follows the treatment for **5 days** and is **35 years old** has a **98.2% chance** of recovery, indicating that longer treatment significantly improves the probability of recovery.
- The negative estimate of β_2 suggests that **older patients** might have slightly lower recovery chances, but the effect is minimal since $\beta_2 = -0.2$ with an HDI that includes 0.

2. Bayesian logistic regression is used to model binary outcomes, such as predicting whether a patient has a disease based on their age and cholesterol level. Here is a sample dataset:

Disease (1 = Yes, 0 = No)	**Age (years)**	**Cholesterol (mg/dL)**
0	45	180
1	52	220
1	60	250
0	35	160
0	40	170
1	55	240
0	30	150
1	65	270
0	42	175
1	58	230

(a) Explain the role of prior distributions in Bayesian logistic regression and discuss their impact on inference.

(b) The logistic function (sigmoid) is used to transform the linear combination of predictors into a probability. Explain why it is appropriate for modeling binary outcomes.

(c) In the Bayesian model given, the posterior estimates of beta's are:

$$\beta_0 = -0.1, \quad \beta_{\text{age}} = -0.25, \quad \text{and} \quad \beta_{\text{cholesterol}} = 0.018.$$

Compute the probability of disease for a 50-year-old patient with a cholesterol level of 200 mg/dL. Interpret the result in the context of disease diagnosis.

Solution

1. Role of Prior Distributions in Bayesian Logistic Regression

In Bayesian logistic regression, prior distributions represent our initial beliefs about the parameters before observing any data. These priors influence the posterior estimates based on the likelihood of the observed data.
Impact on Inference:

- **Weak/Non-informative Priors:** These allow the data to dominate the inference. For example, choosing a normal prior with a large variance (e.g., $\mathcal{N}(0, 10)$) gives a wide range of possible values for the coefficients.
- **Informative Priors:** If we have strong domain knowledge, we can set priors that shrink estimates toward expected values, helping prevent overfitting.
- **Regularization:** Prior distributions, such as Gaussian priors centered around zero, help control the magnitude of coefficients and reduce overfitting, similar to L2 regularization in frequentist logistic regression.

2. Role of the Logistic (Sigmoid) Function

The logistic function (sigmoid) is used to map a linear combination of predictors to a probability between 0 and 1:

$$p(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2)}} \tag{3}$$

Why is it appropriate for binary outcomes?

- The function outputs values between 0 and 1, making it ideal for probability estimation.

- It is **monotonic**, meaning an increase in a predictor (e.g., cholesterol level) will consistently increase or decrease the probability of disease.
- The logistic curve has an **S-shape**, meaning it handles small and large values smoothly without extreme jumps, which is useful for decision boundaries in classification.

3. Computing the Probability of Disease for a 50-Year-Old Patient with Cholesterol = 200 mg/dL

We are given the logistic regression model:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{\text{age}} \cdot X_{\text{age}} + \beta_{\text{cholesterol}} \cdot X_{\text{cholesterol}} \tag{4}$$

Substituting the values:

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= -0.1 + (-0.25 \times 50) + (0.018 \times 200)\\ &= -0.1 - 12.5 + 3.6\\ &= -9\end{aligned}$$

Now, solving for p:

$$\begin{aligned}p &= \frac{1}{1+e^{-(-9)}}\\ &= \frac{1}{1+e^{9}}\\ &= \frac{1}{1+8103.08}\\ &\approx \frac{1}{8104.08}\\ &\approx 0.00012\end{aligned}$$

Thus, the probability of disease for a 50-year-old patient with a cholesterol level of 200 mg/dL is **0.00012 (or 0.012%)**.

Interpretation

- The estimated probability of disease is **extremely low**. This suggests that, under this model, a patient of this age and cholesterol level is very unlikely to have the disease.
- The negative coefficient for age ($\beta_{\text{age}} = -0.25$) suggests that older age is associated with a lower probability of disease in this model, which may indicate an unusual trend in the dataset or a need to include interaction terms.
- The positive coefficient for cholesterol ($\beta_{\text{cholesterol}} = 0.018$) means higher cholesterol levels increase the risk of disease, which aligns with medical intuition.
- The model should be validated with **more data** to ensure robustness before making clinical predictions.