Probability and Statistical Techniques BCNS 1206C
Unit 4: Logistic Regression

1 Logistic Regression Model

Logistic regression is a method in statistics used for modeling binary outcome variables; situations where the result falls into one of two categories (e.g., yes/no, purchase/no purchase, 1/0).

Unlike linear regression, which predicts numeric outcomes, logistic regression estimates the probability that an event of interest occurs, based on input features.

In logistic regression, the natural logarithm of the odds of the event is modeled as a linear function of the explanatory variables:

$$\log(\text{odds}) = \beta_0 + \beta_1 x,$$

where β_0 is the intercept and β_1 is the coefficient for predictor x. The odds are expressed as:

$$\text{odds} = \frac{p}{1-p},$$

where p is the probability of the event. When x is a categorical variable, it is typically converted to a numerical form using indicator variables.

1.1 Example: Online Shopping Behavior by Age

Scenario: A survey investigates whether individuals from different age groups shop online. The collected data are summarized in the table below:

	Shops Online	Does Not Shop Online	Total
Under 30	180	120	300
30 and Above	150	210	360
Total	330	330	660

Step 1: Compute proportions. We calculate the proportions of online shoppers in each group:

$$p_{¡30} = \frac{180}{300} = 0.6, \quad p_{\geq 30} = \frac{150}{360} = 0.4167.$$

Step 2: Convert to odds. Next, we compute the odds of online shopping for each group:

$$\text{odds}_{¡30} = \frac{0.6}{1-0.6} = 1.5, \quad \text{odds}_{\geq 30} = \frac{0.4167}{1-0.4167} \approx 0.7143.$$

Step 3: Define indicator variable for age group. To encode age as a binary variable:

$$x = \begin{cases} 1, & \text{if age ¡ 30,} \\ 0, & \text{if age } \geq 30. \end{cases}$$

Step 4: Calculate log-odds.

$$\log(\text{odds}_{¡30}) = \log(1.5) \approx 0.4055, \quad \log(\text{odds}_{\geq 30}) = \log(0.7143) \approx -0.3365.$$

Step 5: Estimate the model. Using the logistic regression structure:

$$\log(\text{odds}_{¡30}) = \beta_0 + \beta_1(1) = 0.4055,$$
$$\log(\text{odds}_{\geq 30}) = \beta_0 + \beta_1(0) = -0.3365.$$

Solving gives:

$$\beta_0 = -0.3365, \quad \beta_1 = 0.742.$$

Hence, the fitted logistic regression model is:

$$\log(\text{odds}) = -0.3365 + 0.742x.$$

Step 6: Interpreting the Odds Ratio. Exponentiating the slope provides the odds ratio:

$$\frac{\text{odds}_{\text{i}30}}{\text{odds}_{\geq 30}} = e^{0.742} \approx 2.1.$$

This means that individuals under 30 are approximately twice as likely to shop online as those 30 and above.

If we had encoded the age group the other way (under 30 = 0, 30+ = 1), the sign of the slope would reverse, yielding:

$$e^{-0.742} \approx 0.476,$$

indicating that the older group has less than half the odds compared to the younger group.

1.2 Statistical Inference in Logistic Regression

For a logistic regression with one predictor:

- Find estimates of β_0 and β_1, and write the regression equation.
- Construct a 95% confidence interval for the slope and test whether the slope differs from zero.
- Calculate and interpret the odds ratio and its confidence interval.

Inference for logistic regression follows a structure similar to that in linear regression. We estimate parameters and compute standard errors, then construct confidence intervals using the standard normal distribution (z-values).

A confidence interval for the slope β_1 is:

$$\beta_1 \pm z^* \text{SE}_{\beta_1},$$

where SE_{β_1} is the standard error, typically obtained using statistical software.
The confidence interval for the odds ratio e^{β_1} is:

$$\left(e^{\beta_1 - z^* \text{SE}_{\beta_1}},\, e^{\beta_1 + z^* \text{SE}_{\beta_1}}\right).$$

1.3 Tutorial

Problem 1. *A survey was conducted to assess whether individuals from rural and urban areas use online banking services. The results are summarised below:*

	Online Banking User	*Non-User*	*Total*
Urban	250	150	400
Rural	180	220	400
Total	430	370	800

(i) Identify the response variable and compute the proportions of online banking users in each group.

(ii) Calculate the odds of being an online banking user for both groups.

(iii) Compute the natural logarithm of the odds (log-odds) for both groups.

(iv) Define a binary indicator variable to represent the explanatory variable (urban vs. rural).

(v) Use the results to fit a logistic regression model of the form: $\log(odds) = \beta_0 + \beta_1 x$*, and determine the values of* β_0 *and* β_1*.*

(vi) Interpret the odds ratio obtained from the fitted model.

Problem 2. *Suppose the fitted logistic regression model for a study on gym membership (coded as 1 for member and 0 for non-member) by gender is:*

$$\log\left(\frac{p}{1-p}\right) = -0.27 + 0.59x,$$

where $x = 1$ *if the person is female and 0 otherwise.*

(i) What is the estimated odds ratio of being a gym member for females compared to males?

(ii) Explain how the coding of the indicator variable affects the sign and interpretation of the slope.

(iii) If the odds of a male being a gym member are 0.75, what are the odds for a female?

Problem 3. *A logistic regression model is used to predict whether students pass an online course based on whether they participated in weekly discussion forums. The slope coefficient* β_1 *was estimated to be 1.25, with a standard error of 0.4.*

(i) Construct a 95% confidence interval for β_1*.*

(ii) Construct the corresponding 95% confidence interval for the odds ratio.

(iii) Interpret both intervals in the context of student success.

Problem 4. *For a study on the use of public transport based on possession of a driver's license, the logistic regression model output gives:*

- *Estimated slope:* $\beta_1 = -0.78$
- *Standard error:* $SE_{\beta_1} = 0.22$

(i) Conduct a significance test to determine if the slope is significantly different from zero at the 5% level.

(ii) Interpret the meaning of the negative slope.

(iii) Compute and interpret the odds ratio.

Problem 5. *The following table records whether individuals with or without pets tend to work remotely:*

	Remote Worker	*On-Site Worker*	*Total*
Has Pet	210	90	300
No Pet	120	180	300
Total	330	270	600

(i) Calculate the proportion, odds, and log-odds of working remotely for both groups.

(ii) Define a dummy variable for pet ownership.

(iii) Estimate the logistic regression coefficients and write the fitted model.

(iv) Interpret the odds ratio.

Problem 6. *A logistic regression model is used to predict the likelihood of attending a webinar based on whether an individual received a reminder email. The model is:*

$$\log\left(\frac{p}{1-p}\right) = -0.45 + 1.15x,$$

where $x = 1$ if the person received a reminder.

(i) Compute the probability that someone who received the reminder will attend.

(ii) Compute the probability that someone who did not receive the reminder will attend.

(iii) Find and interpret the odds ratio.

Problem 7. *A study finds that the log-odds of participating in recycling programs are modeled as:*

$$\log(odds) = -1.1 + 0.9x,$$

where $x = 1$ if the person owns a house, and 0 otherwise.

(i) What is the odds ratio of homeowners vs. non-homeowners participating in recycling?

(ii) What does a positive coefficient tell us in this context?

(iii) If the odds for non-homeowners are 0.33, find the odds for homeowners.

Problem 8. *In a study on food delivery app usage, a logistic regression model was fitted with a slope estimate of $\beta_1 = 0.76$ and a standard error of 0.28.*

(i) Test the null hypothesis $H_0 : \beta_1 = 0$ at the 5% significance level.

(ii) Calculate the 95% confidence interval for the slope.

(iii) Transform this interval into a confidence interval for the odds ratio.