ADS4293 LAB6

Chapter 9: Logistic Regression using R

TA: Malak Al-Hanafi

October 16 , 2025

LAB SETUP

- Excel file **Ch9** dataset
- We'll go over one question at a time
- Submission:
  - Submit your updated **Excel file** on Brightspace (Owl)
  - For each question sheet, paste your **R code** below your answers
- Today's lab has **2 questions**

QUESTION 1

13. FILE ***STEM.*** Several studies have reported lower participation in the science, technology, engineering, and mathematics (STEM) careers among female and minority ethnic group students. The accompanying data file includes information on whether the student has applied to a STEM field (1 if STEM, 0 otherwise), whether or not the student is female (1 if female, 0 otherwise), white (1 if white, 0 otherwise), and Asian (1 if Asian, 0 otherwise). Also included in the survey is the information on the student's high school GPA and the SAT scores.

    a. Estimate and interpret the logistic regression model using STEM as the response variable, and GPA, SAT, White, Female, and Asian as the predictor variables.

    b. Find the predicted probability that a white male student will apply to a STEM field with GPA = 3.4 and SAT = 1400. Find the corresponding probabilities for an Asian male and a male who is neither white nor Asian.

    c. Find the predicted probability that a white female student will apply to a STEM field with GPA = 3.4 and SAT = 1400. Find the corresponding probabilities for an Asian female and a female who is neither white nor Asian.

Q 1 A

a. Estimate and interpret the logistic regression model using STEM as the response variable, and GPA, SAT, White, Female, and Asian as the predictor variables.

Using R

1. **Create a new R script in RStudio, place your Excel file in the same folder, and make sure the readxl package is installed**

2. **Import the data file into a data frame (table) and label it myData.**

```
# Replace "your_path_here" with your own file location
myData <- read_excel("your_path_here/Lab6_StudentName.xlsx", sheet = "Q9.13 STEM")
```

3. **Fit the Logistic Regression and display summary results**

```
Logistic_Model <- glm(STEM ~ GPA + SAT + White + Female + Asian, family = binomial, data = myData)
summary(Logistic_Model)
```

4. **Interpret the significant variables (p < 0.05) & the direction (positive or negative) coefficients***

   ***hint: In regression analysis, we only interpret the direction of a variable if it's statistically significant.*

5. **Write the Logistic Regression Equation using the coefficients**

   General logistic regression form

   - **a** is the intercept
   - **bX** represents all the predictors multiplied by their coefficients

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Q 1 B , C

b. Find the predicted probability that a white male student will apply to a STEM field with GPA = 3.4 and SAT = 1400. Find the corresponding probabilities for an Asian male and a male who is neither white nor Asian.

c. Find the predicted probability that a white female student will apply to a STEM field with GPA = 3.4 and SAT = 1400. Find the corresponding probabilities for an Asian female and a female who is neither white nor Asian.

Using R, you'll need 6 calls to predict() — one for each student profile:

predict(Logistic_Model, data.frame(GPA = ___, SAT = ___, White = ___, Female = ___, Asian = ___), type = "response")

- **Fill in the ___ with the correct numeric values for each case**
  - **GPA and SAT stay fixed**
  - **White, Female, Asian: Set to 1 or 0 depending on the student profile**
- **Always include *type = "response":* tells R to return a probability (between 0 and 1)**

QUESTION 2

17. **FILE** ***Membership.*** The manager is concerned about the attrition rate at her gym. She would like to identify the profile of loyal members who retain the membership for at least one year. The accompanying data file contains information on whether the member has been loyal (Loyal = 1 if the member stayed at the gym for at least one year, 0 otherwise). Also included are the member's age and income (in $1,000s) and whether he/she joined on a single or a family plan.

    a. Estimate the logistic regression model using Loyal as the response variable and Age, Income, and Single (equals 1 if on a single plan, 0 otherwise) as the predictor variables.

    b. Predict the probability of being loyal for a 40- and a 60-year-old member with income of $80,000 and on a single plan.

    c. Predict the corresponding probabilities for a member on a family plan.

QUESTION 2A

a. Estimate the logistic regression model using Loyal as the response variable and Age, Income, and Single (equals 1 if on a single plan, 0 otherwise) as the predictor variables.

1. **Import the data file into a data frame (table) and label it myData.**

2. Create a **dummy variable** for **plan** because Logistic regression can only use numeric inputs for categorical variables
   myData$Single <- ifelse(myData$Plan == "Single", 1, 0)
   - **Converts text ("Single"/"Family") into numeric form**
   - **1 = Single plan, 0 = Family plan**

3. **Fit the Logistic Regression model and display summary results**
   - **Loyal is the dependent variable (Y)**
   - **Age, Income, and Single are independent variables (X)**

4. **Interpret the significant variables (p < 0.05) and their directions.**

5. **Write the Logistic Regression Equation using the coefficients**
   General logistic regression form
   - a is the intercept
   - bX represents all the predictors multiplied by their coefficients

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

QUESTION 2B,C

b. Predict the probability of being loyal for a 40- and a 60-year-old member with income of $80,000 and on a single plan.

Use the predict() function to find probabilities

You'll calculate two predictions: one for Age = 40 and one for Age = 60, ***hint: income is in thousands***

predict(Logistic_Model, data.frame(Age = ___, Income = ___, Single = ___), type = "response")

c. Predict the corresponding probabilities for a member on a family plan.

Use the predict() function to find probabilities

You'll calculate two predictions: one for Age = 40 and one for Age = 60, ***hint: Single: 0 (family plan)***

predict(Logistic_Model, data.frame(Age = ___, Income = ___, Single = ___), type = "response")

FOR NEXT LAB

Please Read about the Holdout Method & Odds in Chapter 9 of the textbook.

skim ch9 :)

Thank you!