ADS4293 LAB7

Chapter 9: Logistic Regression using R

TA: Malak Al-Hanafi

October 23 , 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

21. FILE *Complication.* A logistic regression model is estimated to analyze the probability of complications for male patients resulting from a serious infection. Predictor variables include the patient's weight and age and whether he is diabetic (Diabetes equals 1 if diabetic, 0 otherwise). The accompanying data file includes information on 260 male patients who had tested positive for a serious infection.

   a. Estimate the logistic regression model to find the odds of complications for a 60-year-old diabetic patient with a weight of 180 pounds.

   b. Find the corresponding odds if the patient is not diabetic.

   c. What is the percentage difference in the odds for a diabetic patient compared to a nondiabetic patient, holding the other variables constant?

1. **Create a new R script in RStudio, place your Excel file in the same folder, and make sure the readxl package is installed**
   library(readxl)
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", sheet = "Q9.21 Complication")

3. **Fit the Logistic Regression and display summary results**
   Logistic_Model <- glm(?? ~ ?? + ?? + ??, family = binomial, data = myData)
   - **Complication is the dependent variable (Y)**
   - **Weight, Age, and Diabetes are independent variables (X)**

   summary(Logistic_Model)

Q 1 A , B

a. Estimate the logistic regression model to find the odds of complications for a 60-year-old diabetic patient with a weight of 180 pounds.

1. **Compute the predicted probability for a 60-year-old, 180 lb diabetic patient**

predicted_prob <- predict(Logistic_Model, data.frame(Weight=??, Age=??, Diabetes=??), type = "response")

2. Odds are computed as $\frac{p}{1-p}$, where p is the probability of success.

predicted_odds <- ??

3. **Print the values**

b. Find the corresponding odds if the patient is not diabetic.

1. **Repeat for b but keep diabetes = 0**

Q 1 C

c. What is the percentage difference in the odds for a diabetic patient compared to a nondiabetic patient, holding the other variables constant?

THE PARTIAL EFFECT OF A PREDICTOR VARIABLE ON ODDS

Consider a logistic regression model with predictor variables $x_1, x_2, \ldots, x_k$, where b_j is the estimated coefficient for x_j. We interpret $\left(\exp\left(b_j\right) - 1\right) \times 100$ as the percentage change in the odds when x_j increases by one unit, holding the other predictor variables constant.

1. **Extract the coefficients**
   coefficients <- coef(Logistic_Model)
2. **Use the formula to convert the Diabetes coefficient to a percentage change in odds**
   (exp(coefficients["Diabetes"]) - 1) * 100
3. **Interpret the result**
   hint: +ve or -ve value? does it mean diabetes increase or decrease the odds of complications?

QUESTION 2

41. FILE ***Admit.*** Consider the accompanying data file to estimate the logistic regression model for predicting college admission (Admit equals 1 if admitted, 0 otherwise). Predictor variables include the applicant's grade point average (GPA) and scores on the SAT test. Use the holdout method, using the first 90 observations for training and the remaining 30 observations for validation, to calculate and interpret the accuracy, sensitivity, and specificity measures.

Holdout Cross-Validation concept

1. Split the dataset into two parts:
   - Training set: used to build (fit) the model.
   - Validation set: used to test how well the model generalizes.
2. Train the model using the training data only.
3. Predict outcomes for the validation data.
4. Compare predictions vs. real outcomes to measure accuracy, sensitivity, and specificity.

QUESTION 2

41. FILE ***Admit.*** Consider the accompanying data file to estimate the logistic regression model for predicting college admission (Admit equals 1 if admitted, 0 otherwise). Predictor variables include the applicant's grade point average (GPA) and scores on the SAT test. Use the holdout method, using the first 90 observations for training and the remaining 30 observations for validation, to calculate and interpret the accuracy, sensitivity, and specificity measures.

1. **Split into training and validation sets**
   TData <- head(myData, 90)
   VData <- tail(myData, 30)
2. **Fit the logistic regression model**
   Model <- glm(?? ~ ?? + ??, family = binomial, data = TData)
   - **Admit is the dependent variable (Y)**
   - **GPA and SAT are independent variables (X)**
3. **Predict on the validation data**
   pred <- predict(Model, VData, type = "response")
   we apply the trained model to the validation set, pred is a list of predicted probabilities for each student

QUESTION 2

41. FILE ***Admit.*** Consider the accompanying data file to estimate the logistic regression model for predicting college admission (Admit equals 1 if admitted, 0 otherwise). Predictor variables include the applicant's grade point average (GPA) and scores on the SAT test. Use the holdout method, using the first 90 observations for training and the remaining 30 observations for validation, to calculate and interpret the accuracy, sensitivity, and specificity measures.

4. **Classify the predictions (0 or 1)**

We convert probabilities into binary predictions:

If probability ≥ 0.5 → predict 1 (admitted), Otherwise → predict 0 (not admitted)

classify <- ifelse(pred >= 0.5, 1, 0)

5. **Compare predictions to actual results**

- True Positives → predicted admitted and actually admitted
- True Negatives → predicted not admitted and actually not admitted

TP <- ifelse(classify == 1 & VData$Admit == 1, 1, 0)

TN <- ifelse(????)

$ operator is used to select a specific column from a data frame.

6. **Calculate performance metrics**

Accuracy → overall % of correct predictions **100 * mean(VData$Admit == classify)**

Sensitivity → % of actual admits correctly predicted **100 * (sum(TP) / sum(VData$Admit == 1))**

Specificity → % of actual non-admits correctly predicted **100 * (sum(TN) / sum(VData$Admit == 0))**

Thank you!