Logistic Regression, Classification & Generalized Linear Models

S.M.Bendre
Indian Statistical Institute, Chennai Centre

March 27, 2023

Objectives

- Modeling the Available Information
- Model Selection and Interpretation
- Concept of Generalized Linear Models
- Logistic Regression Model
- Logistic Regression as a Classifier

Modeling the Available Information

- The body fat is an important component in determining the health of an individual. However, the percentage bady fat is not as easy to measure as some of the other components such as age, height, weight, abdominal circumference etc. Of interest is to investigate whether there is any relationship between the percentage body fat and these other components. If yes, the question is whether the percentage body fat can be represented by a function of these components.
- It is claimed that there is gender discrimination with respect to wages and men get higher wages than women with the same characteristics such as educational qualification, age etc. In order to investigate such a claim the data is collected on the wage, age, gender, educational qualification and the years of education. This can further lead to investigating whether it is worthwhile to have higher educational qualification to get a higher wage, after accounting for the gender difference and age.

Modeling the Available Information - contd

- In a NHS study to find out the success of a surgery, the interest is in investigating the effect of the gender, age, health condition of the patient, hereditary factors etc on the success or the failure of the surgery.
- The banks and other finance companies which allocate credit/debit cards or loans have to take decision regarding allocation of card or increasing credit limit, sanctioning of loan based on the capacity of the customer to replay the credit. The capacity to repay depends on various factors such as steady income, source of income, other commitments - personal and financial, credit score, customer background etc. based on the available information on a customer, the decision is whether to sanction card/loan.

Modeling the Available Information - contd

- Businesses like banks, mobile service providers, commercial establishments have to worry about problem of 'churning' i.e. customers leaving and joining another service provider. It is important to understand which aspects of the service influences a customer's decision in this regard. Management can concentrate efforts on improvement of service, keeping in mind these priorities.
- Employee attrition (loss of employees due to resignation) is one of the major problems all businesses face. The resignation of an employee typically affects the particular work/project the employee was engaged in at the time of leaving, apart from the cost to organization. As a result, the organizations prefer to know how various factors such as age, gender, qualification, location, type of job, type of incentives etc affect the attrition.

Modeling the Available Information - contd

- For an insurance company, the interest is in predicting the number of claims that an insurer will make in one year from the third party automobile insurance, given the amount insured, the make of the car, the non-claim bonus received last year, age of the insurer and so on.

The Problem

- Note that in all these examples, there is a 'random variable' of interest which is typically called 'the response variable' to be denoted by Y and is one of the following
  1. a numerical variable such as percentage body fat of an individual or amount of wage earned.
  2. a binary or dichotomous variable such as whether a surgery is successful or not, whether a customer is given loan or not, whether a customer is churning or not, which represents success or failure or more generally, presence or absence of an characteristic of interest.
  3. a count variable such as number of claims or number of accidents, i.e. the number of times an event of interest occures (either by itself or over a specified period of exposure such as one year or one month), which is a discrete integer valued random variable.

The Problem - contd

- Additional information is provided on other characteristics such as
  1. age, height, weight, abdominal circumference of an individual or the gender, age, health
  2. age, weight, hereditary factors of a patient or steady income, source of income, other commitments - personal and financial, credit score, customer background etc of a customer.
  3. the amount insured, the make of the car, the non-claim bonus received last year, age of the insurer and so on.

  These characteristics are believed to have additional knowledge on the response variable of interest and can affect the value of the response variable.
  The information on these characteristics act as regressors or explanatory variables $X_1, X_2, \ldots, X_p$.
  In literature, various other terms are used for $X_1, X_2, \ldots, X_p$ such as the term regressors for linear regression models and explanatory variables for logistic regression models.
  Note that the regressors can be numerical or categorical and need to be handled appropriately.

The Aim

The aim is to build an appropriate model to use the information available on the regressors or explanatory variables to make better prediction about the response variable Y.

Thus the interest is to provide a mathematical function

$$E(Y) = \text{function}(X_1, X_2, \ldots, X_p)$$

based on the given values of $X_1, X_2, \ldots, X_p$ and helps in predicting the expected value of the response variable Y.

Generalized Linear Model

Note that the function involves the unknown parameters, typically denoted by Greek letters α, β etc.

For many situations, the model further assumes that function$(X_1, X_2, \ldots, X_p)$ is a function of the linear function of the unknown parameters $\beta_0, \beta_1, \ldots, \beta_p$ given by $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$ which is called the linear predictor.

Thus the function is expressed as

$$E(Y) = \text{function}(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p).$$

which is called a **Generalized Linear Model** (GLM).

Why?

1. To predict the model, i.e. the expected value of the response variable for a new case, based on the data on the explanatory variables for that case (Supervised learning).
2. To study the impact of one explanatory variable on the response variable keeping other explanatory variables fixed. For example, with same gender and same educational qualification, one can explore how the age affects the wages received.
3. To verify whether the data support certain beliefs (hypotheses) - for example, whether the age affects the success of surgery and if so, how does it affect with one year increase in the age.
4. To use the model for classification of an unit based on the explanatory variables.

Different Models

Set Up:
Response variable Y,
Explanatory variables or Regressors $X_1, X_2, \ldots, X_p$
Objective:
To model **expectation** of Y in terms of a function of $X_1, X_2, \ldots, X_p$

• The model ideally depends on the type of response variable under consideration.

1. If Y is continuous: Multiple Linear Regression **if** Y can be assumed to be Normal
2. If Y is continuous: other models such as Gamma Regression etc if Y can not be assumed to be Normal (GLM)
3. If Y is binary: Logistic Regression (GLM)
4. If Y is count data: Poisson, Negative Binommial Regression (GLM)
5. If Y is multicategory: Multilogit Regression (ordinal or nominal)

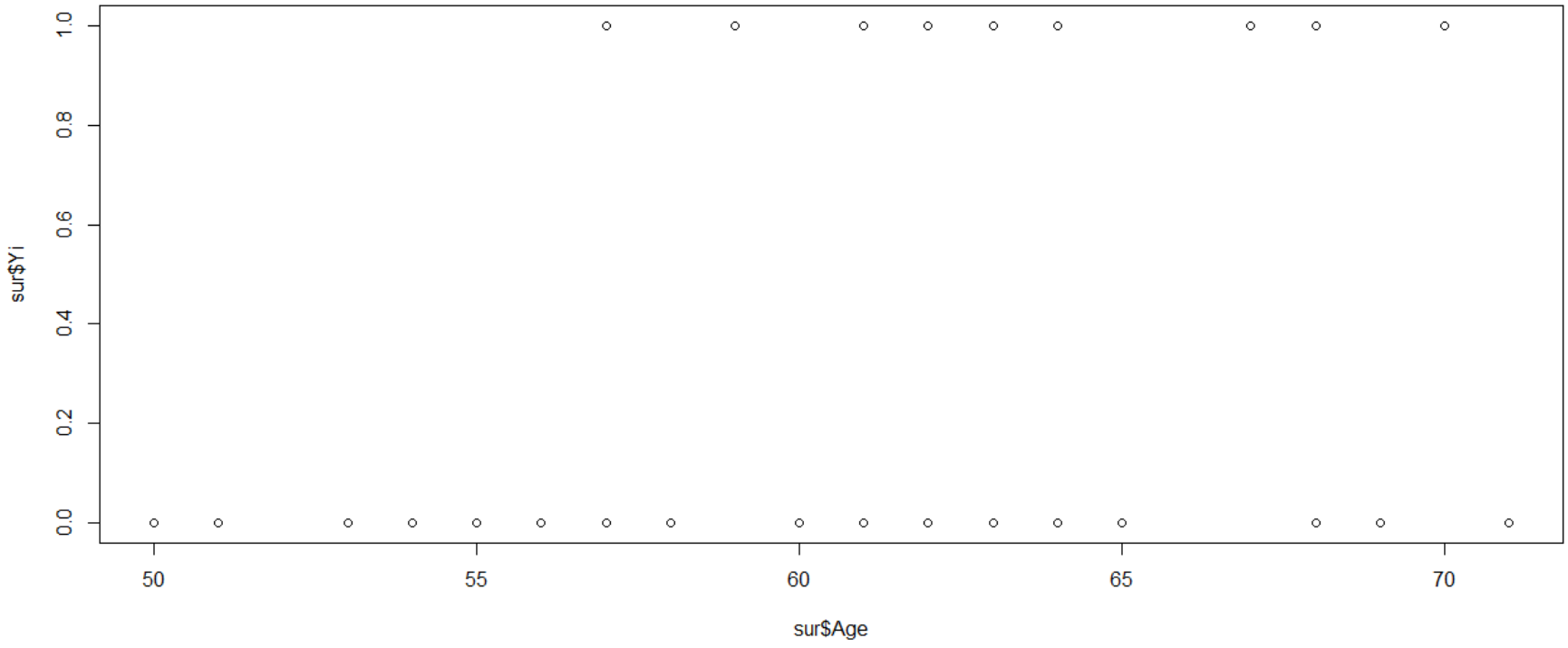
Example: Surgery Data

The records are from surgery on 40 patients where $y_i = 1$ if the patient died within 30 days of surgery and zero otherwise. Age, in years, is recorded for each patient.

Patient	Age	yi	Patient	Age	yi
1	50	0	21	61	0
2	50	0	22	61	1
3	51	0	23	61	1
4	51	0	24	62	1

Of interest is the effect of age on the chances of survival. We would be particularly interested in knowing whether the age is associated with the survival rate and if so, what is the relation between age and chance of survival.
Note that $y_i = 1$ stands for death here so 'success' is death and needs to be interpreted appropriately.

Scatter Plot of Surgery Data



Modeling Binary Response

With the dichotomous variable Y, the outcome of interest to the experimenter is called a 'success' and we define

$$Y = \begin{cases} 1 & \text{if the outcome is a success} \\ 0 & \text{if the outcome is a failure} \end{cases}$$

Thus we have a binary response variable taking one of the two values in $\{0, 1\}$. Let π denote the probability of success, $0 < \pi < 1$. Then

$$P[Y = 1] = \pi, \quad \text{and} \quad P[Y = 0] = 1 - \pi.$$

Further, we can express

$$P[Y = y] = \pi^y (1 - \pi)^{1-y}, \quad y = 0, 1, \ \pi \in (0, 1)$$

and hence $Y \sim$ Bernoulli(π). We also know that

$$E(Y) = \pi, Var(Y) = \pi(1 - \pi).$$

Corresponding to every Y_i we will have one π_i, $i = 1, \ldots, n$.

Modelling- contd

Alternatively, corresponding to each π_i, we may have outcomes of n_i trials for $i = 1, \ldots, n$. In such a situation, we define n independent random variables Y_i where

$$Y_i = \text{number of successes in } n_i \text{ independent trials,}$$

with the probability of success $\pi \in (0, 1)$.
Then $Y_i \sim$ Binomial(n_i, π_i), $\pi_i \in (0, 1)$ with probability mass function

$$P[Y_i = y_i] = \binom{n_i}{y_i} \pi^{y_i} (1 - \pi_i)^{n_i - y_i}$$

and

$$E(Y_i) = \mu_i = n_i \pi_i, \quad \text{Var}(Y_i) = n_i \pi_i (1 - \pi_i).$$

Logistic Regression Model

For the binary response, we equate the expected value of Y to be a function of the linear predictor $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$. In particular, we consider the Logistic Regression Model, where

$$E(Y) = \pi = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}$$

or equivalently

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

which ensures that π lies between $(0, 1)$.

Fitting Logistic Regression Model

- The Logistic Regression Model is a particular case of Generalized Linear Models (GLM).
- The model fitting involves estimation of the model parameters $\beta_0, \beta_1, \ldots, \beta_p$ and these are estimated using an iterative procedure (Iterative Re-Weighted Least Squares:IRWLS) for the Generalized Linear Models.
- Most of the statistical packages provide IRWLS computation procedures.
- It can be fitted using the built-in function glm() in R package and can be used for classification.

Fitted Model

```
glm(formula = Yi ~ Age, family = binomial, data = sur)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6601  -0.8099  -0.5839   1.0491   1.7079

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -10.48174    4.30409  -2.435   0.0149
Age           0.16295    0.07018   2.322   0.0202

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 51.796  on 39  degrees of freedom
Residual deviance: 45.301  on 38  degrees of freedom
AIC: 49.301

Number of Fisher Scoring iterations: 3
```

Interpretation of Results

The explanaory variable ‘age’ significantly contributes to the ‘success’ or ‘failure’ of the surgery, since the p-value is small (0.0202).

Note that the data is on patients with age greater than 50 and hence the interpretation holds only for such patients.

Fitted Values

We need to check what are the fitted values of the response variable Y which is either 0 or 1.
Note that the model specified

$$E(Y) = \pi = \frac{\exp(\beta_0 + \beta_1 \text{Age})}{1 + \exp(\beta_0 + \beta_1 \text{Age})}$$

and with estimated values of parameter given by

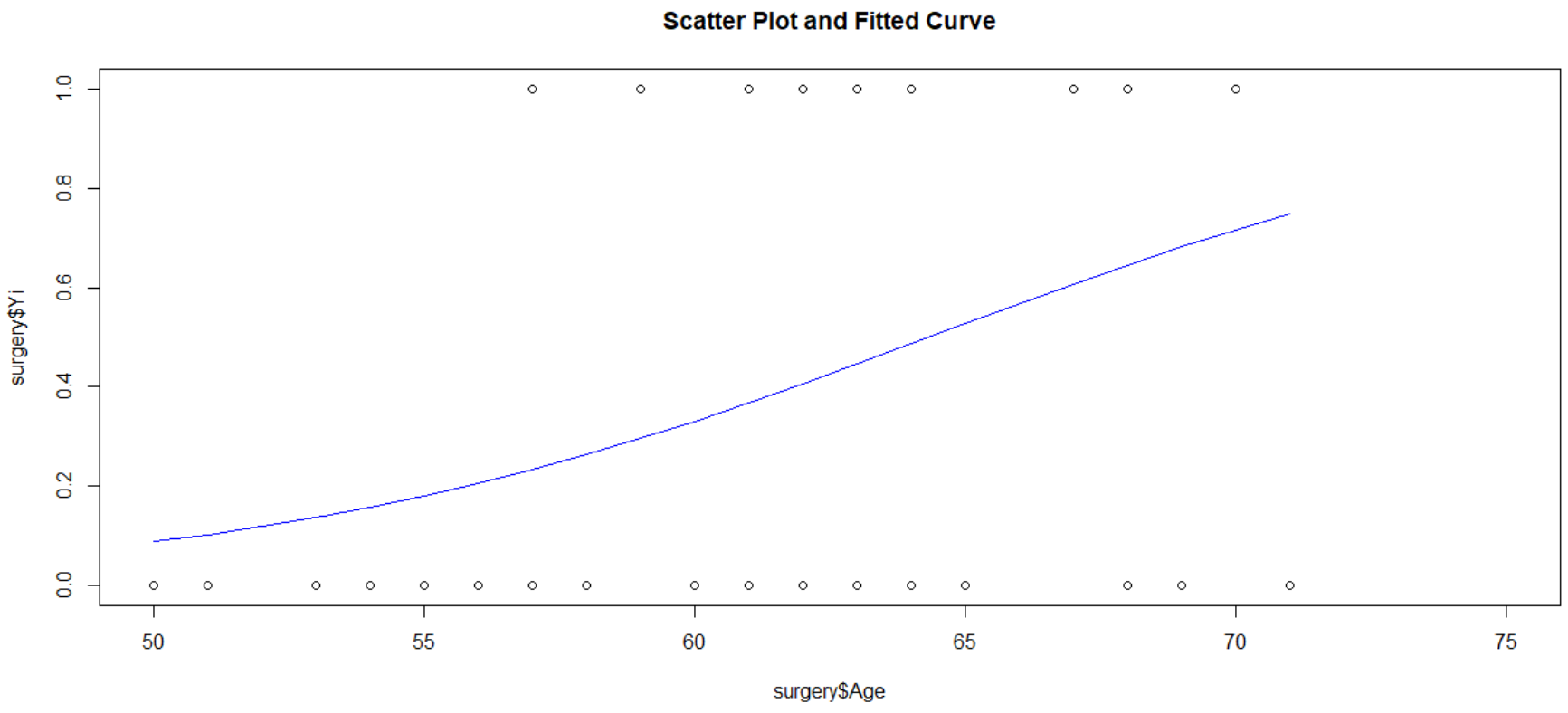
$$\hat{\beta}_0 = 10.48174, \quad \hat{\beta}_1 = 0.16295$$

we get the fitted values of Y_i, given by

$$\hat{Y}_i = \hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \text{Age}_i)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 \text{Age}_i)}$$

which is between $(0, 1)$.

Fitted Curve and Scatter Plot of Surgery Data



Logistic Regression Model as a Classifier

Now, the aim is to decide based on the fitted model whether given the age of the patient, the patient will survive or not.

i.e, given the age, to let the model predict whether the corresponding response is '0' or '1'

This leads to classifying the patient in to two classes: survival or death

Classification

For any classification technique, the given data is spilt into

- training data
- validation data
- test data

For logistic regression, based on the training data we need to decide which values of predicted probability can be considered as 0 and which as 1.
i.e. we need a threshold for classification which is often decided using the validation and test data.

Classification: Confusion Matrix

Confusion Matrix is the standard tool used for any classification problem.
For the surgery data, the confusion matrix will be

Fact	Model Based Prediction	
$\downarrow$	Does Not Survive	Survive
Does Not Survive	$*$	error
Survive	error	$*$

There are various measures based on the confusion matrix to decide the appropriateness of a classifier.

Logistic regression Model as a Classifier

For any given set of explanatory variables, the model gives the predicted response. Using the predetermined threshold value one can classify the predicted response into one of the two classes.

Using Logistic Regression as a classifier has certain advantages such as

- choice of threshold
- statement about the probability
- statement about the probability of misclassification
- statement about the odds ratio etc

References

1. S. Chatterjee G. Hadi (2013) Regression Analysis by Examples
2. J. Fox (2015)Applied Regression Analysis and Generalized Linear Models. 3rd edn, John Wiley.
3. D.C Montgomery, E. Peck G. Vinning (2012) Introduction to Linear Regression Analysis. 5th edn, John Wiley.
4. R.H. Myers, D.C Montgomery, G.G. Vining T.J. Robinson (2010) Generalized Linear Models: With Applications in Engineering and the Sciences. 2nd edn, John Wiley.
5. D.W. Hosmer S. Lemeshow (2000) Applied Logisitic Regression. 2nd edn, John Wiley.