MA 214: Applied Statistics

Instructor: Ashis Gangopadhyay

Logistic Regression

Where we have been….

- We were discussing how to analyze a categorical response variable based on one categorical covariate (factor).
- Analysis involved comparing the frequencies of occurrences of the response variable within each categories of the covariate.

Where we are going….

- We wish to consider modeling a categorical response variable based on multiple covariates, either continuous or categorical.
- Particularly useful situation is when the response variable is dichotomous, i.e., it takes two possible values (have a certain disease/do not have the disease, defective/non-defective, passed a class/didn't pass a class, etc.).
- We will be discussing how to model this type of response variable.

Low birth weight study

- The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams).
- Data were collected on 189 women, 59 of them gave birth to a baby weighing less than 2500 grams.

Low birth weight study

Y_i : Indicator variable for low birth weight (1=Low, 0=Normal)

X_{1i} : Mothers weight at the time of conception.

X_{2i} : Indicator variable for smoking habit of the mother (1=Smoker, 0=Nonsmoker)

Hosmer and Lemeshow (2000) Applied Logistic Regression: Second Edition.

Data were collected at Baystate Medical Center, Springfield, Massachusetts during 1986. Original data has several other covariates.

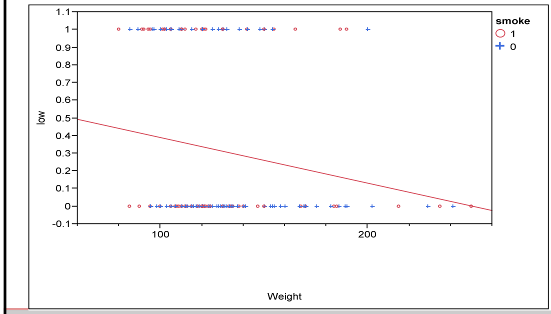
Low birth weight study

Data:

ID	Low (Y)	Weight (X_1)	Smoke (X_2)
1	0	182	0
2	0	155	0
3	0	105	1
4	0	108	1
⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮
188	1	142	0
189	1	130	1

- In the column Low, 1 indicates low birth weight.
- In the column Smoke, 1 indicates smoker

Plot of Low Birth Weight vs. Weight



Logistic Regression

Y_i is a Bernoulli random variable with probability of success p_i, i.e.,

$$P(Y_i=1)=p_i \text{ and } P(Y_i=0)=1-p_i$$

Or equivalently,

$$E(Y_i)=p_i$$

The probability of low birth weight of babies may be different due to the characteristics of their mothers, such as weight, age, smoking habit, etc. Since these characteristics are unique for each mother, the probability of low birth weights for babies are also distinct.

Logistic Regression

Why don't we fit the usual multiple regression model?

- Distribution of Y_i is Bernoulli and not Normal.
- The mean of Y_i is p_i , where $0 \le p_i \le 1$. Thus the model must estimate the mean of Y between 0 and 1. A linear regression may not do that.
- $Var(Y_i) = p_i(1-p_i)$, thus the variance is not a constant.

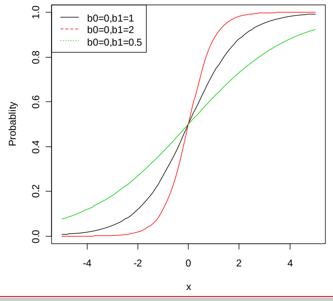
Logistic Function

A logistic function is given by

$$f(x) = \frac{e^x}{1+e^x}, \text{for all x in R}$$

The function f(x) is bounded between 0 and 1. Different shapes of the function can be achieved by incorporating parameter as follows:

$$f(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}, \text{for all x in R}$$



Logistic regression

In logistic regression, Y_i is modeled via a logistic function. First, consider the case with only one covariate X_i

$$p_i = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} \quad \text{or:} \quad \frac{p_i}{1-p_i} = \exp(\beta_0+\beta_1 x_i)$$

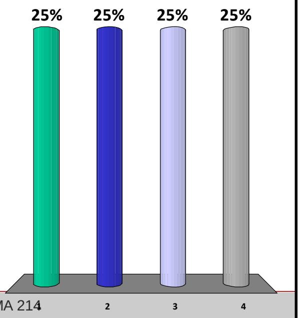
or:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0+\beta_1 x_i$$

- $\frac{p_i}{1-p_i} = \frac{P(Success)}{P(Failure)}$ is called "odds".
- Thus the model states that the log-odds is linearly related to covariates.

In a logistic regression model given by log(odds)=b_0+b_1X, the interpretation of the slope b_1 is……

1. Average change in Y for unit increase in X.
2. Average change in the probability of success for unit increase in X.
3. Average change in the probability of failure for unit increase in X.
4. ✓ Odds changes by e^{b_1} for unit increase in X.



Interpretation of β_1

Model: $log\,\frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_i$

This means:

- For every unit increase in X, *log(odds)* changes by β_1
- Equivalently, for every unit increase in X, *odds* changes by e^{β_1} .

Low Birth Weight Example

Fitted Model: $log\,\frac{p_i}{1-p_i} = 1.0 - 0.014X_i$

- Thus, for each unit increase in mother's weight, the odds of low birth weight changes by $e^{-0.014} \approx 0.986$.
- That is, the likelihood of low birth weight decreases by 1.4% for unit increase of mother's weight.

Logistic regression with multiple covariates

If the model has more than one covariate, it can be incorporated following the same approach as in multiple regression, i.e.,

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_{p-1} x_{i,p-1}$$

For parameter estimation, the least square method is inappropriate. The parameters are estimated using an approach called Maximum Likelihood.

Ordinal Regression Model

Multinomial variable

• Let $Y = j$ be a random variable with more than two possible outcomes $(j = 1,2,\cdots J)$

• The probability $P(Y = j) = p_j$ for $j = 1,2,\cdots J$

•The cumulative probability

$$P(Y \le j) = p_1 + p_2 + \cdots p_j$$

Example

Suppose we formulate a question with 5 categories of response (Y)

Q: The current health plan is awesome…..

1. Completely disagree
2. Disagree
3. Neutral
4. Agree
5. Strongly agree

Example (con't)

If the probability of the categories are p_1, p_2, p_3, p_4, p_5, then the cumulative probabilities are

$$P(Y \le 1) = p_1$$

$$P(Y \le 2) = p_1 + p_2$$

$$P(Y \le 3) = p_1 + p_2 + p_3$$

$$P(Y \le 4) = p_1 + p_2 + p_3 + p_4$$

$$P(Y \le 5) = p_1 + p_2 + p_3 + p_4 + p_5 = 1$$

Ordinal Regression

Model:

$$P(Y_i \le j \mid x_i) = \frac{\exp(\alpha_j + \beta x_i)}{1 + \exp(\alpha_j + \beta x_i)}$$

$$\Leftrightarrow \frac{P(Y_i \le j \mid x_i)}{1 - P(Y_i \le j \mid x_i)} = \exp(\alpha_j + \beta x_i)$$

$$\Leftrightarrow \log\left(\frac{P(Y_i \le j \mid x_i)}{1 - P(Y_i \le j \mid x_i)}\right) = \alpha_j + \beta x_i$$

$$\Leftrightarrow \operatorname{logit}(P(Y_i \le j \mid x_i)) = \alpha_j + \beta x_i$$

$$\Leftrightarrow \log(\text{cumulative odds}) = \alpha_j + \beta x_i$$

Ordinal Regression

- Note that each cumulative logit has its own intercept.

- However, covariate has the same effect on the logit regardless of category.

- In most cases, the coefficient of the covariate(s) β is of primary interest.

Ordinal Regression

- The cumulative logit model satisfies

$$\operatorname{logit}(P(Y_i \le j \mid x_1)) - \operatorname{logit}(P(Y_i \le j \mid x_2)) = \beta(x_1 - x_2)$$

- The odds of having $Y \le j$ at $X = x_1$ are

$$\exp(\beta(x_1 - x_2))$$

times the odds at $X = x_2$

- Thus, the log of odds is proportional to the distance. This property gives the name "Proportional odds model" .

Ordinal Regression

Also note

$$\begin{aligned} & P(Y_i = j \mid x_i) \\ & = P(Y_i \le j \mid x_i) - P(Y_i \le j-1 \mid x_i) \\ & = \frac{\exp(\alpha_j + \beta x_i)}{1 + \exp(\alpha_j + \beta x_i)} - \frac{\exp(\alpha_{j-1} + \beta x_i)}{1 + \exp(\alpha_{j-1} + \beta x_i)} \end{aligned}$$

Example

Agresti (2002): It is a study of mental health for a random sample of adults residents of Alachua County, FL. It relates mental impairment to two covariates.

- Mental (response): Mental impairment is an ordinal response with categories (well, mild, moderate, impaired)

- Life (covariate 1): The life event index is a measure of number of important life events: birth of a child, new job, divorce, death in family, etc within the last 3 years.

- SES (covariate 2): Socio-economic statues (0=low; 1=high)