## Logistic Regression[1]

Learning Objectives

1. Introduce logistic regression
2. Discuss odds
3. Define the logistic regression model
4. Explain how to interpret regression coefficients
5. Provide some additional comments about logistic regression

Introduction

- The MLR model we have examined in this course has considered many different types of IVs (continuous, categorical, interactions, polynomial) but the DV, $Y$, has always been assumed to be continuous
- Logistic regression is a regression technique for binary DV (e.g., college completion, divorce)

Odds

- Before introducing the logistic regression model, we will review odds
- Odds are a numerical expression of the likelihood that an event of interest will occur
    - "Odds for" or "in favor of" event occurrence
    - Ex: the odds of winning the lottery or the odds of surviving a heart attack
- Odds can also be expressed as the likelihood that an event of interest will not occur
    - "Odds against" event occurrence
    - Ex: the odds of not winning the lottery or the odds of not surviving a heart attack
- If the odds of event occurrence are 3, that is written as the odds of that particular event occurring are 3 to 1 or 3:1. For every three individuals who will experience an event, one will not.
- Odds are sometimes expressed as a probability to help with interpretation. Instead of saying the odds of an event are 3 to 1, one might say
    - the odds are 3 in 4 that the event will occur (75% probability of occurrence) or
    - the odds are 1 in 4 that the event will not occur (25% probability of non-occurrence)
- Example: The odds a child lives with at least one sibling is 3.8 to 1. *For every 3.8 children who have at least one sibling, one child will not have a sibling.*

[1] Portions of these notes are adapted from the course notes of Dr. Robert MacCallum (with permission) and from Cohen, J., Cohen, P., West, S.G., & Aiken, L.S. (2003). Do not distribute without the permission of Dr. Shane Hutton.

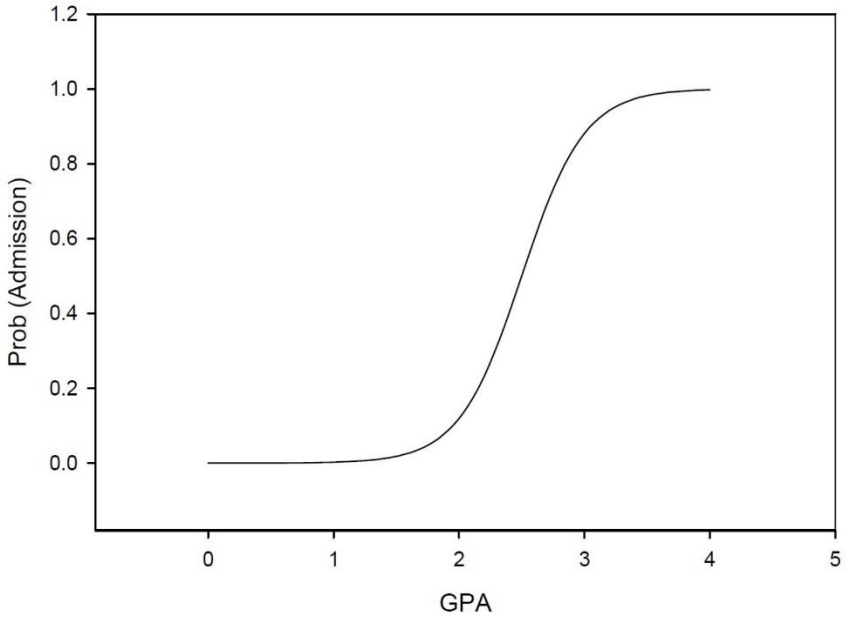
- Odds are computed by dividing the probability of the event occurring by the probability of the event not occurring (i.e., $\frac{p}{1-p}$) and range from 0 to infinity, the greater the probability of event occurrence, the larger the odds

| Odds of Event Occurrence | Probability of Event Occurrence |
|---|---|
| 0 | 0 |
| .2 | .167 |
| .4 | .286 |
| .6 | .375 |
| .8 | .444 |
| 1 | .5 |
| 2 | .667 |
| 5 | .833 |
| 10 | .909 |
| 20 | .952 |
| 30 | .968 |
| ∞ | 1 |

Logistic Regression

- Example:
  - $Y$: Admission to a particular university (1 = admitted, 0 = not admitted)
  - $X$: High school GPA
- For a binary $Y$, traditional linear regression is not appropriate because a binary $Y$ is not a linear function of a continuous $X$
- Traditional linear regression would not produce meaningful predicted values of $Y$
- To overcome this problem, consider that for any value of $X$ there is a proportion of individuals who score a "1" on $Y$
  - Call this proportion $\pi$
- Ex: For applicants with a GPA of 2.5, a particular proportion (e.g., .5) would be admitted, i.e., the probability that an individual with a 2.5 GPA would be admitted is .5

- Typically, the proportions follow a common form across the range of $X$




- This shape is represented by a logistic function
- This common form can be used as a basis for a regression model
    - For a given value of $X$, the model will not predict $Y$, but will predict the probability that $Y = 1$
- The logistic function is represented as

$$\hat{p}_i = \frac{1}{1 + e^{-(b_1 X_i + b_0)}} = \frac{e^{(b_1 X_i + b_0)}}{1 + e^{(b_1 X_i + b_0)}}$$

    - $\hat{p}_i$ is the predicted probability that $Y = 1$
    - Recall $e = 2.71828$ is a constant
- However, the probability $p_i$ is bounded between 0 and 1 and this model is nonlinear
    - We would prefer a continuous dependent variable and a linear model

Logistic Model using Odds

- The model can be reformulated so that, instead of predicting probability $p_i$, it predicts odds
- The odds that $Y = 1$ is defined as $p_i / (1 - p_i)$
- Using the logistic function, the model can be expressed so as to predict odds of $Y = 1$

$$\frac{\hat{p}_i}{1 - \hat{p}_i} = e^{(b_1 X_i + b_0)}$$

- However, odds range from 0 and infinity
- By taking the natural log of both sides, the model can then be expressed so as to predict the logarithm of odds (also called the logit)

$$\ln\left(\frac{\hat{p}_i}{1 - \hat{p}_i}\right) = b_1 X_i + b_0$$

  - This now creates a continuous, unbounded DV and the model in linear

| odds | log odds |
|---|---|
| .001 | -6.91 |
| .01 | -4.61 |
| .50 | -0.69 |
| 1 | 0 |
| 3 | 1.10 |
| 10 | 2.30 |
| 30 | 3.40 |

Interpretation of Coefficients

- For the model

$$\ln\left(\frac{\hat{p}_i}{1 - \hat{p}_i}\right) = b_1 X_i + b_0$$

  - $B_1$ represents the predicted increment in the logit (log odds) for a one-unit increase in $X$
  - $B_0$ represents the predicted logit when $X = 0$

- A useful interpretation of $B_1$ is based on the notion of the odds ratio (OR)
- The odds ratio is the ratio of the odds for one group to the odds for the other group (e.g., the odds of college completion vs. the odds of not completing college)
- For a continuous $X$ in the logistic regression model, consider a one unit increase in $X$, this can be represented as

$$OR = \frac{e^{b_0 + b_1(X+1)}}{e^{b_0 + b_1 X}} = \frac{e^{b_0 + b_1 X + b_1(1)}}{e^{b_0 + b_1 X}} = \frac{e^{b_0 + b_1 X} e^{b_1}}{e^{b_0 + b_1 X}} = e^{b_1}$$

    - For every one unit increase in $X$, the odds that $Y = 1$ is multiplied by $e^{b_1}$
    - Note:
        - if $b_1 > 0$, then $e^{b_1} > 1$ meaning that odds increase for a one unit increase in $X$
        - if $b_1 < 0$, then $e^{b_1} < 1$ meaning that odds decrease for a one unit increase in $X$
        - if $b_1 = 0$, then $e^0 = 1$ meanging there is no effect of $X$ on odds
- Another useful interpretive device involves the value of $-b_0/b_1$
    - This quantity represents the value of $X$ at which the predicted probability that $Y = 1$ will be .50.
- GPA Example:

$$logit(\hat{p}_i) = 4\,GPA - 10$$

    - For every one unit increase in GPA, the odds of being admitted is multiplied by $e^4 = 54.56$
    - A GPA of $-(-10)/4 = 2.50$ corresponds to a .5 chance of admission

Notes on Logistic Regression

- OLS regression is not application for logistic regression and there is no algebraic solution for the coefficients
    - Maximum likelihood is implemented
        - Conceptually, the idea of maximum likelihood estimation is to determine estimates of population parameters that most likely would have given rise to the observed sample data
- The logistic model can be extended to incorporate multiple IVs

$$\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = b_1X_1 + b_2X_2 + \cdots + b_kX_k + b_0$$

- In OLS regression we typically evaluate the quality of the regression model by focusing on the squared multiple correlation but in logistic regression the is no such measure available
    - ML estimation produces a deviance statistic
        - It is a measure of deviation or lack of correspondence between the model and the observed data
    - Deviance statistics are used to test the difference between our model with $k$ predictors and a null model (i.e., a model with no predictors) or a saturated model (i.e., a model that would explain the data perfectly)
        - The difference in deviance between the two models is a test statistic (distributed as chi-square with $k$ degrees of freedom)
- The logistic model can be extended to include more than two categories for $Y$
    - This is called multinomial logistic regression