# Logistic Regression

Sravani Vadlamani

# Agenda

- Overview of Classification
- Logistic Regression
- Multiple Logistic Regression
- Multinomial Logistic Regression

# Introduction

- Linear regression assumes quantitative response variable
- In many situations, response variable is qualitative
- Process for predicting qualitative responses is called classification
  - Assigning an observation to a category and hence predicting a qualitative response to an observation is referred as classifying that observation
  - Methods used for classification predict the probability that the observation belongs to each of the categories of a qualitative variable as the basis for making the classification

# Classification Techniques

- Logistic regression
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Naïve Bayes
- K-nearest neighbors
- Generalized additive models
- Decision Trees, Random forests, Boosting
- Support Vector Machines

# Classification Examples

- Determine whether a student will pass STA 3241 or not
- How do banks determine if a credit card transaction is fraudulent or not
- Determine whether a patient has cancer or not
- A patient has symptoms attributed to one of four medical conditions. Which of the four conditions does this patient have?
- You have three routes to choose from for your morning commute. Which one would you choose?

# Why not Linear Regression for Qualitative Variables

- The coding of qualitative variables are coded as 1, 2, 3 …implies a natural ordering on the outcomes and insists that the difference between any two outcomes is the same when this need not be true.
- A binary qualitative variable can be coded using the dummy variable approach (0 and 1) and a linear regression model can be used to predict response variable ($\hat{Y} > 0.5$). However, some predictions may be outside the 0, 1 interval making it hard to interpret probabilities.

# Logistic Regression

- Logistic regression models the probability that Y belongs to a particular category
- The following linear equation cannot be used to model the probability because

$$p = \beta_0 + \beta_1 X_1$$

  - Probabilities can take values between 0 and 1 (bounded constraint)
  - Non-linear relation between probability and X variables

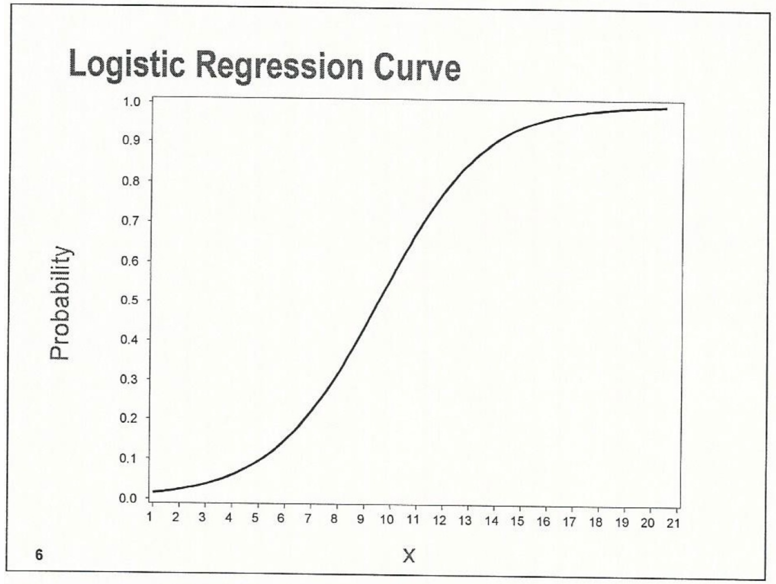
# Logistic Regression

- We use a logistic function to model p(x) such that the output is always between 0 and 1 for all values of X

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1+e^{\beta_0 + \beta_1 x}}$$

- The above equation can be rewritten as

$$\frac{p(x)}{1-p(x)} = e^{\beta_0 + \beta_1 x}$$



# Logistic Regression

- $\frac{p(x)}{1-p(x)}$ is called the odds and takes values between 0 and $\infty$

- For example, a probability of 1 in 5 will yield an odds of 0.25

$$\mathrm{p(x)} = 1/5 = 0.2$$

$$\frac{\mathrm{p(x)}}{1-\mathrm{p(x)}} = \frac{0.2}{1-0.2} = \frac{0.2}{0.8} = 0.25$$

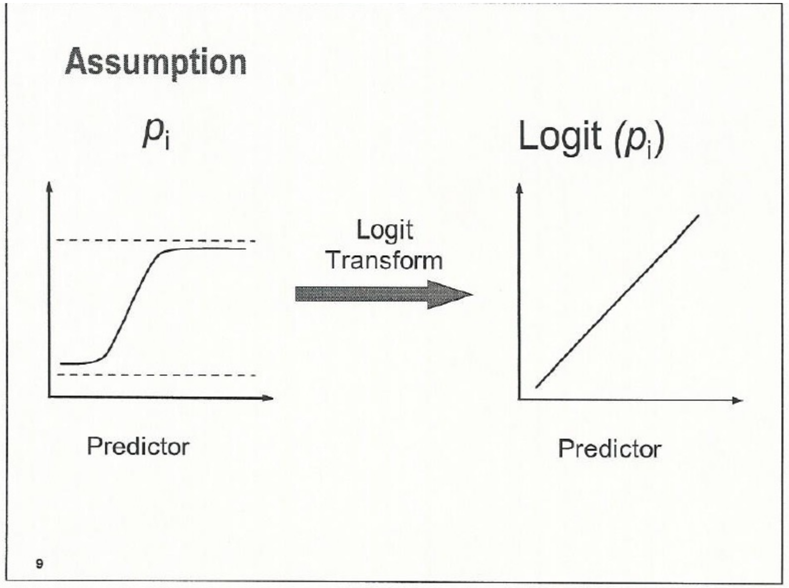
# Logistic Regression

$$\frac{p(x)}{1-p(x)} = e^{\beta_0 + \beta_1 x}$$

- Taking the log of both sides of the above equation will result in log odds or logit.

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x$$

- Logistic regression has a logit that is linear in X



# Important Definitions

- P = probability of event
- Odds is the probability of event divided by the probability of non-event

$$Odds = \frac{P}{1-P}$$

- Logit – Natural log of odds

$$\ln(Odds) = \ln\left(\frac{P}{1-P}\right)$$

- In logistic regression, the log of the odds (logit) is linearly related to the predictor

$$logit(P) = b_0 + b_i x_i$$

# Logistic Regression Parameter Interpretation

- In linear regression, $\beta_1$ gives the average change in Y associated with a one-unit increase in X
- In logistic regression, a one-unit change in X yields a $\beta_1$ change in the log-odds
    - This is equivalent to multiplying the odds by $e^{\beta_1}$
- If $\beta_1$ is positive, increasing X will increase $p(x)$
- If $\beta_1$ is negative, increasing X will decrease $p(x)$
- The rate of change in $p(x)$ per unit change in X depends on the value of X

# Estimating Parameters

- The coefficients $\beta_0 \; and \; \beta_1$ must be estimated based on the training
- Logistic regression uses maximum likelihood to estimate $\beta_0 \; and \; \beta_1$ such that the predicted probability is as close to the observed classes using the following likelihood function

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_i))$$

- The estimates $\beta_0 \; and \; \beta_1$ maximize the above function

# Estimating Parameters

- The accuracy of coefficient estimates is measured using $Z -$ statistic

$$Z(\widehat{\beta_1}) = \frac{\widehat{\beta_1}}{SE(\widehat{\beta_1})}$$

- A large z-statistic offers evidence against the null hypothesis $H_0$: $\beta_1 = 0$
- After estimating the coefficients, $\beta_0$ *and* $\beta_1$ we can make predictions by plugging them into the model equation

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1+e^{\beta_0 + \beta_1 x}}$$

- $\beta_0$ captures the ratio of positive and negative classifications in the given dataset

# Making Predictions

| | Coefficient | Std. error | $z$-statistic | $p$-value |
|---|---|---|---|---|
| `Intercept` | −10.6513 | 0.3612 | −29.5 | <0.0001 |
| `balance` | 0.0055 | 0.0002 | 24.9 | <0.0001 |

**TABLE 4.1.** *For the* `Default` *data, estimated coefficients of the logistic regression model that predicts the probability of* `default` *using* `balance`. *A one-unit increase in* `balance` *is associated with an increase in the log odds of* `default` *by* 0.0055 *units.*

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1,000}}{1 + e^{-10.6513+0.0055\times 1,000}} = 0.00576,$$

# Making Predictions

| | Coefficient | Std. error | $z$-statistic | $p$-value |
|---|---|---|---|---|
| Intercept | −3.5041 | 0.0707 | −49.55 | <0.0001 |
| student[Yes] | 0.4049 | 0.1150 | 3.52 | 0.0004 |

**TABLE 4.2.** *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using student status. Student status is encoded as a dummy variable, with a value of 1 for a student and a value of 0 for a non-student, and represented by the variable* student[Yes] *in the table.*

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

# Multiple Logistic Regression

- Predicting binary response variable with multiple predictors
- The logistic function is given by the following equation

$$p(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p}}$$

- The log odds is given by the following equation

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

- Maximum likelihood is used to estimate $\beta_0, \beta_1, \ldots. \beta_p$

# Multinomial Logistic Regression

- Logistic regression allows for only k=2 classes of the response variable
- For K>2 classes we use multinomial logistic regression
  - One class is chosen to serve as the baseline and left out of the model
  - The log odds between any pair of classes is linear in the features
  - Interpretation of coefficients is tied to the choice of baseline

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}} \tag{4.10}$$

for $k = 1, \ldots, K-1$, and

$$\Pr(Y = K|X = x) = \frac{1}{1 + \sum_{l=1}^{K-1} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}. \tag{4.11}$$

It is not hard to show that for $k = 1, \ldots, K-1$,

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = K|X = x)}\right) = \beta_{k0} + \beta_{k1}x_1 + \cdots + \beta_{kp}x_p. \tag{4.12}$$

# Multinomial Logistic Regression

- Softmax coding
  - Treat all K classes symmetrically and estimate coefficients for all classes

$$\Pr(Y = k|X = x) = \frac{e^{\beta_{k0}+\beta_{k1}x_1+\cdots+\beta_{kp}x_p}}{\sum_{l=1}^{K} e^{\beta_{l0}+\beta_{l1}x_1+\cdots+\beta_{lp}x_p}}. \tag{4.13}$$

Thus, rather than estimating coefficients for $K - 1$ classes, we actually estimate coefficients for all $K$ classes. It is not hard to see that as a result of (4.13), the log odds ratio between the $k$th and $k'$th classes equals

$$\log\left(\frac{\Pr(Y = k|X = x)}{\Pr(Y = k'|X = x)}\right) = (\beta_{k0} - \beta_{k'0}) + (\beta_{k1} - \beta_{k'1})x_1 + \cdots + (\beta_{kp} - \beta_{k'p})x_p. \tag{4.14}$$

- Discriminant analysis is the preferred means of handling multiple- class classification