

# Logistic Regression

## Logistic Regression

- Consider the Multiple Linear Regression Model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_i$$

where the response variable  $y$  is dichotomous,  
taking on only one of two values:

$$\begin{aligned} y &= 1 && \text{if a success} \\ &= 0 && \text{if a failure} \end{aligned}$$

## Logistic Regression

- The response variable,  $y$ , is really just a Bernoulli trial, with

$$E(y) = \pi$$

where

$\pi$  = probability of a success on any given trial

- $\pi$  can only take on values between 0 and 1

## Logistic Regression

- Thus, the Multiple Regression Model

$$\pi = \mu_{y|x} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}$$

is not appropriate for a dichotomous response variable, since this model assumes  $\pi$  can take on any value, but in fact it can only take on values between 0 and 1.

## Logistic Regression

- When the response variable is dichotomous, a more appropriate linear model is the

### **Logistic Regression Model:**

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

## Logistic Regression

- The ratio

$$\frac{\pi}{1-\pi}$$

is called the odds, and is a function of the probability  $\pi$

## Logistic Regression

- Consider the model with one predictor (k=1):

Logit  $\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$

Odds  $\frac{\pi}{1-\pi} = e^{\alpha + \beta x} = e^{\alpha} (e^{\beta})^x$

## Logistic Regression

Odds  $\frac{\pi}{1-\pi} = e^{\alpha} (e^{\beta})^x$

- Every one-unit increase in x increases the odds by a factor of  $e^{\beta}$ .

## Logistic Regression

$$x = 1 \quad \frac{\pi}{1 - \pi} = e^{\alpha} (e^{\beta})$$

$$x = 2 \quad \frac{\pi}{1 - \pi} = e^{\alpha} (e^{\beta})^2 = e^{\alpha} (e^{\beta})(e^{\beta})$$

$$x = 3 \quad \frac{\pi}{1 - \pi} = e^{\alpha} (e^{\beta})^3 = e^{\alpha} (e^{\beta})(e^{\beta})(e^{\beta})$$

## Logistic Regression

- Thus,  $e^{\beta}$  is the odds ratio, comparing the odds at  $x+1$  with the odds at  $x$
- An odds ratio equal to 1 (i.e.,  $e^{\beta} = 1$ ) occurs when  $\beta = 0$ , which describes the situation where the predictor  $x$  has no association with the response  $y$ .

## Logistic Regression

- As with regular linear regression, we obtain a sample of  $n$  observations, with each observation measured on all  $k$  predictor variables and on the response variable.
- We use these sample data to fit our model and estimate the parameters

## Logistic Regression

- Using the sample data, we obtain the model:

$$\log\left(\frac{p}{1-p}\right) = a + bx$$

for a single predictor

# Logistic Regression

- The estimate of  $\pi$  is

$$p = \frac{e^{a+bx}}{1+e^{a+bx}}$$

## Example: Coronary Heart Disease

<u>Age</u>	<u>Age Group</u>	<u>CHD Present</u>	<u>N</u>	<u>p</u>	<u>Odds p/(1-p)</u>	<u>Log(Odds) log(p/(1-p))</u>
25	1	1	10	0.10	0.11111	-2.19722
30	2	2	15	0.13	0.15385	-1.87180
35	3	3	12	0.25	0.33333	-1.09861
40	4	5	15	0.33	0.50000	-0.69315
45	5	6	13	0.46	0.85714	-0.15415
50	6	5	8	0.63	1.66667	0.51083
55	7	13	17	0.76	3.25000	1.17865
60	8	8	10	0.80	4.00000	1.38629

## Example: Coronary Heart Disease

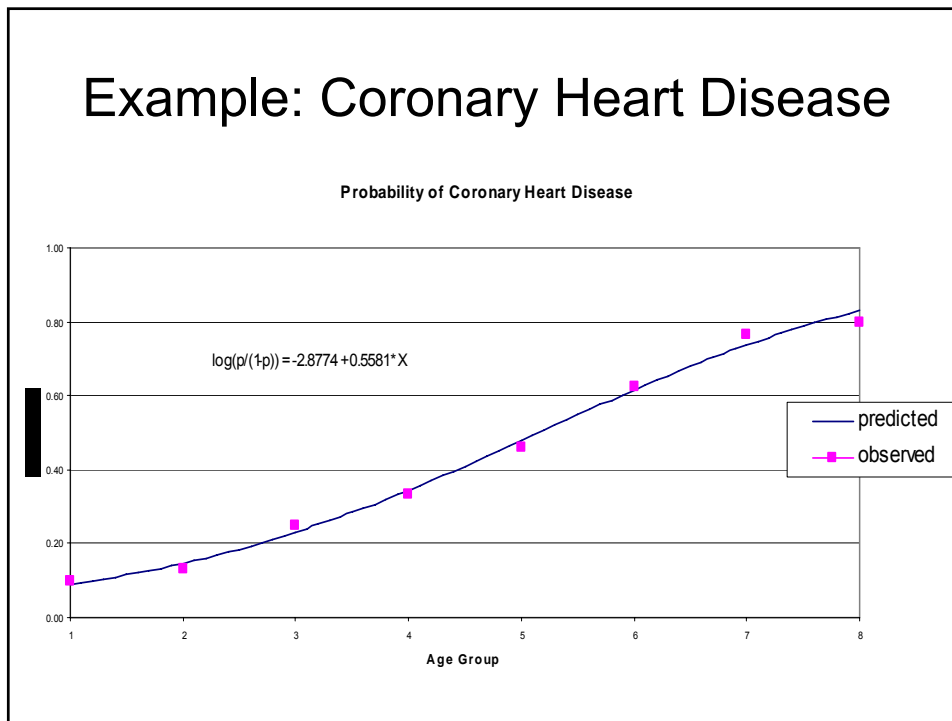
Age	Age Group x	CHD Present	N	p	Predicted Odds $\exp(a+b*x)$	Predicted Log(Odds) $a+b*x$	Predicted p-hat $1/(1+\exp(-a-b*x))$
25	1	1	10	0.10	0.09834	-2.31929	0.08954
30	2	2	15	0.13	0.17184	-1.76117	0.14664
35	3	3	12	0.25	0.30028	-1.20305	0.23093
40	4	5	15	0.33	0.52470	-0.64493	0.34413
45	5	6	13	0.46	0.91685	-0.08681	0.47831
50	6	5	8	0.63	1.60209	0.47131	0.61569
55	7	13	17	0.76	2.79947	1.02943	0.73681
60	8	8	10	0.80	4.89175	1.58755	0.83027

$\text{Exp}(b) =$   
 Odds Ratio  
 $= 1.74738$

$b = 0.55812$   
 $a = -2.87741$

Multiplicative      Additive

## Example: Coronary Heart Disease



## Parameter Estimation

- A 100(1- $\alpha$ )% Confidence Interval for  $\beta$  is:

$$\hat{\beta} \pm z_{\alpha/2} \times \text{a.s.e.}(\hat{\beta})$$

where  $z_{\alpha/2}$  is a critical value from the standard normal distribution, and a.s.e. stands for the asymptotic standard error

## Parameter Estimation

- A 100(1- $\alpha$ )% Confidence Interval for the odds ratio  $e^{\beta}$  is:

$$e^{\hat{\beta} \pm z_{\alpha/2} \times \text{a.s.e.}(\hat{\beta})}$$

## Hypothesis Testing

$$H_0: \beta = 0$$

$$H_1: \beta \neq 0$$

The test statistics for testing this hypothesis are  $\chi^2$  statistics. There are 3 that are commonly used:

## Hypothesis Testing

Wald Test  $Q_W = \left( \frac{\hat{\beta}}{\text{a.s.e.}(\hat{\beta})} \right)^2 \sim \chi_1^2$

Likelihood Ratio Test  $Q_{LR} = -2(L_\alpha - L_{\alpha,\beta}) \sim \chi_1^2$

Score Test  $Q_S \sim \chi_1^2$

## Goodness of Fit

- Let  $m$  = # of levels of  $x$  ( $m=8$  for CHD example)
- Let  $n_i$  = number of observations in the  $i^{\text{th}}$  level of  $x$
- Let  $k$  = number of parameters ( $k=2$  for CHD example)

$$X_{\text{Pearson}}^2 = \sum_{i=1}^m n_i \frac{(p_i - \hat{\pi}_i)^2}{\hat{\pi}_i} \sim \chi_{m-k}^2$$

$$X_{\text{Deviance}}^2 = 2 \sum_{i=1}^m n_i p_i \log\left(\frac{p_i}{\hat{\pi}_i}\right) \sim \chi_{m-k}^2$$

## Hosmer-Lemeshow Goodness of Fit Test

- Used when multiple observations do not occur for each value of some predictor variable(s).
- Fit the model, and estimate  $\pi_i$  with  $\hat{\pi}_i$
- Divide the observations into (approximately) 10 groups, of approximately equal size, based on the percentiles of  $\hat{\pi}_i$

## Hosmer-Lemeshow Goodness of Fit Test

$$Q_{HL} = \sum_{i=1}^g \frac{(O_i - N_i \bar{\pi}_i)^2}{N_i \bar{\pi}_i (1 - \bar{\pi}_i)} \sim \chi_{g-2}^2$$

where

$g$  = # of groups

$O_i$  = observed # of events in  $i^{\text{th}}$  group

$N_i$  = # of subjects in  $i^{\text{th}}$  group

$\bar{\pi}_i$  = average estimated probability of an event in the  $i^{\text{th}}$  group

### Example: Coronary Heart Disease

- This example consists of  $n=100$  observations
- When these data are written one observation per line, the data looks like:

Logistic Regression Example  
Coronary Heart Disease vs. Age  
(first 25 observations, out of 100)

Obs	Age	Age Group	CHD
1	25	1	1
2	25	1	0
3	25	1	0
4	25	1	0
5	25	1	0
6	25	1	0
7	25	1	0
8	25	1	0
9	25	1	0
10	25	1	0
11	30	2	1
12	30	2	1
13	30	2	0
14	30	2	0
15	30	2	0
16	30	2	0
17	30	2	0
18	30	2	0
19	30	2	0
20	30	2	0
21	30	2	0
22	30	2	0
23	30	2	0
24	30	2	0
25	30	2	0

## SAS Code: Proc Logistic

```
Proc Logistic Data=CHD100 Descending;  
  Model CHD=AgeGroup / Scale=none Aggregate;  
Run;
```

Pearson and Deviance Goodness of Fit statistics are calculated and printed when

SCALE=option

is specified.

SCALE=NONE

requests statistics with no adjustment of overdispersion.

The AGGREGATE option specifies the subpopulations on which these statistics are calculated. Observations with common values in the list of variables define each subpopulation.

## Proc Logistic Output: CHD Example

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	28.4851	1	<.0001
Score	26.0782	1	<.0001
Wald	21.4281	1	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.8770	0.6233	21.3024	<.0001
AgeGroup	1	0.5580	0.1206	21.4281	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
AgeGroup	1.747	1.380	2.213

## Proc Logistic Output: CHD Example

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > ChiSq
Deviance	6	0.2164	0.0361	0.9998
Pearson	6	0.2178	0.0363	0.9998

Number of unique profiles: 8

## Coronary Heart Disease Example

- The predictor variable, or input variable, in this example is AgeGroup, and it takes on values 1 – 8.
- AgeGroup can be regarded as measured on an interval scale, and each one-unit increase in AgeGroup represents a 5-year increase in Age.
- If we had the actual ages of each of the 100 individuals, the data set would look like:

AgeCutpoint	Age	AgeGroup	CHD
25	22	1	1
25	21	1	0
25	23	1	0
25	21	1	0
25	22	1	0
25	25	1	0
25	23	1	0
25	24	1	0
25	24	1	0
25	25	1	0
30	28	2	1
30	26	2	1
30	30	2	0
30	29	2	0
30	28	2	0
30	27	2	0
30	30	2	0
30	28	2	0
30	27	2	0
30	29	2	0
30	26	2	0
30	30	2	0
30	27	2	0
30	26	2	0
30	29	2	0

## SAS Code for Logistic Regression

```
Proc Logistic Data=Miner.CHD100 Descending;  
Model CHD=Age / Lackfit;  
Units Age=5;  
Output Out=C1 p=pi_hat;  
Run;
```

The LACKFIT option requests that the Hosmer-Lemeshow Goodness of Fit test is calculated and printed.

## Proc Logistic Output: CHD Example

The LOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	28.6373	1	<.0001
Score	26.1600	1	<.0001
Wald	21.4538	1	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.8815	1.0374	22.1433	<.0001
Age	1	0.1114	0.0240	21.4538	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
Age	1.118	1.066 1.172

## Proc Logistic Output: CHD Example

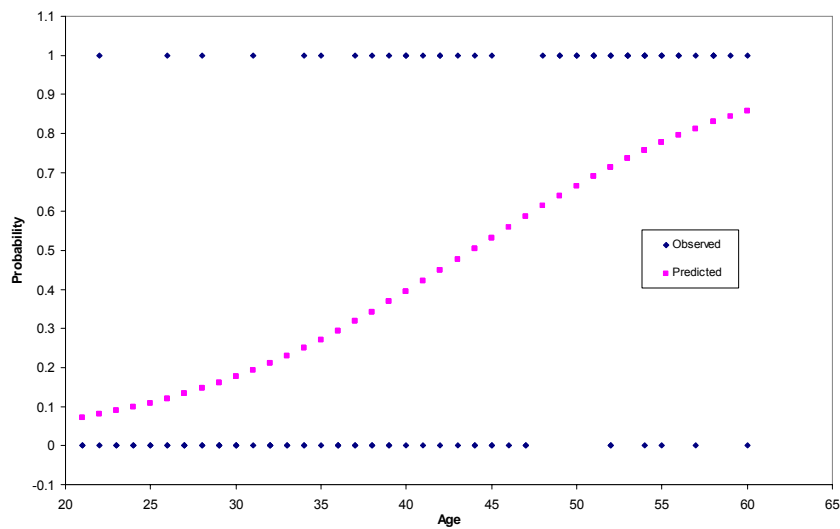
### Adjusted Odds Ratios

Effect	Unit	Estimate
Age	5.0000	1.745

### Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
1.6102	7	0.9782

Logistic Regression Model predicting CHD from Age



## Challenger Shuttle Disaster

- For the 23 space shuttle flights that occurred before the Challenger mission disaster in 1986, the following data set shows the temperature, in °F, at the time of the flight, and whether or not at least one primary O-ring suffered thermal distress.

<b>Flight</b>	<b>Temp</b>	<b>Thermal Distress</b>	<b>Flight</b>	<b>Temp</b>	<b>Thermal Distress</b>
1	66	0	13	67	0
2	70	1	14	53	1
3	69	0	15	67	0
4	68	0	16	75	0
5	67	0	17	70	0
6	72	0	18	81	0
7	73	0	19	76	0
8	70	0	20	79	0
9	57	1	21	75	1
10	63	1	22	76	0
11	70	1	23	58	1
12	78	0			

## Predicting Thermal Distress from Temperature

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	7.9520	1	0.0048
Score	7.2312	1	0.0072
Wald	4.6008	1	0.0320

Analysis of Maximum Likelihood Estimates

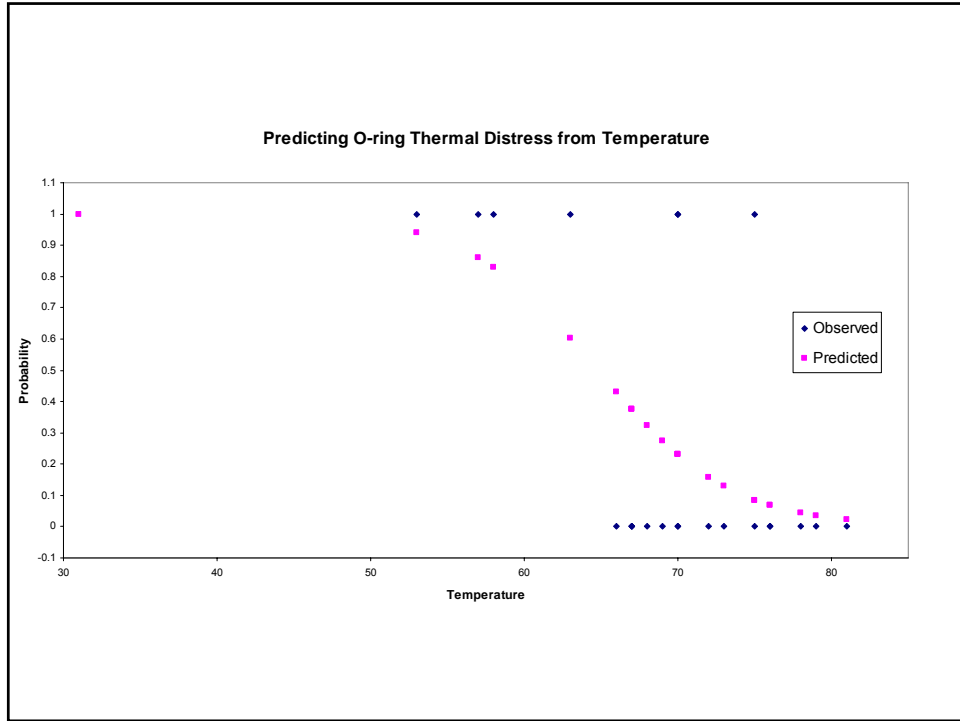
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	15.0429	7.3786	4.1563	0.0415
Temp	1	-0.2322	0.1082	4.6008	0.0320

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Temp	0.793	0.641	0.980

## Predicting Thermal Distress from Temperature

- The temperature at the time of the final Challenger flight was 31°.
- From the logistic regression model, the predicted probability of thermal distress in one or more of the O-rings was 0.99961.



## Multiple Logistic Regression

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

## Testing Hypotheses about the $\beta$ 's

Let  $k=4$

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

```
Proc Logistic Data=A Descending;  
Model Y=X1 X2 X3 X4;  
Test X1=0;          *Tests H0:  $\beta_1=0$ ;  
Test X1=X2=0;      *Tests H0:  $\beta_1=\beta_2=0$ ;  
Test X1=X2;        *Tests H0:  $\beta_1=\beta_2$ ;  
Run;
```

## Residuals

Suppose there are  $g$  groups, with  $n_i$  observations in the  $i^{\text{th}}$  group.

Pearson residuals

$$e_i = \frac{y_i - n_i \hat{\pi}_i}{\sqrt{n_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$$

Deviance residuals

$$d_i = \text{sgn}(y_i - n_i \hat{\pi}_i) \left[ 2y_i \log\left(\frac{y_i}{n_i \hat{\pi}_i}\right) + 2(n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - n_i \hat{\pi}_i}\right) \right]^{\frac{1}{2}}$$

## Influence Diagnostic Statistics

$h_i$  = Diagonal Elements of the Hat Matrix,

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Adjusted (Standardized) Pearson Residuals

$$e_{s_i} = \frac{e_i}{\sqrt{1-h_i}}$$

## Influence Diagnostic Statistics

DFBETA<sub>ij</sub> = standardized difference in the  $j^{\text{th}}$  estimated  $\beta$ ,  $b_j$ , due to deleting the  $i^{\text{th}}$  observation

Confidence Interval displacement (analogous to Cook's D in linear regression)

$$C_i = e_i^2 \frac{h_i}{(1-h_i)^2} \qquad \bar{C}_i = e_i^2 \frac{h_i}{(1-h_i)}$$

## Influence Diagnostic Statistics

Change in Goodness of Fit Statistics due to eliminating the  $i^{\text{th}}$  observation

$$\text{DIFCHISQ} = \frac{\bar{C}_i}{h_i} = \frac{e_i^2}{1-h_i} = e_{S_i}^2$$

$$\text{DIFDEV} = d_i^2 + \bar{C}_i$$

## Multicollinearity

- Since multicollinearity is strictly a condition of the **X** matrix, use PROC REG to assess multicollinearity.
- Ignore all output except the values of the Variance Inflation Factors (VIFs), and the matrix of variance proportions.

```
Proc Reg Data=A;  
  Model Y=X1 X2 X3 X4 / VIF Collin ColliNoInt;  
Run;
```

# Variable Selection Techniques

## Best Subset Regression

SELECTION = SCORE

- Finds specified number of models with highest score statistic  $\chi^2$  for all possible model sizes (1, 2, ..., k variable models)

BEST = number

Can specify minimum model size and/or maximum model size

START = minimum

STOP = maximum

# Variable Selection Techniques

## Stepwise Regression Algorithms

### 1. Forward Selection

Adds variables, one at a time, based on score  $\chi^2$  statistic (smallest p-value < SLENTY=level)

or

until residual  $\chi^2$  becomes nonsignificant STOPRES

## Variable Selection Techniques

### Stepwise Regression Algorithms

#### 2. Backward Elimination

Eliminates variables, one at a time,  
based on Wald  $\chi^2$  statistic  
(largest p-value < SLSTAY=level)

or

until residual  $\chi^2$  becomes significant      STOPRES

## Variable Selection Techniques

### Stepwise Regression Algorithms

#### 3. Stepwise Selection

Forward Selection, with Backward Elimination  
also done at each step after 2 or more  
variables are in the model.

## Generalized Coefficient of Multiple Determination

Model 1:  $\text{logit}(\pi) = \alpha$

Model 2:  $\text{logit}(\pi) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$

$$R^2 = 1 - \left( \frac{L_1}{L_2} \right)^{\frac{2}{n}}$$

Where L is the value  
of the log-likelihood

## Generalized Coefficient of Multiple Determination

It can be shown that

$$\max(R^2) = R_{\max}^2 = 1 - (L_1)^{2/n}$$

For example, if 50% of the observations have  $y=1$ ,  
and 50% have  $y=0$ , then  $\max(R^2)=0.75$

To correct this problem,  
an adjusted  $R^2$  has been proposed:

$$R_{\text{adj}}^2 = \frac{R^2}{R_{\max}^2}$$

## Existence of Maximum Likelihood Estimates

- The likelihood equation for a logistic regression model does not always have a finite solution.
- Sometimes there is a nonunique maximum on the boundary of the parameter space, at infinity.
- The existence, finiteness, and uniqueness of maximum likelihood estimates for the logistic regression model depend on the patterns of data points in the observation space.

## Existence of Maximum Likelihood Estimates

- There are three mutually exclusive and exhaustive types of data configurations:
  - complete separation,
  - quasi-complete separation, and
  - overlap
- Let  $Y$  denote a binary response.
- Let  $\mathbf{X}$  denote the vector of explanatory variables, including the intercept.

## Complete Separation

There is a complete separation of data points if there exists a vector  $\mathbf{b}$  that correctly allocates all observations to their response groups, that is:

$$\mathbf{b}'\mathbf{X} > 0 \quad Y = 1$$

$$\mathbf{b}'\mathbf{X} < 0 \quad Y = 0$$

This configuration gives nonunique infinite estimates. If the iterative process of maximizing the likelihood function is allowed to continue, the log likelihood diminishes to zero, and the dispersion matrix becomes unbounded.

## Quasi-Complete Separation

The data are not completely separable but there is a vector  $\mathbf{b}$  such that

$$\mathbf{b}'\mathbf{X} \geq 0 \quad Y = 1$$

$$\mathbf{b}'\mathbf{X} \leq 0 \quad Y = 0$$

and equality holds for at least one subject in each response group. This configuration also yields nonunique infinite estimates. If the iterative process of maximizing the likelihood function is allowed to continue, the dispersion matrix becomes unbounded and the log likelihood diminishes to a nonzero constant.

## Overlap

If neither complete nor quasi-complete separation exists in the sample points, there is an overlap of sample points.

In this configuration, the maximum likelihood estimates exist and are unique.

## Conditions for Separation

- Complete separation and quasi-complete separation are problems typically encountered with small data sets.
- Although complete separation can occur with any type of data, quasi-complete separation is not likely with truly continuous explanatory variables.
- The most common cause of quasi-complete separation in predictive modeling is categorical inputs with rare categories. The best remedy for sparseness is collapsing levels of the categorical variable.

## Logistic Discriminant Analysis with 2 or more Populations

Example: 3 populations

$$\log\left(\frac{\pi_2}{\pi_1}\right) = \alpha_1 + \boldsymbol{\beta}'_1 \mathbf{x}$$

$$\log\left(\frac{\pi_3}{\pi_1}\right) = \alpha_2 + \boldsymbol{\beta}'_2 \mathbf{x}$$

## Logistic Discriminant Analysis with 2 or more Populations

$$\pi_1 = \frac{1}{1 + e^{\alpha_1 + \boldsymbol{\beta}'_1 \mathbf{x}} + e^{\alpha_2 + \boldsymbol{\beta}'_2 \mathbf{x}}}$$

$$\pi_2 = \frac{e^{\alpha_1 + \boldsymbol{\beta}'_1 \mathbf{x}}}{1 + e^{\alpha_1 + \boldsymbol{\beta}'_1 \mathbf{x}} + e^{\alpha_2 + \boldsymbol{\beta}'_2 \mathbf{x}}}$$

$$\pi_3 = \frac{e^{\alpha_2 + \boldsymbol{\beta}'_2 \mathbf{x}}}{1 + e^{\alpha_1 + \boldsymbol{\beta}'_1 \mathbf{x}} + e^{\alpha_2 + \boldsymbol{\beta}'_2 \mathbf{x}}}$$

## Logistic Discriminant Analysis with 2 or more Populations

Estimate  $\pi_1, \pi_2, \dots$

Classify  $\mathbf{x}$  into the population corresponding  
to the largest  $\hat{\pi}$