Math 5305 Notes
Logistic Regression and Discriminant Analysis

Jesse Crawford

Department of Mathematics
Tarleton State University
1. Logistic Regression

2. Discriminant Analysis
Output variable $Y$ is dichotomous ($Y_i = 0$ or $Y_i = 1$)

$$g_i = X_i \beta = \beta_1 X_{i1} + \cdots + \beta_p X_{ip}, \text{ for } i = 1, \ldots, n.$$  

$$P(Y_i = 1) = \pi_i = \frac{e^{g_i}}{1 + e^{g_i}}, \text{ for } i = 1, \ldots, n.$$
Likelihood function

\[ L = \prod_{i=1}^{n} \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i} \]

Likelihood equations

\[ \sum_{i=1}^{n} X_{ij}(Y_i - \pi_i) = 0, \text{ for } j = 1, \ldots, p. \]
Example in R

True Model

\[ g_i = -3 + 0.06 X_i, \text{ for } i = 1, \ldots, 100000. \]

\[
X = \text{runif}(100000, 0, 100)
\]

\[
g = -3 + 0.06 \times X
\]

\[
\Pi = \frac{\exp(g)}{1 + \exp(g)}
\]

\[
U = \text{runif}(100)
\]

\[
Y = (U < \Pi) \times 1
\]
True Model

\[ g_i = -3 + 0.06X_i, \text{ for } i = 1, \ldots, 100000. \]

```r
model = glm(Y ~ X, family = binomial)
summary(model)
```

```
Call:
  glm(formula = Y ~ X, family = binomial)

Deviance Residuals:
     Min      1Q  Median      3Q     Max
-2.4339 -0.6851 -0.3054  0.6772  2.5390

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.1834684  0.0205484 -154.9 <2e-16 ***
X             0.0609352  0.0003656   166.7 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

  Null deviance: 138436  on 99999  degrees of freedom
  Residual deviance:  92361  on 99998  degrees of freedom
  AIC: 92365

Number of Fisher Scoring iterations: 4
```
Plots

$Y$ vs. $X$ (Not very useful).
Plots

\( \hat{\pi} \) vs. \( X \)
Plots

$\hat{g} \text{ vs. } X$ (Best plot for assessing functional form)
Plots

\hat{g} \text{ vs. } X \text{ (Best plot for assessing functional form)}
Deviance and AIC

Deviance \(= -2 \ln(L) = -2 \sum_{i=1}^{n} Y_i \ln(\hat{\pi}_i) + (1 - Y_i) \ln(1 - \hat{\pi}_i)\)

AIC \(= 2p - 2 \ln(L)\)

\(\hat{g}_i = X_i\hat{\beta} = \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}, \text{ for } i = 1, \ldots, n.\)

\(\hat{\pi}_i = \frac{e^{\hat{g}_i}}{1 + e^{\hat{g}_i}}, \text{ for } i = 1, \ldots, n.\)
Consider the logistic regression model

\[ P(Y_i = 1) = \frac{e^{g_i}}{1 + e^{g_i}}, \text{ where} \]

\[ g_i = X_i \beta. \]

Let \( V_0 \leq V \leq \mathbb{R}^p \), and consider the testing problem

\[ H_0 : \beta \in V_0 \text{ vs. } H : \beta \in V. \]

The test statistic is \( G = D_0 - D \), where \( D_0 \) and \( D \) are the deviances under \( H_0 \) and \( H \), respectively.

Under \( H_0 \), the approximate distribution of \( G \) is chi-square with \( \text{dim}(V) - \text{dim}(V_0) \) degrees of freedom, so

\[ \text{reject } H_0 \text{ if } G > \chi^2_{\alpha}(\text{dim}(V) - \text{dim}(V_0)). \]
Assessing the Model

- Functional form:
  - Group plots
  - Likelihood ratio tests

- Overall Performance
  - Classification Accuracy
  - Area under ROC Curve
Classification Accuracy

- Choose a cutoff value, and use the classification rule
  - If $\hat{\pi}_i > \text{cutoff}$, then $\hat{Y}_i = 1$
  - If $\hat{\pi}_i < \text{cutoff}$, then $\hat{Y}_i = 0$.

- The classification accuracy is percentage of observations that were correctly classified (percentage of cases where $Y_i = \hat{Y}_i$).

\[
\text{Classification Accuracy} = P(Y_i = \hat{Y}_i)
\]

- To optimize classification accuracy, a reasonable cutoff to use is 0.5.
Sensitivity and Specificity

- Sensitivity = $P(\hat{Y}_i = 1 \mid Y_i = 1)$
- Specificity = $P(\hat{Y}_i = 0 \mid Y_i = 0)$
Variable Selection

- Manually
- Stepwise
- Best subsets
Outline

1. Logistic Regression

2. Discriminant Analysis
Discriminant Analysis

- Used when output variable $Y$ is categorical.
- Assume $Y$ is categorical with possible values $0, \ldots, k$.
- Let $X$ be a vector of input variables.
- Given an observation $X = x$, we want to predict the value of $Y$.
- Can also be viewed as a classification problem.

Example

- $Y =$ grade in Biol 120 ($Y = 1$ or $Y = 0$)
- $X =$ student’s high school rank ($0 \leq X \leq 1$)
• Y is a *discrete random variable*.
• It has a p.m.f.

\[ f(y) = P(Y = y), \text{ for } y = 0, \ldots, k. \]

• For each value of Y, the vector X has a conditional distribution given by

\[ f(x \mid y) \]

• The conditional p.m.f. of Y given \( X = x \) is

\[
P(Y = y \mid X = x) = f(y \mid x) = \frac{f(x, y)}{f(x)} = \frac{f(y)f(x \mid y)}{\sum_{y=0}^{k} f(y)f(x \mid y)}
\]
The conditional p.m.f. of $Y$ given $X = x$ is

\[
P(Y = y \mid X = x) = f(y \mid x) = \frac{f(x, y)}{f(x)} = \frac{f(y)f(x \mid y)}{\sum_{y=0}^{k} f(y)f(x \mid y)}
\]

Given the observation $X = x$, we predict $Y$ will be equal to the value of $y$ maximizing $f(y)f(x \mid y)$. 
Given $Y = y$, $X \sim N(\mu_y, \Sigma_y)$, for $y = 0, \ldots, k$.

$$f(x \mid y) = (2\pi)^{-p/2}|\Sigma_y|^{-1/2} \exp\left\{-\frac{1}{2}(x - \mu_y)'\Sigma_y^{-1}(x - \mu_y)\right\}$$

If $X = x$, we predict $Y$ will be equal to the value of $y$ minimizing

$$d_y^2(x) = -2 \ln[f(y)] + \ln|\Sigma_y| + (x - \mu_y)'\Sigma_y^{-1}(x - \mu_y)$$

In practice, we would use

$$\hat{d}_y^2(x) = -2 \ln[\hat{f}(y)] + \ln|S_y| + (x - \bar{x}_y)'S_y^{-1}(x - \bar{x}_y)$$
In practice, we would use

\[ \hat{d}_y^2(x) = -2 \ln[\hat{f}(y)] + \ln |S_y| + (x - \bar{x}_y)'S_y^{-1}(x - \bar{x}_y) \]

How can we estimate these quantities?

Assume we have observations for \( Y_i \) and \( X_i \), for \( i = 1, \ldots, n \).

\[ \hat{f}(y) = \frac{\text{Number of times } Y_i = y}{n} \]
Let \( x_1, \ldots, x_n \in \mathbb{R}^p \) be observations from \( N(\mu, \Sigma) \).

The estimate for the mean \( \mu \) is the sample mean \( \bar{x} \).

\[
\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

The estimate for the covariance matrix \( \Sigma \) is the empirical covariance matrix \( S \).

\[
\hat{\Sigma} = S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})'
\]

If \( X \) is a matrix whose rows are \( x_1, \ldots, x_n \) then \( \bar{x} \) and \( S \) can be obtained with the R commands \( \text{colMeans}(X) \) and \( \text{cov}(X) \).
In practice, we would use

$$\hat{d}_y^2(x) = -2 \ln[f(y)] + \ln |S_y| + (x - \bar{x}_y)'S_y^{-1}(x - \bar{x}_y)$$

For each $y$, set aside all rows of data where $Y_i = y$.

$\bar{x}_y$ and $S_y$ are the sample mean and covariance matrix for the vectors $x_i$ from these rows of data.

For each $y$, let $x_{y1}, \ldots, x_{yn_y}$ be the values of $X_i$ for those subjects with $Y_i = y$. 
Highschool Rank and Biol 120 Grade

trank=rank[1:2000]
tgrade=grade[1:2000]
vrank=rank[2001:3146]
vgrade=grade[2001:3146]

L=groupplot(trank,tgrade,20)
plot(L$x,L$g)
model = glm(tgrade ~ trank, family = binomial)
betahat = coef(model)

x = (1:100)/100
lines(x, betahat[1] + betahat[2]*x, col = 'red')
model2=glm(tgrade~trank+I(trank^2), family=binomial)

betahat2=coef(model2)

lines(x, betahat2[1]+betahat2[2]*x +betahat2[3]*x^2, col='red')
model3 = glm(tgrade ~ trank + I(trank^2) + I(trank^3), family = binomial)
betahat3 = coef(model3)

lines(x, betahat3[1] + betahat3[2]*x + betahat3[3]*x^2 + betahat3[4]*x^3, col='red')
Model with 4th Order Term

![Graph of Model with 4th Order Term]

- L$g$ is plotted on the y-axis.
- L$x$ is plotted on the x-axis.

(Tarleton State University)
Logistic Reg. and Discr. Analysis
Likelihood Ratio Tests

```r
> LRtest(model, model2)
[1] 5.34558e-11
> LRtest(model2, model3)
[1] 0.0006829123
> LRtest(model3, model4)
[1] 0.9089255
```
For $i = 2001, \ldots, 3146$, we predict $\hat{Y}_i = 1$ if $\hat{\pi}_i \geq \frac{1}{2}$.

$\hat{\pi}_i \geq \frac{1}{2}$ iff $g(x_i) \geq 0$.

$$g(x) = -2.89 + 11.63x - 24.19x^2 + 18.92x^3$$

$g(x_i) \geq 0$ iff $x \geq .71929$
Classification Accuracy

\[ \text{vrank} = \text{rank}[2001:3146] \]
\[ \text{vgrade} = \text{grade}[2001:3146] \]
\[ \text{vgradehat} = (\text{vrank} \geq 0.71929) \times 1 \]
\[ \text{mean(} \text{vgradehat} == \text{vgrade}) \]

Classification Accuracy = \( P(Y_i = \hat{Y}_i) = 0.699 \)
If $X = x$, we predict $Y$ will be equal to the value of $y$ minimizing

$$
\hat{d}_y^2(x) = -2 \ln[f(y)] + \ln |S_y| + (x - \bar{x}_y)'S_y^{-1}(x - \bar{x}_y)
$$

```
rank0=trank[tgrade==0]
rank1=trank[tgrade==1]
n0=length(rank0)
n1=length(rank1)
f0=n0/(n0+n1)
f1=n1/(n0+n1)
```
If $X = x$, we predict $Y$ will be equal to the value of $y$ minimizing

$$
\hat{d}_y^2(x) = -2 \ln[\hat{f}(y)] + \ln |S_y| + (x - \bar{x}_y)' S_y^{-1} (x - \bar{x}_y)
$$

```python
rank0 = trank[tgrade==0]
rank1 = trank[tgrade==1]

xbar0 = mean(rank0)
xbar1 = mean(rank1)

s0 = sd(rank0)
s1 = sd(rank1)
```
If \( X = x \), we predict \( Y \) will be equal to the value of \( y \) minimizing

\[
\hat{d}_y^2(x) = -2 \ln[\hat{f}(y)] + \ln |S_y| + (x - \bar{y}_y)' S_y^{-1} (x - \bar{y}_y)
\]

allranks=(1:1000)/1000

d0=-2*log(f0)+log(s0^2)+(allranks-xbar0)^2/s0^2
d1=-2*log(f1)+log(s1^2)+(allranks-xbar1)^2/s1^2
cbind(allranks, d0, d1, d0<d1)

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>643</td>
<td>0.643</td>
<td>-1.9274647498</td>
<td>-1.868919021</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>644</td>
<td>0.644</td>
<td>-1.9235675598</td>
<td>-1.874871637</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>645</td>
<td>0.645</td>
<td>-1.9196068876</td>
<td>-1.880758028</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>646</td>
<td>0.646</td>
<td>-1.9155827331</td>
<td>-1.886578196</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>647</td>
<td>0.647</td>
<td>-1.9114950963</td>
<td>-1.892332140</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>648</td>
<td>0.648</td>
<td>-1.9073439772</td>
<td>-1.898019860</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>649</td>
<td>0.649</td>
<td>-1.9031293759</td>
<td>-1.903641356</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>650</td>
<td>0.650</td>
<td>-1.8988512923</td>
<td>-1.909196628</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>651</td>
<td>0.651</td>
<td>-1.8945097264</td>
<td>-1.914685677</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>652</td>
<td>0.652</td>
<td>-1.8901046782</td>
<td>-1.920108501</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>653</td>
<td>0.653</td>
<td>-1.8856361478</td>
<td>-1.925465101</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>654</td>
<td>0.654</td>
<td>-1.8811041351</td>
<td>-1.930755478</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

Discr. Analysis Optimal Cutoff $= 0.649$
vgradehat = (vrank >= .649) * 1
mean (vgradehat == vgrade)

Classification Accuracy = P(Y_i = \hat{Y_i}) = 0.702
"Brute Force" Approach

```r
allranks = (1:1000)/1000

classacc = 1:1000

for(i in 1:1000){
  tempgradehat = (trank >= allranks[i]) * 1
  classacc[i] = mean(tempgradehat == tgrade)
}

max(classacc)

allranks[classacc == max(classacc)]
```

Optimal Cutoffs = (.717, .719, .720, .721)

Optimal Cutoffs = .7195
