Math 5305 Notes

Logistic Regression and Discriminant Analysis

Jesse Crawford

Department of Mathematics
Tarleton State University

Outline

1. Logistic Regression

2. Discriminant Analysis

Logistic Regression Models

Output variable Y is dichotomous ($Y_i = 0$ or $Y_i = 1$)

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

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

Maximum Likelihood Estimation

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

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

```
X=runif(100000,0,100)

g=-3+.06*X
Pi=(exp(g)/(1+exp(g)))

U=runif(100)
Y=(U<Pi)*1
```

True Model

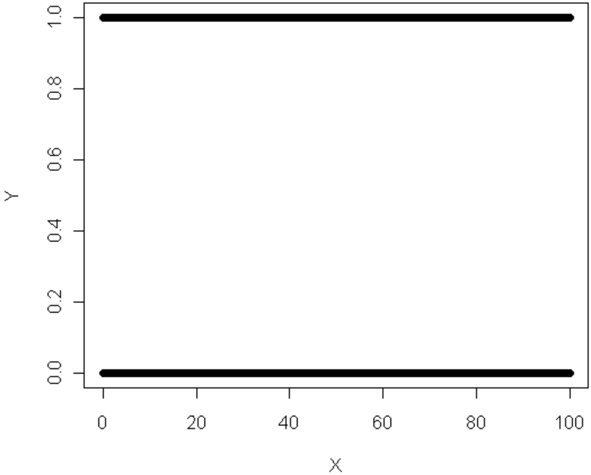
$$g_i = -3 + 0.06X_i, \text{ for } i = 1, \ldots, 100000.$$

```
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
```

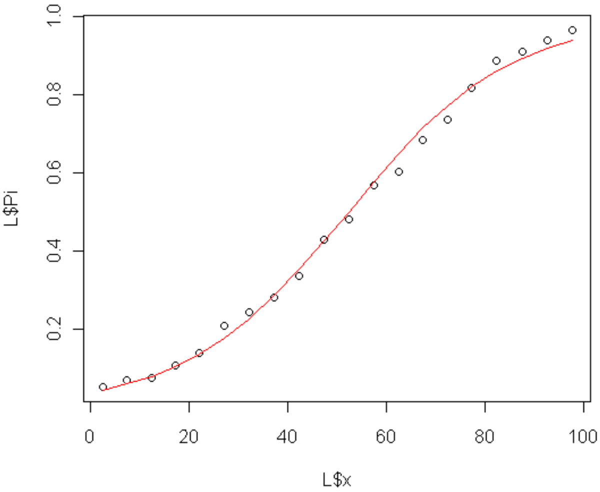
Plots

Y vs. *X* (Not very useful).



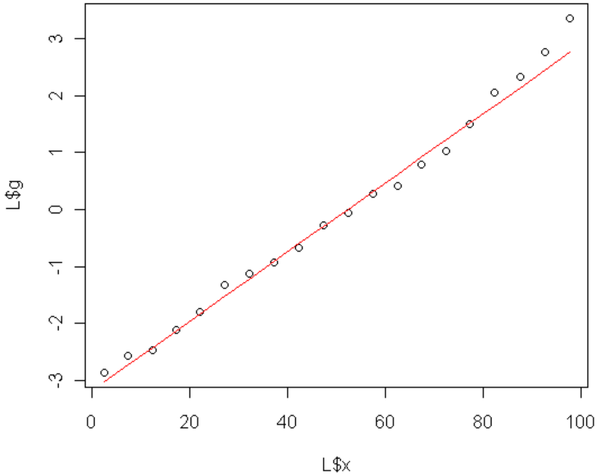
Plots

$\hat{\pi}$ vs. X



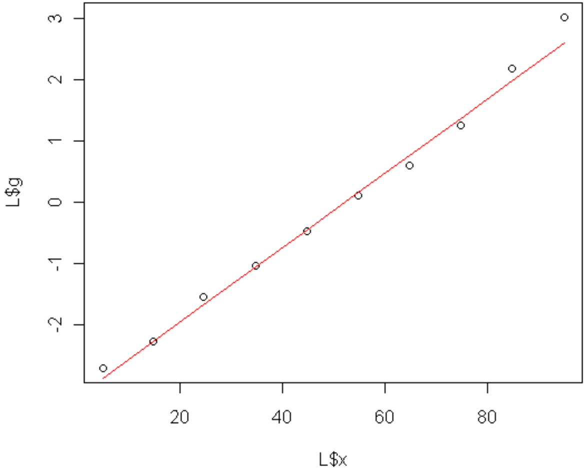
Plots

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



Plots

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



Deviance and AIC

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

$$\text{AIC} = 2p - 2\ln(L)$$

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

$$\hat{\pi}_i = \frac{e^{\hat{g}_i}}{1 + e^{\hat{g}_i}}, \text{ for } i = 1, \ldots, n.$$

Hypothesis Testing

- 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

$$\mathrm{H}_0 : \beta \in V_0 \text{ vs. } \mathrm{H} : \beta \in V.$$

- 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 $\dim(V) - \dim(V_0)$ degrees of freedom, so

$$\text{reject } \mathrm{H}_0 \text{ if } G > \chi^2_\alpha(\dim(V) - \dim(V_0)).$$

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 $\hat{\pi}_i >$ cutoff, then $\hat{Y}_i = 1$
  - If $\hat{\pi}_i <$ 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 $f(y)f(x \mid y)$.

Discriminant Analysis with Multivariate Normal Predictor

- Given $Y = y$, $X \sim N(\mu_y, \Sigma_y)$, for $y = 0, \ldots, k$.

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

- 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

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

- In practice, we would use

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

- How can we estimate these quantities?
- Assume we have observations for Y_i and X_i, for $i = 1, \ldots, n$.
- $$\widehat{f(y)} = \frac{\text{Number of times } Y_i = y}{n}$$

Sample Mean and Covariance Matrix

- Let $x_1, \ldots, x_n \in \mathbb{R}^p$ be observations from $N(\mu, \Sigma)$.
- The estimate for the mean μ is the sample mean $\overline{x}$.

$$\hat{\mu} = \overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- The estimate for the covariance matrix Σ is the empirical covariance matrix S.

$$\hat{\Sigma} = S = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})(x_i - \overline{x})'$$

- If X is a matrix whose rows are $x_1, \ldots, x_n$ then $\overline{x}$ and S can be obtained with the R commands `colMeans(X)` and `cov(X)`.

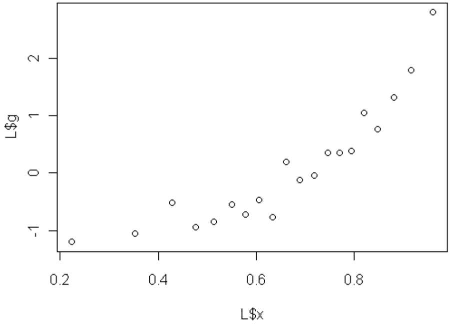
- In practice, we would use

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

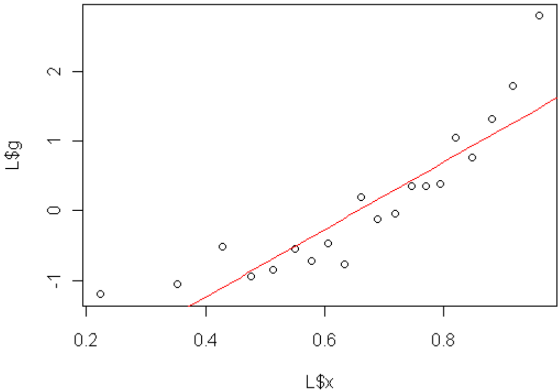
- For each y, set aside all rows of data where $Y_i = y$.
- $\overline{x}_y$ and S_y are the sample mean and covariance matrix for the vectors x_i from these rows of data.
- For each y, let $x_{y1}, \ldots, x_{yn_y}$ be the values of X_i for those subjects with $Y_i = y$.

Highschool Rank and Biol 120 Grade

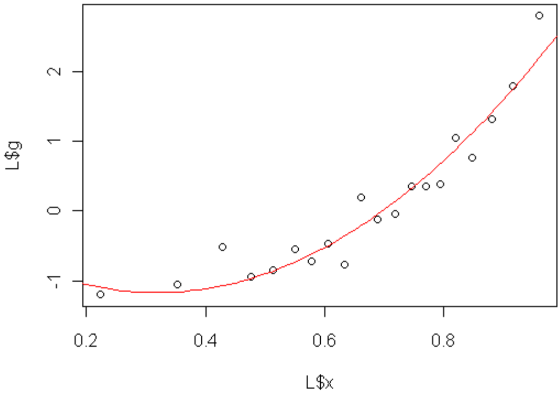
```
trank=rank[1:2000]
tgrade=grade[1:2000]
vrank=rank[2001:3146]
vgrade=grade[2001:3146]

L=groupplot(trank,tgrade,20)
plot(L$x,L$g)
```



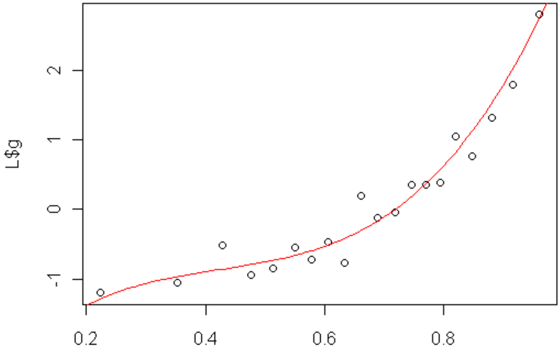
```
model=glm(tgrade~trank,family=binomial)
betahat=coef(model)

x=(1:100)/100
lines(x,betahat[1]+betahat[2]*x,col='red')
```



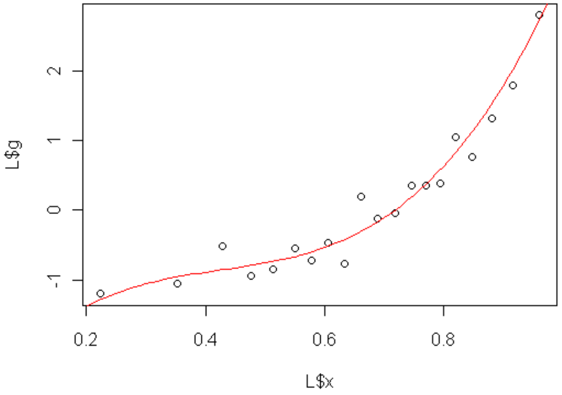
```
model2=glm(tgrade~trank+I(trank^2),family=binomial)
betahat2=coef(model2)

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



Likelihood Ratio Tests

Console ~/

```
> LRtest(model,model2)
[1] 5.34558e-11
> LRtest(model2,model3)
[1] 0.0006829123
> LRtest(model3,model4)
[1] 0.9089255
>
```

Model Validation

- 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

```
vrank=rank[2001:3146]
vgrade=grade[2001:3146]

vgradehat=(vrank>=.71929)*1
mean(vgradehat==vgrade)
```

$$\text{Classification Accuracy} = P(Y_i = \hat{Y}_i) = 0.699$$

HS Rank and Biol 120 Grade with Discriminant Analysis

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

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

```
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

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

```
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

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

```
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)
```

```
[643,]     0.643 -1.9274647498 -1.868919021 1
[644,]     0.644 -1.9235675598 -1.874871637 1
[645,]     0.645 -1.9196068876 -1.880758028 1
[646,]     0.646 -1.9155827331 -1.886578196 1
[647,]     0.647 -1.9114950963 -1.892332140 1
[648,]     0.648 -1.9073439772 -1.898019860 1
[649,]     0.649 -1.9031293759 -1.903641356 0
[650,]     0.650 -1.8988512923 -1.909196628 0
[651,]     0.651 -1.8945097264 -1.914685677 0
[652,]     0.652 -1.8901046782 -1.920108501 0
[653,]     0.653 -1.8856361478 -1.925465101 0
[654,]     0.654 -1.8811041351 -1.930755478 0
```

Discr. Analysis Optimal Cutoff $= 0.649$

Cross-validation for Discriminant Analysis

```
vgradehat=(vrank>=.649)*1
mean(vgradehat==vgrade)
```

$$\text{Classification Accuracy} = P(Y_i = \hat{Y}_i) = 0.702$$

“Brute Force” Approach

```
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)]
```

$$\text{Optimal Cutoffs} = (.717, .719, .720, .721)$$

$$\text{Optimal Cutoffs} = .7195$$

Additional Reading

- Hosmer, D.W. (2000). *Applied Logistic Regression, 2nd ed.* Wiley-Interscience, New York, N.Y.
- Khattree, R. and Naik, D.N. (1999) *Applied Multivariate Statistics with SAS Software, 2nd ed.* SAS Institute Inc., Cary, N.C.
- Khattree, R. and Naik, D.N. (2000) *Multivariate Data Reduction and Discrimination with SAS Software* SAS Institute Inc., Cary, N.C.