Ch5. Building and Applying Logistic Regression Models

Namhyoung Kim

Dept. of Applied Statistics

Gachon University

nhkim@gachon.ac.kr

Building and Applying Logistic Regression Models

- Model selection
- Model checking
- Be careful with “sparse” categorical data(estimators may take value ∞ or $-\infty$)

5.1 Strategies in Model Selection

- The model should be ***complex enough*** to fit the data well, but ***simpler models are easier to interpret***.

How Many Predictors Can you Use?

- Data are **unbalanced** on Y if y=1(or y=0) occurs few times.
  - This limits the number of predictors
  - One guideline suggests there should ideally be at least 10 outcomes of each type for every predictor.
- When the guideline is violated, ML estimates may be quite biased and estimates of standard errors may be poor

How Many Predictors Can you Use?

- A model with several predictors has the potential for ***multicollinearity***
  - Strong correlations among predictors making it seem that no one variable is important when all others are in the model.
  - Deleting such a redundant predictor can be helpful.

Example: Horseshoe Crabs

- 4 predictors : **color**(4 categories), **spine** condition(3 categories), **weight, width** of the shell.
- Consider a model with all the main effects.
  - $\{c_1, c_2, c_3\}$: indicator variables for the first three colors
  - $\{s_1, s_2\}$: indicator variables for the first two spine conditions

$$\text{logit}[P(Y=1)] = \alpha + \beta_1 \text{wieght} + \beta_2 \text{width} + \beta_3 c_1 \\ + \beta_4 c_2 + \beta_5 c_3 + \beta_6 s_1 + \beta_7 s_2$$

treats color and spine condition as **nominal-scale** factors

SAS program

```
/*Treat color and spine condition as nominal scale*/
PROC GENMOD data=crab DESC;
CLASS color spine;
MODEL y= color spine weight width/dist=bin link=logit LRCI TYPE1;
RUN;
```

SAS results

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-9.2702	3.8378	-17.0134	-1.8581	5.83	0.0157
color	1	1	1.6092	0.9356	-0.1679	3.5577	2.96	0.0854
color	2	1	1.5061	0.5667	0.4229	2.6665	7.06	0.0079
color	3	1	1.1203	0.5933	-0.0163	2.3303	3.56	0.0590
color	4	0	0.0000	0.0000	0.0000	0.0000	.	.
spine	1	1	-0.4003	0.5027	-1.3864	0.6000	0.63	0.4259
spine	2	1	-0.4964	0.6292	-1.7511	0.7473	0.62	0.4301
spine	3	0	0.0000	0.0000	0.0000	0.0000	.	.
weight		1	0.8263	0.7035	-0.5352	2.2713	1.38	0.2402
width		1	0.2629	0.1953	-0.1239	0.6502	1.81	0.1781
Scale		0	1.0000	0.0000	1.0000	1.0000		

LR Statistics For Type 1 Analysis				
Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	225.7585			
color	212.0608	3	13.70	0.0033
spine	208.8338	2	3.23	0.1992
weight	186.9937	1	21.84	<.0001
width	185.1990	1	1.79	0.1804

Example: Horseshoe Crabs

- A likelihood-ratio test
  - H_0: $\beta_1 = \cdots = \beta_7 = 0$
  - The test statistic is $-2(L_0 - L_1)$=40.6 with df=7(P<0.0001)
  - This shows extremely strong evidence that at least one predictor has an effect.
  - Although this overall test is highly significant, the Table 5.1 results are discouraging.

Example: Horseshoe Crabs

Table 5.1. Parameter Estimates for Main Effects Model with Horseshoe Crab Data

Parameter	Estimate	*SE*
Intercept	−9.273	3.838
Color(1)	1.609	0.936
Color(2)	1.506	0.567
Color(3)	1.120	0.593
Spine(1)	−0.400	0.503
Spine(2)	−0.496	0.629
Weight	0.826	0.704
Width	0.263	0.195

Example: Horseshoe Crabs

- The small P-value for the overall test, yet the lack of significance for individual effects is a warning sign of ***multicollinearity(다중공선성)***
- There is a *strong linear relationship between width and weight* with a correlation of 0.887
- It does not make much sense to analyze an effect of width while controlling for weight, since weight naturally increases as width does.
- Further analysis uses width(W) with color(C) and spine condition(S) as predictors

Example: Horseshoe Crabs

- For simplicity, we symbolize models by their highest-order terms, regarding C and S as factors.

- For instance, (C+S+W) denotes the model with main effects

$$\text{logit}[P(Y=1)] = \alpha + \beta_1 c_1$$
$$+\beta_2 c_2 + \beta_3 c_3 + \beta_4 s_1 + \beta_5 s_2 + \beta_6 w$$

- (C+S*W) denotes the model with those main effects plus an S*W interaction.

$$\text{logit}[P(Y=1)] = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3$$
$$+\beta_4 s_1 + \beta_5 s_2 + \beta_6 w + \beta_7 s_1 * w + \beta_8 s_2 * w$$

Stepwise Variable Selection Algorithms

- To select a model, we can select or delete predictors from a model in a stepwise manner.
- **Forward selection** adds terms sequentially until further additions do not improve the fit.
- **Backward elimination** begins with a complex model and sequentially removes terms.
  - At each stage, we eliminate the term in the model that has the largest p-value when we test that its parameters equal to zero
  - We test only the highest order terms for each variable.

Stepwise Variable Selection Algorithms

- For categorical predictors with more than two categories, the process should consider the entire variable
  - Add or drop the entire variable rather than just one of its indicators

Example: Backward Elimination for Horseshoe Crabs

- We can test the null hypothesis that the simpler model is adequate against the alternative hypothesis that the more complex model fits better.

Example: Backward Elimination for Horseshoe Crabs

- Recall that the *deviance* (ch.3.4)

$$\text{Deviance} = -2[L_M - L_S]$$

- L_M: maximized log-likelihood value for a model M of interest
- L_S: maximized log-likelihood value for the most complex model possible(saturated)

- For some GLMs, the deviance has approximately a chi-squared distribution with *df=(number of observation-number of model parameter)*

Example: Backward Elimination for Horseshoe Crabs

Table 5.2. Results of Fitting Several Logistic Regression Models to Horseshoe Crab Data

Model	Predictors	Deviance	*df*	AIC	Models Compared	Deviance Difference
1	$C * S + C * W + S * W$	173.7	155	209.7	–	
2	$C + S + W$	186.6	166	200.6	(2)–(1)	12.9 ($df = 11$)
3a	$C + S$	208.8	167	220.8	(3a)–(2)	22.2 ($df = 1$)
3b	$S + W$	194.4	169	202.4	(3b)–(2)	7.8 ($df = 3$)
3c	$C + W$	187.5	168	197.5	(3c)–(2)	0.9 ($df = 2$)
4a	C	212.1	169	220.1	(4a)–(3c)	24.6 ($df = 1$)
4b	W	194.5	171	198.5	(4b)–(3c)	7.0 ($df = 3$)
5	$C = \text{dark} + W$	188.0	170	194.0	(5)–(3c)	0.5 ($df = 2$)
6	None	225.8	172	227.8	(6)–(5)	37.8 ($df = 2$)

Note: C = color, S = spine condition, W = width.

아름다운 인재의 샘 가천대학교

Example: Backward Elimination for Horseshoe Crabs

- Table 5.2 summarizes results of fitting and comparing several logistic regression models.
- To select a model, we use a backward elimination procedure.
  - We start with a complex model (1)
  - We test all the interactions simultaneously by comparing it to model (2)
  - The likelihood-ratio statistic is 186.6-173.7=12.9(df=166-155=11, P=0.30)
  - This does not suggest that the interactions terms are needed.→ remove interactions terms

Example: Backward Elimination for Horseshoe Crabs

- The next stage consider dropping a term from the main effect model.
- little consequence from removing spine condition S(model 3c) → remove S
- Both remaining variables are significant.
- The simpler model that has a single indicator variable for color fits essentially as well.

Example: Backward Elimination for Horseshoe Crabs

```
PROC LOGISTIC data=crab DESC;
CLASS color spine;
MODEL y=color spine width color*spine color*width
        spine*width/selection=backward slstay=0.1;
RUN;
```

SELECTION=BACKWARD | B | FORWARD | F | NONE | N | STEPWISE | S | SCORE

specifies the method used to select the variables in the model.

SLENTRY=*value*

specifies the significance level of the score chi-square for entering an effect into the model in the FORWARD or STEPWISE method.

SLSTAY=*value*

specifies the significance level of the Wald chi-square for an effect to stay in the model in the BACKWARD or STEPWISE method.

AIC, Model Selection, and the "Correct" Model

- Other criteria besides significance test can help select a good model.
- The best known is the *Akaike information criterion* (AIC)
  - It judges a model by how close its fitted values tend to be to the true expected values, as summarized by a certain expected distance between the two
  - AIC=-2(log likelihood-number of parameters in model)

AIC, Model Selection, and the "Correct" Model

- For the model C+W, a -2log likelihood value is 187.5.
- The model has five parameters (an intercept and a width effect and three coefficients of indicator variables for color)
- Thus, AIC=187.5+2*5=197.5
- The AIC penalizes a model for having many parameters.

Summarizing Predictive Power

- Classification Tables
- ROC Curves
- A Correlation

Summarizing Predictive Power: Classification Tables

Table 5.3. Classification Tables for Horseshoe Crab Data

	Prediction, $\pi_0 = 0.64$		Prediction, $\pi_0 = 0.50$		
Actual	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	Total
$y = 1$	74	37	94	17	111
$y = 0$	20	42	34	28	62

- $\hat{y} = 1$ when $\hat{\pi}_i > \pi_0$ and $\hat{y} = 0$ when $\hat{\pi}_i \leq \pi_0$, for some cutoff π_0.

Summarizing Predictive Power: Classification Tables

- Two useful summaries of predictive power are
  - sensitivity=P($\hat{y} = 1 \mid y = 1$)
  - specificity=P($\hat{y} = 0 \mid y = 0$)
- When π_0=0.642, the estimated sensitivity=74/111=0.667 and specificity=42/62=0.677

Summarizing Predictive Power: Classification Tables

- Another summary of predictor power is the overall proportion of correct classifications.

P(correct classification)

$=P(\hat{y} = 1 \text{ and } y = 1)+P(\hat{y} = 0 \text{ and } y = 0)$

$=P(\hat{y} = 1 \mid y = 1)P(y = 1)+P(\hat{y} = 0 \mid y = 0)P(y = 0)$

- When π_0=0.642, the proportion of correct classifications is (74+42)/173=0.671

Summarizing Predictive Power: ROC Curves

- A receiver operating characteristic (ROC) curve is a **plot of sensitivity as a function of (1-specificity)** for the possible cutoffs π_0.
- When $\boldsymbol{\pi_0}$ **gets near 0**, almost all predictions are $\hat{y} = 1$; then, sensitivity is near 1, specificity is near 0→(1-specificity, sensitivity)=**(1, 1)**
- When $\boldsymbol{\pi_0}$ **gets near 1**, almost all predictions are $\hat{y} = 0$; then, sensitivity is near 0, specificity is near 1→(1-specificity, sensitivity)=**(0, 0)**

Summarizing Predictive Power: ROC Curves

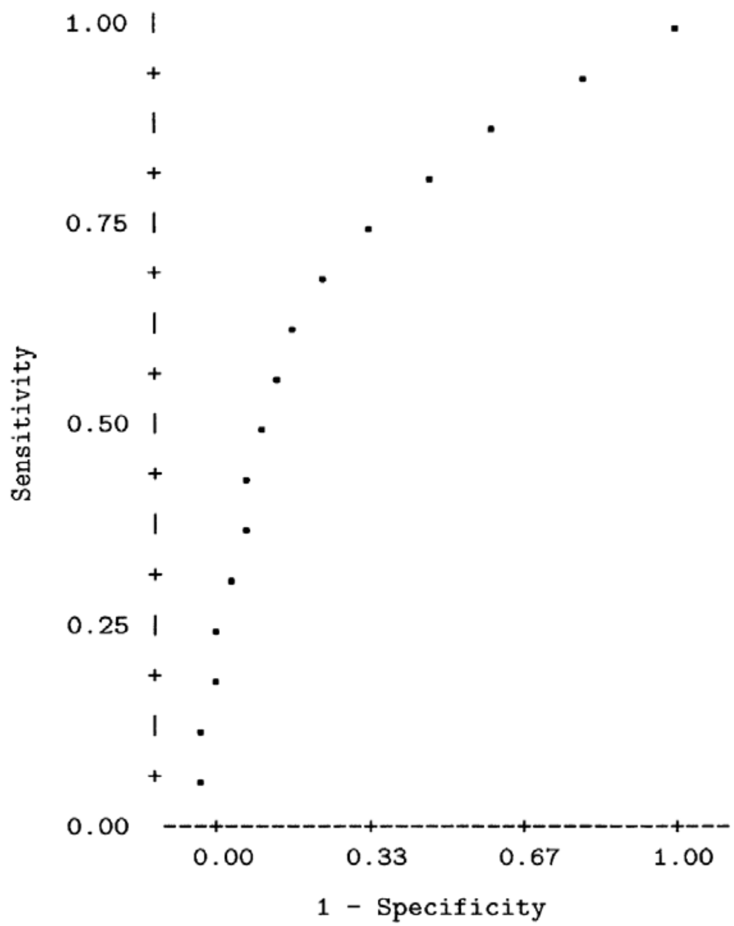

Figure 5.1. ROC curve for logistic regression model with horseshoe crab data.

아름다운 인재의 샘
가천대학교

Summarizing Predictive Power: ROC Curves

- For a given specificity, better predictive power correspond to higher-sensitivity.
- The area under the ROC curve is identical to the value of a measure of predictive power called the *concordance index*.

Summarizing Predictive Power: A Correlation

- For a GLM, a way to summarize prediction power is by the correlation R between the observed responses $\{y_i\}$ and the model's fitted values $\{\hat{\mu}_i\}$.

아름다운 인재의 샘
가천대학교