Georgia State University
Digital Archive @ GSU

Mathematics Theses | Department of Mathematics and Statistics

8-5-2011

Discrimination of High Risk and Low Risk Populations for the Treatment of STDs

Hui Zhao
GSU, hzhao@student.gsu.edu

Follow this and additional works at: http://digitalarchive.gsu.edu/math_theses

Recommended Citation
Zhao, Hui, "Discrimination of High Risk and Low Risk Populations for the Treatment of STDs" (2011). *Mathematics Theses.* Paper 104.

This Thesis is brought to you for free and open access by the Department of Mathematics and Statistics at Digital Archive @ GSU. It has been accepted for inclusion in Mathematics Theses by an authorized administrator of Digital Archive @ GSU. For more information, please contact digitalarchive@gsu.edu.

DISCRIMINATION OF HIGH RISK AND LOW RISK POPULATIONS FOR THE TREATMENT OF STDS

by

HUI ZHAO

Under the Direction of Dr. Jiawei Liu

ABSTRACT

It is an important step in clinical practice to discriminate real diseased patients from healthy persons. It would be great to get such discrimination from some common information like personal information, life style, and the contact with diseased patient. In this study, a score is calculated for each patient based on a survey through generalized linear model, and then the diseased status is decided according to previous sexually transmitted diseases (STDs) records. This study will facilitate clinics in grouping patients into real diseased or healthy, which in turn will affect the method clinics take to screen patients: complete screening for possible diseased patient and some common screening for potentially healthy persons.

INDEX WORDS: Sexually transmitted diseases (STDs), Generalized linear model, Receiver operating characteristic (ROC) curve

DISCRIMINATION OF HIGH RISK AND LOW RISK POPULATIONS FOR THE TREATMENT OF STDS

by

HUI ZHAO

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2011

Copyright by
Hui Zhao
2011

DISCRIMINATION OF HIGH RISK AND LOW RISK POPULATIONS FOR THE TREATMENT OF STDS

by

HUI ZHAO

Committee Chair: Dr. Jiawei Liu

Dr. Yichuan Zhao

Committee: Dr. Xu Zhang

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2011

DEDICATION

To my parents, my husband, and my children

ACKNOWLEDGEMENTS

This thesis would not have been possible without the support from many people who gave their supports in different ways. To them I would like to express my deepest gratitude and sincere appreciation.

First of all, I would like to thank my advisor Dr. Jiawei Liu for her encouragement and insightful guidance.

Second, I must thank my co-advisor Dr. Yichuan Zhao. He gave me critical suggestions in my thesis research.

My appreciation also goes to my committee Dr. Xu Zhang. She helped me lot on my programming.

I would also thank all other faculty in our department for their support and encourage when I was on my way to get my degree.

Thanks to all my colleagues helped me. Without their help, I could by no means achieve the goal of this research.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iv

LIST OF TABLES viii

LIST OF FIGURES ix

Chapter

1 Introduction 1

2 General Setting and Mathematical Model 5

2.1 Classical Linear Model 5

2.2 Exponential Family of Distributions 6

2.3 Generalized Linear Model (GLM) 7

2.4 Logistic Regression 9

2.5 Risk Function 10

2.6 ROC Curve 11

3 Logistic Regression Result 16

3.1 Dataset Description 16

3.2 Logistic Regression Result 16

3.2.1 Step0: Full Logistic Regression Model 16

3.2.2 Step1: Remove Variable “MarriageStatus” 19

3.2.3 Step2: Remove Variable “Age” 21

3.2.4 Step3: Remove Variable “Gender” 23

3.2.5 Step4: Remove Variable “Race” 25

3.2.6 Step5: Remove Variable “InjectionDrugUse” 27

3.2.7 Step6: Remove Variable “FinancialBarrier” 29

3.2.8 Step7: Remove Variable “Income” 31

3.2.9 Step8: Remove Variable “EmployedLastWeek” 33

3.2.10 Summary of Backward Selection 35

4 ROC Curve Result 37

4.1 ROC Curves of Backward Selection 37

4.1.1 ROC Curve of Full Model 37

4.1.2 ROC Curve of Backward Model Selection Step1 39

4.1.3 ROC Curve of Backward Model Selection Step2 39

4.1.4 ROC Curve of Backward Model Selection Step3 40

4.1.5 ROC Curve of Backward Model Selection Step4 40

4.1.6 ROC Curve of Backward Model Selection Step5 41

4.1.7 ROC Curve of Backward Model Selection Step6 41

4.1.8 ROC Curve of Backward Model Selection Step7 42

4.1.9 ROC Curve of Backward Model Selection Step8 42

4.2 Cut-Off Point Selection 43

5 Conclusion 46

REFERENCES 48

APPENDIX: SAS Code 49

LIST OF TABLES

Table 2.1 The four outcomes from a binary predition problem 11

Table 3.1 Repersentative points on ROC curve of full model 18

Table 3.2 Some repersentative points on ROC curve at step1 20

Table 3.3 Some repersentative points on ROC curve at step2 22

Table 3.4 Some repersentative points on ROC curve at step3 24

Table 3.5 Some repersentative points on ROC curve at step4 26

Table 3.6 Some repersentative points on ROC curve at step5 28

Table 3.7 Some repersentative points on ROC curve at step6 30

Table 3.8 Some repersentative points on ROC curve at step7 32

Table 3.9 Some repersentative points on ROC curve at step8 34

Table 3.10 Summary of the backward selection steps 35

LIST OF FIGURES

Figure 1.1 *Chlamydia trachomatis* bacteria 2

Figure 1.2 *Neisseria gonorrhoeae* bacteria 2

Figure 1.3 *Syphilis* bacteria 2

Figure 2.1 The cut point and its four outcomes 12

Figure 2.2 A typical ROC curve 13

Figure 2.3 ROC curve and AUC 14

Figure4.1 ROC curve of the full model 37

Figure4.2 ROC curve comparison of the full model to non-informative model 38

Figure4.3 ROC curve comparison of step1 model to non-informative model 39

Figure4.4 ROC curve comparison of step2 model to non-informative model 39

Figure4.5 ROC curve comparison of step3 model to non-informative model 40

Figure4.6 ROC curve comparison of step4 model to non-informative model 40

Figure4.7 ROC curve comparison of step5 model to non-informative model 41

Figure4.8 ROC curve comparison of step6 model to non-informative model 41

Figure4.9 ROC curve comparison of step7 model to non-informative model 42

Figure4.10 ROC curve comparison of step8 model to non-informative model 42

Figure4.11 ROC curves for all model building steps 43

Chapter 1

Introduction

The sexually transmitted diseases (STDs) are a variety of clinical syndromes caused by pathogens that can be acquired and transmitted through sexual activity. Chlamydia, Gonorrhea, and Syphilis are the three most often detected STDs.

Discussion focused on four principal outcomes of STD therapy for each individual disease: 1) treatment of infection based on microbiologic eradication; 2) alleviation of signs and symptoms; 3) prevention of sequelae; and 4) prevention of transmission.

Chlamydia trachomatis infection is the most commonly reported STD in the United States. Chlamydial infections in women can cause pelvic inflammatory disease (PID), which is a major reason of infertility, ectopic pregnancy, and chronic pelvic pain. Chlamydia can also facilitate the transmission of human immunodeficiency virus (HIV) infection. Pregnant women infected with Chlamydia can pass the infection to their infants during delivery, potentially resulting in neonatal ophthalmia and pneumonia. In 2009, a total of 1,244,180 cases of sexually transmitted *Chlamydia trachomatis* infection were reported to the Centers for Disease Control and Prevention (CDC). This case count means a rate of 409.2 cases per 100,000 population, an increase of 2.8% compared with the rate in 2008 (398.1 cases per 100,000 population) (STDsurv2009-Complete).

Figure 1.1 *Chlamydia trachomatis* bacteria

Gonorrhea is the second most commonly reported STD in the United States. It is an infection due to *Neisseria gonorrhoeae*, and it is also a major cause of PID. It has also been proved by epidemiologic and biologic studies that gonococcal infections can assist the HIV infection. In 2009, a total of 301,174 cases of gonorrhea were reported in the United States, which corresponds to a rate of 99.1 cases per 100,000 population. The 2009 rate is a 10.5% decrease from the rate of 110.7 cases per 100,000 population in 200 (STDsurv2009-Complete).

Figure 1.2 *Neisseria gonorrhoeae* bacteria

Syphilis is a genital ulcerative disease, which causes significant complications if untreated and facilitates the transmission of HIV infection. Untreated early syphilis in pregnant women results in prenatal death in up to 40% of cases, and, if acquired during the 4 years before preg-

nancy, can lead to infection of the fetus in 80% of cases. In 2009, a total of 13,997 cases of primary and secondary syphilis were reported to CDC. This case count is the highest number of cases reported since 1995 and corresponds to a rate of 4.6 cases per 100,000 population, which means a 5% increase from 2008 (4.4 cases per 100,000 population). Since 2005, the rate of primary and secondary syphilis has increased 59% (from 2.9 cases to 4.6 cases per 100,000 population) (STDsurv2009-Complete).

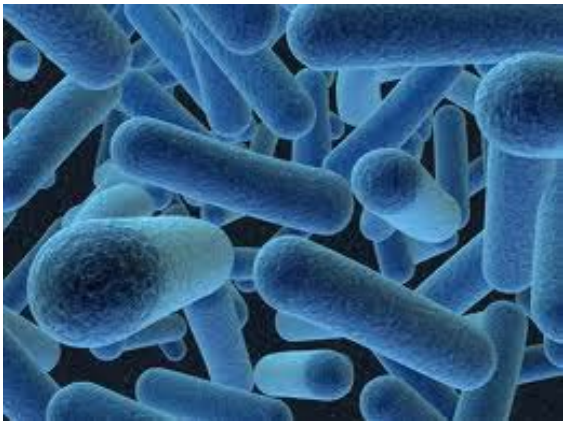

Figure 1.3 *Syphilis* bacteria

As a standard treatment of STD, the client is asked to finish a survey when s/he comes in a clinical office. At that time point, this client is either sick or healthy. There is not a third option. If the client is sick, s/he should be classified into high risk group; if the client is healthy, s/he should be grouped into low risk group. But we don't know the true status of the patient at that time point, so we must postulate the status from the result of the survey. If the patient shows low risk symptom of disease from the survey, s/he is guided to normal screening method; if the patient has high risk of disease from the survey, s/he must take a complete screening test. For each patient, there are two possible disease situations: not disease or disease; when we choose a cut point of risk, the same patient may have one of the two risk status: low risk or high risk. Ideally, if the classification method is powerful enough and we choose right cut point, all disease patients

will show high risk and all normal patients will show low risk. But, in real life, there are always some disease patients show low risk (false negative), and some normal patients show high risk (false positive). To false positive patients, we will do the complete screening, and then find out that they are normal and let them go. There is not too much harm in this situation, except for spending more money on the complete screening test. But to false negative clients, it is a totally different story. If we do normal screen to false negative patients and not correctly classify them as disease patients, they will go out freely without any treatment, and they may spread disease to more people. This is much more severe harmful to our society. In our study, we set the goal as to decide the best cut point in risk, which will lower both false negative rate and false positive rate, especially the false negative rate.

Chapter 2

General Setting and Mathematical Model

2.1 Classical Linear Model

In classical linear regression model, the dependent variable is assumed to be continuous, normally distributed, with constant variance (Montgomery, Peck, and Vining, 2006). The dependent variable is a linear function of a set of regressors. Assume there are *n* observations and *p* parameters, the model of classical linear regression is:

$$\boldsymbol{Y=X\beta+\varepsilon}$$

Where $Y = \begin{bmatrix} Y_1 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{bmatrix}$ is a vector of responses of size $n\times1$,

$X = \begin{bmatrix} X_{11} & \cdot & \cdot & \cdot & X_{1p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ X_{n1} & \cdot & \cdot & \cdot & X_{np} \end{bmatrix}$ is an $n\times p$ matrix of known constants,

$\beta = \begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix}$ is a $p\times1$ vector of parameters which elements are to be estimated,

and $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$ is the error vector, the elements in which follow an independent and identical normal distribution with mean 0.

The expected value of Y is Xβ since the error term ε has mean 0.

$$E(Y) = \mu = X\beta$$

The restrictions on dependent variable Y in classic linear model is strict (continuous, equal variance, and normally distributed). The assumption of independent identical distributed data is rarely satisfied in practice. In real applications, the dependent variable may be a categorical variable, or a count variable, or other continuous variables other than normal. In these cases, all three assumptions do not hold.

2.2 Exponential Family of Distributions

The normal distribution is a member of a big distribution family called exponential family, which includes normal, exponential, gamma, chi-square, beta, Dirichlet, Bernoulli, binomial, multinomial, Poisson, and many others (Forbes et al. 2011).

The distributions in exponential family share a general form:

$$f(y_i;\theta) = h(y_i)\exp(d(\theta)T(y_i) - A(\theta)),$$

where h(.), d(.), T(.), and A(.) are all known functions that have the same form for all y_i, i=1, 2, …, n.

In exponential family, binomial distribution is an important one with only two kinds of outcome for each subject: success or failure. There are many examples of this distribution, like the number of person who are sick or not at some time point in a clinic (the possible outcomes for

each person are sick or healthy), the number of candidates who pass an exam or fails it (the possible outcomes for each candidate being to pass or to fail), the number of phone calls a person made in last month is larger than 500 or less or equal to 500 (the possible outcomes for each person are >500 or ≤500). We define success as '1', with the probability π; and failure as '0', with the probability 1-π. Let the random variable *Y* be the number of 'successes' in *n* independent trials. Then Y has the binomial distribution with probability density function

$$f(y;\pi) = \binom{n}{y}\pi^{y}(1-\pi)^{n-y},$$

where *y* takes the values 0, 1, 2, …, *n*. This is denoted by Y~binomial (n,π). The probability function can be rewritten as

$$f(y;\pi) = \exp\left[y\log\pi - y\log(1-\pi) + n\log(1-\pi) + \log\binom{n}{y}\right]$$ to match with the general form of exponential family distribution.

2.3 Generalized Linear Model (GLM)

In 1972, Nelder and Wedderburn introduced generalized linear model which expends the linear regression from normal distribution into the family of exponential distributions (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989). There is a transformation of $\boldsymbol{\mu_i}$ such that

$$g(\mu_i) = x_i^T\beta .$$

In this equation, *g* is a monotone, differentiable function called the link function; $\boldsymbol{x_i}$ is a *p*×1 vector of explanatory variables

$$Xi = \begin{bmatrix} x_{i1} \\ \cdot \\ \cdot \\ \cdot \\ xi_p \end{bmatrix}$$, which is the *i*th column of the matrix X,

and β is the $p \times 1$ vector of parameters $\beta = \begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix}$.

There are three components is a generalized linear model:

1. Random component: response variables $\boldsymbol{Y_1, \ldots, Y_n}$ which are assumed to share the same distribution from the exponential family;
2. Systematic component: a set of parameters $\boldsymbol{\beta}$ and explanatory variables

$$X = \begin{bmatrix} X_1^T \\ \cdot \\ \cdot \\ \cdot \\ X_n^T \end{bmatrix} = \begin{bmatrix} X_{11} & \cdot & \cdot & \cdot & X_{1p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ X_{n1} & \cdot & \cdot & \cdot & X_{np} \end{bmatrix};$$

3. Link function: there exists a transformation function g of μ_i such that $g(\mu_i) = x_i^T \beta$

   Where $\mu_i = E(Y_i)$. The link function connects the random and systematic components.

In the studies where the response variable is binary, the mean $\mu_i = E(Y_i)$ is usually the proportion of successes; we define this proportion of successes as π.

1. The simplest case for link function to describe the proportion of successes is the linear model

$$\pi = X^T \beta .$$

This is used in some practical applications but it has the disadvantage that although π is a probability, which should be a number between zero and one, the fitted values $X^T\beta$ may be less than zero or greater than one.

2. Another link function can be used is probit model

$$\pi = \Phi\left(\frac{x-\mu}{\delta}\right) .$$

where Φ denotes the cumulative probability function for the standard normal distribution N (0,1). Thus,

$$\Phi^{-1}(\pi) = X\beta .$$

Probit models are used in several areas of biological and social sciences in which there are natural interpretations of the model.

3. The third link function is logistic regression function, as listed below.

2.4 Logistic Regression model

The most used generalized linear model for analyzing data involving binary or binomial response and several explanatory variables (X) is logistic regression model, with the link function is:

$$\log\left(\frac{\pi}{1-\pi}\right) = X\beta ,$$

where π is the probability of success,

$$X = \begin{bmatrix} X_{11} & \cdot & \cdot & \cdot & X_{1p} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ X_{n1} & \cdot & \cdot & \cdot & X_{np} \end{bmatrix}$$ is the matrix of known constants,

and $\beta = \begin{bmatrix} \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_p \end{bmatrix}$ is a vector of parameters to be estimated.

This link function models the log of an odds ratio. There is an important interpretation of logistic model parameters: the odds ratio multiply by $e^{\beta i}$ for every unit increase in x_i, when other parameters are controlled (Hosmer and Lemeshow, 2000).

2.5 Risk Function

The risk function is important in deciding if a client is sick or not. But, there is no mature formula for it. Since we have some patient data, we can calculate our formula from these known data. From this empirical formula, we can construct the risk function to reflect the cost of screening, false positive, false negative, etc.

Let's define risk function as:

$$Ri = f(x_{i1}, x_{i2}, \ldots x_{ip})$$

Where *x's* are the survey results, *i* is patient ID, *p* is the number of survey questions.

2.6 ROC Curve

A receiver operating characteristic (ROC) curve was developed by electrical engineers and radar engineers during World War II for detecting enemy objects in battle field. It has widely applied in psychology, medicine, radiology, machine learning, and data mining since then.

In a two-class prediction problem, for example to determine whether a person has hypertension based on blood pressure measurements, the outcomes are either positive (p (+)) or negative (n (-)). There are four possible outcomes from a binary classifier. If the outcome from a prediction is *p* and the actual value is also *p*, then it is called a *true positive*, and its counts/frequency is TP; if the prediction is *p* but the actual value is *n*, then it is called a *false positive*, and its counts/frequency is FP. If the prediction is *n* and actual value is also *n*, it is called *true negative*, and its counts/frequency is TN; if the prediction is *n* and actual value is *p*, it is called *false negative*, and its counts/frequency is FN.

Table 2.1 The four outcomes from a binary prediction problem

		actual value	
		p (+)	**n (-)**
prediction outcome	**p' (+)**	true positive	false positive
	n' (-)	false negative	true negative

A diagnostic test for a particular disease is a typical two-class prediction problem, which classifies clients into two groups: the positive and the negative. The goodness of a test is assessed by how well the test can discriminate positive from negative correctly.

The sensitivity of a diagnostic test is the proportion of patients for whom the outcome is positive that are correctly identified by the test, i.e. sensitivity equals to the ratio between true positive and total real positive (TP/p). The specificity is the proportion of patients for whom the outcome is negative that are correctly identified by the test, that is, specificity equals to the ratio between true negative and total negative (TN/n). The perfect test has 100% sensitivity and 100%

specificity; which means this test can group all positive patients as positive and all negative patients as negative. Of course it is very difficult to find such good test for all diagnosis.

$$sensitivity = \frac{turePositive}{total\,\mathrm{Re}\,alPositive} = \frac{truePositive}{truePositive + falseNegative}$$

$$specificity = \frac{tureNegative}{total\,\mathrm{Re}\,alNegative} = \frac{trueNegative}{trueNegative + falsePositive}$$

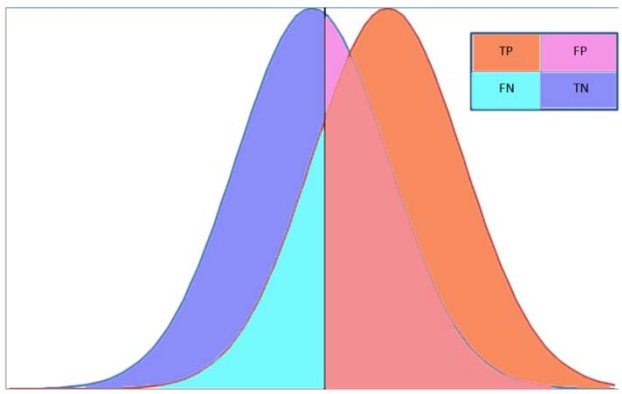

Figure 2.1 The cut point and its four outcomes

A ROC curve is defined by sensitivity and 1-specificity as *y* and *x* axes for every possible cut-off value respectively. The ROC curve for a perfect test would start at the origin (0, 0) and go vertically up the *y* axis to (0, 1), and then horizontally across to (1, 1). A good test should be close to this situation (Metz, 1978).

Assume the continuous diagnostic variable for positive and negative patients are from different distributions, and assume the larger value means an increased chance of a positive result, when the cut-off value is increased, the sensitivity will decrease and the specificity will increase. When the cut-off value slides across a reasonable range of the diagnostic variable, we will have different pairs of sensitivity and specificity. If we plot them in a graph, we get the ROC curve of this test.

A good diagnostic test is one that has small false positive and false negative rates across a reasonable range of cut off values.

ROC analysis provides a useful means to assess the diagnostic accuracy of a test and to compare the performance of more than one test for the same outcome.

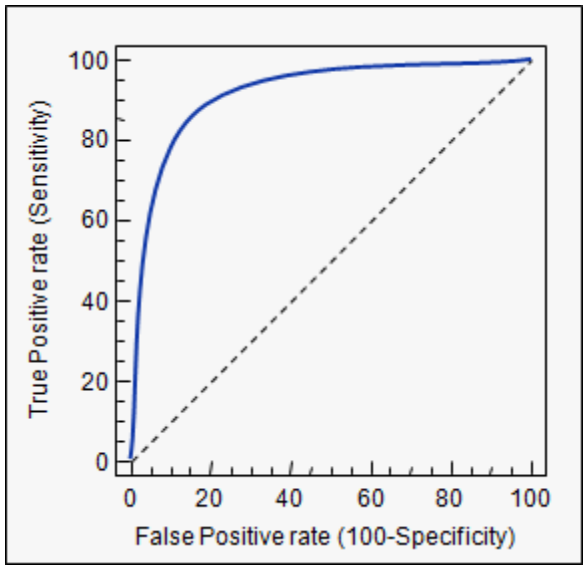

Figure 2.2 A typical ROC curve

The performance of a diagnostic variable can be quantified by calculating the area under the ROC curve (AUC). The area measures discrimination, that is, the ability of the test to correctly classify those with and without the disease. The larger the area, the better the diagnostic test. The ideal test would have an AUC of 1, because it achieves both 100% sensitivity and 100% specificity; whereas a random guess would have an AUC of 0.5, which has effectively 50% sensitivity and 50% specificity. This is a test that is no better than flipping a coin. In practice, a diagnostic test is going to have an area somewhere between these two extremes. The closer the area is to 1.0, the better the test is, and the closer the area is to 0.5, the worse the test is. Figure 2.3 shows ROC of three different tests.

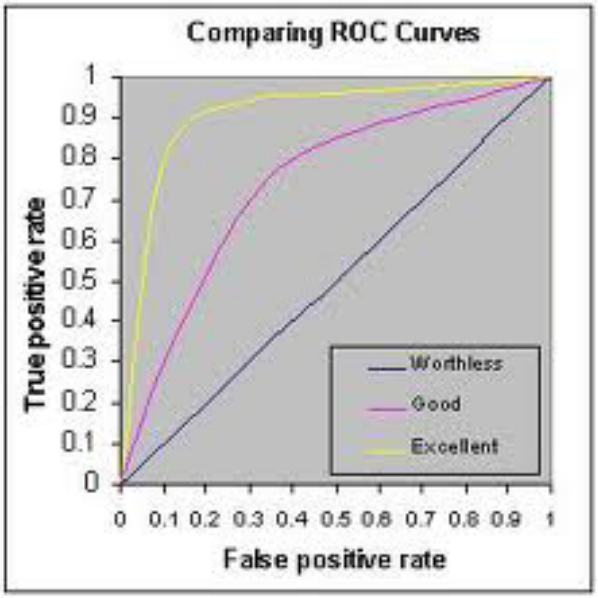

Figure 2.3 ROC curves and AUC

A rough guide for classifying the accuracy of a diagnostic test is the traditional academic point system:

- 0.90-1.00 = excellent (A)
- 0.80-0.90 = good (B)
- 0.70-0.80 = fair (C)
- 0.60-0.70 = poor (D)
- 0.50-0.60 = fail (F)

Two methods are commonly used in computing the AUC:

1. a non-parametric method based on constructing trapezoids under the curve as an approximation of area;
2. a parametric method using a maximum likelihood estimator to fit a smooth curve to the data points.

Area under the curve does have one direct interpretation. If you take a random healthy person and get a score of X and a random diseased patient and get a score of Y, then the area un-

der the curve is an estimate of P[Y>X] (assuming that large values of the test are indicative of disease).

Chapter 3

Logistic Regression Result

3.1 Dataset Description

There are 653 observations in total in our dataset. The dependent variable is the medical result of STD tests, including Chlamydia, Gonorrhea, and/or Syphilis, with two outcomes: yes or no. The regression factors include some personal, habitual, financial survey information. The survey questions are gender, age, race, highest education level, marriage status, income, if employed last week, any financial barrier to care in past year, likelihood of getting HIV, STD-related symptoms, exchange of sex for money or drugs, injection drug usage, and contact to known STD patient. All survey questions are used in logistic regression to build an empirical formula to be used in predicting the probability of a new client getting STD.

3.2 Logistic Regression Result

3.2.1 Step 0: Full Logistic Regression Model

Of all the thirteen regression variables, STD-related symptoms, contact to STD patient, education level, and exchange of sex for money or drug are the four factors significant at 0.05 level.

From the logistic regression result, the formula of log ratio between the possibility of having STD (π) and not having STD ($1-\pi$) is:

$$\log\left(\frac{\pi}{1-\pi}\right) = 15.6728$$

$$+\begin{cases} 0.0994 & if\ gender = 0 \\ 0.0000 & if\ gender = 1 \end{cases}$$

$$+ 0.00339 * age$$

$$+\begin{cases} 5.9889 & if\ race = 1 \\ 5.8351 & if\ race = 2 \\ 6.0881 & if\ race = 3 \\ 4.8517 & if\ race = 4 \\ 0.0000 & if\ race = 5 \end{cases}$$

$$+\begin{cases} -0.812 & if\ education = 1 \\ 0.4890 & if\ education = 2 \\ -0.325 & if\ education = 3 \\ -0.884 & if\ education = 4 \\ 0.0000 & if\ education = 5 \end{cases}$$

$$+\begin{cases} 0.0572 & if\ marriage = 1 \\ -6.755 & if\ marriage = 2 \\ 0.0000 & if\ marriage = 3 \end{cases}$$

$$+ 0.0757 * income$$

$$+\begin{cases} -0.5276 & if\ employedLastWeek = 0 \\ 0.00000 & if\ employedLastWeek = 1 \end{cases}$$

$$+\begin{cases} 0.3577 & if\ \ financialBarrier = 0 \\ 0.0000 & if\ \ financialBarrier = 1 \end{cases}$$

$$+\begin{cases} -24.5252 & if\ likelihoodHIV = 1 \\ -25.2692 & if\ likelihoodHIV = 2 \\ 0.000000 & if\ likelihoodHIV = 3 \end{cases}$$

$$+\begin{cases} 3.6966 & if\ STDsymptoms = 0 \\ 0.0000 & if\ STDsymptoms = 1 \end{cases}$$

$$+\begin{cases} 0.9062 & if\ exchangeSex = 0 \\ 0.0000 & if\ exchangeSex = 1 \end{cases}$$

$$+\begin{cases} 0.2915 & if\ injectionDrugUse = 0 \\ 0.0000 & if\ injectionDrugUse = 1 \end{cases}$$

$$+\begin{cases} 1.6327 & if\ contactSTD = 0 \\ 0.0000 & if\ contactSTD = 1 \end{cases}$$

In table 3.1, some representative points are listed.

Table 3.1 Representative points on ROC curve of full model

STEP	_PROB_	_POS_	_NEG_	_FALPOS_	_FALNEG_	_SENSIT_	_1MSPEC_
0	1	1	357	0	295	0.003378	0
0	0.99111	15	357	0	281	0.050676	0
0	0.98538	29	357	0	267	0.097973	0
0	0.98041	43	357	0	253	0.14527	0
0	0.97192	57	357	0	239	0.192568	0
0	0.96156	70	356	1	226	0.236486	0.0028011
0	0.94399	84	356	1	212	0.283784	0.0028011
0	0.93396	97	355	2	199	0.327703	0.0056022
0	0.92669	111	355	2	185	0.375	0.0056022
0	0.91855	123	353	4	173	0.415541	0.0112045
0	0.90513	136	352	5	160	0.459459	0.0140056
0	0.88947	150	352	5	146	0.506757	0.0140056
0	0.87865	162	350	7	134	0.547297	0.0196078
0	0.85395	175	349	8	121	0.591216	0.022409
0	0.82635	187	347	10	109	0.631757	0.0280112
0	0.80124	201	347	10	95	0.679054	0.0280112
0	0.76865	209	341	16	87	0.706081	0.0448179
0	0.74324	217	335	22	79	0.733108	0.0616246
0	0.70478	226	330	27	70	0.763514	0.0756303
0	0.66597	236	326	31	60	0.797297	0.0868347
0	0.63097	247	323	34	49	0.834459	0.0952381
0	0.59227	256	318	39	40	0.864865	0.1092437
0	0.55454	267	315	42	29	0.902027	0.1176471
0	0.49526	274	308	49	22	0.925676	0.1372549
0	0.42015	275	295	62	21	0.929054	0.1736695
0	0.31041	278	284	73	18	0.939189	0.2044818
0	0.14097	281	273	84	15	0.949324	0.2352941
0	0.11581	283	261	96	13	0.956081	0.2689076
0	0.09046	283	247	110	13	0.956081	0.3081232
0	0.0756	284	234	123	12	0.959459	0.3445378
0	0.06971	286	221	136	10	0.966216	0.3809524
0	0.06137	288	209	148	8	0.972973	0.4145658
0	0.05534	290	197	160	6	0.97973	0.4481793
0	0.05113	290	183	174	6	0.97973	0.487395
0	0.04779	291	170	187	5	0.983108	0.5238095
0	0.04141	291	156	201	5	0.983108	0.5630252
0	0.03522	292	143	214	4	0.986486	0.5994398
0	0.03238	292	129	228	4	0.986486	0.6386555
0	0.02875	292	115	242	4	0.986486	0.6778711
0	0.02443	293	102	255	3	0.989865	0.7142857
0	0.02191	293	87	270	3	0.989865	0.7563025
0	0.01911	295	75	282	1	0.996622	0.789916
0	0.01684	295	61	296	1	0.996622	0.8291317
0	0.01448	295	47	310	1	0.996622	0.8683473
0	0.01188	295	33	324	1	0.996622	0.907563
0	0.009	296	20	337	0	1	0.9439776

3.2.2 Step 1: Remove Variable "MarriageStatus"

After removing the most insignificant variable, "MarriageStatus", the significant variables in the remaining independent variables are: STD-related symptoms, contact to STD patient, education level, and exchange of sex for money or drug.

From the logistic regression result when "MarriageStatus" variable is moved out, the formula of log ratio between the possibility of having STD (π) and not having STD (1-π) is:

$$
\begin{aligned}
&\log\left(\frac{\pi}{1-\pi}\right) = 9.9836 + \begin{cases} 0.1051 & if\ gender = 0 \\ 0.0000 & if\ gender = 1 \end{cases} \\
&+ 0.00335 * age \\
&+ \begin{cases} 7.0446 & if\ race = 1 \\ 6.8887 & if\ race = 2 \\ 7.1519 & if\ race = 3 \\ 5.9180 & if\ race = 4 \\ 0.0000 & if\ race = 5 \end{cases} + \begin{cases} -0.807 & if\ education = 1 \\ 0.4930 & if\ education = 2 \\ -0.326 & if\ education = 3 \\ -0.879 & if\ education = 4 \\ 0.0000 & if\ education = 5 \end{cases} \\
&+ 0.0754 * income \\
&+ \begin{cases} -0.5376 & if\ employedLastWeek = 0 \\ 0.00000 & if\ employedLastWeek = 1 \end{cases} \\
&+ \begin{cases} 0.3575 & if\ \ financialBarrier = 0 \\ 0.0000 & if\ \ financialBarrier = 1 \end{cases} \\
&+ \begin{cases} -19.8384 & if\ likelihoodHIV = 1 \\ -20.5796 & if\ likelihoodHIV = 2 \\ 0.000000 & if\ likelihoodHIV = 3 \end{cases} \\
&+ \begin{cases} 3.7015 & if\ STDsymptoms = 0 \\ 0.0000 & if\ STDsymptoms = 1 \end{cases} \\
&+ \begin{cases} 0.9034 & if\ exchangeSex = 0 \\ 0.0000 & if\ exchangeSex = 1 \end{cases} \\
&+ \begin{cases} 0.2853 & if\ injectionDrugUse = 0 \\ 0.0000 & if\ injectionDrugUse = 1 \end{cases} \\
&+ \begin{cases} 1.6401 & if\ contactSTD = 0 \\ 0.0000 & if\ contactSTD = 1 \end{cases}
\end{aligned}
$$

Table 3.2 shows some representative points on ROC curve after step1.

Table 3.2 Some representative points on ROC curve at step1

STEP	_PROB_	_POS_	_NEG_	_FALPOS_	_FALNEG_	_SENSIT_	_1MSPEC_
0	1	1	357	0	295	0.003378	0
0	0.9911	15	357	0	281	0.050676	0
0	0.9854	29	357	0	267	0.097973	0
0	0.9804	43	357	0	253	0.14527	0
0	0.9721	57	357	0	239	0.192568	0
0	0.9616	70	356	1	226	0.236486	0.0028011
0	0.9438	84	356	1	212	0.283784	0.0028011
0	0.9335	97	355	2	199	0.327703	0.0056022
0	0.9264	111	355	2	185	0.375	0.0056022
0	0.9186	123	353	4	173	0.415541	0.0112045
0	0.9049	136	352	5	160	0.459459	0.0140056
0	0.8892	150	352	5	146	0.506757	0.0140056
0	0.8779	162	350	7	134	0.547297	0.0196078
0	0.8544	175	349	8	121	0.591216	0.022409
0	0.8268	187	347	10	109	0.631757	0.0280112
0	0.8002	201	347	10	95	0.679054	0.0280112
0	0.7683	208	340	17	88	0.702703	0.047619
0	0.742	217	335	22	79	0.733108	0.0616246
0	0.7148	226	330	27	70	0.763514	0.0756303
0	0.668	236	326	31	60	0.797297	0.0868347
0	0.6284	247	323	34	49	0.834459	0.0952381
0	0.5924	255	317	40	41	0.861486	0.1120448
0	0.5538	267	315	42	29	0.902027	0.1176471
0	0.494	274	308	49	22	0.925676	0.1372549
0	0.4188	275	295	62	21	0.929054	0.1736695
0	0.3084	278	284	73	18	0.939189	0.2044818
0	0.1409	281	273	84	15	0.949324	0.2352941
0	0.1155	283	261	96	13	0.956081	0.2689076
0	0.0908	284	248	109	12	0.959459	0.3053221
0	0.0751	284	234	123	12	0.959459	0.3445378
0	0.0689	285	220	137	11	0.962838	0.3837535
0	0.061	287	208	149	9	0.969595	0.4173669
0	0.055	290	197	160	6	0.97973	0.4481793
0	0.0513	291	184	173	5	0.983108	0.4845938
0	0.0482	291	170	187	5	0.983108	0.5238095
0	0.041	291	156	201	5	0.983108	0.5630252
0	0.0365	292	143	214	4	0.986486	0.5994398
0	0.0324	292	129	228	4	0.986486	0.6386555
0	0.0289	292	115	242	4	0.986486	0.6778711
0	0.0242	293	102	255	3	0.989865	0.7142857
0	0.0217	293	87	270	3	0.989865	0.7563025
0	0.0192	295	75	282	1	0.996622	0.789916
0	0.017	295	61	296	1	0.996622	0.8291317
0	0.0144	295	47	310	1	0.996622	0.8683473
0	0.0119	295	33	324	1	0.996622	0.907563
0	0.0095	296	20	337	0	1	0.9439776

3.2.3 Step2: Remove Variable "Age"

The next been removed most insignificant variable is "Age", the significant variables in the remaining independent variables are: STD-related symptoms, contact to STD patient, education level, and exchange of sex for money or drug.

The logistic regression result in selection step2 shows that when "Age" variable is moved out, the formula of log ratio between the possibility of having STD (π) and not having STD (1-π) is:

$$\log\left(\frac{\pi}{1-\pi}\right) = 10.0202 + \begin{cases} 0.1003 & if\ gender = 0 \\ 0.0000 & if\ gender = 1 \end{cases}$$

$$+ \begin{cases} 6.9745 & if\ race = 1 \\ 6.8320 & if\ race = 2 \\ 7.1299 & if\ race = 3 \\ 5.8863 & if\ race = 4 \\ 0.0000 & if\ race = 5 \end{cases} + \begin{cases} -0.805 & if\ education = 1 \\ 0.4935 & if\ education = 2 \\ -0.321 & if\ education = 3 \\ -0.866 & if\ education = 4 \\ 0.0000 & if\ education = 5 \end{cases}$$

$$+ 0.0795 * income$$

$$+ \begin{cases} -0.5272 & if\ employedLastWeek = 0 \\ 0.00000 & if\ employedLastWeek = 1 \end{cases}$$

$$+ \begin{cases} 0.3482 & if\ \ financialBarrier = 0 \\ 0.0000 & if\ \ financialBarrier = 1 \end{cases}$$

$$+ \begin{cases} -19.7233 & if\ likelihoodHIV = 1 \\ -20.4734 & if\ likelihoodHIV = 2 \\ 0.000000 & if\ likelihoodHIV = 3 \end{cases}$$

$$+ \begin{cases} 3.7099 & if\ STDsymptoms = 0 \\ 0.0000 & if\ STDsymptoms = 1 \end{cases}$$

$$+ \begin{cases} 0.9137 & if\ exchangeSex = 0 \\ 0.0000 & if\ exchangeSex = 1 \end{cases}$$

$$+ \begin{cases} 0.2843 & if\ injectionDrugUse = 0 \\ 0.0000 & if\ injectionDrugUse = 1 \end{cases}$$

$$+ \begin{cases} 1.6497 & if\ contactSTD = 0 \\ 0.0000 & if\ contactSTD = 1 \end{cases}$$

Table 3.3 Some representative points on ROC curve at step2

STEP	_PROB_	_POS_	_NEG_	_FALPOS_	_FALNEG_	_SENSIT_	_1MSPEC_
0	1	1	357	0	295	0.003378	0
0	0.9911	15	357	0	281	0.050676	0
0	0.986	28	357	0	268	0.094595	0
0	0.9816	41	357	0	255	0.138514	0
0	0.9751	54	357	0	242	0.182432	0
0	0.9637	67	357	0	229	0.226351	0
0	0.9531	79	356	1	217	0.266892	0.0028011
0	0.9363	92	356	1	204	0.310811	0.0028011
0	0.9298	104	355	2	192	0.351351	0.0056022
0	0.9232	116	354	3	180	0.391892	0.0084034
0	0.9132	127	352	5	169	0.429054	0.0140056
0	0.8975	140	352	5	156	0.472973	0.0140056
0	0.8852	153	352	5	143	0.516892	0.0140056
0	0.8718	164	350	7	132	0.554054	0.0196078
0	0.8476	176	349	8	120	0.594595	0.022409
0	0.8291	187	347	10	109	0.631757	0.0280112
0	0.8025	200	347	10	96	0.675676	0.0280112
0	0.7678	208	341	16	88	0.702703	0.0448179
0	0.749	216	336	21	80	0.72973	0.0588235
0	0.7206	224	331	26	72	0.756757	0.0728291
0	0.6744	233	327	30	63	0.787162	0.0840336
0	0.6478	242	323	34	54	0.817568	0.0952381
0	0.6172	254	322	35	42	0.858108	0.0980392
0	0.5675	262	317	40	34	0.885135	0.1120448
0	0.5428	269	311	46	27	0.908784	0.1288515
0	0.4622	274	302	55	22	0.925676	0.1540616
0	0.3873	276	291	66	20	0.932432	0.1848739
0	0.2016	280	282	75	16	0.945946	0.210084
0	0.127	281	270	87	15	0.949324	0.2436975
0	0.115	283	259	98	13	0.956081	0.2745098
0	0.0907	284	247	110	12	0.959459	0.3081232
0	0.0743	284	232	125	12	0.959459	0.3501401
0	0.0685	286	216	141	10	0.966216	0.394958
0	0.0582	289	205	152	7	0.976351	0.4257703
0	0.0539	290	193	164	6	0.97973	0.4593838
0	0.0503	291	179	178	5	0.983108	0.4985994
0	0.0446	291	166	191	5	0.983108	0.535014
0	0.0408	292	154	203	4	0.986486	0.5686275
0	0.0349	292	141	216	4	0.986486	0.605042
0	0.0315	292	127	230	4	0.986486	0.6442577
0	0.0271	292	111	246	4	0.986486	0.6890756
0	0.0237	293	96	261	3	0.989865	0.7310924
0	0.0216	293	83	274	3	0.989865	0.767507
0	0.0186	295	70	287	1	0.996622	0.8039216
0	0.0157	295	53	304	1	0.996622	0.8515406
0	0.0132	295	38	319	1	0.996622	0.8935574

3.2.4 Step3: Remove Variable "Gender"

The next been removed variable is "Gender", the significant variables in the remaining independent variables are still: STD-related symptoms, contact to STD patient, education level, and exchange of sex for money or drug.

The logistic regression result in selection step3 shows that when "Gender" variable is moved out, the formula of log ratio between the possibility of having STD (π) and not having STD (1-π) is:

$$
\begin{aligned}
&\log\left(\frac{\pi}{1-\pi}\right) = 10.0009 \\
&+\begin{cases} 6.9012 & if\ race = 1 \\ 6.7612 & if\ race = 2 \\ 7.0434 & if\ race = 3 \\ 5.8141 & if\ race = 4 \\ 0.0000 & if\ race = 5 \end{cases} + \begin{cases} -0.785 & if\ education = 1 \\ 0.4901 & if\ education = 2 \\ -0.308 & if\ education = 3 \\ -0.859 & if\ education = 4 \\ 0.0000 & if\ education = 5 \end{cases} \\
&+ 0.0774 * income \\
&+\begin{cases} -0.5312 & if\ employedLastWeek = 0 \\ 0.00000 & if\ employedLastWeek = 1 \end{cases} \\
&+\begin{cases} 0.3559 & if\ financialBarrier = 0 \\ 0.0000 & if\ financialBarrier = 1 \end{cases} \\
&+\begin{cases} -19.6060 & if\ likelihoodHIV = 1 \\ -20.3657 & if\ likelihoodHIV = 2 \\ 0.000000 & if\ likelihoodHIV = 3 \end{cases} \\
&+\begin{cases} 3.7126 & if\ STDsymptoms = 0 \\ 0.0000 & if\ STDsymptoms = 1 \end{cases} \\
&+\begin{cases} 0.9208 & if\ exchangeSex = 0 \\ 0.0000 & if\ exchangeSex = 1 \end{cases} \\
&+\begin{cases} 0.2854 & if\ injectionDrugUse = 0 \\ 0.0000 & if\ injectionDrugUse = 1 \end{cases} \\
&+\begin{cases} 1.6434 & if\ contactSTD = 0 \\ 0.0000 & if\ contactSTD = 1 \end{cases}
\end{aligned}
$$

Table 3.4 Some representative points on ROC curve at step3

STEP	_PROB_	_POS_	_NEG_	_FALPOS_	_FALNEG_	_SENSIT_	_1MSPEC_
0	1	1	357	0	295	0.003378	0
0	0.9904	16	357	0	280	0.054054	0
0	0.9855	29	357	0	267	0.097973	0
0	0.9817	42	357	0	254	0.141892	0
0	0.9753	55	357	0	241	0.185811	0
0	0.9629	68	357	0	228	0.22973	0
0	0.9518	80	356	1	216	0.27027	0.0028011
0	0.9372	93	356	1	203	0.314189	0.0028011
0	0.9314	105	355	2	191	0.35473	0.0056022
0	0.9207	117	354	3	179	0.39527	0.0084034
0	0.9127	128	352	5	168	0.432432	0.0140056
0	0.898	141	352	5	155	0.476351	0.0140056
0	0.8857	154	352	5	142	0.52027	0.0140056
0	0.8704	165	350	7	131	0.557432	0.0196078
0	0.8451	177	349	8	119	0.597973	0.022409
0	0.8209	188	347	10	108	0.635135	0.0280112
0	0.7976	200	345	12	96	0.675676	0.0336134
0	0.7685	208	340	17	88	0.702703	0.047619
0	0.7477	218	336	21	78	0.736486	0.0588235
0	0.7192	225	330	27	71	0.760135	0.0756303
0	0.6702	235	327	30	61	0.793919	0.0840336
0	0.6378	243	322	35	53	0.820946	0.0980392
0	0.6135	253	319	38	43	0.85473	0.1064426
0	0.5687	263	316	41	33	0.888514	0.1148459
0	0.5284	271	309	48	25	0.915541	0.1344538
0	0.4447	275	299	58	21	0.929054	0.162465
0	0.3433	278	289	68	18	0.939189	0.1904762
0	0.1728	280	277	80	16	0.945946	0.2240896
0	0.1259	282	266	91	14	0.952703	0.254902
0	0.1009	283	254	103	13	0.956081	0.2885154
0	0.0807	284	242	115	12	0.959459	0.3221289
0	0.0704	285	222	135	11	0.962838	0.3781513
0	0.0615	288	210	147	8	0.972973	0.4117647
0	0.0553	289	197	160	7	0.976351	0.4481793
0	0.0524	290	182	175	6	0.97973	0.4901961
0	0.0468	291	169	188	5	0.983108	0.5266106
0	0.0412	292	157	200	4	0.986486	0.5602241
0	0.035	292	143	214	4	0.986486	0.5994398
0	0.0314	292	128	229	4	0.986486	0.6414566
0	0.0276	292	112	245	4	0.986486	0.6862745
0	0.024	293	97	260	3	0.989865	0.7282913
0	0.0218	293	83	274	3	0.989865	0.767507
0	0.0184	295	67	290	1	0.996622	0.8123249
0	0.0149	295	50	307	1	0.996622	0.859944
0	0.0119	295	35	322	1	0.996622	0.9019608
0	0.0085	296	18	339	0	1	0.9495798

3.2.5 Step4: Remove Variable "Race"

The variable been removed in step 4 is "Race", the significant variables in the remaining independent variables are still: STD-related symptoms, contact to STD patient, education level, and exchange of sex for money or drug.

The logistic regression result in selection step4 shows that, the formula of log ratio between the possibility of having STD (π) and not having STD (1-π), without "Race" variable, is:

$$
\begin{aligned}
&\log\left(\frac{\pi}{1-\pi}\right) = 11.0974 \\
&+\begin{cases} -1.115 & if\ education = 1 \\ 0.4358 & if\ education = 2 \\ -0.297 & if\ education = 3 \\ -0.930 & if\ education = 4 \\ 0.0000 & if\ education = 5 \end{cases} \\
&+ 0.0814 * income \\
&+\begin{cases} -0.504 & if\ employedLastWeek = 0 \\ 0.00000 & if\ employedLastWeek = 1 \end{cases} \\
&+\begin{cases} 0.3171 & if\ \ financialBarrier = 0 \\ 0.0000 & if\ \ financialBarrier = 1 \end{cases} \\
&+\begin{cases} -13.8245 & if\ likelihoodHIV = 1 \\ -14.5621 & if\ likelihoodHIV = 2 \\ 0.000000 & if\ likelihoodHIV = 3 \end{cases} \\
&+\begin{cases} 3.7467 & if\ STDsymptoms = 0 \\ 0.0000 & if\ STDsymptoms = 1 \end{cases} \\
&+\begin{cases} 0.9062 & if\ exchangeSex = 0 \\ 0.0000 & if\ exchangeSex = 1 \end{cases} \\
&+\begin{cases} 0.2606 & if\ injectionDrugUse = 0 \\ 0.0000 & if\ injectionDrugUse = 1 \end{cases} \\
&+\begin{cases} 1.5688 & if\ contactSTD = 0 \\ 0.0000 & if\ contactSTD = 1 \end{cases}
\end{aligned}
$$

Table 3.5 Some representative points on ROC curve at step4

STEP	_PROB_	_POS_	_NEG_	_FALPOS_	_FALNEG_	_SENSIT_	_1MSPEC_
0	1	1	357	0	295	0.003378	0
0	0.9892	16	357	0	280	0.054054	0
0	0.9841	29	357	0	267	0.097973	0
0	0.9799	42	357	0	254	0.141892	0
0	0.9709	55	357	0	241	0.185811	0
0	0.9607	68	357	0	228	0.22973	0
0	0.9459	80	356	1	216	0.27027	0.0028011
0	0.9365	93	356	1	203	0.314189	0.0028011
0	0.9285	105	355	2	191	0.35473	0.0056022
0	0.9191	117	354	3	179	0.39527	0.0084034
0	0.91	129	353	4	167	0.435811	0.0112045
0	0.8999	141	352	5	155	0.476351	0.0140056
0	0.8878	154	352	5	142	0.52027	0.0140056
0	0.8651	165	350	7	131	0.557432	0.0196078
0	0.8358	177	349	8	119	0.597973	0.022409
0	0.8232	188	347	10	108	0.635135	0.0280112
0	0.7932	199	345	12	97	0.672297	0.0336134
0	0.7718	208	340	17	88	0.702703	0.047619
0	0.7503	218	337	20	78	0.736486	0.0560224
0	0.7213	225	330	27	71	0.760135	0.0756303
0	0.6818	235	327	30	61	0.793919	0.0840336
0	0.6434	243	322	35	53	0.820946	0.0980392
0	0.6056	251	317	40	45	0.847973	0.1120448
0	0.5568	262	312	45	34	0.885135	0.1260504
0	0.4962	270	307	50	26	0.912162	0.140056
0	0.4315	276	297	60	20	0.932432	0.1680672
0	0.3205	279	287	70	17	0.942568	0.1960784
0	0.1544	280	275	82	16	0.945946	0.2296919
0	0.1103	282	264	93	14	0.952703	0.2605042
0	0.0889	284	253	104	12	0.959459	0.2913165
0	0.0775	284	239	118	12	0.959459	0.3305322
0	0.067	284	222	135	12	0.959459	0.3781513
0	0.0609	287	205	152	9	0.969595	0.4257703
0	0.0541	290	194	163	6	0.97973	0.4565826
0	0.0503	290	178	179	6	0.97973	0.5014006
0	0.0462	291	164	193	5	0.983108	0.5406162
0	0.0395	291	150	207	5	0.983108	0.5798319
0	0.0339	292	134	223	4	0.986486	0.6246499
0	0.0289	292	120	237	4	0.986486	0.6638655
0	0.0262	293	105	252	3	0.989865	0.7058824
0	0.0227	293	90	267	3	0.989865	0.7478992
0	0.0203	295	72	285	1	0.996622	0.7983193
0	0.0175	295	55	302	1	0.996622	0.8459384
0	0.0155	295	40	317	1	0.996622	0.8879552
0	0.0126	296	23	334	0	1	0.9355742
0	0.009	296	8	349	0	1	0.977591

3.2.6 Step5: Remove Variable "InjectionDrugUse"

The variable been removed in step 5 is "InjectionDrugUse", the significant variables in the remaining independent variables are still: STD-related symptoms, contact to STD patient, education level, and exchange of sex for money or drug.

The logistic regression result in selection step5 shows that, the formula of log ratio between the possibility of having STD (π) and not having STD (1-π), without "InjectionDrugUse" variable, is:

$$\log\left(\frac{\pi}{1-\pi}\right) = 11.354$$
$$+\begin{cases} -1.091 & if\ education = 1 \\ 0.4173 & if\ education = 2 \\ -0.314 & if\ education = 3 \\ -0.955 & if\ education = 4 \\ 0.0000 & if\ education = 5 \end{cases}$$
$$+ 0.0748 * income$$
$$+\begin{cases} -0.5293 & if\ employedLastWeek = 0 \\ 0.00000 & if\ employedLastWeek = 1 \end{cases}$$
$$+\begin{cases} 0.3251 & if\ \ financialBarrier = 0 \\ 0.0000 & if\ \ financialBarrier = 1 \end{cases}$$
$$+\begin{cases} -13.6711 & if\ likelihoodHIV = 1 \\ -14.4287 & if\ likelihoodHIV = 2 \\ 0.000000 & if\ likelihoodHIV = 3 \end{cases}$$
$$+\begin{cases} 3.7227 & if\ STDsymptoms = 0 \\ 0.0000 & if\ STDsymptoms = 1 \end{cases}$$
$$+\begin{cases} 0.8778 & if\ exchangeSex = 0 \\ 0.0000 & if\ exchangeSex = 1 \end{cases}$$
$$+\begin{cases} 1.5739 & if\ contactSTD = 0 \\ 0.0000 & if\ contactSTD = 1 \end{cases}$$

Table 3.6 Some representative points on ROC curve at step5

STEP	_PROB_	_POS_	_NEG_	_FALPOS_	_FALNEG_	_SENSIT_	_1MSPEC_
0	0.99999	1	357	0	295	0.003378	0
0	0.98845	14	357	0	282	0.047297	0
0	0.98457	27	357	0	269	0.091216	0
0	0.97866	40	357	0	256	0.135135	0
0	0.97101	52	357	0	244	0.175676	0
0	0.96377	64	357	0	232	0.216216	0
0	0.94969	77	357	0	219	0.260135	0
0	0.94274	88	356	1	208	0.297297	0.0028011
0	0.93287	99	355	2	197	0.334459	0.0056022
0	0.92514	110	354	3	186	0.371622	0.0084034
0	0.91506	121	353	4	175	0.408784	0.0112045
0	0.90141	133	353	4	163	0.449324	0.0112045
0	0.8962	144	352	5	152	0.486486	0.0140056
0	0.88974	155	351	6	141	0.523649	0.0168067
0	0.86344	165	349	8	131	0.557432	0.022409
0	0.84507	176	348	9	120	0.594595	0.0252101
0	0.82169	187	347	10	109	0.631757	0.0280112
0	0.79499	196	344	13	100	0.662162	0.0364146
0	0.78121	206	342	15	90	0.695946	0.0420168
0	0.75774	213	334	23	83	0.719595	0.0644258
0	0.72863	224	331	26	72	0.756757	0.0728291
0	0.68209	234	328	29	62	0.790541	0.0812325
0	0.65999	243	325	32	53	0.820946	0.0896359
0	0.61642	250	320	37	46	0.844595	0.1036415
0	0.58659	259	315	42	37	0.875	0.1176471
0	0.52143	267	308	49	29	0.902027	0.1372549
0	0.43994	277	300	57	19	0.935811	0.1596639
0	0.34809	278	289	68	18	0.939189	0.1904762
0	0.23314	280	279	78	16	0.945946	0.2184874
0	0.11285	280	267	90	16	0.945946	0.2521008
0	0.10047	282	257	100	14	0.952703	0.280112
0	0.08926	284	247	110	12	0.959459	0.3081232
0	0.07467	284	234	123	12	0.959459	0.3445378
0	0.06691	286	222	135	10	0.966216	0.3781513
0	0.06136	288	198	159	8	0.972973	0.4453782
0	0.05231	290	185	172	6	0.97973	0.4817927
0	0.04585	290	170	187	6	0.97973	0.5238095
0	0.04158	291	153	204	5	0.983108	0.5714286
0	0.03521	291	138	219	5	0.983108	0.6134454
0	0.03275	292	124	233	4	0.986486	0.6526611
0	0.02643	292	111	246	4	0.986486	0.6890756
0	0.02388	294	95	262	2	0.993243	0.7338936
0	0.0216	295	79	278	1	0.996622	0.7787115
0	0.01796	295	59	298	1	0.996622	0.8347339
0	0.01577	295	40	317	1	0.996622	0.8879552
0	0.01195	296	29	328	0	1	0.9187675

3.2.7 Step6: Remove Variable "FinancialBarrier"

The variable been removed in step 6 is "FinancialBarrier", the significant variables in the remaining independent variables are still: STD-related symptoms, contact to STD patient, education level, and exchange of sex for money or drug.

The logistic regression result in selection step6 shows that, the formula of log ratio between the possibility of having STD (π) and not having STD (1-π), without "FinancialBarrier" variable, is:

$$\log\left(\frac{\pi}{1-\pi}\right) = 11.14$$
$$+\begin{cases} -1.083 & if\ education = 1 \\ 0.3818 & if\ education = 2 \\ -0.342 & if\ education = 3 \\ -0.955 & if\ education = 4 \\ 0.0000 & if\ education = 5 \end{cases}$$
$$+0.0752 * income$$
$$+\begin{cases} -0.5286 & if\ employedLastWeek = 0 \\ 0.00000 & if\ employedLastWeek = 1 \end{cases}$$
$$+\begin{cases} -13.4879 & if\ likelihoodHIV = 1 \\ -14.2443 & if\ likelihoodHIV = 2 \\ 0.000000 & if\ likelihoodHIV = 3 \end{cases}$$
$$+\begin{cases} 3.7200 & if\ STDsymptoms = 0 \\ 0.0000 & if\ STDsymptoms = 1 \end{cases}$$
$$+\begin{cases} 0.9029 & if\ exchangeSex = 0 \\ 0.0000 & if\ exchangeSex = 1 \end{cases}$$
$$+\begin{cases} 1.5531 & if\ contactSTD = 0 \\ 0.0000 & if\ contactSTD = 1 \end{cases}$$

Table 3.7 Some representative points on ROC curve at step6

STEP	_PROB_	_POS_	_NEG_	_FALPOS_	_FALNEG_	_SENSIT_	_1MSPEC_
0	1	1	357	0	295	0.003378	0
0	0.9873	13	357	0	283	0.043919	0
0	0.9847	26	357	0	270	0.087838	0
0	0.9809	38	357	0	258	0.128378	0
0	0.9745	49	357	0	247	0.165541	0
0	0.9616	60	357	0	236	0.202703	0
0	0.9493	71	357	0	225	0.239865	0
0	0.9431	81	356	1	215	0.273649	0.0028011
0	0.9378	93	355	2	203	0.314189	0.0056022
0	0.9273	103	354	3	193	0.347973	0.0084034
0	0.9173	116	354	3	180	0.391892	0.0084034
0	0.9114	127	353	4	169	0.429054	0.0112045
0	0.8976	139	352	5	157	0.469595	0.0140056
0	0.8862	148	350	7	148	0.5	0.0196078
0	0.8769	159	350	7	137	0.537162	0.0196078
0	0.8615	168	348	9	128	0.567568	0.0252101
0	0.8321	178	347	10	118	0.601351	0.0280112
0	0.8115	187	345	12	109	0.631757	0.0336134
0	0.7932	196	343	14	100	0.662162	0.0392157
0	0.773	205	339	18	91	0.692568	0.0504202
0	0.7593	214	337	20	82	0.722973	0.0560224
0	0.7372	221	330	27	75	0.746622	0.0756303
0	0.7005	233	327	30	63	0.787162	0.0840336
0	0.6488	242	324	33	54	0.817568	0.092437
0	0.6295	250	321	36	46	0.844595	0.1008403
0	0.5856	259	316	41	37	0.875	0.1148459
0	0.5234	265	309	48	31	0.89527	0.1344538
0	0.4722	273	300	57	23	0.922297	0.1596639
0	0.4034	278	292	65	18	0.939189	0.1820728
0	0.2717	280	282	75	16	0.945946	0.210084
0	0.1248	280	270	87	16	0.945946	0.2436975
0	0.102	282	260	97	14	0.952703	0.2717087
0	0.0895	283	250	107	13	0.956081	0.2997199
0	0.081	284	238	119	12	0.959459	0.3333333
0	0.0695	285	224	133	11	0.962838	0.372549
0	0.0615	287	213	144	9	0.969595	0.4033613
0	0.0548	288	191	166	8	0.972973	0.464986
0	0.0462	290	167	190	6	0.97973	0.5322129
0	0.0406	291	156	201	5	0.983108	0.5630252
0	0.0386	292	144	213	4	0.986486	0.5966387
0	0.0321	292	131	226	4	0.986486	0.6330532
0	0.027	295	115	242	1	0.996622	0.6778711
0	0.025	295	101	256	1	0.996622	0.7170868
0	0.0214	295	79	278	1	0.996622	0.7787115
0	0.0191	295	59	298	1	0.996622	0.8347339
0	0.0175	296	36	321	0	1	0.8991597

3.2.8 Step7: Remove Variable "Income"

The variable been removed in step7 is "Income", the significant variables in the remaining independent variables are still: STD-related symptoms, contact to STD patient, education level, and exchange of sex for money or drug.

The logistic regression result in selection step7 shows that, the formula of log ratio between the possibility of having STD (π) and not having STD (1-π), without "Income" variable, is:

$$\log\left(\frac{\pi}{1-\pi}\right) = 11.546$$
$$+\begin{cases} -1.155 & if\ education = 1 \\ 0.3726 & if\ education = 2 \\ -0.348 & if\ education = 3 \\ -1.004 & if\ education = 4 \\ 0.0000 & if\ education = 5 \end{cases}$$
$$+\begin{cases} -0.5083 & if\ employedLastWeek = 0 \\ 0.00000 & if\ employedLastWeek = 1 \end{cases}$$
$$+\begin{cases} -13.8168 & if\ likelihoodHIV = 1 \\ -14.5811 & if\ likelihoodHIV = 2 \\ 0.000000 & if\ likelihoodHIV = 3 \end{cases}$$
$$+\begin{cases} 3.8372 & if\ STDsymptoms = 0 \\ 0.0000 & if\ STDsymptoms = 1 \end{cases}$$
$$+\begin{cases} 0.9713 & if\ exchangeSex = 0 \\ 0.0000 & if\ exchangeSex = 1 \end{cases}$$
$$+\begin{cases} 1.5442 & if\ contactSTD = 0 \\ 0.0000 & if\ contactSTD = 1 \end{cases}$$

Table 3.8 Some representative points on ROC curve at step7

STEP	_PROB_	_POS_	_NEG_	_FALPOS_	_FALNEG_	_SENSIT_	_1MSPEC_
0	1	1	357	0	295	0.003378	0
0	0.9885	13	357	0	283	0.043919	0
0	0.9767	45	357	0	251	0.152027	0
0	0.9702	46	357	0	250	0.155405	0
0	0.9573	49	357	0	247	0.165541	0
0	0.9514	60	357	0	236	0.202703	0
0	0.9491	65	357	0	231	0.219595	0
0	0.9406	87	355	2	209	0.293919	0.0056022
0	0.9381	88	355	2	208	0.297297	0.0056022
0	0.9267	117	353	4	179	0.39527	0.0112045
0	0.91	120	353	4	176	0.405405	0.0112045
0	0.9012	125	353	4	171	0.422297	0.0112045
0	0.8953	151	350	7	145	0.510135	0.0196078
0	0.8806	152	350	7	144	0.513514	0.0196078
0	0.8742	165	350	7	131	0.557432	0.0196078
0	0.8318	172	348	9	124	0.581081	0.0252101
0	0.8226	183	346	11	113	0.618243	0.0308123
0	0.7993	199	341	16	97	0.672297	0.0448179
0	0.7718	208	336	21	88	0.702703	0.0588235
0	0.7423	226	329	28	70	0.763514	0.0784314
0	0.6835	231	328	29	65	0.780405	0.0812325
0	0.6607	249	321	36	47	0.841216	0.1008403
0	0.6371	254	319	38	42	0.858108	0.1064426
0	0.6116	255	319	38	41	0.861486	0.1064426
0	0.5729	262	311	46	34	0.885135	0.1288515
0	0.4865	271	303	54	25	0.915541	0.1512605
0	0.4742	278	295	62	18	0.939189	0.1736695
0	0.4124	278	291	66	18	0.939189	0.1848739
0	0.3296	279	290	67	17	0.942568	0.1876751
0	0.297	280	282	75	16	0.945946	0.210084
0	0.2835	280	279	78	16	0.945946	0.2184874
0	0.1615	280	276	81	16	0.945946	0.2268908
0	0.1193	281	266	91	15	0.949324	0.254902
0	0.0963	282	263	94	14	0.952703	0.2633053
0	0.0909	282	250	107	14	0.952703	0.2997199
0	0.0839	282	248	109	14	0.952703	0.3053221
0	0.0791	285	230	127	11	0.962838	0.3557423
0	0.0585	290	173	184	6	0.97973	0.5154062
0	0.0473	290	172	185	6	0.97973	0.5182073
0	0.042	291	154	203	5	0.983108	0.5686275
0	0.0384	292	142	215	4	0.986486	0.6022409
0	0.0328	292	138	219	4	0.986486	0.6134454
0	0.0281	294	109	248	2	0.993243	0.6946779
0	0.0223	295	70	287	1	0.996622	0.8039216
0	0.0192	296	38	319	0	1	0.8935574
0	0.0149	296	34	323	0	1	0.9047619

3.2.9 Step8: Remove Variable "EmployedLastWeek"

The variable been removed in step8 is "EmployedLastWeek", the left independent variables are: STD-related symptoms, exchange of sex for money or drug, contact to STD patient, education level, and likelihood of getting HIV. All of them are significant, and no more independent variable can be moved from our model. This is our final model.

The logistic regression result in selection step8 shows that, the formula of log ratio between the possibility of having STD (π) and not having STD (1-π), without "EmployedLastWeek" variable, is:

$$\log\left(\frac{\pi}{1-\pi}\right) = 11.0273$$
$$+\begin{cases} -1.161 & if\ education = 1 \\ 0.4537 & if\ education = 2 \\ -0.284 & if\ education = 3 \\ -1.010 & if\ education = 4 \\ 0.0000 & if\ education = 5 \end{cases}$$
$$+\begin{cases} -13.7062 & if\ likelihoodHIV = 1 \\ -14.4843 & if\ likelihoodHIV = 2 \\ 0.000000 & if\ likelihoodHIV = 3 \end{cases}$$
$$+\begin{cases} 3.8568 & if\ STDsymptoms = 0 \\ 0.0000 & if\ STDsymptoms = 1 \end{cases}$$
$$+\begin{cases} 1.3376 & if\ exchangeSex = 0 \\ 0.0000 & if\ exchangeSex = 1 \end{cases}$$
$$+\begin{cases} 1.5746 & if\ contactSTD = 0 \\ 0.0000 & if\ contactSTD = 1 \end{cases}$$

Table 3.9 Some representative points on ROC curve at step8

STEP	_PROB_	_POS_	_NEG_	_FALPOS_	_FALNEG_	_SENSIT_	_1MSPEC_
0	1	1	357	0	295	0.003378	0
0	1	2	357	0	294	0.006757	0
0	0.9835	35	357	0	261	0.118243	0
0	0.9782	45	357	0	251	0.152027	0
0	0.9648	46	357	0	250	0.155405	0
0	0.9561	58	357	0	238	0.195946	0
0	0.9538	60	357	0	236	0.202703	0
0	0.9512	80	355	2	216	0.27027	0.0056022
0	0.9493	83	355	2	213	0.280405	0.0056022
0	0.9252	114	353	4	182	0.385135	0.0112045
0	0.9219	118	353	4	178	0.398649	0.0112045
0	0.919	122	353	4	174	0.412162	0.0112045
0	0.903	146	351	6	150	0.493243	0.0168067
0	0.8995	149	350	7	147	0.503378	0.0196078
0	0.8781	151	350	7	145	0.510135	0.0196078
0	0.8504	158	348	9	138	0.533784	0.0252101
0	0.8443	159	348	9	137	0.537162	0.0252101
0	0.8364	179	343	14	117	0.60473	0.0392157
0	0.8308	180	343	14	116	0.608108	0.0392157
0	0.7948	193	341	16	103	0.652027	0.0448179
0	0.7646	213	334	23	83	0.719595	0.0644258
0	0.7097	237	323	34	59	0.800676	0.0952381
0	0.6743	247	321	36	49	0.834459	0.1008403
0	0.6402	249	320	37	47	0.841216	0.1036415
0	0.5986	255	313	44	41	0.861486	0.1232493
0	0.5289	268	304	53	28	0.905405	0.1484594
0	0.5042	278	296	61	18	0.939189	0.1708683
0	0.4873	278	295	62	18	0.939189	0.1736695
0	0.3521	279	292	65	17	0.942568	0.1820728
0	0.3183	280	281	76	16	0.945946	0.2128852
0	0.2916	280	280	77	16	0.945946	0.2156863
0	0.2489	280	278	79	16	0.945946	0.2212885
0	0.1321	280	274	83	16	0.945946	0.232493
0	0.1078	281	271	86	15	0.949324	0.2408964
0	0.1028	281	270	87	15	0.949324	0.2436975
0	0.094	285	248	109	11	0.962838	0.3053221
0	0.087	285	246	111	11	0.962838	0.3109244
0	0.0757	285	241	116	11	0.962838	0.32493
0	0.0642	290	181	176	6	0.97973	0.4929972
0	0.0491	291	157	200	5	0.983108	0.5602241
0	0.0473	292	146	211	4	0.986486	0.5910364
0	0.0362	292	145	212	4	0.986486	0.5938375
0	0.0244	295	81	276	1	0.996622	0.7731092
0	0.0232	296	73	284	0	1	0.7955182
0	0.021	296	38	319	0	1	0.8935574
0	0.0114	296	18	339	0	1	0.9495798

3.2.10 Summary of Backward Selection

Table 3.10 Summary of the backward selection steps

	AIC	SC	-2 LogL	R-Square	Max-rescaled R-Square	Likelihood Ratio	Score	Wald	Residual Chi-Square Test
step0	434.704	533.299	390.704	0.5412	0.7238	<0.0001	<0.0001	<0.0001	
step1	430.879	520.511	390.879	0.5411	0.7236	<0.0001	<0.0001	<0.0001	0.9518
step2	428.941	514.091	390.941	0.5411	0.7236	<0.0001	<0.0001	<0.0001	0.9835
step3	427.056	507.724	391.056	0.5410	0.7234	<0.0001	<0.0001	<0.0001	0.9914
step4	422.350	485.092	394.350	0.5387	0.7203	<0.0001	<0.0001	<0.0001	0.9021
step5	421.294	479.554	395.294	0.5380	0.7194	<0.0001	<0.0001	<0.0001	0.8828
step6	420.823	474.602	396.823	0.5369	0.7180	<0.0001	<0.0001	<0.0001	0.8211
step7	421.646	470.943	399.646	0.5349	0.7153	<0.0001	<0.0001	<0.0001	0.6825
step8	421.998	466.814	401.998	0.5332	0.7131	<0.0001	<0.0001	<0.0001	0.5576

Table 3.10 shows the model evaluation parameters. All models have p value less than 0.0001; final model shows R-square value as 0.7131, close to the full model's R-square value (0.7238). Our final model is:

$$\log\left(\frac{\pi}{1-\pi}\right) = 11.0273$$
$$+\begin{cases} -1.161 & if\ education = 1 \\ 0.4537 & if\ education = 2 \\ -0.284 & if\ education = 3 \\ -1.010 & if\ education = 4 \\ 0.0000 & if\ education = 5 \end{cases}$$
$$+\begin{cases} -13.7062 & if\ likelihoodHIV = 1 \\ -14.4843 & if\ likelihoodHIV = 2 \\ 0.000000 & if\ likelihoodHIV = 3 \end{cases}$$
$$+\begin{cases} 3.8568 & if\ STDsymptoms = 0 \\ 0.0000 & if\ STDsymptoms = 1 \end{cases}$$
$$+\begin{cases} 1.3376 & if\ exchangeSex = 0 \\ 0.0000 & if\ exchangeSex = 1 \end{cases}$$
$$+\begin{cases} 1.5746 & if\ contactSTD = 0 \\ 0.0000 & if\ contactSTD = 1 \end{cases} = M$$

If we let the right part of this equation equal to M, then the risk function for each client is:

$$Ri = \pi_i = \frac{\exp(Mi)}{1+\exp(Mi)}.$$

Chapter 4

ROC Curve Result

4.1 ROC Curves of Backward Selection

4.1.1 ROC Curve of Full Model

After the logistic model is established, we can calculate the probability of real STD for each client. Comparing to known STD status in dataset, we can decide the sensitivity and specificity for each cutoff point. Receiver operating characteristic (ROC) curve is a two dimensional measurement of performance, with sensitivity and 1-specificity as its vertical and horizontal axes, respectively. The ROC curve from our logistic regression is in Fig 4.1.

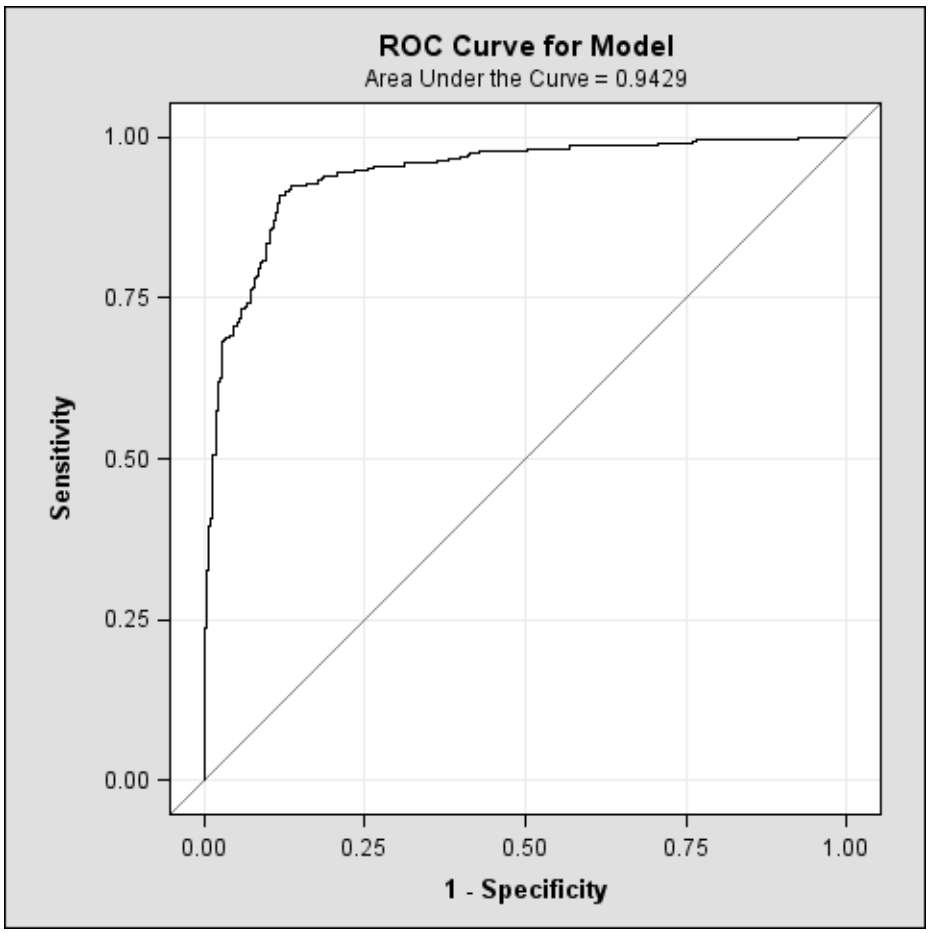

Figure 4.1 ROC curve of the full model

The area under the curve (AUC) of our model is 0.9429, which belongs to excellent level. This AUC shows that our logistic regression model can predict client's STD status quite well.

We compared our model with a non-informative model (p = 0.5, AUC = 0.5). The comparison curve is shown in Fig 4.2. The p value of this comparison is less than 0.0001, which means that our model is much superior to a flip-coin model in predicting if a client has STD.

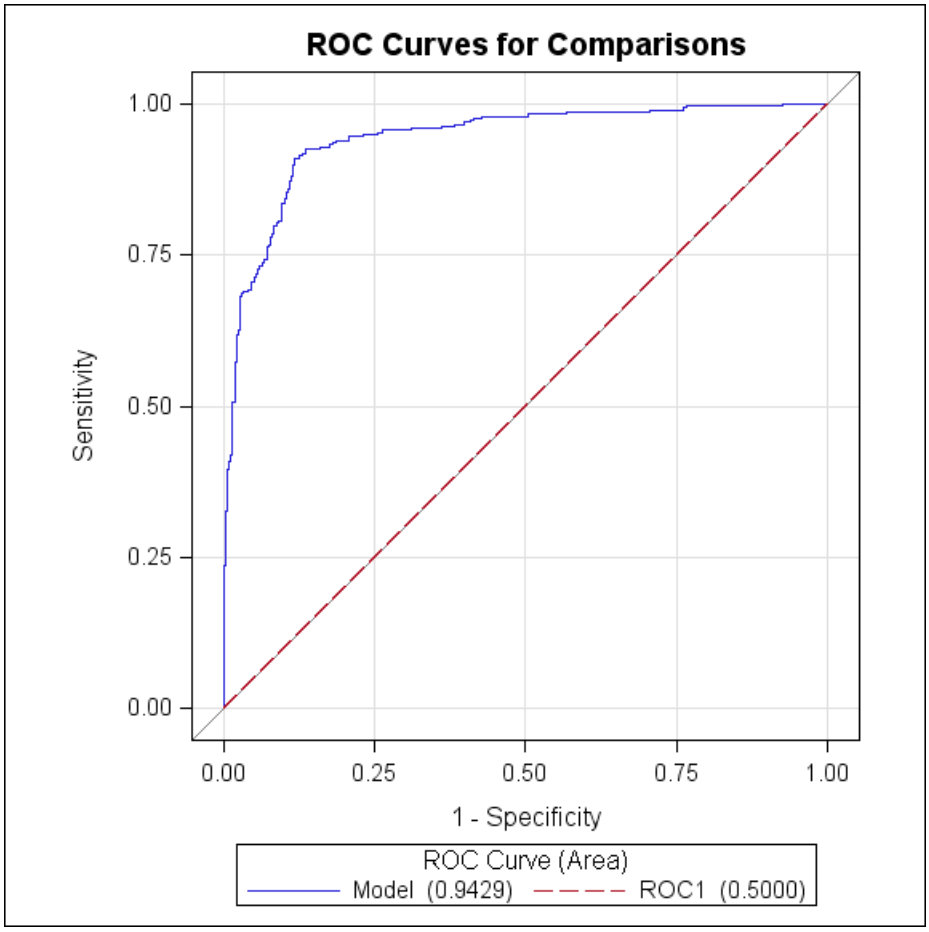

Figure 4.2 ROC curve comparison of full model to non-informative model

4.1.2 ROC Curve of Backward Model Selection Step1

In backward model selection step1, “MarriageStatus” variable is removed from model. The ROC curve from the step1 logistic regression is in Fig 4.3.

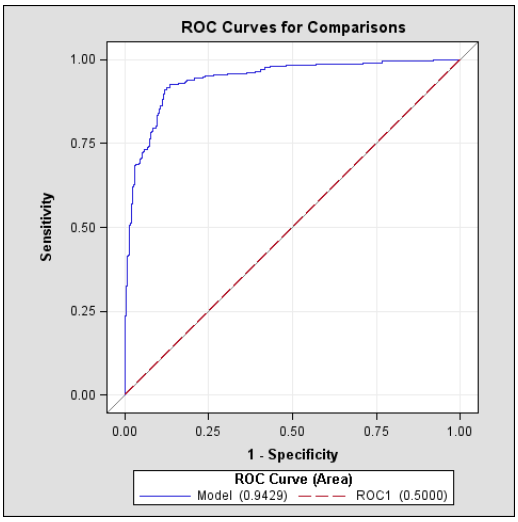

Figure 4.3 ROC curve comparison of step1 model to non-informative model

4.1.3 ROC Curve of Backward Model Selection Step2

In backward model selection step2, “Age” variable is removed from model. The ROC curve from the step2 logistic regression is in Fig 4.4.

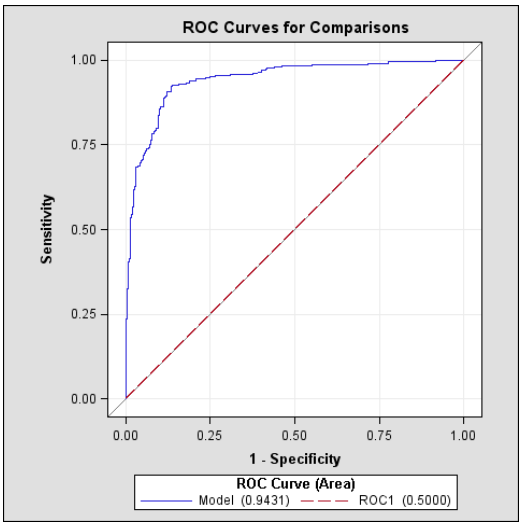

Figure 4.4 ROC curve comparison of step2 model to non-informative model

4.1.4 ROC Curve of Backward Model Selection Step3

In backward model selection step3, “Gender” variable is removed from model. The ROC curve from the step3 logistic regression is in Fig 4.5.

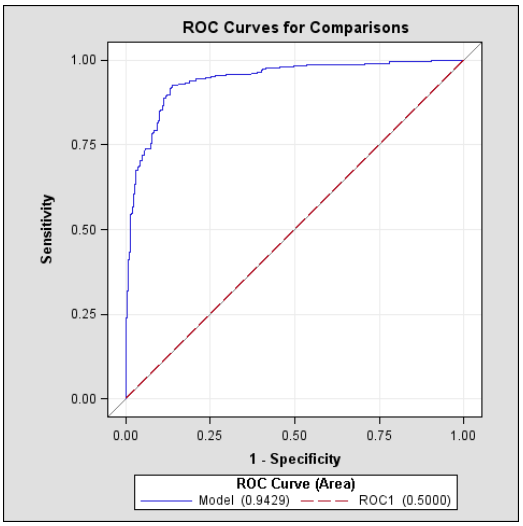

Figure 4.5 ROC curve comparison of step3 model to non-informative model

4.1.5 ROC Curve of Backward Model Selection Step4

In backward model selection step4, “Race” variable is removed from model. The ROC curve from the step4 logistic regression is in Fig 4.6.

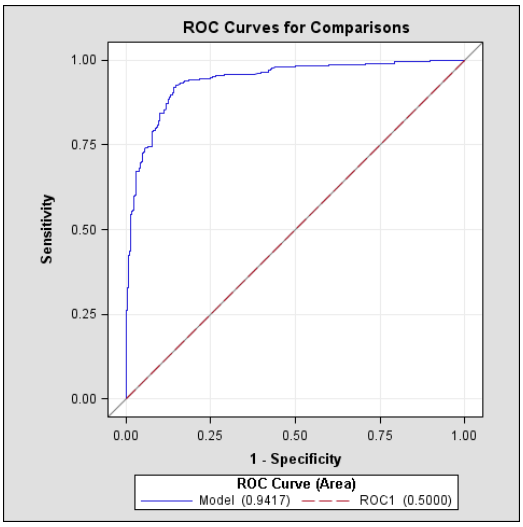

Figure 4.6 ROC curve comparison of step4 model to non-informative model

4.1.6 ROC Curve of Backward Model Selection Step5

In backward model selection step5, “InjectionDrugUse” variable is removed from model. The ROC curve from the step5 logistic regression is in Fig 4.7.

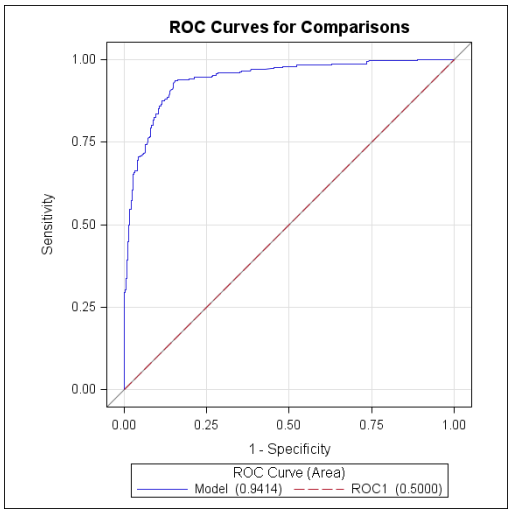

Figure 4.7 ROC curve comparison of step5 model to non-informative model

4.1.7 ROC Curve of Backward Model Selection Step6

In backward model selection step6, “FinancialBarrier” variable is removed from model. The ROC curve from the step6 logistic regression is in Fig 4.8.

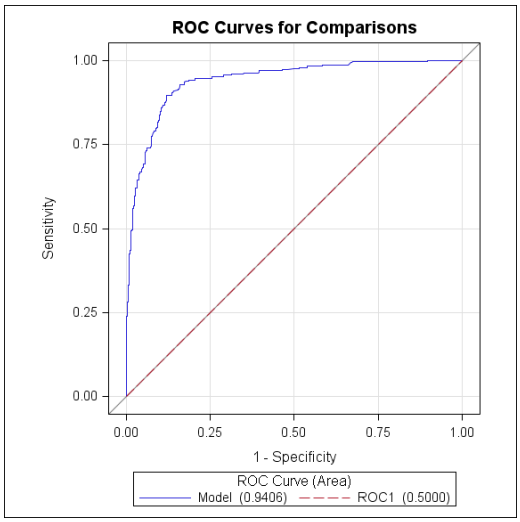

Figure 4.8 ROC curve comparison of step6 model to non-informative model

4.1.8 ROC Curve of Backward Model Selection Step7

In backward model selection step7, “Income” variable is removed from model. The ROC curve from the step7 logistic regression is in Fig 4.9.

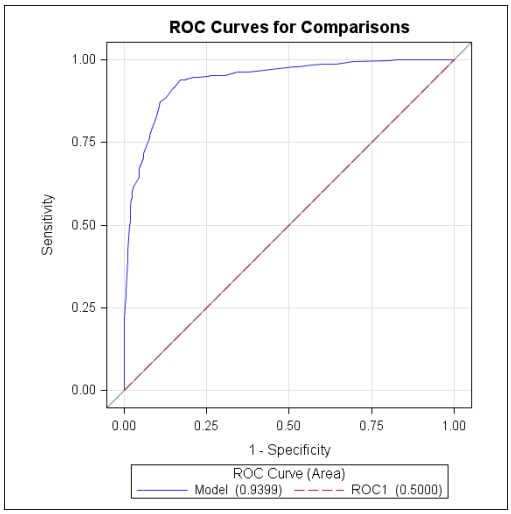

Figure 4.9 ROC curve comparison of step7 model to non-informative model

4.1.9 ROC Curve of Backward Model Selection Step8

In backward model selection step8, “EmployedLastWeek” variable is removed from model. The ROC curve from the step8 logistic regression is in Fig 4.10.

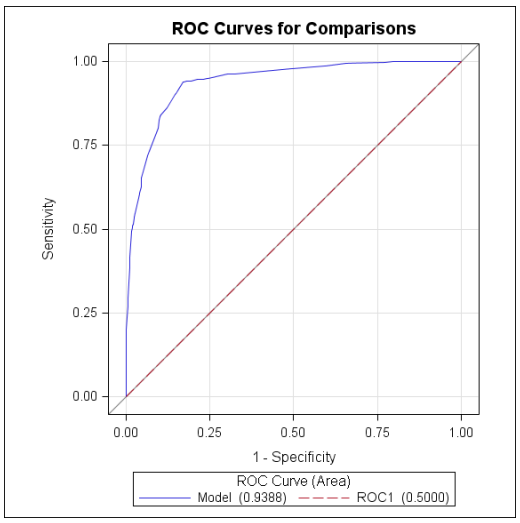

Figure 4.10 ROC curve comparison of step8 model to non-informative model

After eight steps' selection, we have our best model. Fig 4.11 shows ROC curves for all model building steps.

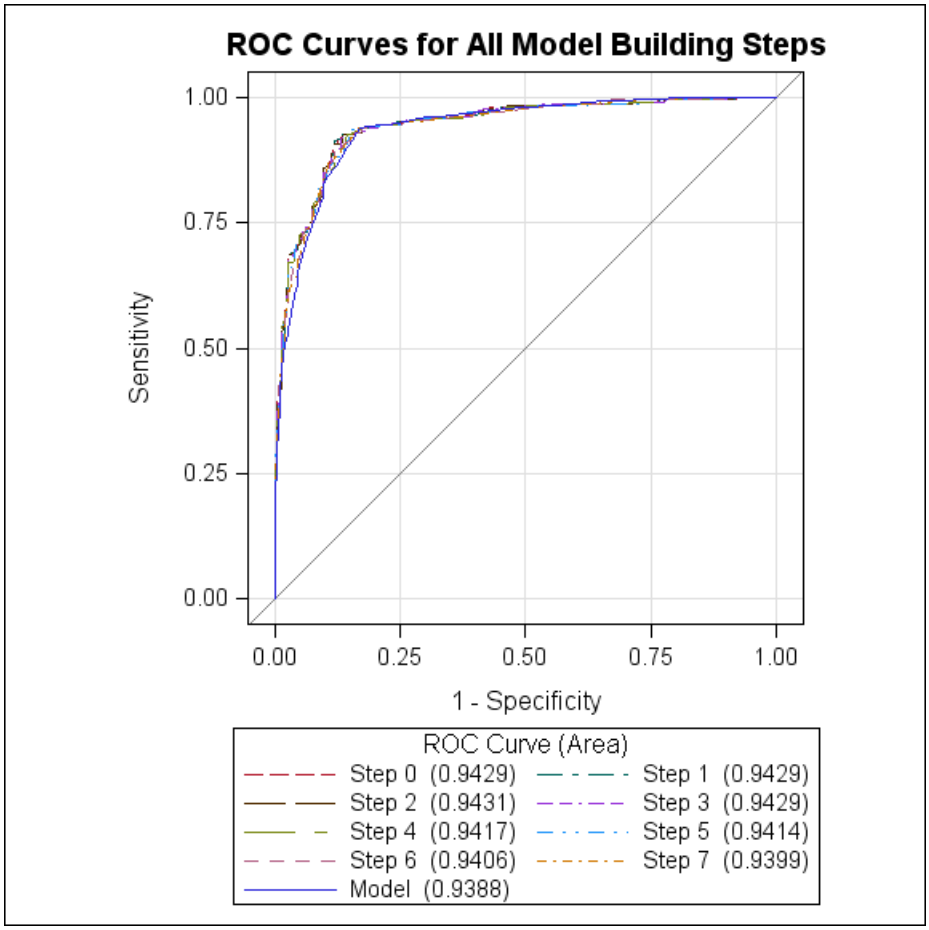

Figure 4.11 ROC curves for all model building steps

4.2 Cut-off Point Selection

After we have the best logistic regression model and ROC curve, the next question in practice is where should be the cutoff point to separate high risk, potential real patient from low risk, healthy client, and therefore which scanning test should they take, the complete screening test or the normal screening test.

There are two kinds of methods in deciding the cutoff point, statistical method and budget-limit method.

In statistical method, there are many options (Gonen, 2007; Lambert and Lipkovich, 2008):

1) choosing the point close to the perfect point
2) choosing the point far away from the non-informative line
3) choosing the point with the highest total accuracy
4) choosing the point with the highest Youden index
5) choosing the point with the highest Matthews correlation coefficient

In the ROC curve plot, the perfect point is (0, 1), which represents the 100% sensitivity and 100% specificity. A cutoff point close to this point has the balanced best sensitivity and the best specificity. With this criteria, the cutoff point is $\pi = 0.541981$. At this cutoff point, the sensitivity is 0.902027, and the specificity is 0.854342.

The non-informative line in ROC curve is the 45 degree diagonal line. The point with the largest distance from this line is $\pi = 0.504158$. At this cutoff point, the sensitivity is 0.939189, and the specificity is 0.829132.

The three other methods show the same cutoff point as far away from the non-informative line method, the cutoff point is $\pi = 0.504158$, with sensitivity as 0.939189 and specificity as 0.829132.

Though there are five methods in statistical area, we have only two cutoff points, because four methods gave the same result. These two cutoff points are close. There shouldn't be much difference between choosing either cutoff point in the point view of statistics. In practice, the disease we are studying is sexually-transmitted disease. It is more harmful to define a patient with disease as healthy person and let s/he go freely than to classify a normal person as patient

and have s/he take complete screening test. With the consideration of our project, we prefer the cutoff point with higher sensitivity (π = 0.504158, sensitivity=0.939189, specificity=0.829132).

For any institute or clinic, there is an annual budget limitation. Usually, the price of complete screening test is higher than that of normal screening test. If the budget is enough, we can have all clients to do complete screening test to minimize the false negative rate. If this is impossible, we need to choose a cutoff point with the possible highest sensitivity with the consideration of budget. To choose cutoff point according to budget, we need more information about the prices of complete screening test and normal screening test, the annual budget, and the estimated diseased patient number and the healthy client number for that year.

Chapter 5

Conclusion

In medical practice of sexually-transmitted diseases, it is important to correctly classify clients into high-risk patients and low-risk clients according to their very first survey result. Logistic regression model is the correct statistical tool for this kind of problems. Our full logistic regression model includes all the survey questions and can calculate the probability of a client has STD at high accuracy. The full logistic regression model is very complicated, and this may limit its usage in practice. To simplify this model, backward model selection was done with the criteria p=0.05. After eight steps of selection, which only includes five survey questions, our final logistic model is:

$$\log\left(\frac{\pi}{1-\pi}\right) = 11.0273$$
$$+\begin{cases} -1.161 & if\ education = 1 \\ 0.4537 & if\ education = 2 \\ -0.284 & if\ education = 3 \\ -1.010 & if\ education = 4 \\ 0.0000 & if\ education = 5 \end{cases}$$
$$+\begin{cases} -13.7062 & if\ likelihoodHIV = 1 \\ -14.4843 & if\ likelihoodHIV = 2 \\ 0.000000 & if\ likelihoodHIV = 3 \end{cases}$$
$$+\begin{cases} 3.8568 & if\ STDsymptoms = 0 \\ 0.0000 & if\ STDsymptoms = 1 \end{cases}$$
$$+\begin{cases} 1.3376 & if\ exchangeSex = 0 \\ 0.0000 & if\ exchangeSex = 1 \end{cases}$$
$$+\begin{cases} 1.5746 & if\ contactSTD = 0 \\ 0.0000 & if\ contactSTD = 1 \end{cases}$$

This final model has p value less than 0.0001 (same as full model) with 0.7131 as the max-rescaled R-square (0.7238 for full model).

The risk function for each client would be:

$$Ri = \pi_i = \frac{\exp(Mi)}{1+\exp(Mi)},$$

if we let the right part of model to M.

The ROC curve shows that our final logistic regression model is significantly better than a random guessing; the *p* value is less than 0.0001. The AUC of our model is 0.9388, which classifies the model to excellent level.

The cutoff point for practice can be either π = 0.541981 (according to close to perfect point rule, sensitivity=0.902027, specificity=0.854342), or π = 0.504158 (according to far-away from non-informative line rule, or highest total accuracy rule, or highest Youden index rule, or highest Matthews correlation coefficient rule, sensitivity=0.939189, specificity=0.829132), or according to budget.

REFERENCES

Forbes, Catherine., Evans, Merran., Hastings, Nicholas., and Peacock, Brian (2011). *Statistical Distributions, Fourth Edition*. Wiley Publication.

Gonen, Mithat (2007). *Analyzing Receiver Operating Characteristic Curves with SAS*. SAS publishing.

Hosmer, David W. and Lemeshow, Stanley (2000). *Applied Logistic Regression, Second edition*. Wiley Publication.

Lambert, Jennifer and Lipkovich, Ilya (2008). *A Macro For Getting More Out Of Your ROC Curve*. http://www2.sas.com/proceedings/forum2008/231-2008.pdf.

McCullagh, P., and Nelder, J.A (1989). *Generalized Linear Models, Second Edition*. Chapman and Hall Publication.

Metz, Charles (1978). *Basic Principles of ROC Analysis*. Seminars in Nuclear Medicine, 8(4): 283-298

Montgomery, Douglas C., Peck, Elizabeth A., and Vining, G. Geoffrey (2006). *Introduction to Linear Regression Analysis, Fourth Edition*. Wiley-Interscience Publication.

Nelder, J.A. and R.W.M Wedderburn (1972). *Generalized Linear Models*. Journal of the Royal Statistical Society, Series A, 135, 370-385

STDsurv2009-Complete, http://www.cdc.gov/std/stats09/surv2009-Complete.pdf

APPENDIX: SAS code

```
      *libname std 'E:';
      libname std 'C:\Users\Hui\Documents\thesis\New Folder\data';
      libname sel 'C:\Users\Hui\Documents\thesis\New Folder\data\selection';

      data std.data;
            infile 'C:\Users\Hui\Documents\thesis\New Folder\data\data.txt'
            DSD dlm=',';
            input A1a $ A2 A3 A4a $ A5a $ A6a $ A7a $ A8 A9a $ A10a $ A11
A12a $ A13a $ A14 A15 A16a $;
      run;

      ods graphics on;
      proc logistic data=std.data;
            class A1 A4 A6a A7a A9 A10 A11b A12 A13 A14b A15b / param=ref;
            model A16 (event='1') =A1 A2 A4 A6a A7a A8 A9 A10 A11b A12 A13
A14b A15b /ctable rsquare outroc=std.roc;
            roc;
            roccontrast;
      run;
      ods graphics off;

      *add 5 more accuracy parameters into std.roc: Distance to Perfect
Point(DtoPerfect), Distance to Non-informative line (DtoNoninf),
      Total Accuracy (TA), Youden Index (J), and Matthews Correlation Coeffi-
cient (MCC);
      data std.cutoff;
            set std.roc (obs=649);
            index=_n_;
            DtoPerfect=sqrt((_1mspec_)**2+(1-_sensit_)**2);
            DtoNoninf=sqrt((_sensit_-_1mspec_)**2/2);
            TA=(_pos_+_neg_)/(_pos_+_neg_+_falpos_+_falneg_);
            J=_sensit_-_1mspec_;
            MCC=(_pos_*_neg_-
_falpos_*_falneg_)/(sqrt((_pos_+_falneg_)*(_pos_+_falpos_)*(_neg_+_falpos_)*(
_neg_+_falneg_)));
      run;

      *model selection: backward selection;
      ods graphics on;
      proc logistic data=std.data;
            class A1 A4 A6a A7a A9 A10 A11b A12 A13 A14b A15b / param=ref;
            model A16 (event='1')=A1 A2 A4 A6a A7a A8 A9 A10 A11b A12 A13
A14b A15b
            /ctable rsquare selection=backward outroc=std.rocSelection;
            roc;
            roccontrast;
      run;
      ods graphics off;

      data std.cutoffSelection;
            set std.rocSelection;
            index=_n_;
            DtoPerfect=sqrt((_1mspec_)**2+(1-_sensit_)**2);
```

```
            DtoNoninf=sqrt((_sensit_-_1mspec_)**2/2);
            TA=(_pos_+_neg_)/(_pos_+_neg_+_falpos_+_falneg_);
            J=_sensit_-_1mspec_;
            MCC=(_pos_*_neg_-
_falpos_*_falneg_)/(sqrt((_pos_+_falneg_)*(_pos_+_falpos_)*(_neg_+_falpos_)*(
_neg_+_falneg_)));
      run;

      *model selection: step1--remove A13;
      ods html;
      ods graphics on;
      proc logistic data=std.data;
            class A1 A4 A6a A7a A9 A10 A11b A12 A13 A14b A15b / param=ref;
            model A16 (event='1')=A1 A2 A4 A6a A7a A8 A9 A10 A11b A12 A14b
A15b
            /ctable rsquare outroc=std.step1roc3;
            roc;
            roccontrast;
      run;
      ods graphics off;

      *model selection: step2--remove A2;
      ods html;
      ods graphics on;
      proc logistic data=std.data;
            class A1 A4 A6a A7a A9 A10 A11b A12 A13 A14b A15b / param=ref;
            model A16 (event='1')=A1 A4 A6a A7a A8 A9 A10 A11b A12 A14b A15b
            /ctable rsquare outroc=std.step2roc3;
            roc;
            roccontrast;
      run;
      ods graphics off;

      *model selection: step3--remove A1;
      ods html;
      ods graphics on;
      proc logistic data=std.data;
            class A1 A4 A6a A7a A9 A10 A11b A12 A13 A14b A15b / param=ref;
            model A16 (event='1')=A4 A6a A7a A8 A9 A10 A11b A12 A14b A15b
            /ctable rsquare outroc=std.step3roc3;
            roc;
            roccontrast;
      run;
      ods graphics off;

      *model selection: step4--remove A7a;
      ods html;
      ods graphics on;
      proc logistic data=std.data;
            class A1 A4 A6a A7a A9 A10 A11b A12 A13 A14b A15b / param=ref;
            model A16 (event='1')=A4 A6a A8 A9 A10 A11b A12 A14b A15b
            /ctable rsquare outroc=std.step4roc3;
            roc;
            roccontrast;
      run;
      ods graphics off;
```

```
*model selection: step5--remove A12;
ods html;
ods graphics on;
proc logistic data=std.data;
      class A1 A4 A6a A7a A9 A10 A11b A12 A13 A14b A15b / param=ref;
      model A16 (event='1')=A4 A6a A8 A9 A10 A11b A14b A15b
      /ctable rsquare outroc=std.step5roc3;
      roc;
      roccontrast;
run;
ods graphics off;

*model selection: step6--remove A12;
ods html;
ods graphics on;
proc logistic data=std.data;
      class A1 A4 A6a A7a A9 A10 A11b A12 A13 A14b A15b / param=ref;
      model A16 (event='1')=A4 A6a A8 A9 A10 A11b A15b
      /ctable rsquare outroc=std.step6roc3;
      roc;
      roccontrast;
run;
ods graphics off;

*model selection: step7--remove A8;
ods html;
ods graphics on;
proc logistic data=std.data;
      class A1 A4 A6a A7a A9 A10 A11b A12 A13 A14b A15b / param=ref;
      model A16 (event='1')=A4 A6a A9 A10 A11b A15b
      /ctable rsquare outroc=std.step7roc3;
      roc;
      roccontrast;
run;
ods graphics off;

*model selection: step8--remove A10;
ods html;
ods graphics on;
proc logistic data=std.data;
      class A1 A4 A6a A7a A9 A10 A11b A12 A13 A14b A15b / param=ref;
      model A16 (event='1')=A4 A6a A9 A11b A15b
      /ctable rsquare outroc=std.step8roc3;
      roc;
      roccontrast;
run;
ods graphics off;

*prepare step datasets for tables;
data sel.step0;
      set std.cutoff;
run;

data sel.step1;
      set sel.step1roc3 (obs=650);
      index=_n_;
      DtoPerfect=sqrt((_1mspec_)**2+(1-_sensit_)**2);
```

```
            DtoNoninf=sqrt((_sensit_-_1mspec_)**2/2);
            TA=(_pos_+_neg_)/(_pos_+_neg_+_falpos_+_falneg_);
            J=_sensit_-_1mspec_;
            MCC=(_pos_*_neg_-
_falpos_*_falneg_)/(sqrt((_pos_+_falneg_)*(_pos_+_falpos_)*(_neg_+_falpos_)*(
_neg_+_falneg_)));
      run;
      data sel.step2;
            set sel.step2roc3 (obs=619);
            index=_n_;
            DtoPerfect=sqrt((_1mspec_)**2+(1-_sensit_)**2);
            DtoNoninf=sqrt((_sensit_-_1mspec_)**2/2);
            TA=(_pos_+_neg_)/(_pos_+_neg_+_falpos_+_falneg_);
            J=_sensit_-_1mspec_;
            MCC=(_pos_*_neg_-
_falpos_*_falneg_)/(sqrt((_pos_+_falneg_)*(_pos_+_falpos_)*(_neg_+_falpos_)*(
_neg_+_falneg_)));
      run;
      data sel.step3;
            set sel.step3roc3 (obs=599);
            index=_n_;
            DtoPerfect=sqrt((_1mspec_)**2+(1-_sensit_)**2);
            DtoNoninf=sqrt((_sensit_-_1mspec_)**2/2);
            TA=(_pos_+_neg_)/(_pos_+_neg_+_falpos_+_falneg_);
            J=_sensit_-_1mspec_;
            MCC=(_pos_*_neg_-
_falpos_*_falneg_)/(sqrt((_pos_+_falneg_)*(_pos_+_falpos_)*(_neg_+_falpos_)*(
_neg_+_falneg_)));
      run;
      data sel.step4;
            set sel.step4roc3 (obs=592);
            index=_n_;
            DtoPerfect=sqrt((_1mspec_)**2+(1-_sensit_)**2);
            DtoNoninf=sqrt((_sensit_-_1mspec_)**2/2);
            TA=(_pos_+_neg_)/(_pos_+_neg_+_falpos_+_falneg_);
            J=_sensit_-_1mspec_;
            MCC=(_pos_*_neg_-
_falpos_*_falneg_)/(sqrt((_pos_+_falneg_)*(_pos_+_falpos_)*(_neg_+_falpos_)*(
_neg_+_falneg_)));
      run;
      data sel.step5;
            set sel.step5roc3 (obs=562);
            index=_n_;
            DtoPerfect=sqrt((_1mspec_)**2+(1-_sensit_)**2);
            DtoNoninf=sqrt((_sensit_-_1mspec_)**2/2);
            TA=(_pos_+_neg_)/(_pos_+_neg_+_falpos_+_falneg_);
            J=_sensit_-_1mspec_;
            MCC=(_pos_*_neg_-
_falpos_*_falneg_)/(sqrt((_pos_+_falneg_)*(_pos_+_falpos_)*(_neg_+_falpos_)*(
_neg_+_falneg_)));
      run;
      data sel.step6;
            set sel.step6roc3 (obs=519);
            index=_n_;
            DtoPerfect=sqrt((_1mspec_)**2+(1-_sensit_)**2);
            DtoNoninf=sqrt((_sensit_-_1mspec_)**2/2);
            TA=(_pos_+_neg_)/(_pos_+_neg_+_falpos_+_falneg_);
```

```
            J=_sensit_-_1mspec_;
            MCC=(_pos_*_neg_-
_falpos_*_falneg_)/(sqrt((_pos_+_falneg_)*(_pos_+_falpos_)*(_neg_+_falpos_)*(
_neg_+_falneg_)));
      run;
      data sel.step7;
            set sel.step7roc3 (obs=83);
            index=_n_;
            DtoPerfect=sqrt((_1mspec_)**2+(1-_sensit_)**2);
            DtoNoninf=sqrt((_sensit_-_1mspec_)**2/2);
            TA=(_pos_+_neg_)/(_pos_+_neg_+_falpos_+_falneg_);
            J=_sensit_-_1mspec_;
            MCC=(_pos_*_neg_-
_falpos_*_falneg_)/(sqrt((_pos_+_falneg_)*(_pos_+_falpos_)*(_neg_+_falpos_)*(
_neg_+_falneg_)));
      run;
      data sel.step8;
            set sel.step8roc3 (obs=59);
            index=_n_;
            DtoPerfect=sqrt((_1mspec_)**2+(1-_sensit_)**2);
            DtoNoninf=sqrt((_sensit_-_1mspec_)**2/2);
            TA=(_pos_+_neg_)/(_pos_+_neg_+_falpos_+_falneg_);
            J=_sensit_-_1mspec_;
            MCC=(_pos_*_neg_-
_falpos_*_falneg_)/(sqrt((_pos_+_falneg_)*(_pos_+_falpos_)*(_neg_+_falpos_)*(
_neg_+_falneg_)));
      run;

      data sel.step0short;
            set sel.step0 (where=(index in (1,15,29,43,57,71,85,99,113,127,
            141,155,169,183,197,211,225,239,253,267,281,295,309,323,337,351,
            365,379,393,407,421,435,449,463,477,491,505,519,533,547,561,575,
            589,603,617,631)));
      run;

      data sel.step1short;
            set sel.step1 (where=(index in (1,15,29,43,57,71,85,99,113,127,
            141,155,169,183,197,211,225,239,253,267,281,295,309,323,337,351,
            365,379,393,407,421,435,449,463,477,491,505,519,533,547,561,575,
            589,603,617,631)));
      run;

      data sel.step2short;
            set sel.step2 (where=(index in (1,14,27,40,53,66,79,92,105,118,
            131,144,157,170,183,196,209,222,235,248,261,274,287,300,313,326,
            339,352,365,378,391,404,417,430,443,456,469,482,495,508,521,534,
            547,560,573,586)));
      run;

      data sel.step3short;
            set sel.step3 (where=(index in (1,14,27,40,53,66,79,92,105,118,
            131,144,157,170,183,196,209,222,235,248,261,274,287,300,313,326,
            339,352,365,378,391,404,417,430,443,456,469,482,495,508,521,534,
            547,560,573,586)));
      run;

      data sel.step4short;
```

```
            set sel.step4 (where=(index in (1,14,27,40,53,66,79,92,105,118,
            131,144,157,170,183,196,209,222,235,248,261,274,287,300,313,326,
            339,352,365,378,391,404,417,430,443,456,469,482,495,508,521,534,
            547,560,573,586)));
      run;

      data sel.step5short;
            set sel.step5 (where=(index in
(1,13,25,37,49,61,73,85,97,109,121,
            133,145,157,169,181,193,205,217,229,241,253,265,277,289,301,313,
            325,337,349,361,373,385,397,409,421,433,445,457,469,481,493,505,
            517,529,541)));
      run;

      data sel.step6short;
            set sel.step6 (where=(index in
(1,12,23,34,45,56,67,78,89,100,111,
            122,133,144,155,166,177,188,199,210,221,232,243,254,265,276,287,
            298,309,320,331,342,353,364,375,386,397,408,419,430,441,452,463,
            474,485,496)));
      run;

      data sel.step7short;
            set sel.step7 (where=(index in
(1,3,5,6,8,10,12,14,15,17,19,21,23,
            24,26,28,30,32,33,35,37,39,41,42,44,46,48,50,51,53,55,57,59,60,62
,
            64,66,68,69,71,73,75,77,78,80,82)));
      run;

      data sel.step8short;
            set sel.step8 (where=(index in
(1,2,4,5,6,8,9,10,11,13,14,15,17,18,
            19,21,22,23,24,26,27,28,30,31,32,34,35,36,37,39,40,41,43,44,45,47
,
            48,49,50,52,53,54,56,57,58,59)));
      run;
```