Logistic Regression and Artificial Neural Networks for Classification of Ovarian Tumors

C. Lu[1], J. De Brabanter[1], S. Van Huffel[1], I. Vergote[2], D. Timmerman[2]
[1]Department of Electrical Engineering,
Katholieke Universiteit Leuven, Leuven, Belgium

[2]Department of Obstetrics and Gynecology,
University Hospitals Leuven, Leuven, Belgium

Technical Report
May 2001

Abstract

Ovarian masses are a common problem in gynaecology. A reliable test for preoperative discrimination between benign and malignant ovarian tumors is of considerable help for clinicians in choosing appropriate treatments for patients. This study was carried out to generate and evaluate both logistic regression models and artificial neural network (ANN) models to predict malignancy of ovarian tumors, using patient data collected at the University Hospitals of Leuven between 1994 and 1997. The first part of the report details the statistical analysis of the ovarian tumor dataset, including explorative univariate and multivariate analysis, and the development of the logistic regression models. The input variable selection was conducted via logistic regression as well. In the second part of the report, we describe the development of several types of feed forward neural networks such as multi-layer perceptrons (MLPs) and generalized regression networks (GRNNs). The issue of model validation is also addressed. Our adapted strategy for model evaluation is to perform Receiver Operating Charateristic (ROC) curve analysis, using both a temporal holdout cross validation (CV) and multiple runs of K-fold CV. The experiments confirm that neural network classifiers have the potential to give a more reliable prediction of the malignancy of ovarian tumors based on patient data.

Contents

PART I Statistical Analysis

0. Introduction …….. 1

0.1 Research question ……..… 1

0.2 Data Acquisition and feature selection ……..… 1

1. Data Exploration 3
1.1 Response variable 3
1.2 Explanatory variable 3
1.2.1 Univariate Analysis 4
1.2.2 Univariate Comparison between classes 22
1.3 Multivariate analysis 23
1.3.1 Principal Components 23
1.3.2 Factor Analysis 24
1.3.3 Discriminant Analysis 33
2 Logistic Regression 40
2.1 Odds Ratio and univariate logistic regression 40
2.2 Full Model 45
2.3 Variable Selection 47
2.3.1 Backward Elimination 48
2.3.2 Forward Selection 50
2.3.3 Stepwise Selection 51
2.3.4 Best Subset Selection 51
2.4 Model Fitting, Diagnosing and Classification 56
2.4.1 Model Fitting and Goodness-of-fit Test 56
2.4.2 Influence Measure and Diagnostics 58
2.4.3 Classification Table and ROC-curve 62
2.5 Model Validation 64
2.6 Conclusion 69

PART II Artificial Neural Network Modeling
3 Artificial Neural Network Models 71
3.1 Network Design and Training 73
3.1.1 Multi-layer Perceptrons 73
3.1.2 Generalized Regression Neural Networks 74
3.2 Simulation Results 75
4 Performance Measure and K-Fold Cross-Validation 77
4.1 ROC Analysis on different Subsets 77
4.2 ROC Analysis with K-fold Cross Validation 80
4.2.1 AUC from Cross-Validation 80
4.2.2 Combining ROC Curves 80
4.2.3 Experimental Results 81
5 Conclusions 85
5.1 Conclusions from This Study 85
5.2 Related Research and Publication 85
5.2 Future Work 86
References 87

0. Introduction

0.1 Research question

Adnexal masses are a common problem in gynaecology (occurrence: 1/70 women). This study was carried out to generate and evaluate both logistic regression models and artificial neural network (ANN) models to predict malignancy of adnexal masses in patients visiting the University Hospital of Leuven.

0.2 Data Acquisition and feature selection

The data were collected from 525 consecutive patients who were scheduled to undergo surgical investigation at the University Hospitals, Leuven. Table 0.1 lists the different indicators which were considered, together with their description [2][3]. 525 observations (patients) with 25 independent variables (measurements) are considered in this data set. The index variables (26, 27 and 28) are used for calculating the RMI (Risk of Malignancy Index) which is the index for predicting malignancy and which was developed by Jacobs et al. The outcome variables include the pathology result of the tumor, the expert's opinion and the staging of the tumor. For this study, we will take only the pathology result as the observed response to do the classification analysis.

Table 0.1 Description of Indicators

Type		Variable	Type	Indicator Description
Demographic	1	Age	continuous	
	2	meno[1]	Binary	menopausal status: (0 – premenopausal; 1 – postmenopausal)
serum marker	3	CA 125[2]	continuous	Serum CA 124 levels: the tumour marker with the highest sensitivity for ovarian cancer
Index obtained with color Doppler imaging Sonography	4	Col score	{1, 2, 3, 4}	subjective semiquantitive assessment of the amount of blood flow in (1- no blood flow; 2 - weak blood flow; 3 - normal blood flow; 4 - strong blood flow)
	5	PI	continuous	Pulsatility Index: (PI)=(S - D) / A. *S=the peak Doppler shifted frequency (PSV), *D=the minimum Doppler shifted frequency (or end-diastolic velocity), *A=the mean Doppler shifted frequency (or TAMX))
	6	RI	continuous	Resistance Index: (RI)=(S - D) / S
	7	PSV	continuous	Peak Doppler freq. PSV=S
	8	TAMX	continuous	Mean Doppler freq. TMAX=A

[1] The variable *meno* in the original data set is 1- premenopausal; 3-postmenopausal; 2-don't know. For computational reasons, variable meno is recoded as 0, 1.

[2] The variable ca 125 in the original dataset contains value '-1', we will take it as missing value. The same applies to the variables PI, RI, PSV, TAMX, Irreg.

B-mode Ultrasono Graphy	9	Asc	binary	Ascites (0 - absence; 1 - presence)
	10	Un	binary	Unilocular cyst (1 - yes)
	11	UnSol	binary	Unilocular solid (1 - yes)
	12	Mul	binary	Multilocular cyst (1 - yes)
	13	MulSol	binary	Multilocular solid (1 - yes)
	14	Sol	binary	Solid tumor (1 - yes)
Morphology	15	Bilat	binary	Bilateral mass (1 - yes)
	16	Smooth	binary	Smooth internal wall (1 – yes) *In cases of solid tumors, the description of the internal wall being Smooth or Irregular was usually not applicable, but the outline of the tumor is described as smooth or irregular.
	17	Irreg	binary	Irregular internal wall or outline of the tumor(1 – yes)
	18	Pap	binary	papillarities : (0 - <=3mm, 1 - >3mm)
	19	Sept	binary	Septa : (0 - <=3mm; 1 - >3mm)
	20	Shadows	binary	Acoustic shadows (1 - presence)
Echogenecity	21	Lucent	binary	Anechoic cystic content (1 - presence)
	22	Low level	binary	Low level echogenicity (1 - yes)
	23	Mixed	binary	Mixed echogenicity (1 - yes)
	24	G.Glass	binary	Ground glass cyst (1 - yes)
	25	Haem	binary	Hemorrhagic cyst (1 - yes)
Index	26	Morph	nominal	Asc + UnSol + Mul + 2*MulSol + Sol + Bilat
	27	Jacobs	nominal	0 - if Morph = 0; 1 - if 0 < Morph <= 1; 3 - if Morph > 1.
	28	RMI	continuous	Risk of malignancy index = (Jacobs * Meno * CA125 - if CA125 > 0; -1 - if CA125 <= 0)
Expert opinion	29	DT	binary	0 - benign; 1 - malignant. (Dirk Timmerman's opinion about the pathology)
Pathology result	30	Path	binary	Pathology result: 0 - benign; 1 – malignant.
Staging	31	outcome	{0, 1, 2, 3}	0 - benign; 1 - borderline; 2 - primary invasive; 3 - metastatic invasive. * 1, 2, 3 are all considered malignant.

1. Data Exploration

1.1 Response variable

The response or dependent variable is *PATH*; this is the pathology result of the tumor. Of the 525 observations, 141 cases are malignant, 384 are benign. The histogram of PATH is below.

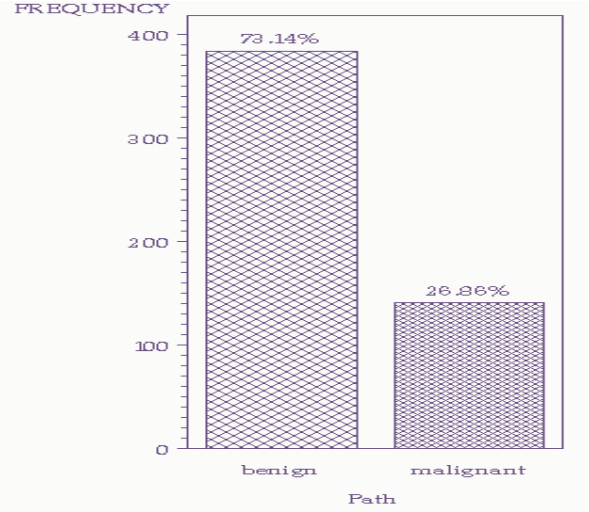


Fig. 1.1 Histogram of Pathology result

1.2 Explanatory variables

The 25 explanatory (or independent) variables can be divided into five classes.

1) Demographic: *Age*, Post Menopausal (*Meno*).
2) Tumor Marker: Serum CA 125 Levels (*CA_125*)
3) Color Doppler Imaging (CDI) Index: Amount of blood flow (*Color_Score*), and *PI, RI, PSV, TAMX*.
4) B-Mode morphologic: Ascites (*Asc*), Unilocular (*Un*), Unilocular Solid cyst (*UnSol*), Multilocular cyst (*Mul*), Multilocular Solid cyst (*MulSol*), Solid cyst (*Sol*), Bilateral mass (*Bilat*), Smooth internal wall (*Smooth*), Irregular internal wall (*Irreg*), Papillarities (>=3mm) (*Pap*), Septa (>3mm) (*Sept*), Acoustic Shadows (*Shadows*)
5) Echogenecity: Low level echogenecity (*Low_Level*), Mixed echogenicity (*Mixed*), Ground Glass Cyst (*G_Glass*), hemorrhagic cyst (*Haem*).

Age, Serum CA 125 level, and the 4 CDI index: PI, RI, PSV, TAMX are continuous variables. The Color Score about the amount of blood is categorical variable whose value can be integer from 1 to 4. The rest of the indicators are all binary variables that indicate if the specific character be present or not (if present, ‘1’ is assigned to it, otherwise ‘0’).

Before starting data exploration, we should notice that ideally some obvious constraints or interrelations exist among some variables, which implies that some variables can be represented as other variables' linear combination. Within a single observation,

1) Un, UnSol, Mul, MulSol, Sol must sum to 1, which means only one of the 5 characters can be present within a single observation;
2) If the Color_Score is 1, which means no blood flow, the other 4 CDI indices will then be unavailable, i.e. missing.
3) The four echogenecity variables Low_level, Mixed, G_glass, Haem should sum to one, which means only one of the characters should be present within one observation. However, we found some violation of this rule due to the difficulties in visual identification. But it will not be a serious problem in our study because they are found to be unimportant variables for prediction of malignancy.

1.2.1 Univariate Analysis:

We want to look at the distribution of the variables. For each variable, we'll give a list of descriptive statistics: the number of missing values, minimum and maximum value, mean, and standard deviation. The histogram is also shown. And for continuous variables, their boxplots will be presented as well.

1) Age

Variable	N	NMiss	Minimum	Maximum	Mean	Std Dev
Age	525	0	14.0000000	93.0000000	48.6038095	15.8742429

The total number of observations, N, in the data set is 525. *Nmiss* = 0 shows that no value is missing for variable *Age*. The *mean* age of the 525 patients is 48.6, the minimum value of the Age is 14, and the maximum is 93. The standard deviation, Std Dev of this variable is 15.87.

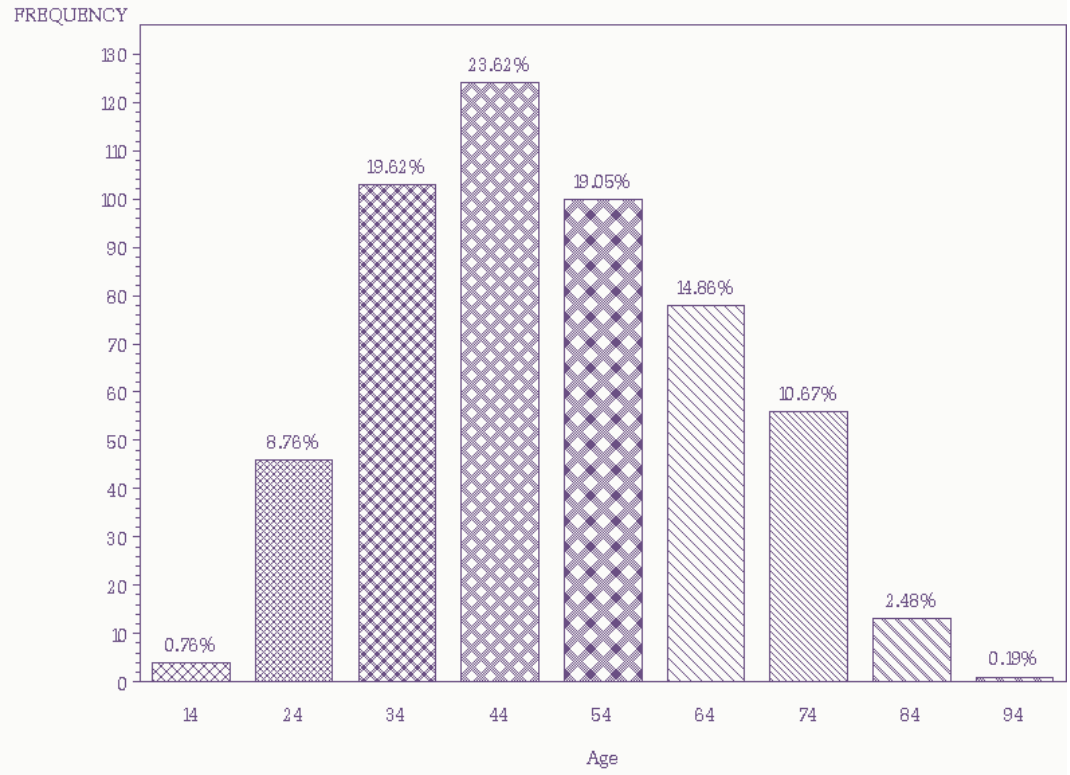


Fig.1.2 (a) Histogram of variable *Age*

The vertical axis on the histogram shows the frequency, which is the count of the observations in each bar. The horizontal axis gives the midpoint of variable 'Age' for each bar. The label above each bar is the percentage of observations of each bar over the whole data set.

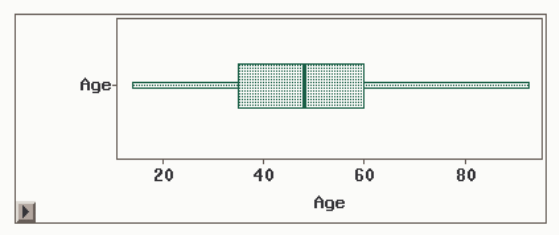


Fig 1.2 (b) Box-plot: variable Age

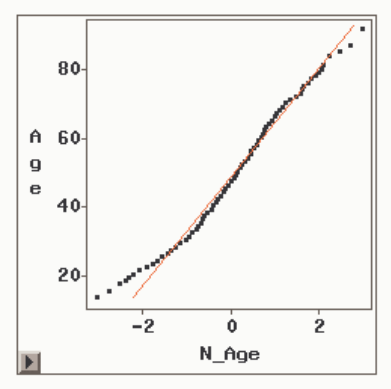


Fig.1.2(c) QQ-plot: variable Age.

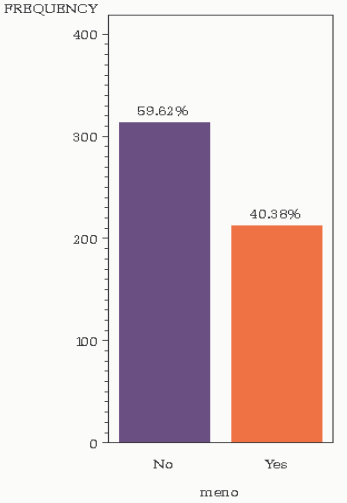
The left and right ends of the box plot indicate the 25th percentile and the 75th percentile, which is 35 and 60 respectively for variable *Age*. The length of the box is one interquartile range (i.e. the difference between the 75th percentile and the 25th percentile). The central horizontal lines are called whiskers. Each whisker extends up to 1.5 interquartile range from the end of the box. Values outside the whiskers are marked with a small bold rectangle. Values that are far away from the rest of the data, >3 interquartile range of the box, are called outliers. Fig 1.2.b shows that the variable *Age* contains no outliers.

The Quantile-Quantile plots (QQ-plots) aims to check if a random variable follows a particular distribution. Here we use a normal quantile plot to check the normality of the variable Age. The linearity between the variable quantile and its corresponding normal quantile with correlation 0.9917 suggests that the variable age follows a normal distribution.

2) Meno

Variable	N	NMiss	Minimum	Maximum	Mean	Std Dev
Meno	525	0	0	1.0000000	0.4038095	0.4911281

The ‘Meno’ variable contains no missing value. 212 patients out of the 525 ones are post menopausal which is represented as “1”.



3) CA_125:

Variable	N	NMiss	Minimum	Maximum	Mean	Std Dev
CA_125	432	93	1.0000000	31090.00	430.3078704	2145.46
L_CA125	432	93	0	10.3446415	3.7098912	1.800240

The total number of observations in the data set is 525, but 93 observations don't have the measurement of variable 'CA_125'. The minimum value for 'CA_125' is 1, and the maximum is 31090.0. The mean for CA_125 is 430.3. The histogram is shown below.

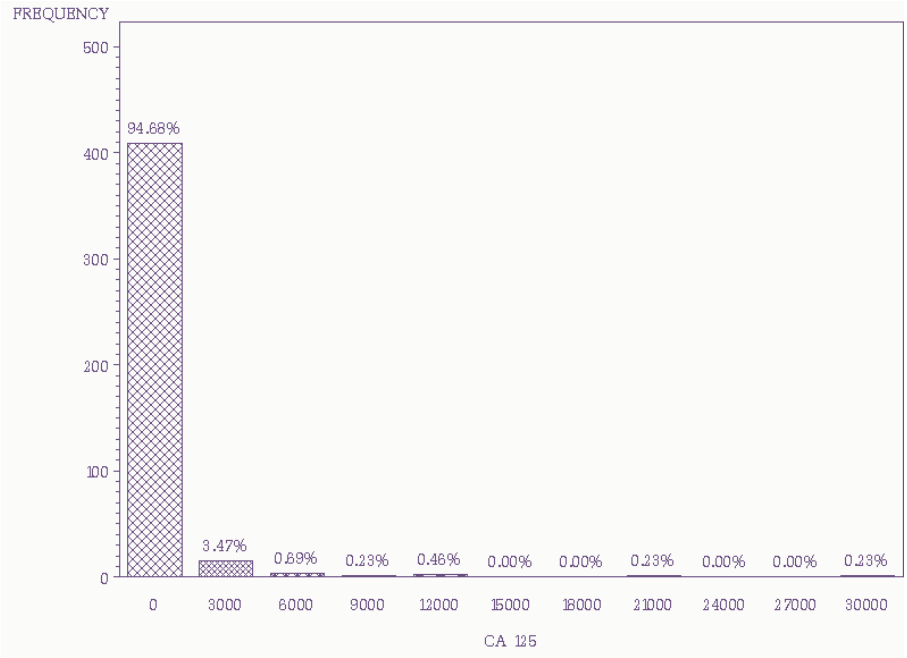


Fig.1.3 (a) Histogram of CA_125

Fig.1.3 (b) Box-Plot of CA_125

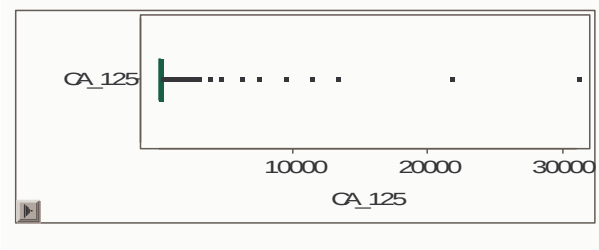
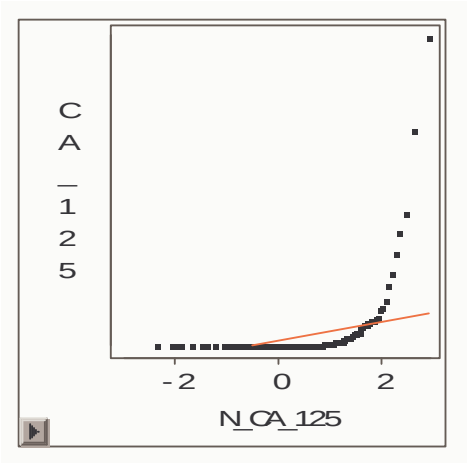


Fig.1.3(c) QQ-plot of CA_125



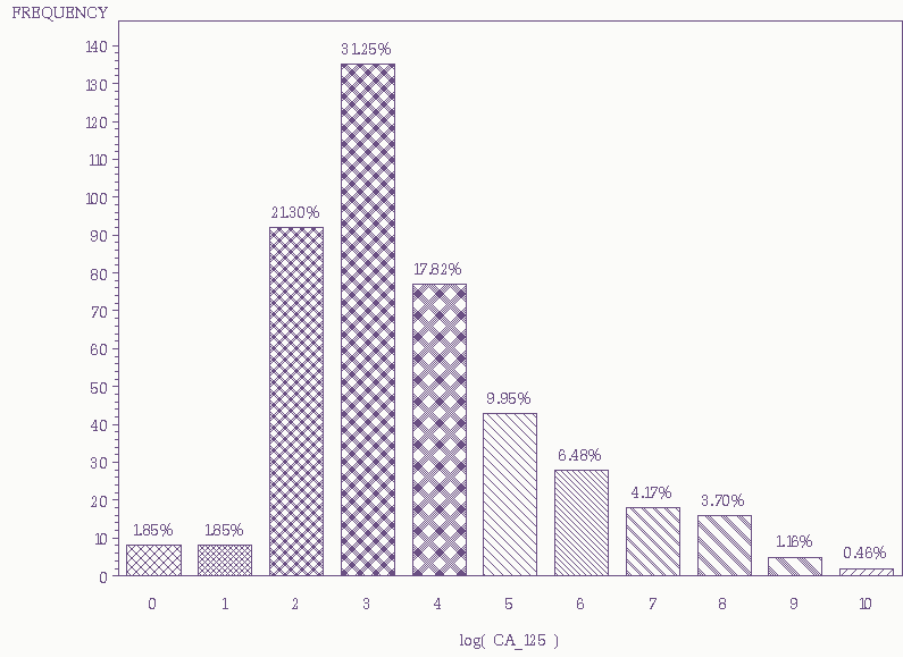


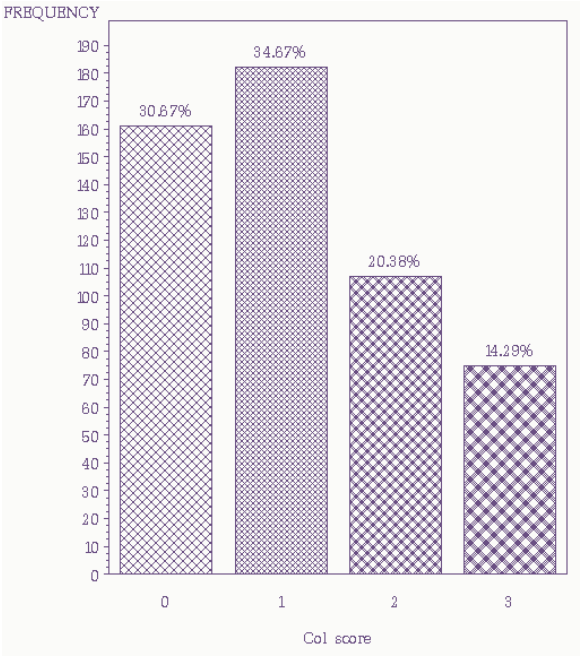
Fig.1.3(d) Histogram of l_ca125 (log(CA_125))

As almost all values were situated in a small interval between 0 and 100, and only a few cases took values up to 30,000, this variable was rescaled by taking its logarithm. The normal quantile plot shown in Fig1.3.c also indicates that the variable 'CA_125' does not follow a normal distribution. A new variable l_ca125 = log(CA_125) will be used in the following study instead of the original CA_125. The rescaled variable L_ca125 then falls into the interval of 0 to 10, with mean 3.7.

4) Color_Score

Variable	N	NMiss	Minimum	Maximum	Mean	Std Dev
Col_score	525	0	1.0000000	4.0000000	2.1828571	1.0249239

The nominal scaled variable color_score has been coded as 1=`no blood flow', 2='weak blood flow', 3='normal blood flow', or 4='strong blood flow'. The counts for each score value are 161, 182, 107 and 75 respectively. There is no missing value for this variable.



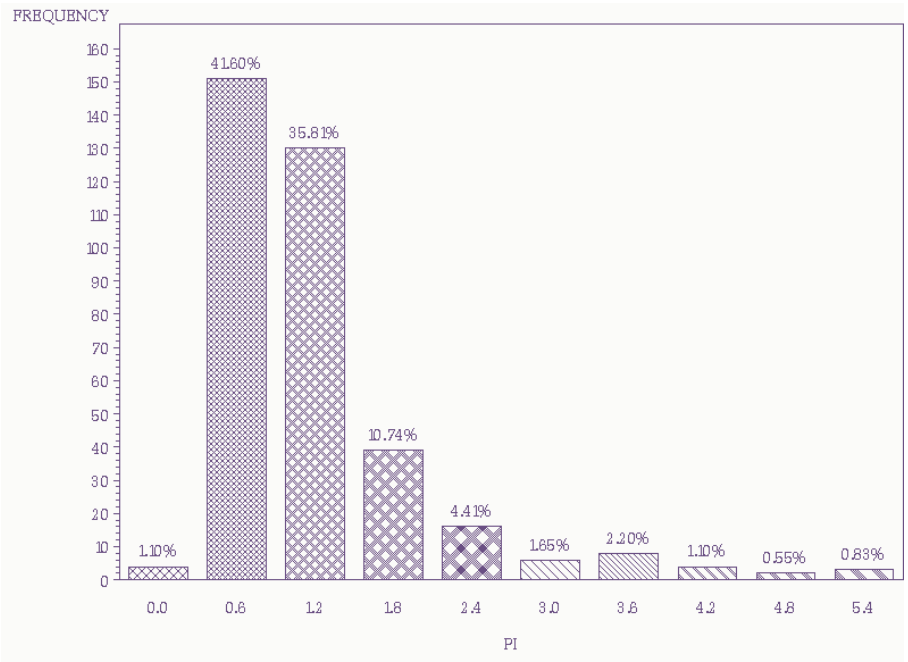
As a nominal variable is not suitable for our later analysis, Col_score has been transformed into 3 design (dummy) variables (colsc2, colsc3 and colsc4). Table illustrates the coding of the design variables.

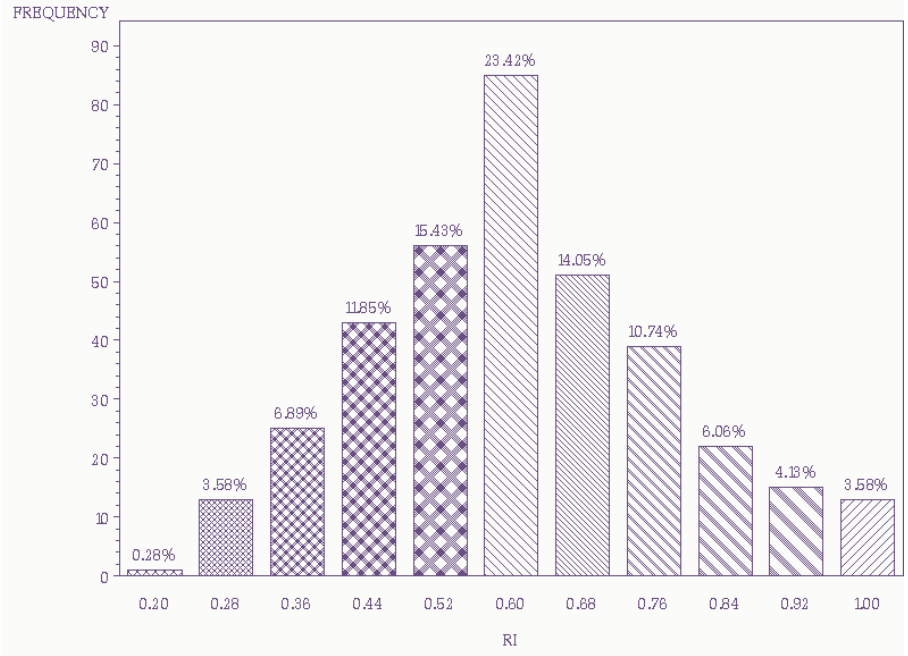
Col_score	Design Variable		
	Colsc2	Colsc3	Colsc4
1= no blood flow	0	0	0
2=weak blood flow	1	0	0
3=normal blood flow	0	1	0
4=strong blood flow	0	0	1

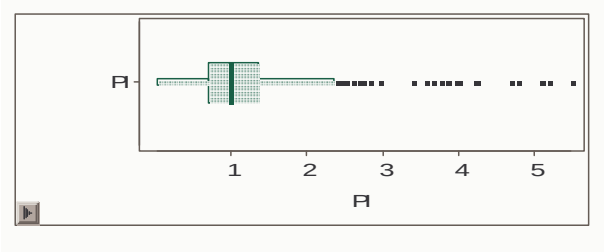
5) PI, RI, PSV and TAMX

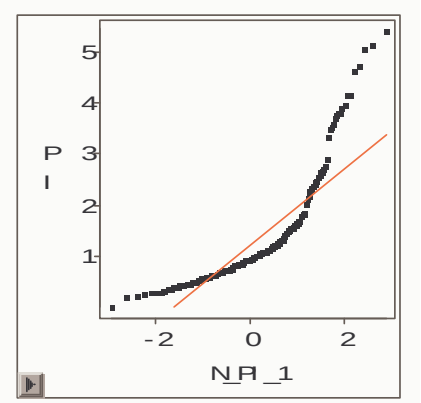
Variable	N	NMiss	Minimum	Maximum	Mean	Std Dev
pi	363	162	0.0200000	5.4900000	1.2052066	0.8567658
ri	363	162	0.2100000	1.0000000	0.6056474	0.1714250
psv	363	162	4.0000000	104.0000000	22.4655647	15.7083501
tamx	363	162	1.0000000	72.0000000	13.5041322	10.7475738

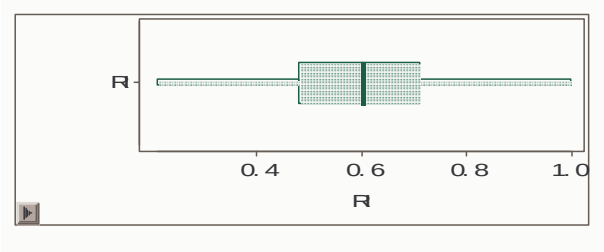
There are 162 missings for these 4 variables. Notice that all the corresponding color_score of these 162 observations are 1, which indicates no blood flow observed, hence the other CDI variables will also be unavailable.

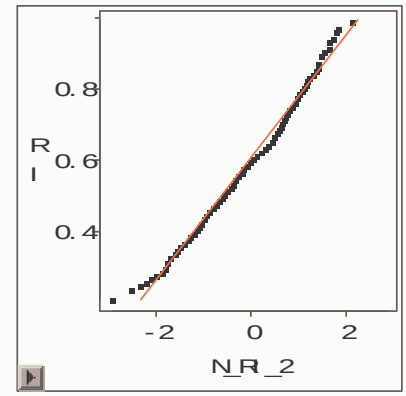




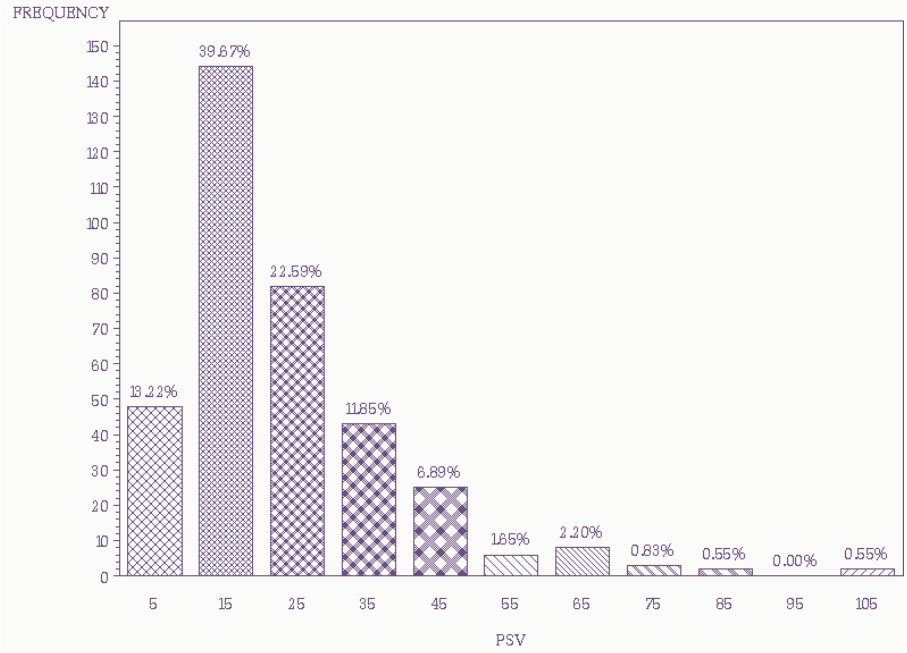


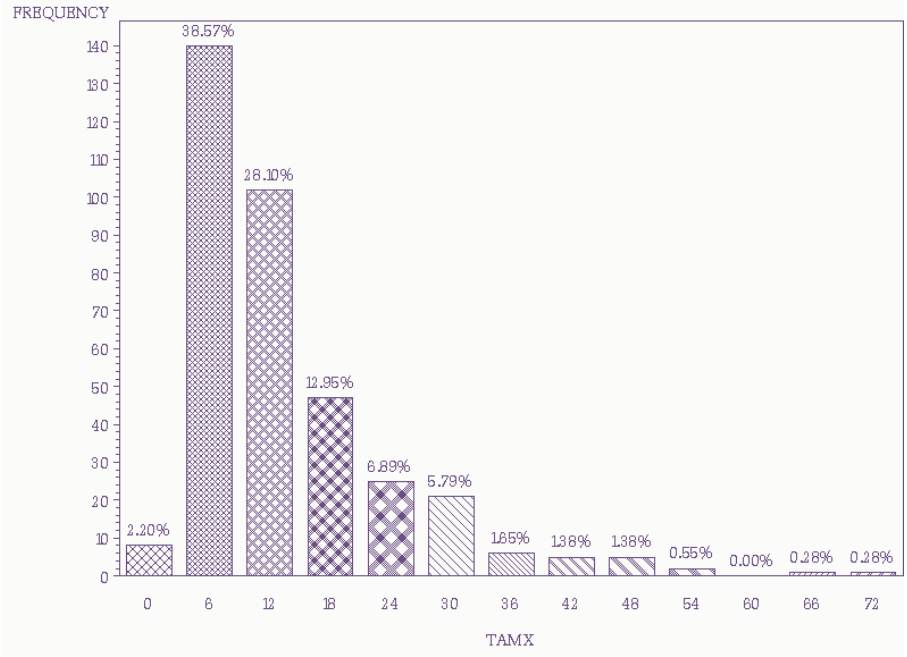


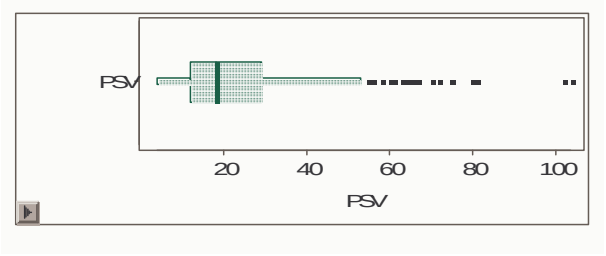


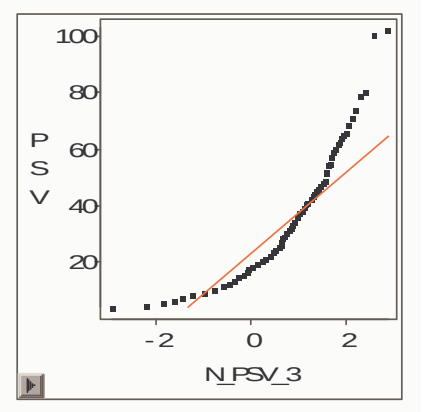


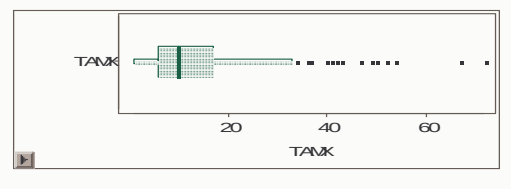
The box plot of PI shows the existence of some outliers with PI greater than 3. And PI doesn't follow a normal distribution as shown in the QQ-plot. Instead, RI is quite normal, which has the mean value 0.21 and maximum value 1.

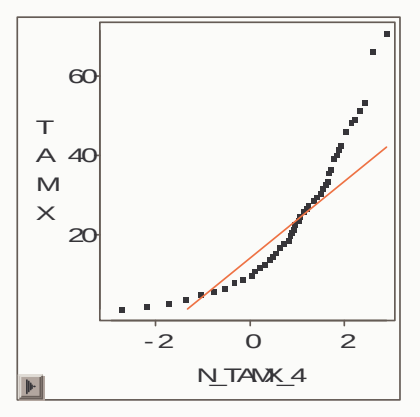

FREQUENCY
150
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
13.22%
39.67%
22.59%
11.85%
6.89%
1.65%
2.20%
0.83%
0.55%
0.00%
0.55%
5
15
25
35
45
55
65
75
85
95
105
PSV


FREQUENCY
140
130
120
110
100
90
80
70
60
50
40
30
20
10
0
2.20%
38.57%
28.10%
12.95%
6.89%
5.79%
1.65%
1.38%
1.38%
0.55%
0.00%
0.28%
0.28%
0
6
12
18
24
30
36
42
48
54
60
66
72
TAMX


PSV
20
40
60
80
100
PSV


100
80
P
S
V
60
40
20
-2
0
2
N_PSV_3



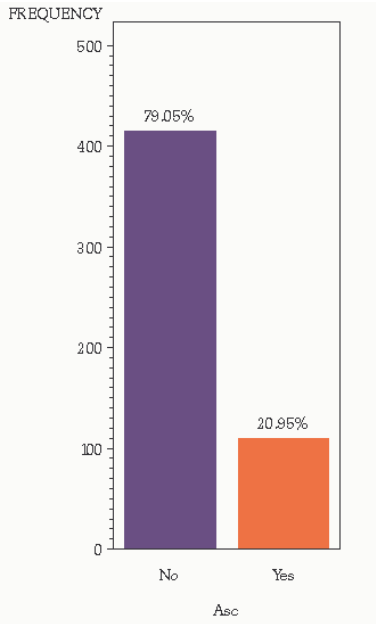


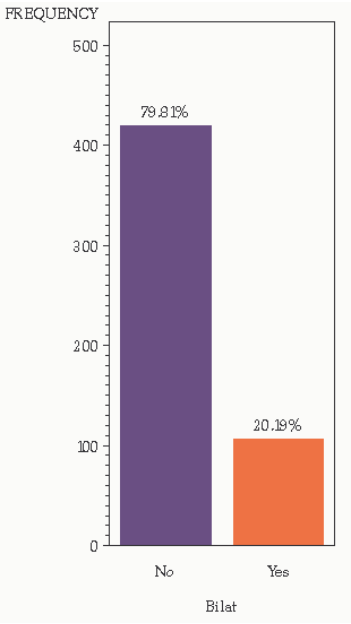
The box-plots of PSV and TAMX tell us the presence of the outliers. And the two normal quantile plots show that these two variables PSV and TAMX do not follow a normal distribution.

6) Asc and Bilat

Variable	N	NMiss	Minimum	Maximum	Mean	Std Dev
Asc	525	0	0	1.0000000	0.2095238	0.4073569
Bilat	525	0	0	1.0000000	0.2019048	0.4018044

There are no missing values for these two variables. Of the 525 patients, 21% had ascites; 20% had bilateral masses.

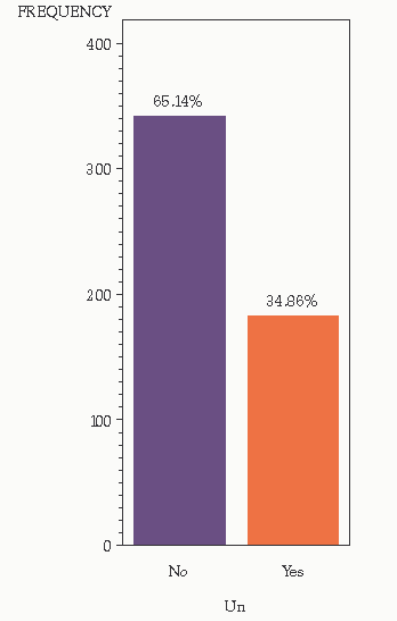


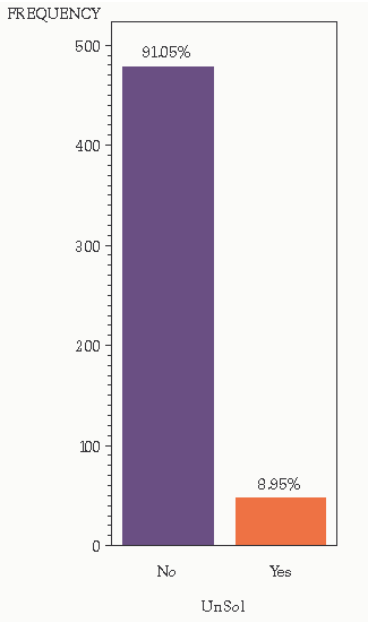


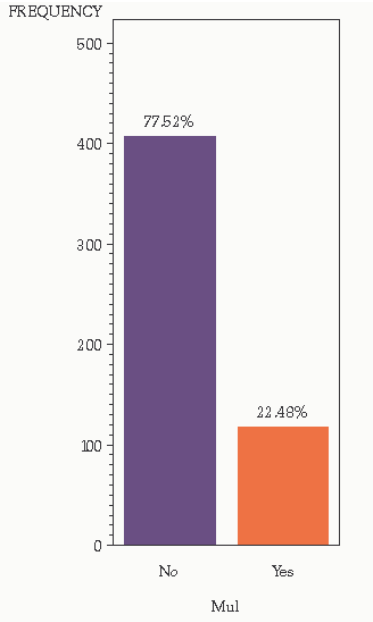
7) Un, UnSol, Mul, MulSol, and Sol:

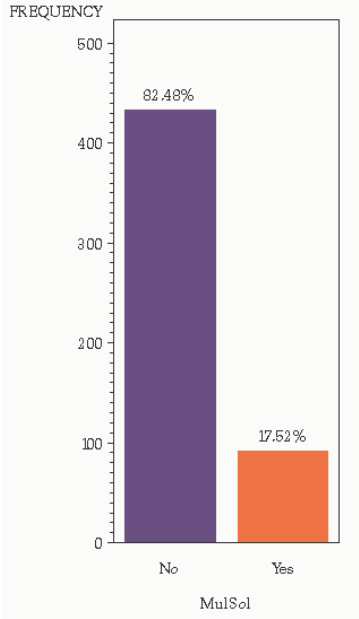
Variable	N	NMiss	Minimum	Maximum	Mean	Std Dev
Un	525	0	0	1.0000000	0.3485714	0.4769725
Unsold	525	0	0	1.0000000	0.0895238	0.2857706
Mul	525	0	0	1.0000000	0.2247619	0.4178235
Mulsol	525	0	0	1.0000000	0.1752381	0.3805332
Sol	525	0	0	1.0000000	0.1619048	0.3687147

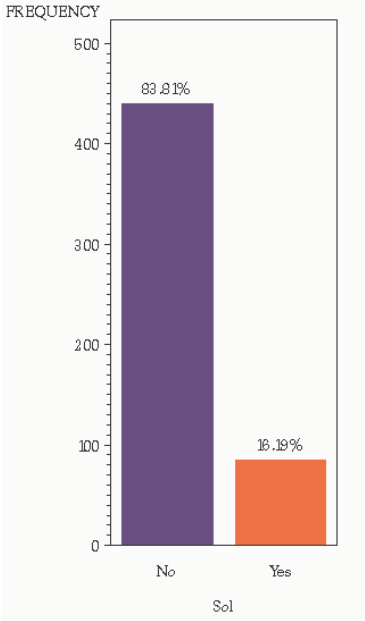
Around 35% of the patients have unilocular mass, 9% of the patients have unilocular solid mass, 22% have multiple locular masses, 18% have multiple locular solid masses, while only 16% have complete solid masses.







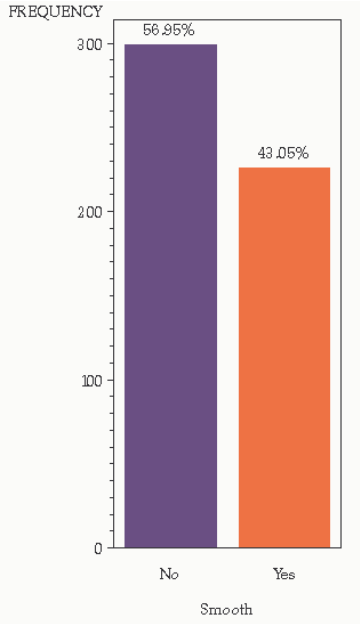

FREQUENCY
500
400
300
200
100
0
82.48%
17.52%
No
Yes
MulSol

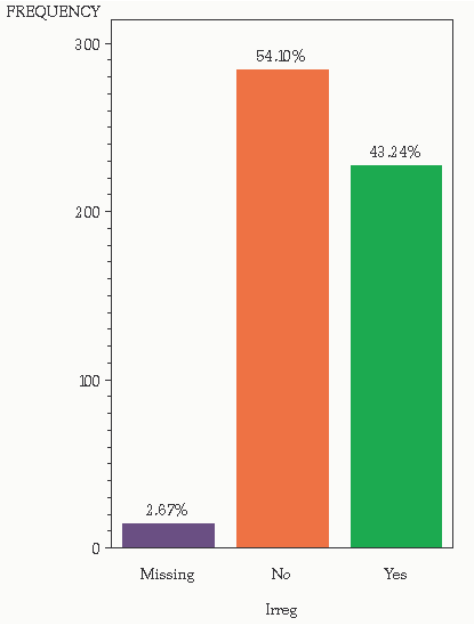

FREQUENCY
500
400
300
200
100
0
83.81%
16.19%
No
Yes
Sol

7) Smooth and Irreg

Variable	N	NMiss	Minimum	Maximum	Mean	Std Dev
Smooth	525	0	0	1.0000000	0.4304762	0.495615
Irreg	511	14	0	1.0000000	0.4442270	0.4973665

Among 525 observations, 43% have smooth internal walls. Concerning Irreg, 14 of which are missing, 43% have irregular internal walls. Also remarkable is the fact that in case of solid tumors, the description of Smooth or Irregular was usually not applicable.

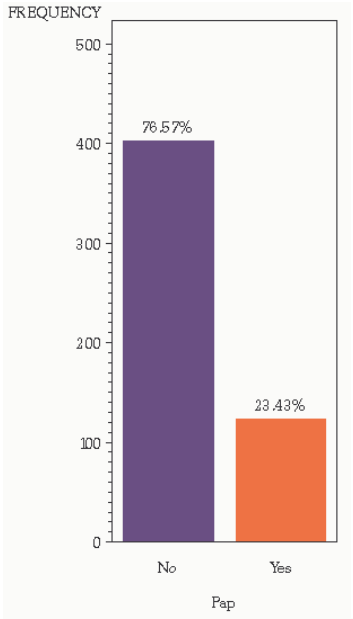


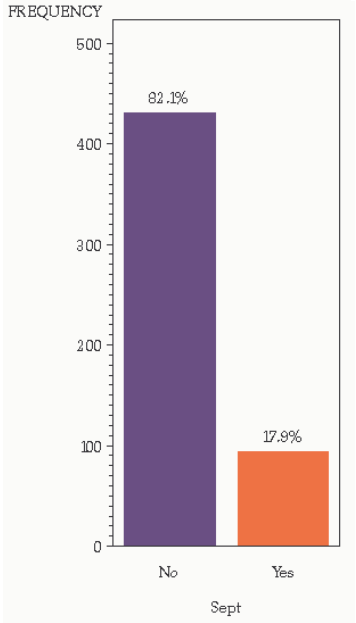


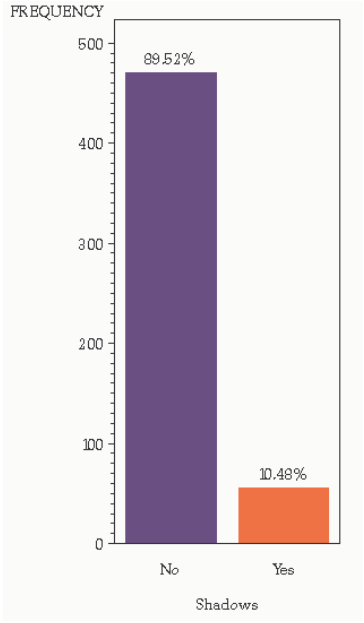
8) Pap, Sept, Shadows

Variable	N	NMiss	Minimum	Maximum	Mean	Std Dev
pap	525	0	0	1.0000000	0.2342857	0.4239555
sept	525	0	0	1.0000000	0.1790476	0.3837578
shadows	525	0	0	1.0000000	0.1047619	0.3065385

There is no missing value in these variables. 23% of the patients have papillarities >3mm, 18% have septa, 10.5% have acoustic shadows.



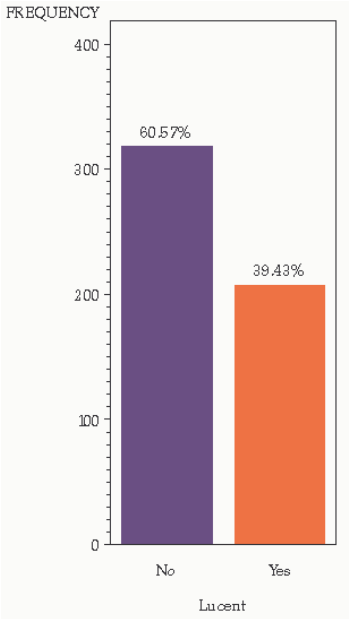


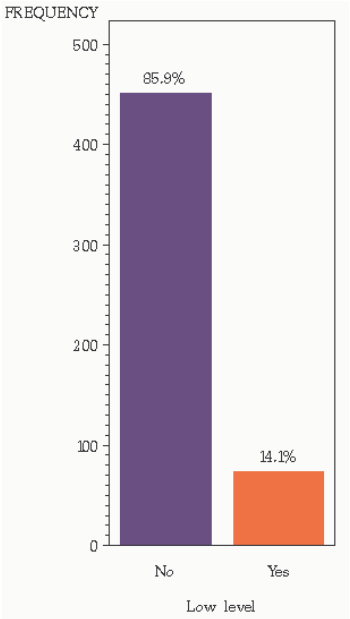


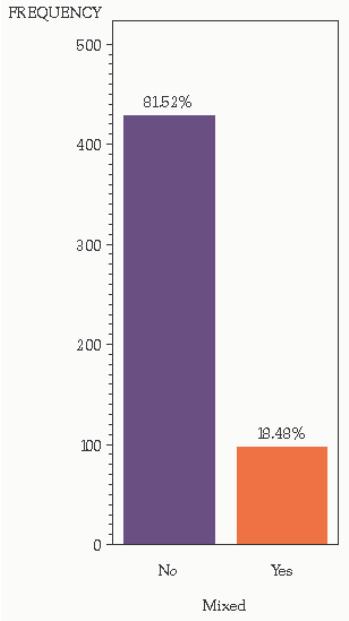
9) Lucent, Low_level, mixed, g_glass, haem

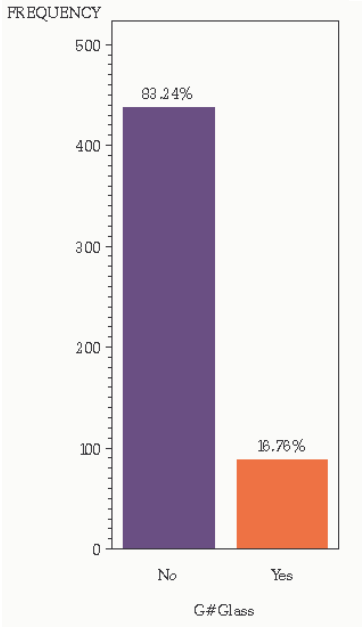
Variable	N	NMiss	Minimum	Maximum	Mean	Std Dev
lucent	525	0	0	1.0000000	0.3942857	0.4891628
low_level	525	0	0	1.0000000	0.1409524	0.3483043
mixed	525	0	0	1.0000000	0.1847619	0.3884744
g_glass	525	0	0	1.0000000	0.1676190	0.3738839
haem	525	0	0	1.0000000	0.0285714	0.1667575

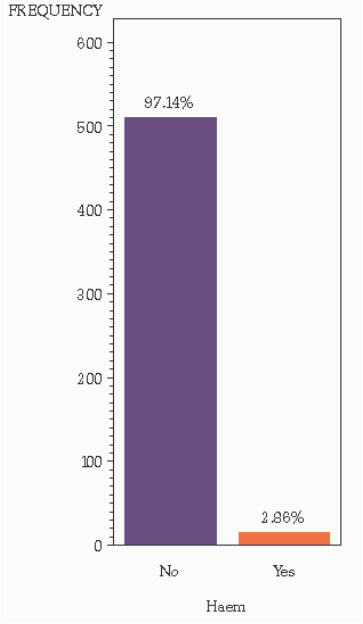
There are no missing values for these five echogenicity variables. 39% of the cases have Anechoic cystic contents which is indicated by *Lucent, 14*% have low level echogenicity, 18.5% have mixed echogenicity, 17% have ground glass cyst, only 3% of the cases have hemorrhagic cyst (which shows that *haem* is almost a constant variable).










FREQUENCY
600
500
400
300
200
100
0
97.14%
2.86%
No
Yes
Haem

1.2.2 Univariate Comparison between Two Classes

The means of the unpaired groups (benign versus malignant) in age were compared with the student t test. For the other continuous variable *pi, ri, psv, tamx, l_ca125*, which are not normally distributed, a nonparametric Wilcoxon Rank Sum test was performed. The proportions of benign and malignant cases with various morphologic or doppler features were compared with the continuity adjusted chi-square test. The descriptive statistics and comparison result are shown in the table below.

Variable	Benign Mass	Malignant Mass	Statistical Significance
Age (y, mean ± SD)	45.6 ± 15.2	56.9 ± 14.6	P < .0001 ***
Postmenopausal (%)	30.99	65.96	P < .0001 **
High col_score (3 or 4) (%)	19.01	77.30	P < .0001 *
PI (mean ± SD)	1.34 ± 0.94	0.96 ± 0.61	P < .0001 **
RI (mean ± SD)	0.64 ± 0.16	0.55 ± 0.17	P < .0001 **
PSV (mean ± SD)	19.8 ± 14.6	27.3 ± 16.6	P < .0001 **
TAMX (mean ± SD)	11.4 ± 9.7	17.4 ± 11.5	P < .0001 **
L_CA125 (mean ± SD)	3.03 ± 1.22	5.17 ± 1.50	P < .0001 **
Ascites (%)	32.73	67.27	P < .0001 *
Un (%)	45.83	4.96	P < .0001 *
UnSol (%)	6.51	15.60	P = .0022 *
Mul (%)	28.65	5.67	P < .0001 *
MulSol (%)	10.68	36.17	P < .0001 *
Sol (%)	8.33	37.59	P < .0001 *
Bilat (%)	13.28	39.01	P < .0001 *
Smooth (%)	56.77	5.67	P < .0001 *
Irreg (%)	33.78	73.19	P < .0001 *
Pap (%)	12.50	53.19	P < .0001 *
Sept (%)	13.02	31.21	P < .0001 *
Shadows (%)	12.24	5.67	P = .0437 *
Lucent (%)	43.23	29.08	P = .0045 *
Low_level (%)	11.98	19.86	P = .0309 *
Mixed (%)	20.31	13.48	P = .0965 *
G_glass (%)	19.79	8.51	P = .0033 *
Haem (%)	3.91	0.00	P = .0370 *

*** Based on student t-test ** Based on Wilcoxon Rank Sum test

* Based on Chi-Square tests of association

We find 19 variables which have a statistically significant (P<0.01) difference between benign and malignant cases. The *p*-values of the variables *Low_level*, *Mixed*, *Haem*, and *Shadows, Lucent, G_glass* indicate that, the difference between the benign and malignant cases in these variables are not significant at level 0.01. It suggests that these variables might not be important indicators for predicting the malignancy of the tumor.

1.3 Multivariate analysis

1.3.1 Principal Components

Principal component analysis seeks linear combinations of the variables with maximal (or minimal) variance. Note that the principal components depend on the scaling of the original variables, and this will be undesirable except perhaps if they are in comparable units. Otherwise, it's conventional to take the principal components of the correlation matrix, which implicitly rescales all the variables to have unit sample variance.
Applied principal component analysis is applied to a sample of 525 observations with 27 explanatory variables:

```
age meno colsc2 colsc3 colsc4 l_ca125 pi ri psv tamx
asc un unsol mul mulsol sol bilat smooth irreg pap sept shadows
lucent low_level mixed g_glass haem
```

The eigenvalues and eigenvectors of the correlation matrix are given below respectively.

Initial Factor Method: Principal Components

Prior Communality Estimates: ONE

Eigenvalues of the Correlation Matrix: Total = 27 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	*5.12610648*	*2.34196830*	*0.1899*	*0.1899*
2	*2.78413818*	*0.61071952*	*0.1031*	*0.2930*
3	*2.17341866*	*0.34628478*	*0.0805*	*0.3735*
4	*1.82713387*	*0.21575447*	*0.0677*	*0.4411*
5	*1.61137940*	*0.10263213*	*0.0597*	*0.5008*
6	*1.50874727*	*0.08558612*	*0.0559*	*0.5567*
7	*1.42316116*	*0.06241946*	*0.0527*	*0.6094*
8	*1.36074169*	*0.16787436*	*0.0504*	*0.6598*
9	*1.19286733*	*0.06304430*	*0.0442*	*0.7040*
10	*1.12982302*	*0.16684284*	*0.0418*	*0.7458*
11	0.96298018	0.06880220	0.0357	0.7815
12	0.89417798	0.02849912	0.0331	0.8146
13	0.86567886	0.10815735	0.0321	0.8467
14	0.75752152	0.09023446	0.0281	0.8747
15	0.66728706	0.10337641	0.0247	0.8995
16	0.56391065	0.02196817	0.0209	0.9203
17	0.54194248	0.10092260	0.0201	0.9404
18	0.44101988	0.05821993	0.0163	0.9567
19	0.38279995	0.09046001	0.0142	0.9709
20	0.29233994	0.11948662	0.0108	0.9817
21	0.17285332	0.06079733	0.0064	0.9881
22	0.11205598	0.00754526	0.0042	0.9923
23	0.10451073	0.01953160	0.0039	0.9962
24	0.08497913	0.06655387	0.0031	0.9993
25	0.01842526	0.01842526	0.0007	1.0000
26	0.00000000	0.00000000	0.0000	1.0000
27	0.00000000		0.0000	1.0000

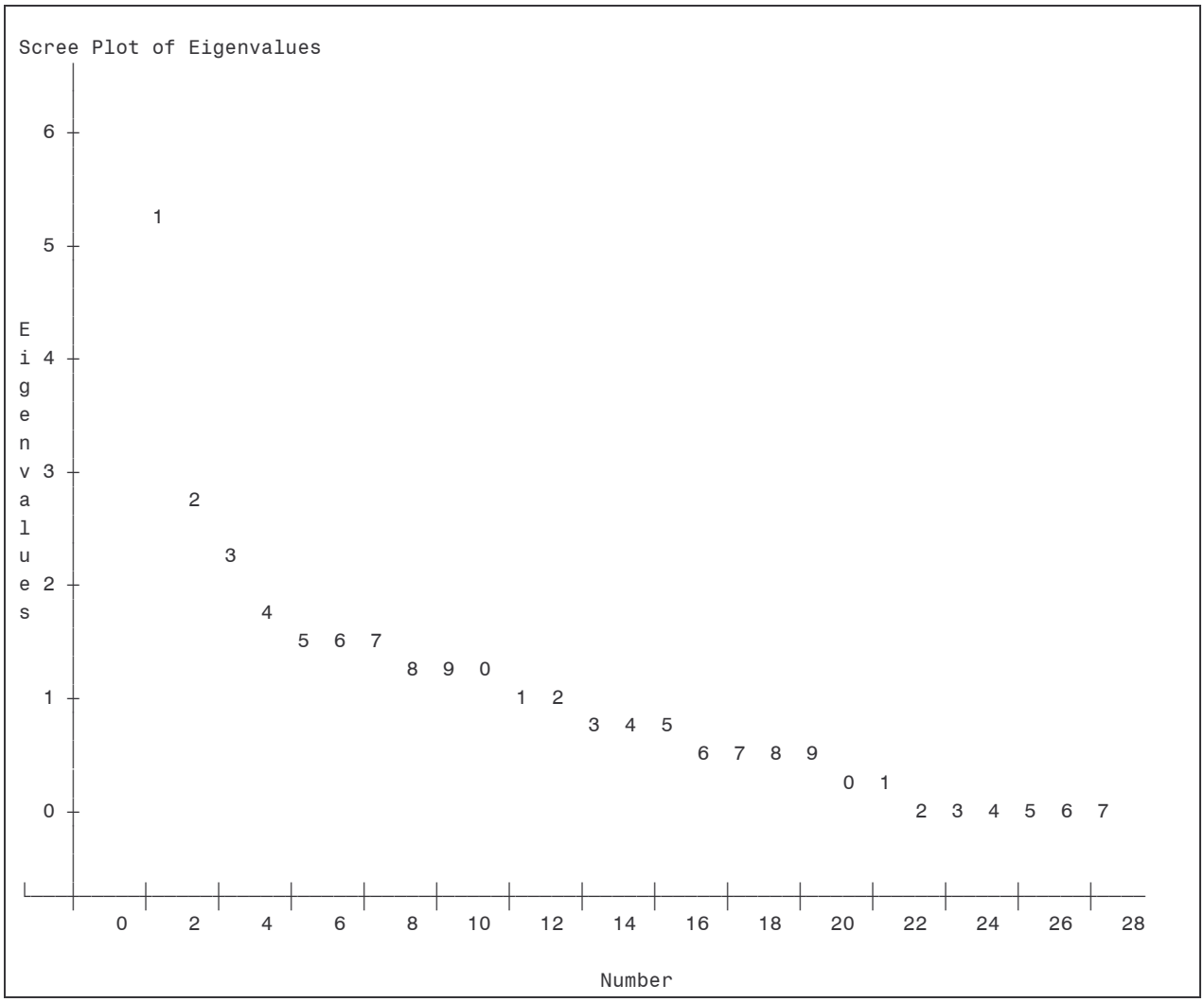
10 factors will be retained by the MINEIGEN criterion.

220 of 525 observations in data set omitted due to missing value

First have a look at the eigenvalues of the components. Since the eigenvalues serve as variance of the principal components, the last few princpal components have small variance; the last two eigenvalues are zeros, indicating that the components represent a linear relationship among the variables that is essentially constant. That is, there exist multicollinearities in the data set.

The last two zero eigenvalues may result in: (1) the sum of the five B-mode ultrasonography variables: un, unsold, mul, mulso and sol, is one; (2) the colsc_2, colsc_3 and colsc_4 are summed to one if the observations with color score 1('no blood flow') is omitted due to the missing value of the corresponding color Doppler imaging variables: pi, ri, psv and tamx. Applying the PCA using only 25 variables in which case 'colsc2' and 'sol' are excluded, we can see that the collinearities disappear and none of the 25 eigenvalues are zero.

In order to summarize the data effectively, according to Mineigen criterion, only the first 10 components with eigenvalues greater than 1 are kept. These components account for 74.58% of the total variance. The scree plot of Eigenvalues shows the PCA selection.



1.3.2 Factor Analysis

We will perform a factor analysis using the principal components as factors. The first 10 principal components are taken as the common factors and used to explain all the variables. And we will focus on the first two factors which contribute most to the discrimination between the benign and malignant tumors. In the following factor pattern table, we see that for the first two factors, only 18 variables (highlighted) have loadings greater than 0.4, while the rest of the variables can't be well explained with only these two variables.

Factor Pattern

	Factor1	Factor2	Factor3	Factor4	Factor5
age	0.33581	0.67597	0.21663	0.19244	0.01109
meno	0.36536	0.66974	0.20027	0.20588	0.05694
colsc2	-0.73504	0.05941	0.22454	-0.17516	0.37330
colsc3	0.19733	-0.13929	-0.06752	0.14937	-0.80249
colsc4	0.64749	0.08178	-0.18992	0.04301	0.43477
l_ca125	0.58621	0.24881	0.02590	-0.03112	0.01368
pi	-0.49319	0.28914	0.34559	0.17251	0.02455
ri	-0.54187	0.26111	0.30453	0.12661	-0.00885
psv	0.47840	-0.20305	-0.48756	0.29660	0.02023
tamx	0.57968	-0.26512	-0.51994	0.25925	0.04164
asc	0.50413	0.37859	-0.12848	-0.09239	-0.05816
un	-0.50891	-0.05948	-0.15843	-0.12732	0.00367
unsol	0.11491	-0.07329	0.20342	-0.50506	-0.37810
mul	-0.41834	-0.16535	0.05783	0.42807	-0.00503
mulsol	0.54830	-0.38258	0.29321	0.23311	0.21363
sol	0.24313	0.69307	-0.37770	-0.20000	0.05623
bilat	0.28198	0.23340	0.02856	0.16309	-0.10869
smooth	-0.64580	-0.10923	-0.31143	0.37186	-0.01304
irreg	0.53522	-0.29100	0.48547	-0.21652	-0.00890
pap	0.52652	-0.20315	0.49879	-0.11604	-0.12238
sept	0.43362	-0.43600	0.32474	0.39335	0.26479
shadows	0.06114	-0.10417	0.04701	-0.45713	0.30695
lucent	-0.14389	0.09999	0.28344	0.48622	-0.02160
low_level	0.16917	-0.27647	0.24276	-0.00967	-0.27064
mixed	0.06177	-0.31714	-0.12757	-0.26281	0.35118
g_glass	-0.24295	-0.35196	-0.10449	0.04157	0.04048
haem	-0.10362	-0.11550	-0.32102	-0.19315	-0.21771

	Factor6	Factor7	Factor8	Factor9	Factor10
age	0.12350	-0.03040	0.04911	-0.21562	-0.21484
meno	0.13443	-0.02240	-0.01642	-0.22746	-0.21660
colsc2	-0.09069	0.06729	0.04907	0.13994	-0.06653
colsc3	0.10756	-0.26277	-0.14640	-0.27518	0.24139
colsc4	-0.01060	0.20680	0.10164	0.13509	-0.18445
l_ca125	0.02043	0.18671	-0.06932	0.13533	0.28240
pi	0.56757	-0.14024	0.05389	0.19691	0.11536
ri	0.57593	-0.19628	0.02210	0.19729	0.18548

psv	0.39324	-0.21122	0.24162	0.29747	-0.05385
tamx	0.19455	-0.15212	0.24163	0.24451	-0.10724
asc	-0.10730	0.20362	0.05795	0.00107	0.33088
un	0.32188	0.60736	0.16557	-0.05009	0.03855
unsol	-0.06305	-0.14123	0.45991	0.18237	-0.07924
mul	-0.42195	-0.39745	-0.05887	0.11398	-0.15783
mulsol	0.20291	0.13638	-0.17398	-0.31582	0.16489
sol	-0.03039	-0.19438	-0.25186	0.12910	0.01192
bilat	-0.19628	0.19319	-0.07532	0.18243	0.41703
smooth	-0.00804	0.19231	0.09005	-0.19917	0.09006
irreg	0.01282	-0.05906	0.11577	0.22192	-0.05746
pap	0.04207	0.17349	0.18040	0.10224	0.06073
sept	0.01418	0.02869	-0.08969	-0.14515	0.05602
shadows	0.01797	-0.33915	0.03370	-0.14527	0.23816
lucent	-0.30203	0.05853	0.64296	0.00497	0.13283
low_level	0.20316	0.19654	-0.40563	0.11669	-0.48218
mixed	0.23323	-0.36056	0.11107	-0.38987	0.17600
g_glass	0.02077	0.14618	-0.36005	0.39479	0.27975
haem	0.20290	0.28774	0.26578	-0.27443	-0.14299

Variance Explained by Each Factor

Factor1	Factor2	Factor3	Factor4	Factor5
5.1261065	2.7841382	2.1734187	1.8271339	1.6113794

Factor6	Factor7	Factor8	Factor9	Factor10
1.5087473	1.4231612	1.3607417	1.1928673	1.1298230

Final Communality Estimates: Total = 20.137517

age	meno	colsc2	colsc3	colsc4	l_ca125	pi
0.76502366	0.78526925	0.80343456	0.96524162	0.75835273	0.54551906	0.87342352

ri	psv	tamx	asc	un	unsol	mul
0.91468026	0.94521015	0.93627078	0.59171044	0.80774729	0.73298315	0.76634398

mulsol	sol	bilat	smooth	irreg	pap	sept
0.84992220	0.84422645	0.46192986	0.75736314	0.72338778	0.67427501	0.74169891

shadows	lucent	low_level	mixed	g_glass	haem
0.51428672	0.87363007	0.72788062	0.69277265	0.58273452	0.50219869

Correlations

	age	meno	colsc2	colsc3	colsc4	l_ca125	pi
age	1.00000	0.82311	-0.17321	-0.00167	0.20496	0.26280	0.06810
meno	0.82311	1.00000	-0.19840	-0.01307	0.24689	0.26197	0.07380
colsc2	-0.17321	-0.19840	1.00000	-0.61077	-0.50868	-0.37855	0.32893
colsc3	-0.00167	-0.01307	-0.61077	1.00000	-0.37103	0.03757	-0.10129
colsc4	0.20496	0.24689	-0.50868	-0.37103	1.00000	0.40311	-0.27563
l_ca125	0.26280	0.26197	-0.37855	0.03757	0.40311	1.00000	-0.17842
pi	0.06810	0.07380	0.32893	-0.10129	-0.27563	-0.17842	1.00000
ri	0.04846	0.01243	0.38504	-0.05848	-0.38799	-0.16367	0.87268
psv	0.01597	0.02973	-0.42957	0.14777	0.34311	0.16726	-0.14073
tamx	0.00010	0.01339	-0.48232	0.12625	0.42837	0.19646	-0.35623
asc	0.25199	0.25124	-0.32671	0.05577	0.32252	0.47786	-0.18627
un	-0.20864	-0.23204	0.34134	-0.16531	-0.22057	-0.24097	0.20900
unsol	-0.00112	-0.04884	-0.05550	0.09587	-0.03915	0.04264	-0.08151
mul	-0.16664	-0.19999	0.24995	-0.02992	-0.26060	-0.27119	0.11376
mulsol	0.06426	0.11704	-0.34567	0.12403	0.27053	0.22253	-0.19561
sol	0.30241	0.33852	-0.17077	-0.01791	0.21975	0.24014	-0.04503
bilat	0.15115	0.16253	-0.17761	0.06489	0.13773	0.25933	-0.10182
smooth	-0.22980	-0.23043	0.30021	-0.05171	-0.29585	-0.28936	0.18834
irreg	0.00821	0.00277	-0.24229	0.04808	0.23188	0.19047	-0.14522
pap	0.14319	0.11637	-0.26580	0.09805	0.20511	0.34921	-0.16319
sept	-0.02175	0.01991	-0.26129	0.03784	0.26529	0.12174	-0.15586
shadows	-0.11966	-0.06618	0.07487	-0.08958	0.00959	-0.00608	-0.06177
lucent	0.07765	0.05556	0.11149	-0.04719	-0.07944	-0.09908	0.15075
low_level	-0.05640	-0.02082	-0.13476	0.11279	0.03541	0.04794	-0.03549
mixed	-0.12247	-0.14625	-0.01825	-0.03603	0.06057	-0.02722	-0.08063
g_glass	-0.29980	-0.29033	0.14605	-0.02704	-0.14188	-0.10346	0.01457
haem	-0.12613	-0.13814	-0.03577	0.06131	-0.02471	-0.10814	-0.03617

	ri	psv	tamx	asc	un	unsol	mul
age	0.04846	0.01597	0.00010	0.25199	-0.20864	-0.00112	-0.16664
meno	0.01243	0.02973	0.01339	0.25124	-0.23204	-0.04884	-0.19999
colsc2	0.38504	-0.42957	-0.48232	-0.32671	0.34134	-0.05550	0.24995
colsc3	-0.05848	0.14777	0.12625	0.05577	-0.16531	0.09587	-0.02992
colsc4	-0.38799	0.34311	0.42837	0.32252	-0.22057	-0.03915	-0.26060
l_ca125	-0.16367	0.16726	0.19646	0.47786	-0.24097	0.04264	-0.27119
pi	0.87268	-0.14073	-0.35623	-0.18627	0.20900	-0.08151	0.11376
ri	1.00000	-0.11744	-0.36689	-0.22377	0.21863	-0.05443	0.12921
psv	-0.11744	1.00000	0.94549	0.11306	-0.15362	-0.02521	-0.14085
tamx	-0.36689	0.94549	1.00000	0.15392	-0.20912	0.00183	-0.14457
asc	-0.22377	0.11306	0.15392	1.00000	-0.20728	0.06266	-0.31284
un	0.21863	-0.15362	-0.20912	-0.20728	1.00000	-0.16059	-0.26547
unsol	-0.05443	-0.02521	0.00183	0.06266	-0.16059	1.00000	-0.19378
mul	0.12921	-0.14085	-0.14457	-0.31284	-0.26547	-0.19378	1.00000
mulsol	-0.23386	0.19365	0.24056	0.11438	-0.27731	-0.20242	-0.33463
sol	-0.04960	0.10337	0.08894	0.35282	-0.24410	-0.17817	-0.29455
bilat	-0.08570	0.05584	0.07044	0.28357	-0.12757	-0.06052	-0.06326
smooth	0.21468	-0.11139	-0.14670	-0.25627	0.40863	-0.17465	0.32833
irreg	-0.19639	0.11385	0.16658	0.11586	-0.25680	0.23806	-0.15416
pap	-0.19229	0.08505	0.12029	0.13957	-0.17099	0.32086	-0.25184
sept	-0.19117	0.18265	0.22811	0.02715	-0.27968	-0.20415	0.02765
shadows	-0.02902	-0.02246	-0.01101	-0.02687	-0.10540	0.13836	-0.13700

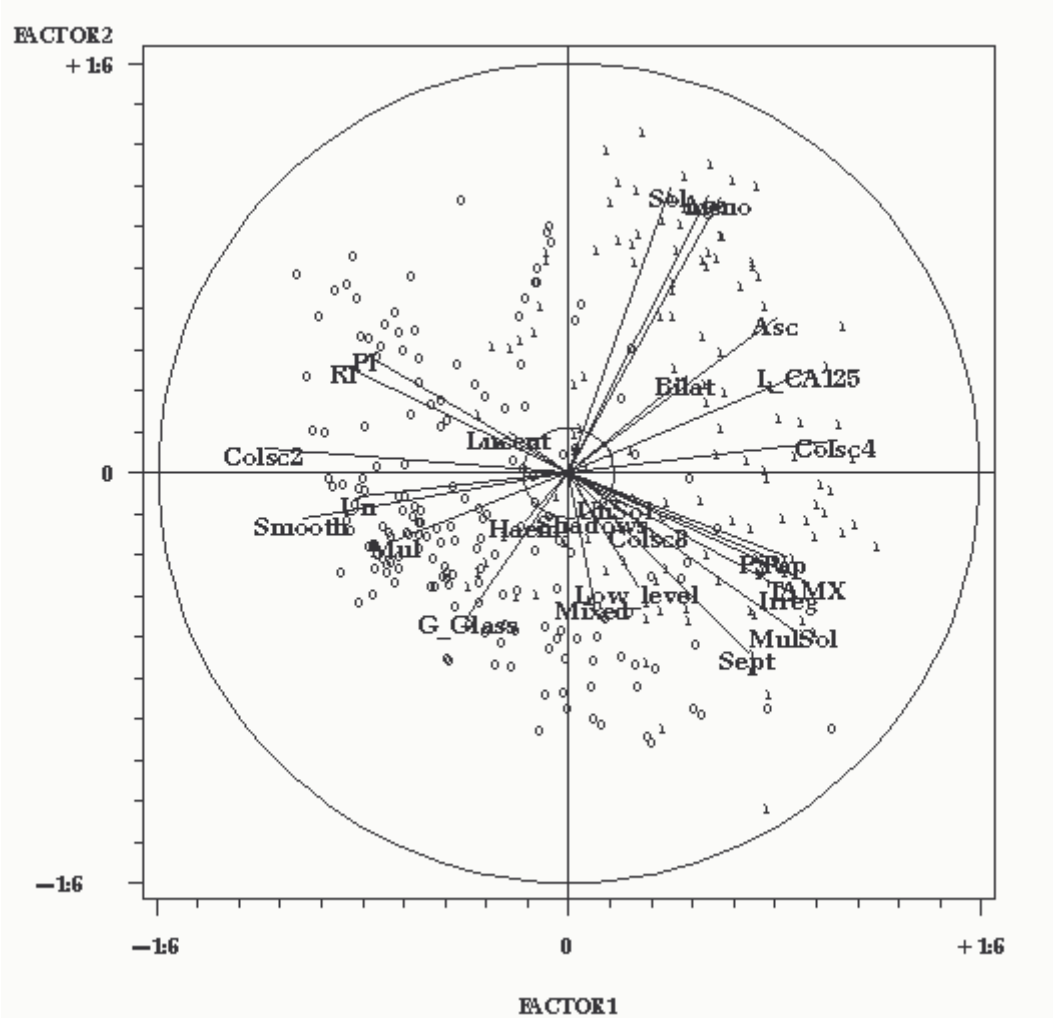
lucent	0.09155	-0.06444	-0.05250	-0.02256	0.05050	0.03190	0.21932
low_level	-0.06015	0.03849	0.04882	-0.09923	-0.03123	0.07238	-0.03287
mixed	-0.02215	0.13365	0.12175	-0.07596	-0.04255	0.06204	-0.02674
g_glass	0.06568	0.00131	-0.02099	-0.18683	0.16087	-0.06491	0.12125
haem	-0.03396	0.05814	0.03726	0.03057	0.27518	0.04854	-0.06846

	mulsol	sol	bilat	smooth	irreg	pap	sept
age	0.06426	0.30241	0.15115	-0.22980	0.00821	0.14319	-0.02175
meno	0.11704	0.33852	0.16253	-0.23043	0.00277	0.11637	0.01991
colsc2	-0.34567	-0.17077	-0.17761	0.30021	-0.24229	-0.26580	-0.26129
colsc3	0.12403	-0.01791	0.06489	-0.05171	0.04808	0.09805	0.03784
colsc4	0.27053	0.21975	0.13773	-0.29585	0.23188	0.20511	0.26529
l_ca125	0.22253	0.24014	0.25933	-0.28936	0.19047	0.34921	0.12174
pi	-0.19561	-0.04503	-0.10182	0.18834	-0.14522	-0.16319	-0.15586
ri	-0.23386	-0.04960	-0.08570	0.21468	-0.19639	-0.19229	-0.19117
psv	0.19365	0.10337	0.05584	-0.11139	0.11385	0.08505	0.18265
tamx	0.24056	0.08894	0.07044	-0.14670	0.16658	0.12029	0.22811
asc	0.11438	0.35282	0.28357	-0.25627	0.11586	0.13957	0.02715
un	-0.27731	-0.24410	-0.12757	0.40863	-0.25680	-0.17099	-0.27968
unsol	-0.20242	-0.17817	-0.06052	-0.17465	0.23806	0.32086	-0.20415
mul	-0.33463	-0.29455	-0.06326	0.32833	-0.15416	-0.25184	0.02765
mulsol	1.00000	-0.30769	0.08506	-0.22874	0.37925	0.37281	0.63424
sol	-0.30769	1.00000	0.14028	-0.35193	-0.18141	-0.21479	-0.29212
bilat	0.08506	0.14028	1.00000	-0.12728	0.07796	0.11997	0.02960
smooth	-0.22874	-0.35193	-0.12728	1.00000	-0.73560	-0.42056	-0.20150
irreg	0.37925	-0.18141	0.07796	-0.73560	1.00000	0.49954	0.35490
pap	0.37281	-0.21479	0.11997	-0.42056	0.49954	1.00000	0.36624
sept	0.63424	-0.29212	0.02960	-0.20150	0.35490	0.36624	1.00000
shadows	0.07737	0.05606	-0.03788	-0.12973	0.08125	0.03791	0.01951
lucent	-0.03436	-0.26411	0.08167	0.21002	0.00956	0.04815	0.10307
low_level	0.22948	-0.23594	0.00375	-0.09961	0.20172	0.19961	0.16571
mixed	0.18988	-0.18162	-0.14688	-0.00711	0.06777	-0.00572	0.08867
g_glass	0.00099	-0.23044	-0.04569	0.21172	-0.07794	-0.05221	0.01782
haem	-0.11436	-0.10066	-0.07650	0.12379	-0.06363	-0.09323	-0.11534

	shadows	lucent	low_level	mixed	g_glass	haem
age	-0.11966	0.07765	-0.05640	-0.12247	-0.29980	-0.12613
meno	-0.06618	0.05556	-0.02082	-0.14625	-0.29033	-0.13814
colsc2	0.07487	0.11149	-0.13476	-0.01825	0.14605	-0.03577
colsc3	-0.08958	-0.04719	0.11279	-0.03603	-0.02704	0.06131
colsc4	0.00959	-0.07944	0.03541	0.06057	-0.14188	-0.02471
l_ca125	-0.00608	-0.09908	0.04794	-0.02722	-0.10346	-0.10814
pi	-0.06177	0.15075	-0.03549	-0.08063	0.01457	-0.03617
ri	-0.02902	0.09155	-0.06015	-0.02215	0.06568	-0.03396
psv	-0.02246	-0.06444	0.03849	0.13365	0.00131	0.05814
tamx	-0.01101	-0.05250	0.04882	0.12175	-0.02099	0.03726
asc	-0.02687	-0.02256	-0.09923	-0.07596	-0.18683	0.03057
un	-0.10540	0.05050	-0.03123	-0.04255	0.16087	0.27518
unsol	0.13836	0.03190	0.07238	0.06204	-0.06491	0.04854
mul	-0.13700	0.21932	-0.03287	-0.02674	0.12125	-0.06846
mulsol	0.07737	-0.03436	0.22948	0.18988	0.00099	-0.11436
sol	0.05606	-0.26411	-0.23594	-0.18162	-0.23044	-0.10066
bilat	-0.03788	0.08167	0.00375	-0.14688	-0.04569	-0.07650
smooth	-0.12973	0.21002	-0.09961	-0.00711	0.21172	0.12379
irreg	0.08125	0.00956	0.20172	0.06777	-0.07794	-0.06363

pap	0.03791	0.04815	0.19961	-0.00572	-0.05221	-0.09323
sept	0.01951	0.10307	0.16571	0.08867	0.01782	-0.11534
shadows	1.00000	-0.12801	-0.06773	0.24300	-0.06363	-0.05653
lucent	-0.12801	1.00000	-0.31900	-0.29379	-0.23630	-0.13610
low_level	-0.06773	-0.31900	1.00000	-0.18997	-0.05945	-0.04093
mixed	0.24300	-0.29379	-0.18997	1.00000	-0.06948	0.04649
g_glass	-0.06363	-0.23630	-0.05945	-0.06948	1.00000	-0.08565
haem	-0.05653	-0.13610	-0.04093	0.04649	-0.08565	1.00000

We show the biplot in a 2-dimensional space generated by (FACTOR1, FACTOR2). The observations are plotted as points (0=benign, 1=malignant), the variables are plotted as vectors from the origin, i.e taking the respective factor loadings as the coordinates. The angles between the vectors reflect the correlation between the variables. The variables have been standardized to unit variance, so the lengths of the vectors reflect the relative proportion of the variance of each variable that is captured by the two dimensional biplot.



The vectors (Age, Meno, Sol, Asc, L_CA125, Colsc4, PSV, PAP, TAMX, Irreg, MulSol, Sept, Smooth, Colsc2) have approximately the same length, (Bilat, Unsol, Colsc3, Shadow, Haem, Low_level, Mixed, G_Glass, Mul, Un, Lucent) are less well represented in the two dimensional biplot with FACTOR1 and FACTOR2.

In order to get a clearer vision, excluding the variables which are poorly explained by factor 1 and factor 2 in the previous analysis, we redo the principal component factor analysis and show another biplot. This time, only 7 factors will be needed to explain 74% of the variance among the 20 variables. And from the biplot it seems that the distribution of data points changes little without those variables, we might say that not much important information is lost by removing some non important variables.

Initial Factor Method: Principal Components

Eigenvalues of the Correlation Matrix: Total = 20 Average = 1

	Eigenvalue	Difference	Proportion	Cumulative
1	*5.03567359*	*2.40143032*	*0.2518*	*0.2518*
2	*2.63424328*	*0.63435847*	*0.1317*	*0.3835*
3	*1.99988481*	*0.48982649*	*0.1000*	*0.4835*
4	*1.51005832*	*0.14948094*	*0.0755*	*0.5590*
5	*1.36057738*	*0.16091214*	*0.0680*	*0.6270*
6	*1.19966524*	*0.11123999*	*0.0600*	*0.6870*
7	*1.08842524*	*0.28707448*	*0.0544*	*0.7414*
8	0.80135076	0.01088297	0.0401	0.7815
9	0.79046780	0.07007430	0.0395	0.8210
10	0.72039350	0.11484316	0.0360	0.8570
11	0.60555035	0.08090808	0.0303	0.8873
12	0.52464226	0.04123191	0.0262	0.9135
13	0.48341035	0.06584132	0.0242	0.9377
14	0.41756903	0.07502970	0.0209	0.9586
15	0.34253933	0.16594463	0.0171	0.9757
16	0.17659471	0.06519397	0.0088	0.9846
17	0.11140073	0.01243004	0.0056	0.9901
18	0.09897069	0.01930357	0.0049	0.9951
19	0.07966713	0.06075165	0.0040	0.9991
20	0.01891548		0.0009	1.0000

7 factors will be retained by the MINEIGEN criterion.

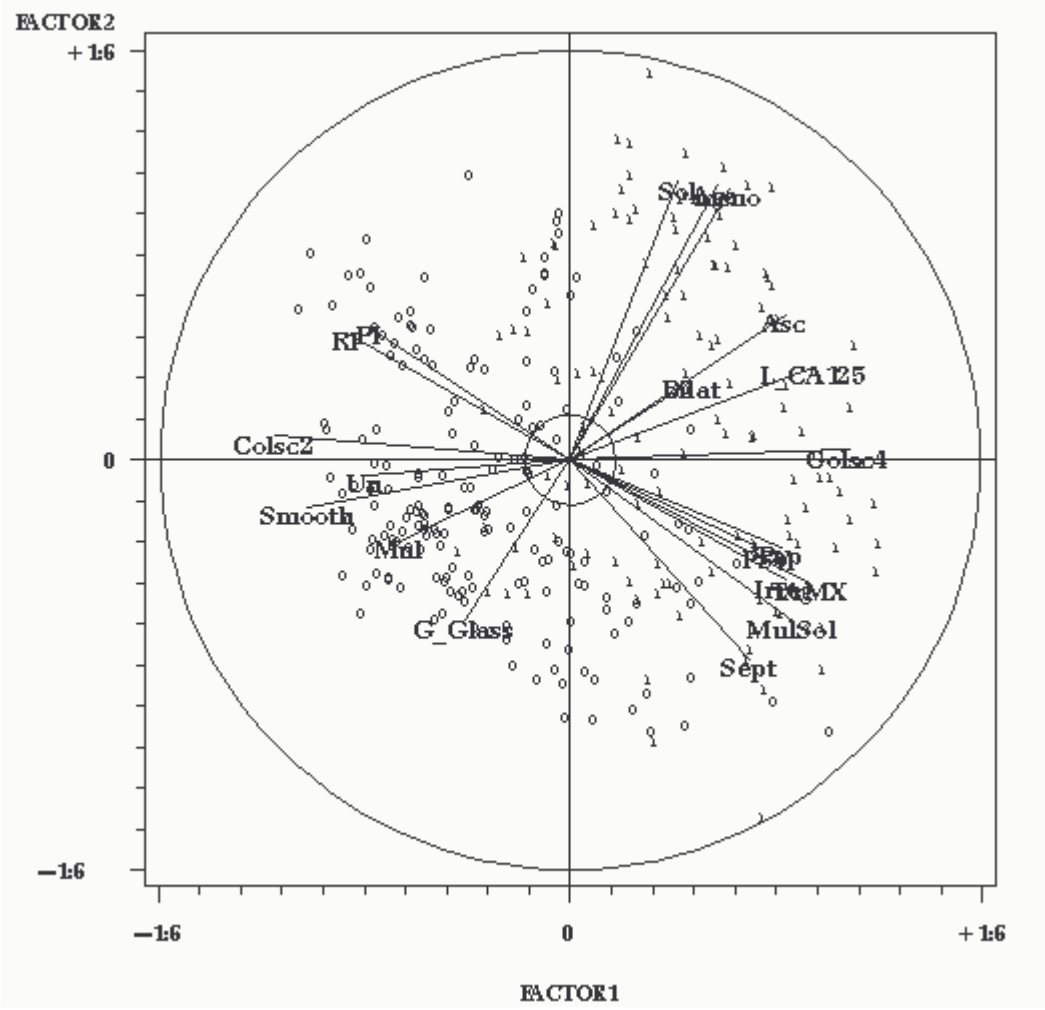
The FACTOR Procedure

Initial Factor Method: Principal Components

Factor Pattern

	Factor1	Factor2	Factor3	Factor4	Factor5	Factor6	Factor7
age	0.35644	0.66875	0.20375	0.21684	-0.17128	0.25149	-0.23393
meno	0.38585	0.66040	0.18746	0.22827	-0.17138	0.27310	-0.24428
colsc2	-0.71748	0.05758	0.20456	-0.10281	0.02444	-0.05966	-0.02094
colsc4	0.67595	0.02352	-0.19394	0.06513	0.05082	0.08715	-0.05482
l_ca125	0.59287	0.22803	0.03444	-0.04434	0.17942	0.15744	0.36950

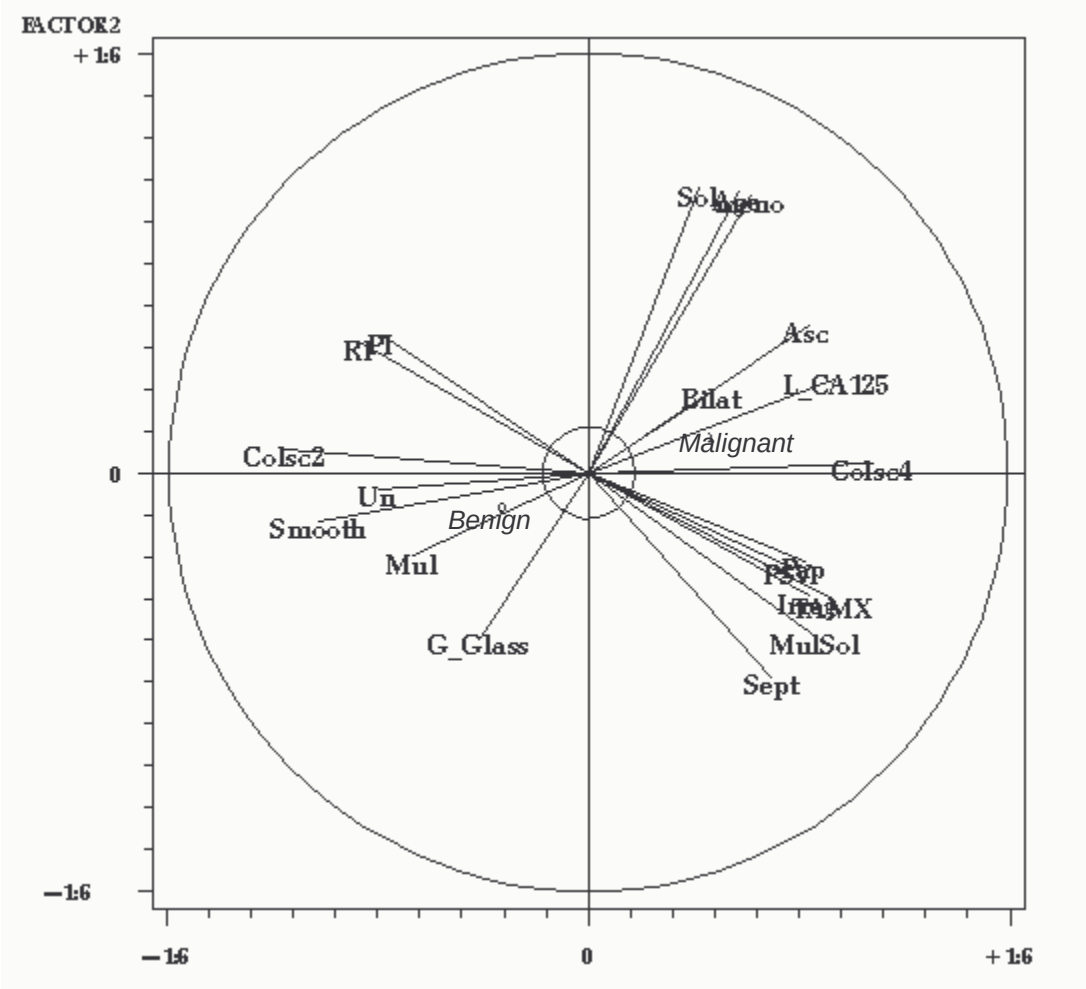
pi	-0.48832	0.32570	0.35482	0.58182	0.12245	-0.17296	0.19032
ri	-0.54290	0.31052	0.32095	0.56874	0.12697	-0.18236	0.24505
psv	0.47928	-0.22419	-0.50777	0.59436	0.05173	-0.21302	0.03155
tamx	0.58194	-0.30316	-0.55451	0.41081	-0.00313	-0.14917	-0.03417
asc	0.52139	0.35215	-0.10932	-0.19681	0.17290	0.11533	0.30887
un	-0.49823	-0.03891	-0.10151	0.01096	0.70436	0.15584	-0.18615
mul	-0.41514	-0.19899	-0.02035	0.04443	-0.77282	0.00433	0.25025
mulsol	0.53965	-0.39286	0.36162	0.19558	0.08681	0.32232	-0.09912
sol	0.26176	0.67780	-0.35762	-0.14352	-0.01576	-0.32452	0.03625
bilat	0.29317	0.19335	-0.00126	-0.10636	0.00710	0.23276	0.62647
smooth	-0.63761	-0.11563	-0.35665	0.19146	-0.00999	0.57071	0.02200
irreg	0.52244	-0.29367	0.51066	-0.09047	0.04924	-0.44907	0.02916
pap	0.51326	-0.21125	0.49723	-0.03145	0.18909	0.01449	0.04346
sept	0.43435	-0.48613	0.35525	0.22674	-0.18700	0.22832	-0.01745
g_glass	-0.25550	-0.39184	-0.07494	0.04383	0.16835	0.11999	0.36854



The variables with small angles like (Age, Meno), (Sol, Age), (PSV, TAMX), (PSV, PAP), (PI, RI), … are highly correlated. Variables with right angles (Age, RI), (Meno, RI), (Meno, TAMX), … are nearly uncorrelated. Variables with angles greater than 90

degrees (Colsc2, Colsc4), (Smooth, Colsc4), (RI, PSV), (PI, TAMX), … are negatively correlated.

Below is another biplot in which the mean points of the two data groups (instead of all data points) are plotted: 0-Benign, 1-Malignant. Large scalar products between the group mean points and the variable points indicate that the corresponding group has high values for those variables. The observations of malignant tumors (1) have relatively high values for variables (Sol, Age, Meno, Asc, Bilat, L_CA125, ColSc4, PSV, TAMX, Pap, Irreg, MulSol, Sept), but relatively low values for the variables (PI, RI, ColSc2, Smooth, Un, Mul, G_glass).



1.3.3 Discriminant Analysis

As a final step in the exploratory analysis a canonical discriminant analysis was carried out to obtain a first idea of important predictor variables. Due to the multicollinearities among the whole 27 variables and missing values, we first use the stepwise procedure to select some important variables, then the discriminant analysis will be applied to the selected variables only.

Identifying a set of variables that provide the "best" discriminants between the two groups is the first objective of discriminant analysis. The second objective is to identify a new axis, Z, such that the new variable Z, given by the projection of observations onto this new axis, provides the maximum separation or discrimination between the groups. The third objective is to classify future observations into one of the groups.

In principal component analysis, a new axis is identified such that the projection of the points onto the new axis accounts for maximum variance in the data, which is equivalent to maximizing *SSt* (total sum of squares). In discriminant analysis, the objective is to maximize the between-group to within-group sum of squares ratio(i.e. *SSb/SSw*) that results in the best discrimination between the groups. The new axis or the linear combination, that is identified is called the linear discriminant function. The projection of a point onto the discriminant function is called the discriminant score.

Fisher's linear discriminant function is used to find a linear combination of the p variables, $Z = \gamma' \boldsymbol{x} = \sum_{i=1}^{p} \gamma_i X_i$, such that (1) the scores (Z) of subjects in the same group are very similar, (2) the scores of subjects from different groups are quite different. In this discriminant analysis, we will only use Fisher's linear discriminant function rule to do classification.

We will use a stepwise procedure to choose the important variables next, and then focus on these subset of variables for discriminant analysis. We start from 24 variables:

```
age meno colsc3 colsc4 l_ca125 pi ri psv tamx
asc un mul mulsol bilat smooth irreg pap sept shadows
lucent low_level mixed g_glass
```

Colsc2 and UnSol are removed from the analysis since colsc2 and UnSol are the linear combinations of the other variables when working on only some part of observations as indicated before. Haem is also excluded from the analysis in order to stabilize the analysis; as Haem is nearly a constant 0, among 525 observations only 15 cases have Haem=1 (presence of haemorrhagic cyst), and no malignant cases have Haem=1. This also implies that Haem contributes little to the prediction of malignancy. We will revisit this problem in the next mulitivariate logistic regression. The following is the partial output of the stepwise procedure by SAS.

The STEPDISC Procedure

The Method for Selecting Variables is STEPWISE

Observations	305	Variable(s) in the Analysis	23
Class Levels	2	Variable(s) will be Included	0
		Significance Level to Enter	**0.15**
		Significance Level to Stay	**0.2**

Class Level Information

path	Variable Name	Frequency	Weight	Proportion
0	_0	182	182.0000	0.596721
1	_1	123	123.0000	0.403279

The STEPDISC Procedure
Stepwise Selection: Step 1

Statistics for Entry, DF = 1, 303

Variable	R-Square	F Value	Pr > F	Tolerance
age	0.1247	43.15	<.0001	1.0000
meno	0.1537	55.02	<.0001	1.0000
colsc3	0.0137	4.21	0.0409	1.0000
colsc4	0.2078	79.48	<.0001	1.0000
l_ca125	0.2855	121.08	<.0001	1.0000
pi	0.0535	17.12	<.0001	1.0000
ri	0.0774	25.43	<.0001	1.0000
psv	0.0424	13.41	0.0003	1.0000
tamx	0.0595	19.16	<.0001	1.0000
asc	0.2515	101.80	<.0001	1.0000
un	0.1231	42.53	<.0001	1.0000
mul	0.1500	53.48	<.0001	1.0000
mulsol	0.0684	22.24	<.0001	1.0000
sol	0.1264	43.86	<.0001	1.0000
bilat	0.0770	25.27	<.0001	1.0000
smooth	0.2477	99.79	<.0001	1.0000
irreg	0.1218	42.04	<.0001	1.0000
pap	0.2027	77.04	<.0001	1.0000
sept	0.0176	5.44	0.0204	1.0000
shadows	0.0083	2.55	0.1115	1.0000
lucent	0.0045	1.37	0.2420	1.0000
low_level	0.0051	1.56	0.2123	1.0000
mixed	0.0156	4.80	0.0293	1.0000
g_glass	0.0337	10.56	0.0013	1.0000

Variable l_ca125 will be entered.

.
Step 2 ..
.
.
Step 11..

The STEPDISC Procedure

Stepwise Selection Summary

Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	Average Squared Canonical Correlation	Pr > ASCC
1	1	l_ca125		0.2855	121.08	<.0001	0.71448836	<.0001	0.28551164	<.0001
2	2	smooth		0.1798	66.22	<.0001	0.58599952	<.0001	0.41400048	<.0001
3	3	asc		0.0920	30.50	<.0001	0.53208540	<.0001	0.46791460	<.0001
4	4	pap		0.0651	20.88	<.0001	0.49745825	<.0001	0.50254175	<.0001
5	5	sol		0.0724	23.32	<.0001	0.46146236	<.0001	0.53853764	<.0001
6	6	colsc4		0.0478	14.96	0.0001	0.43941043	<.0001	0.56058957	<.0001
7	7	colsc3		0.0468	14.59	0.0002	0.41883887	<.0001	0.58116113	<.0001
8	8	meno		0.0339	10.37	0.0014	0.40466103	<.0001	0.59533897	<.0001
9	9	shadows		0.0266	8.05	0.0049	0.39391170	<.0001	0.60608830	<.0001
10	10	irreg		0.0204	6.14	0.0138	0.38585669	<.0001	0.61414331	<.0001
11	9		smooth	0.0007	0.20	0.6549	0.38611948	<.0001	0.61388052	<.0001
12	10	mulsol		0.0085	2.53	0.1127	0.38282306	<.0001	0.61717694	<.0001
13	11	sept		0.0158	4.70	0.0309	0.37677748	<.0001	0.62322252	<.0001
14	12	bilat		0.0057	1.67	0.1977	0.37463904	<.0001	0.62536096	<.0001
15	11		bilat	0.0057	1.67	0.1977	0.00000000	<.0001	0.00000000	<.0001

By now, the stepwise procedure selected the following 11 variables for further discriminant analysis: meno, colsc3, colsc4, l_ca125, asc, mulsol, sol, irreg, pap, sept, shadows.

The DISCRIM Procedure

Observations	425	DF Total	424
Variables	11	DF Within Classes	423
Classes	2	DF Between Classes	1

Class Level Information

path	Variable Name	Frequency	Weight	Proportion	Prior Probability
0	_0	291	291.0000	0.684706	0.500000
1	_1	134	134.0000	0.315294	0.500000

Canonical Discriminant Analysis

	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.773539	0.767783	0.019505	0.598363

Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	1.4898		1.0000	1.0000	0.40163705	55.94	11	413	<.0001

Test of H0: The canonical correlations in the current row and all that follow are zero

NOTE: The F statistic is exact.

Canonical Discriminant Analysis

Total-Sample Standardized Canonical Coefficients

Variable	Can1
meno	0.2074622244
colsc3	0.2665193900
colsc4	0.4054744119
l_ca125	0.3412470667
asc	0.3702987404
mulsol	0.1849009582
sol	0.4846065079
irreg	0.3200569512
pap	0.4664035892
sept	-.1492517874
shadows	-.1731299950

Canonical Discriminant Analysis

Raw Canonical Coefficients

Variable	Can1
meno	0.415848558
colsc3	0.641399002
colsc4	1.079660580
l_ca125	0.188407594
asc	0.878097446
mulsol	0.463766872
sol	1.297585040
irreg	0.640469423
pap	1.048599171
sept	-0.367909200
shadows	-0.621045842

Class Means on Canonical Variables

path	Can1
0	-0.826317751
1	1.794466161

Resubstitution Results using Linear Discriminant Function

Generalized Squared Distance Function

$$D^2_j(X) = (X-\bar{X}_j)' \, COV^{-1} \, (X-\bar{X}_j)$$

Posterior Probability of Membership in Each path

$$Pr(j|X) = \exp(-.5\, D^2_j(X)) \,/\, SUM_k \exp(-.5\, D^2_k(X))$$

* Misclassified observation

Posterior Probability of Membership in path

Obs	From path	Classified into path		0	1
46	0	1	*	0.4835	0.5165
48	0	1	*	0.1878	0.8122
55	0	1	*	0.4890	0.5110
61	1	0	*	0.6507	0.3493
74	0	1	*	0.4482	0.5518
75	0	1	*	0.1381	0.8619
76	1	0	*	0.8543	0.1457
78	0	1	*	0.3588	0.6412
96	0	1	*	0.0872	0.9128
105	0	1	*	0.4649	0.5351
132	0	1	*	0.2109	0.7891
141	1	0	*	0.5660	0.4340
166	0	1	*	0.2698	0.7302
173	1	0	*	0.8720	0.1280
192	1	0	*	0.8768	0.1232
203	1	0	*	0.9726	0.0274
205	1	0	*	0.7853	0.2147
215	1	0	*	0.5060	0.4940
218	0	1	*	0.4451	0.5549
250	1	0	*	0.6347	0.3653
261	0	1	*	0.0066	0.9934
277	0	1	*	0.2206	0.7794
287	0	1	*	0.3720	0.6280
304	0	1	*	0.1146	0.8854
305	1	0	*	0.8684	0.1316
341	1	0	*	0.9793	0.0207
346	0	1	*	0.0889	0.9111
348	0	1	*	0.2153	0.7847
351	0	1	*	0.3483	0.6517
361	0	1	*	0.0146	0.9854
367	1	0	*	0.7746	0.2254
368	1	0	*	0.6026	0.3974
377	1	0	*	0.8255	0.1745
387	0	1	*	0.4926	0.5074
403	1	0	*	0.9910	0.0090

409	1	0	*	0.9601	0.0399
411	0	1	*	0.1383	0.8617
418	0	1	*	0.3727	0.6273
429	0	1	*	0.2174	0.7826
443	1	0	*	0.9860	0.0140
451	1	0	*	0.6817	0.3183
467	0	1	*	0.0356	0.9644
472	1	0	*	0.6581	0.3419
487	1	0	*	0.9607	0.0393
500	1	0	*	0.7462	0.2538
519	0	1	*	0.1083	0.8917

* Misclassified observation

Number of Observations and Percent Classified into path

From path	0	1	Total
0	266	25	291
	91.41	8.59	100.00
1	21	113	134
	15.67	84.33	100.00
Total	287	138	425
	67.53	32.47	100.00
Priors	0.5	0.5	

Error Count Estimates for path

	0	1	Total
Rate	0.0859	0.1567	0.1213
Priors	0.5000	0.5000	

We look for the canonical discriminant functions which maximize the difference between the groups with respect to the 11 variables selected from the stepwise discriminant procedure. The **unstandardized (raw) discriminant function** coefficients are given by the SAS procedure "discrim can ":

Z = 0.42Meno + 0.64Colsc3 + 1.08Colsc4 + 0.19L_ca125 + 0.88Asc + 0.46MulSol + 1.30Sol + 0.64Irreg + 1.05Pap – 0.37Sept –0.62Shadows.

The **class mean o**f benign cases are -0.8263, while for malignant cases is 1.7944. The evidence of malignancy is related to a strongly positive mean value on the canonical variable while the benign character is related to a slightly negative mean value on the canonical variable. This suggests that malignancy is characterized by a high ca125 level, the presence of solid mass, high color score (normal or strong blood flow), the presence of papillarities >3mm, ascites, irregular internal wall, postmenopausal; and absence of the acoustic shadows, septa.

The **squared canonical correlation (CR^2)** is equal to $CR^2 = SS_b / SS_t$. It's obvious that CR^2 gives the amount of variation between the groups that is explained by the discriminating variables. Hence CR^2 is a measure of the strength of the discriminant variables. Here we got 0.598 for CR^2, which is moderate.

Posterior probability of being both classes is also output by SAS. The prior probability of each group is the probability of any random observation belonging to that class (i.e. Class 1 or Class 2). In the present case, it is assumed that the priors are equal; that is both prior probabilities of a given patient having a benign or malignant mass are equal to 0.5.

From the classification table, among the 425 cases, we can see that 21 malignant cases are misclassified as benign, 25 benign cases are misclassified as malignant. In a further check of the misclassified cases, we found some extreme cases whose posterior probabilities of the two memberships are very different. The misclassified cases which have a posterior probability less than 0.1, while the other is greater than 0.90, are highlighted. The case number of these ill-fitted observations for this linear discriminant model are listed below:

1) Malignant cases classified as benign: 203, 341, 403, 409, 443, 487;
2) Benign cases classified as malignant: 96, 261, 346, 361, 467.

2. Logistic Regression

In case of a binary response variable, logistic regression can be used to explore the relation between the response variable (Y) and a number of explanatory variables (X). The response variable has only two possible outcomes: event, denoted by 1 and non-event denoted by 0, with $P(Y=1|\boldsymbol{x})=\pi(\boldsymbol{x})$ and $P(Y=0|\boldsymbol{x})=1-\pi(\boldsymbol{x})$. The logistic regression model can be written as follows:

$$0 \le \pi(x) = \frac{e^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}} \le 1$$

with $\beta_0, \beta_1, \ldots \beta_k,$ parameters of the model. This model defines a curvilinear relation between the mean of the response variable and the explanatory variables, which is equivalent to

$$odds(Y=1) = \frac{P(Y=1)}{P(Y=0)} = \frac{\pi(x)}{1-\pi(x)} = e^{\beta_0+\beta_1 X_1+\ldots+\beta_k X_k}$$

We can also write the model as the function of the logit transformation of $\pi(\boldsymbol{x})$:

$$\ln odds = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

The logit of the mean is linear in the parameters and can range from minus to plus infinity. The slope coefficient β_i, associated with an explanatory variable X_i represents the change in log odds for an increase of one unit in X_i. It can also be interpreted as: if X_i increases with one unit, the odds are multiplied by e^{β_i}

$$odds\ ratio = \frac{odds \mid X_i + 1}{odds \mid X_i} = e^{\beta_i}$$

A maximum likelihood estimation procedure can be used to obtain the parameter estimates. In the following sections, we will use PROC LOGISTIC in SAS to do the logistic regression, whose estimates are obtained by an iterative procedure.

2.1 Odds Ratios and Univariate Logistic Regression Analysis

Odds ratio is a fundamental parameter in logistic regression models. To explain this, let's first take a look at the following contingency table:

	Meno		
Path	0	1	Total
0	265	119	384
1	48	93	141
Total	313	212	525

The 2x2 contingency table classifies 525 patients by malignancy and their menopausal status. Path is a response variable, while Meno is an explanatory variable. We then study the conditional distribution of Path, given Meno.

1) For non-postmenopausal patients(meno=0), the proportion of malignant cases is 48/313=0.15, while for benign cases it is 265/313 = 0.85
2) For postmenopausal patients, the proportion of patients with malignant mass is 93/212 = 0.44, while the proportion of patients with benign mass is 119/212=0.56.

In many instances probabilities are stated as odds. From our contingency table the following odds can be computed:

1) odds of a non-postmenopausal patient having a malignant mass are

odds(Path=1|Meno=0) = 48 / 265 = 0.15 / (1 - 0.15) = 0.1811

2) odds of a postmenopausal patient having a malignant mass are

odds(Path=1|Meno=1) = 93/119 = 0.44/(1-0.44) = 0.7815.

The ratio of odds under the two conditions, is then the odds ratio:

$$\theta = odds\ ratio = \frac{odds \mid meno = 1}{odds \mid meno = 0} = \frac{0.7815}{0.1811} = 4.31$$

The odds ratio is a measure of association for 2x2 contingency tables. The value $\theta = 1$ corresponding to independence serves as a baseline for comparison. In our case, $\theta = 4.31$ indicates that the probability for having a malignant mass among postmenopausal patients is 4.31 times larger than the probability for a malignant mass among non postmenopausal patients.

	odds ratio	Area under ROC curve	SE(parameter estimation)
age	1.05	0.708	0.00689
meno	4.31	0.675	0.2092
l_ca125	2.35	0.832	0.0908
colsc2	0.23	0.640	0.2625
colsc3	2.84	0.593	0.2272
colsc4	20.14	0.698	0.3209
pi	0.48	0.648	0.1956
ri	0.03	0.655	0.7262
psv	1.03	0.657	0.00742
tamx	1.06	0.687	0.0114
asc	10.64	0.716	0.2431
un	0.06	0.704	0.4010
unsol	2.65	0.545	0.3109
mul	0.15	0.615	0.3811
mulsol	4.74	0.627	0.2409
sol	1.33	0.646	0.2536
bilat	4.18	0.629	0.2290
smooth	0.046	0.755	0.3783
irreg	5.35	0.697	0.2212
pap	7.94	0.703	0.2287
sept	3.03	0.597	0.2367
shadows	0.43	0.533	0.3959
lucent	0.54	0.571	0.2127
low_level	1.82	0.539	0.2632
mixed	0.61	0.534	0.2773
g_glass	0.38	0.556	0.3279
haem	<0.001	0.520	602.9

The above table gives the odds ratio of each explanatory variable in the univariate logistic regression, the area under the ROC curve and the standard deviation of the coefficient estimation are also shown. Roughly saying, a good predictor should have a high relative odds ratio (either higher than 1 towards infinity or lower than one towards zero), a high area under the ROC curve which is widely used to estimate the performance of a predictor in medical field (explanation see section 2.4.3), and a low standard deviation in parameter estimation.

This univariate study suggests that *l_ca125* should be a good predictor with area under the ROC curve 0.832, odds ratio 2.35, and a rather low variance in parameter estimation. *Smooth, colsc4, un, asc, pap, mul, irreg, mulsol, colsc2, meno, bilat*, might also be good predictors. Also remarkable is that the high variance in the parameter estimation for *Haem* indicates that it won't be a good predictor for malignancy, as already noticed in the previous univariate analysis, hence it will be excluded from our further study. In the following sections, we will use different selection methods to select the important variables, and build the model.

The following is an illustration of the output of the logistic regression by SAS: a univariate logistic regression with response variable path and one independent variable l_ca125.

[1] Model Information

Response Variable	path
Number of Response Levels	2
Number of Observations	432
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	path	Total Frequency
1	1	137
2	0	295

NOTE: 93 observations were deleted due to missing values for the response or explanatory variables.

[2] Model Fit Statistics

	Criterion	Intercept Only	Intercept and Covariates
[2a]	AIC	541.729	400.091
[2b]	SC	545.798	408.228
[2c]	-2 Log L	539.729	396.091

Testing Global Null Hypothesis: BETA=0

	Test	Chi-Square	DF	Pr > ChiSq
[2d]	Likelihood Ratio	143.6381	1	<.0001
[2e]	Score	132.1652	1	<.0001
[2f]	Wald	88.2178	1	<.0001

[3] Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-4.1117	0.3799	117.1631	<.0001
l_ca125	1	0.8528	0.0908	88.2178	<.0001

[4] Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
l_ca125	2.346	1.964	2.803

[5] Association of Predicted Probabilities and Observed Responses

Percent Concordant	83.0	Somers' D	0.665
Percent Discordant	16.5	Gamma	0.668
Percent Tied	0.6	Tau-a	0.289
Pairs	40415	c	0.832

Response Profile [1]

We see from the response profile that 137 out of 432 patients have malignant masses, 295 masses are benign. For this example, Response = 1 has ordered value =1, and is taken as event; while Response = 0 has ordered value 2. Hence it will be taken as nonevent. Therefore, this logistic regression model is modeling the probability that response is equal to 1, i.e. model the malignancy of the tumors, malignant is taken as event, benign as nonevent.

Assessing Model Fit [2]

The likelihood, *L*, of a model is defined as the probability that the estimated hypothesized model represents the input data. *L* is transformed to –2Log*L*, The **–2Log*L*** statistic [2c], sometimes referred to as the likelihood ratio χ^2 statistic, has a χ^2 distribution with *n-q* degrees of freedom where q is the number of parameters in the model. The -2Log*L* statistic can also be used to determine if the addition of the independent variables significantly improves model fit. This is equivalent to testing whether the coefficients of the independent variables are significantly different from zero. From this output we see

that the difference between the two –2Log*L* (one for model with intercept only, the other with intercept and independent variable *l_ca125*) is equal to 143.6 [2d] with 1 df and is statistically significant [2d]. Therefore, the null hypothesis can be rejected, implying that the inclusion of *l_ca125* significantly improves model fit.

The hypotheses pertaining to whether the coefficients are significantly different from zero can also be established using the statistic reported in [2e] labeled **Score**, which has an aymptoptic chi-square distribution under the null hypothesis. **Wald** test [2f] will also be used to test this hypothesis.

The other measures of goodness-of-fit are the Akaike's information criterion (**AIC**) [2a] and Schwartz's criterion (**SC**) [2b], and they are essentially –2Log*L*'s adjusted for degree of freedom. These two statistics are normally used as heuristics for comparing the fit of different models estimated using the same data set. Lower values of these statistics imply a better fit.

Parameter Estimates [3, 4]

The **maximum likelihood estimates** of model parameters are reported in [3]. The logistic regression model can be written as

$$\ln \frac{p}{1-p} = -4.117 + 0.8528 \times l_ca125$$

where p is the probability of a mass being malignant given the value of independent variables.
The **standard error** of the coefficients can be used to compute the wald **chi_square statistic** $\left[\frac{\hat{\beta}}{SE(\hat{\beta})}\right]^2$ for assessing the statistical significance of each independent variable.
As can be seen, both coefficients are statistically significant at an significance level of 0.05.

Odds ratio is computed by exponentiating the parameter estimate for each explanatory variable.

Association of Predicted Probabilities and Observed Responses [5]

The association of predicted probabilities and observed responses can be assessed by a number of statistics, such as Somer's D, Gamma, and Tau-a and c. These statistics assess the rank order correlation between $\hat{P}$ (the predicted response value) and the observed reponses [5]. These correlations are obtained by first determining the total number of pairs, the number of concordant, disconcordant, and tied pairs, and then transforming them into rank order correlations to measure the association between the observed responses for the dependent variable and $\hat{P}$.

The total number of pairs is the product of the number of events and nonevents, i.e. 137×295 = 40415. A concordant pair is defined as that pair formed by an event and a

nonevent such that the $\hat{P}$ of the event is higher than the $\hat{P}$ of the nonevent. A discordant pair is one in which the $\hat{P}$ for an event is less than the $\hat{P}$ for a nonevent. Tied pairs are ones that are neither concordant nor discordant. Obviously, the higher number of concordant pairs the greater the association between the observed responses and the predicted probabilities. Among the 4 rank order correlations, c is the most interesting one for us, since it represents the area under the ROC curve. And it is computed with the following formulas:

$$c = (nc + 0.5(t - nc - nd)) / t$$

where *nc* is the number of concordant pairs, *nd* is the number of disconcordant pairs, *t* is the total number of pairs. The *c* value of the model with independent variable *l_ca125* is 0.832 which is fairly high, so we can say it is doing a fairly good job in predicting the probability of a mass being malignant.

2.2 Full Model

We will use the following 24 variables (*colsc2, unsol* are excluded due to the multicollinearities, *haem* is excluded due to the big variance in parameter estimation) to generate a 'full' model in which we use almost all the possible variables to :

```
age meno colsc3 colsc4 l_ca125 pi ri psv tamx
asc un mul mulsol sol bilat smooth irreg pap sept shadows
lucent low_level mixed g_glass
```

The LOGISTIC Procedure

Model Information

Response Variable	path
Number of Response Levels	2
Number of Observations	305
Link Function	Logit
Optimization Technique	Fisher's scoring

Response Profile

Ordered Value	path	Total Frequency
1	1	123
2	0	182

NOTE: 220 observations were deleted due to missing values for the response or explanatory variables.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
[2a] AIC	413.334	198.971
[2b] SC	417.055	291.979

[2c] -2 Log L	411.334	148.971	

R-Square 0.5769 Max-rescaled R-Square 0.7792

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	262.3631	24	<.0001
Score	195.5450	24	<.0001
Wald	71.8053	24	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-3.8731	2.5155	2.3707	0.1236
age	1	-0.00165	0.0276	0.0036	0.9524
meno	1	1.0049	0.7676	1.7139	0.1905
colsc3	1	0.8228	0.5380	2.3389	0.1262
colsc4	1	1.8399	0.7695	5.7174	0.0168
l_ca125	1	0.2888	0.1396	4.2770	0.0386
pi	1	-0.6019	0.7631	0.6220	0.4303
ri	1	-2.4546	3.9425	0.3876	0.5335
psv	1	0.1283	0.0748	2.9375	0.0865
tamx	1	-0.2005	0.1104	3.3004	0.0693
asc	1	1.4156	0.5317	7.0883	0.0078
un	1	-2.0470	1.0750	3.6258	0.0569
mul	1	-0.9518	0.9422	1.0204	0.3124
mulsol	1	0.1920	0.9370	0.0420	0.8377
sol	1	2.9403	1.3038	5.0854	0.0241
bilat	1	0.7268	0.5490	1.7521	0.1856
smooth	1	1.0824	1.3056	0.6874	0.4071
irreg	1	2.3672	1.1270	4.4115	0.0357
pap	1	2.4202	0.6346	14.5436	0.0001
sept	1	-0.8659	0.7491	1.3360	0.2477
shadows	1	-1.9155	0.8781	4.7582	0.0292
lucent	1	0.4962	0.8400	0.3489	0.5547
low_level	1	0.2666	0.8913	0.0895	0.7648
mixed	1	-0.3218	0.8572	0.1410	0.7073
g_glass	1	0.0223	0.8576	0.0007	0.9792

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age	0.998	0.946	1.054
meno	2.732	0.607	12.297
colsc3	2.277	0.793	6.536
colsc4	6.296	1.393	28.447
l_ca125	1.335	1.015	1.755

pi	0.548	0.123	2.444
ri	0.086	<0.001	194.917
psv	1.137	0.982	1.316
tamx	0.818	0.659	1.016
asc	4.119	1.453	11.678
un	0.129	0.016	1.062
mul	0.386	0.061	2.447
mulsol	1.212	0.193	7.603
sol	18.921	1.469	243.663
bilat	2.068	0.705	6.067
smooth	2.952	0.228	38.145
irreg	10.667	1.171	97.142
pap	11.249	3.243	39.021
sept	0.421	0.097	1.826
shadows	0.147	0.026	0.823
lucent	1.642	0.317	8.522
low_level	1.306	0.228	7.490
mixed	0.725	0.135	3.889
g_glass	1.023	0.190	5.491

Association of Predicted Probabilities and Observed Responses

Percent Concordant	96.3	Somers' D	0.926
Percent Discordant	3.7	Gamma	0.927
Percent Tied	0.0	Tau-a	0.447
Pairs	22386	c	0.963

Overfitting is typically characterized by unrealistically large estimated coefficients and/or estimated standard errors. In this full model, variable *ri* has the greatest standard error estimate.

2.3 Variable selection

In the following sections, we will initially choose the subset variables for further model building via different model selection procedures: forward, backward, stepwise, best subset selection. After that, for each set of initially selected variables by these methods, we will further check the significance of its coefficients.

Forward selection starts with a model with only an intercept and progressively adds variables one at a time. The significance level of the variables are compared to a critical value for entrance. If the significance level of a variable is lower than this critical value, the variable is added. This process continues until no more variables meet the criterion for entrance.

Backward elimination procedure works the other way around. Starting from a model with all variables, backward elimination removes variables from the model until all the remaining variables satisfied the critical staying significance level.

Stepwise procedure is a combination of forward selection and backward elimination. In each step it is checked whether variables not yet entered in the model meet the criterion for entrance and whether variables in the model meet the criterion for removal.

Finally, **best subset selection** can be used as well. This procedure searches for the "best" models (measured by the score statistic) for all possible model sizes, from one, two, three, up to the model with all the variables. After fitting a multivariate model, the importance of each variable in the model has to be verified.

2.3.1 Backward Elimination

The stay significance level is 0.15. We start from a model with 24 variables, working on 305 observations (220 omitted due to missing values), of which 123 cases are malignant, 182 are benign. The 24 candidate variables are:

```
age meno colsc3 colsc4 l_ca125 pi ri psv tamx
asc un mul mulsol sol bilat smooth irreg pap sept shadows
lucent low_level mixed g_glass.
```

The LOGISTIC Procedure

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	g_glass	1	23	0.0007	0.9792
2	age	1	22	0.0045	0.9466
3	mulsol	1	21	0.0440	0.8338
4	low_level	1	20	0.1374	0.7109
5	lucent	1	19	0.3837	0.5356
6	ri	1	18	0.4584	0.4984
7	mixed	1	17	0.6357	0.4253
8	smooth	1	16	1.3802	0.2401
9	psv	1	15	1.7883	0.1811
10	tamx	1	14	0.3709	0.5425
11	pi	1	13	0.4910	0.4835
12	sept	1	12	1.9596	0.1616

In the end, 12 variables are removed from the model, the remained 12 variables are: Meno, colsc3, coslc4, l_ca125, asc, un, Mul, Sol, Bilat, Irreg, Pap, Shadows.

As there are still variables that are not significant, we rebuild the model based on 425 observations of which 134 are malignant and 291 are benign, and check if all the coefficients are significant. The adjusted coefficient (R^2) and the *p*-value of the Hosmer and Lemeshow goodness-of-fit test (pH.L.) will also be checked. First remove *Bilat* since it has the greatest *p*-value 0.2854. After rebuilding the model based on the 11 variables, we check the c-value which is 0.953 i.e. the same as the model with variable Bilat. Furthermore the decrease of AIC and SC values indicate that the simplified model is preferred. The pH.L. test also indicates that we can't reject the null hypothesis that the

model provides a good fit to the data. Since there are still unsignificant variables, the simplification can still carried out, until all the coeffients are significant at significance level 0.05.

Summary of further backward elimination:

After Removing variable	AIC	SC	-2Log*L*	c	R^2	p H.L.	least significant variable (parameter estimation with biggest p-value)		#Var
							Variable Name	Pr>ChiSq	
	239.091	291.769	213.091	0.953	0.7373	0.9264	Bilat	0.2854	12
Bilat	238.227	286.852	214.227	0.953	0.7355	0.5702	un	0.0902	11
Un	239.237	283.810	217.237	0.951	0.7308	0.6411	Mul	0.1731	10
Mul	239.203	279.724	219.203	0.950	0.7277	0.6624	Shadows	0.1113	9
Shadows	240.076	276.544	222.076	0.948	0.7231	0.5249	Meno	0.0008	8

From the above table, we have two candidate subsets of variables, which have a comparatively good area under the ROC curve (high *c*-value) while being as simple as possible (lower AIC and SC).
MODEL 1a: Meno, colsc3, coslc4, l_ca125, asc, Sol, Irreg, Pap, Shadows
MODEL 1b: Meno, colsc3, coslc4, l_ca125, asc, Sol, Irreg, Pap

If we exclude the non significant variables *pi, ri, psv and tamx* detected previously, we are then working on 425 (instead of 305) observations, of which 134 cases are malignant, 291 are benign. The 21 candidate variables are:

```
age meno colsc2 colsc3 colsc4 l_ca125
asc un mul mulsol sol bilat smooth irreg pap sept shadows
lucent low_level mixed g_glass
```

Note that *colsc2* is included this time. Since 120 observations which haven't been considered before due to missing values in *PI, RI, PSV and TAMX*, will be included in the computation now, the multicollinearities among the 3 variables colsc2, colsc3 and colsc4 vanish. In other words, when the observations indicating the absence of blood flow are also considered, the 3 design variables won't sum to one for each observation, instead they sum to zero for the case of 'no blood flow'.

Summary of Backward Elimination

Step	Effect Removed	DF	Number In	Wald Chi-Square	Pr > ChiSq
1	mixed	1	20	0.0044	0.9470
2	smooth	1	19	0.0391	0.8432
3	lucent	1	18	0.0481	0.8263
4	age	1	17	0.1745	0.6762
5	sept	1	16	0.1863	0.6660

6	low_level	1	15	0.2321	0.6300
7	g_glass	1	14	0.3157	0.5742
8	colsc2	1	13	0.4413	0.5065
9	mulsol	1	12	1.3464	0.2459
10	bilat	1	11	1.1413	0.2854

This time 11 variables are left in the model, they are:

Meno, colsc3, colsc4, l_ca125, asc, un, mul, sol, irreg, pap, shadows.

This is just the same model after removal of bilat from the model selected by the backward elimination.

2.3.2 Forward Selection

We then apply the forward selection to the 24 variables based on the 305 observations, setting the entry significance level to 0.15.

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	l_ca125	1	1	131.4972	<.0001
2	smooth	1	2	67.1619	<.0001
3	asc	1	3	29.8466	<.0001
4	meno	1	4	16.3071	<.0001
5	pap	1	5	15.3891	<.0001
6	colsc4	1	6	9.6423	0.0019
7	colsc3	1	7	13.3768	0.0003
8	sol	1	8	4.7411	0.0294
9	shadows	1	9	2.4923	0.1144
10	irreg	1	10	2.5902	0.1075

We end up with 10 variables:
Meno, colsc3, colsc4, l_ca125, asc, sol, smooth, irreg, pap, shadows.
Fitting the model based on 425 observations, we find that there are still some variables not significant at 0.05 significance level, further simplification could be carried out. After removing smooth, we find that we end up with the same subset as computed by the backward selection and further simplification.

After Removing variable	AIC	SC	-2LogL	c	R^2	p H.L.	least significant variable (parameter estimation with biggest p-value)		#Var
							Variable Name	Pr>ChiSq	
	240.813	285.385	218.813	0.951	0.7382	0.8445	Smooth	0.5275	10
smooth	239.203	279.724	219.203	0.950	0.7277	0.6624	Shadows	0.1113	9

2.3.3 Stepwise Logistic Regression

Combining both forward selection and backward selection, we obtain a stepwise selection procedure.

We will work on 305 observations (220 omitted due to missing values) with 24 variables, of which 123 cases are malignant, 182 are benign. The 24 variables are:

```
age meno colsc3 colsc4 l_ca125 pi ri psv tamx
asc un mul mulsol sol bilat smooth irreg pap sept shadows
lucent low_level mixed g_glass
```

The entry significance level is set to 0.15, stay significance level is 0.20.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	l_ca125		1	1	87.0810	.	<.0001
2	smooth		1	2	51.1439	.	<.0001
3	asc		1	3	22.7192	.	<.0001
4	pap		1	4	15.2009	.	<.0001
5	meno		1	5	17.2223	.	<.0001
6	colsc4		1	6	8.2501	.	0.0041
7	colsc3		1	7	10.9217	.	0.0010
8	sol		1	8	9.5302	.	0.0020
9	irreg		1	9	3.7266	.	0.0536
10		smooth	1	8	.	0.1423	0.7060
11	shadows		1	9	3.7544	.	0.0527

During this procedure 10 variables has been entered and one variable smooth is removed again, The 9 selected variables are:
meno, l_ca125, colsc3, colsc4, asc, sol, irreg, pap, shadows.
Note that this subset is the same as the one we got after the backward elimination and further simplification, and forward selection and further simplification.

2.3.4 Best Subset Selection

Best subset selection finds a specified number of models with the highest likelihood score (chi-square) statistics for all possible model size, from one, two variables, and so on, up to the full model that contains all of the explanatory variables.

We apply this best subset selection on 305 observation with 24 variables:

```
age meno colsc3 colsc4 l_ca125 pi ri psv tamx
asc un unsol mul mulsol bilat smooth irreg pap sept shadows
lucent low_level mixed g_glass
```

The subsets whose number of variables is less than 11 are listed below.

The LOGISTIC Procedure

Regression Models Selected by Score Criterion

Number of Variables	Score Chi-Square	Variables Included in Model
1	37.5385	un
1	37.1610	irreg
1	23.6174	ri
1	23.4814	bilat
1	20.8573	mulsol
1	18.1386	tamx
1	16.3121	pi
1	12.9267	psv
1	10.2692	g_glass
1	5.3758	sept
1	4.7524	mixed
1	4.1841	colsc3
1	2.5433	shadows
1	1.5642	low_level
1	1.3769	lucent
2	127.2389	sol pap
2	126.2701	l_ca125 smooth
2	121.6762	asc pap
2	121.2011	asc smooth
2	113.2865	un mul
2	111.2253	l_ca125 pap
2	111.0272	l_ca125 asc
2	108.1378	colsc4 l_ca125
2	107.8824	meno l_ca125
2	107.3923	colsc4 smooth
2	106.4335	l_ca125 irreg
2	106.4297	l_ca125 mul
2	106.1471	colsc4 asc
2	103.8995	colsc4 pap
2	103.7993	l_ca125 sol
2	103.0485	l_ca125 un
2	102.8710	asc irreg
2	101.8958	age l_ca125
2	100.3313	meno smooth
2	99.7387	meno asc
2	98.4922	l_ca125 ri
2	97.4894	meno pap
2	97.0657	smooth pap
2	96.1242	asc un
3	151.6246	asc sol pap
3	147.6668	l_ca125 sol pap
3	146.8294	colsc4 sol pap
3	142.7140	l_ca125 asc smooth
3	142.6022	mulsol sol pap
3	141.6959	l_ca125 un mul
3	140.5540	colsc4 asc pap
3	140.5208	asc smooth pap
3	139.7067	meno asc pap
3	138.9669	meno l_ca125 smooth

```
3    137.6916  colsc4 l_ca125 smooth
3    137.6300  sol irreg pap
3    137.3007  meno sol pap
3    137.0548  colsc4 asc smooth
3    136.8216  l_ca125 asc pap
3    134.7010  meno asc smooth
3    134.6861  asc un mul
3    134.5193  l_ca125 smooth pap
3    134.4159  l_ca125 sol irreg
3    134.3728  age l_ca125 smooth
3    134.2953  ri sol pap
3    134.1450  l_ca125 mul smooth
3    134.0360  un sol pap
3    133.9159  sol bilat pap

4    163.3514  colsc4 asc sol pap
4    160.8951  asc mulsol sol pap
4    159.8294  l_ca125 asc sol pap
4    159.2362  asc sol irreg pap
4    158.7089  meno asc sol pap
4    158.5417  colsc3 colsc4 sol pap
4    158.1386  colsc4 l_ca125 sol pap
4    156.8485  l_ca125 mulsol sol pap
4    156.5868  l_ca125 sol irreg pap
4    156.4961  asc un sol pap
4    156.1206  asc sol pap shadows
4    155.8437  asc sol smooth pap
4    155.2534  tamx asc sol pap
4    155.2107  age asc sol pap
4    154.8505  asc un mul pap
4    154.8429  ri asc sol pap
4    154.5724  meno l_ca125 sol pap
4    154.3692  psv asc sol pap
4    154.2391  asc sol pap low_level
4    154.1261  colsc4 mulsol sol pap
4    153.9087  asc sol bilat pap
4    153.6357  pi asc sol pap
4    153.5857  colsc4 asc smooth pap
4    153.5447  meno asc smooth pap

5    171.3441  colsc3 colsc4 asc sol pap
5    168.4672  meno colsc4 asc sol pap
5    168.2539  colsc4 asc sol irreg pap
5    168.1446  colsc4 asc mulsol sol pap
5    167.8435  colsc4 l_ca125 asc sol pap
5    167.8034  colsc4 asc sol pap shadows
5    167.1584  l_ca125 asc sol irreg pap
5    167.0105  meno asc sol irreg pap
5    166.9274  colsc3 colsc4 l_ca125 sol pap
5    166.8627  asc mulsol sol pap shadows
5    166.7854  colsc4 asc un sol pap
5    166.7608  l_ca125 asc mulsol sol pap
5    166.1482  colsc4 asc sol smooth pap
5    166.0771  age colsc4 asc sol pap
5    165.6290  meno l_ca125 asc sol pap
5    165.5627  asc mulsol sol irreg pap
5    165.5339  meno asc mulsol sol pap
```

```
5    165.3873  colsc4 asc sol bilat pap
5    165.2828  colsc4 asc sol pap low_level
5    164.8215  asc sol irreg pap shadows
5    164.5676  meno colsc3 colsc4 sol pap
5    164.4722  colsc4 asc sol pap mixed
5    164.3682  asc un mul sol pap
5    164.2710  colsc4 asc mul sol pap

6    176.1730  meno colsc3 colsc4 asc sol pap
6    175.3453  colsc3 colsc4 asc sol irreg pap
6    174.7477  colsc3 colsc4 l_ca125 asc sol pap
6    174.6826  colsc3 colsc4 asc sol pap shadows
6    174.0866  meno colsc4 asc sol irreg pap
6    173.9860  age colsc3 colsc4 asc sol pap
6    173.7181  colsc3 colsc4 asc mulsol sol pap
6    173.7001  colsc4 asc mulsol sol pap shadows
6    173.5979  meno l_ca125 asc sol irreg pap
6    173.5932  colsc4 asc sol irreg pap shadows
6    173.5446  colsc3 colsc4 asc sol smooth pap
6    173.1616  asc un mul sol pap shadows
6    172.9597  colsc4 l_ca125 asc sol irreg pap
6    172.9142  colsc3 colsc4 asc sol bilat pap
6    172.9022  colsc3 colsc4 asc un sol pap
6    172.5995  colsc3 colsc4 asc sol pap mixed
6    172.4131  l_ca125 asc sol irreg pap shadows
6    172.3678  meno colsc4 l_ca125 asc sol pap
6    172.3202  colsc3 colsc4 asc sol pap low_level
6    172.2699  l_ca125 asc mulsol sol pap shadows
6    172.2661  asc mulsol sol irreg pap shadows
6    172.2637  l_ca125 un mul sol pap shadows
6    172.0876  colsc4 l_ca125 asc sol pap shadows
6    172.0688  meno colsc4 asc sol pap shadows

7    180.8240  meno colsc3 colsc4 asc sol irreg pap
7    179.9191  colsc4 asc un mul sol pap shadows
7    179.4564  colsc3 colsc4 asc sol irreg pap shadows
7    179.0909  meno colsc3 colsc4 l_ca125 asc sol pap
7    178.9796  colsc3 colsc4 l_ca125 asc sol irreg pap
7    178.8093  meno colsc3 colsc4 asc sol pap shadows
7    178.6549  l_ca125 asc un mul sol pap shadows
7    178.4993  age colsc3 colsc4 asc sol irreg pap
7    178.4800  meno colsc4 asc sol irreg pap shadows
7    178.3769  meno colsc3 colsc4 asc sol smooth pap
7    178.1583  meno colsc4 l_ca125 asc sol irreg pap
7    178.0835  colsc4 l_ca125 asc sol irreg pap shadows
7    177.9961  colsc3 colsc4 l_ca125 asc sol pap shadows
7    177.9527  colsc3 colsc4 asc mulsol sol pap shadows
7    177.8713  meno l_ca125 asc sol irreg pap shadows
7    177.8125  l_ca125 asc mulsol sol irreg pap shadows
7    177.8080  colsc4 asc mulsol sol irreg pap shadows
7    177.7394  colsc4 l_ca125 un mul sol pap shadows
7    177.6705  meno colsc3 colsc4 asc mulsol sol pap
7    177.6118  colsc3 colsc4 asc sol smooth pap shadows
7    177.3906  meno colsc3 colsc4 asc sol bilat pap
7    177.3798  meno colsc3 colsc4 l_ca125 sol irreg pap
7    177.2541  colsc3 colsc4 l_ca125 asc sol smooth pap
7    177.2452  meno colsc3 colsc4 asc un sol pap
```

```
 8    184.1517  meno colsc3 colsc4 asc sol irreg pap shadows
 8    183.9389  meno colsc3 colsc4 l_ca125 asc sol irreg pap          [MODEL1a]
 8    183.2460  colsc4 l_ca125 asc un mul sol pap shadows
 8    183.0089  colsc3 colsc4 l_ca125 asc sol irreg pap shadows
 8    182.4533  colsc3 colsc4 asc un mul sol pap shadows
 8    182.4177  meno colsc4 l_ca125 asc sol irreg pap shadows
 8    182.3307  meno colsc4 asc un mul sol pap shadows
 8    182.3141  colsc4 asc un mul sol irreg pap shadows
 8    182.1854  l_ca125 asc un mul sol irreg pap shadows
 8    181.9884  meno colsc3 colsc4 asc sol bilat irreg pap
 8    181.8642  colsc4 asc un mul sol bilat pap shadows
 8    181.7973  meno colsc3 colsc4 asc sol irreg pap mixed
 8    181.6925  colsc3 colsc4 asc mulsol sol irreg pap shadows
 8    181.6871  age colsc3 colsc4 asc sol irreg pap shadows
 8    181.6860  meno colsc3 colsc4 l_ca125 asc sol pap shadows
 8    181.6812  age colsc3 colsc4 l_ca125 asc sol irreg pap
 8    181.6631  meno colsc3 colsc4 asc sol smooth pap shadows
 8    181.5784  meno colsc3 colsc4 l_ca125 asc sol smooth pap
 8    181.5633  meno colsc3 colsc4 asc mulsol sol irreg pap
 8    181.5310  meno l_ca125 asc mulsol sol irreg pap shadows
 8    181.4724  colsc4 l_ca125 asc mulsol sol irreg pap shadows
 8    181.4717  meno colsc3 colsc4 asc sol irreg pap sept
 8    181.3510  meno colsc3 colsc4 asc sol irreg pap low_level
 8    181.3365  meno colsc3 colsc4 asc mul sol irreg pap

 9    187.2336  meno colsc3 colsc4 l_ca125 asc sol irreg pap shadows   [MODEL1b]
 9    185.8925  colsc4 l_ca125 asc un mul sol irreg pap shadows
 9    185.4265  meno l_ca125 asc un mul sol irreg pap shadows
 9    185.4252  meno colsc3 colsc4 asc mulsol sol irreg pap shadows
 9    185.3731  meno colsc4 l_ca125 asc un mul sol pap shadows
 9    185.3578  colsc3 colsc4 l_ca125 asc un mul sol pap shadows
 9    185.2964  meno colsc4 asc un mul sol irreg pap shadows
 9    185.2013  meno colsc3 colsc4 asc sol bilat irreg pap shadows
 9    185.0935  meno colsc3 colsc4 asc mul sol irreg pap shadows
 9    185.0725  meno colsc3 colsc4 asc un mul sol pap shadows
 9    184.9893  meno colsc3 colsc4 l_ca125 asc sol irreg pap mixed
 9    184.8680  age colsc3 colsc4 l_ca125 asc sol irreg pap shadows
 9    184.8569  meno colsc3 colsc4 l_ca125 asc sol smooth pap shadows
 9    184.8546  colsc3 colsc4 asc un mul sol irreg pap shadows
 9    184.8152  colsc3 colsc4 l_ca125 asc mulsol sol irreg pap shadows
 9    184.7888  meno colsc3 colsc4 asc un sol irreg pap shadows
 9    184.7518  meno colsc3 colsc4 l_ca125 asc sol bilat irreg pap
 9    184.7013  meno colsc3 colsc4 asc sol irreg pap sept shadows
 9    184.6880  meno colsc4 l_ca125 asc mulsol sol irreg pap shadows
 9    184.6826  colsc4 l_ca125 asc un mul sol bilat pap shadows
 9    184.6363  meno colsc3 colsc4 pi asc sol irreg pap shadows
 9    184.5307  meno colsc3 colsc4 asc sol irreg pap shadows low_level
 9    184.5214  meno colsc3 colsc4 l_ca125 asc sol irreg pap sept
 9    184.4793  meno colsc3 colsc4 l_ca125 asc mulsol sol irreg pap

10    188.5713  meno colsc4 l_ca125 asc un mul sol irreg pap shadows
10    188.2390  meno colsc3 colsc4 l_ca125 asc mulsol sol irreg pap shadows
10    188.1853  meno colsc3 colsc4 l_ca125 asc mul sol irreg pap shadows
10    188.0702  meno colsc3 colsc4 asc un mul sol irreg pap shadows
10    187.9945  colsc3 colsc4 l_ca125 asc un mul sol irreg pap shadows
10    187.9525  meno colsc3 colsc4 l_ca125 asc sol bilat irreg pap shadows
```

```
      10    187.7680  meno colsc3 colsc4 l_ca125 asc un sol irreg pap shadows
      10    187.7240  meno colsc3 colsc4 l_ca125 asc sol irreg pap sept shadows
      10    187.7124  meno colsc3 colsc4 l_ca125 pi asc sol irreg pap shadows
      10    187.6830  meno colsc3 colsc4 l_ca125 asc un mul sol pap shadows
      10    187.6234  meno colsc3 colsc4 l_ca125 ri asc sol irreg pap shadows
      10    187.5795  meno colsc3 colsc4 asc mulsol sol irreg pap sept shadows
      10    187.5650  meno colsc3 colsc4 l_ca125 asc sol irreg pap shadows mixed
      10    187.5127  meno colsc3 colsc4 l_ca125 asc sol irreg pap shadows low_level
      10    187.4048  age meno colsc3 colsc4 l_ca125 asc sol irreg pap shadows
      10    187.3459  meno colsc3 colsc4 l_ca125 asc sol irreg pap shadows g_glass
      10    187.3137  meno colsc3 colsc4 l_ca125 asc sol smooth irreg pap shadows
      10    187.2608  meno colsc3 colsc4 l_ca125 asc sol irreg pap shadows lucent
      10    187.2544  colsc4 l_ca125 asc un mul sol bilat irreg pap shadows
      10    187.2521  meno colsc3 colsc4 l_ca125 psv asc sol irreg pap shadows
      10    187.2505  meno colsc3 colsc4 l_ca125 tamx asc sol irreg pap shadows
      10    187.0131  meno l_ca125 ri asc un mul sol irreg pap shadows
      10    186.9748  colsc4 l_ca125 asc un mul sol irreg pap shadows mixed
      10    186.9106  colsc3 colsc4 l_ca125 asc mulsol sol irreg pap sept shadows
```

Let's focus on the models with 8-9 explanatory variables. We notice that, MODEL 1a which is selected by our previous three model selection procedures, has the highest score in the 9 variable subsets, while MODEL 1b has the second highest score in the subsets with 8 variables.

```
8     183.9389  meno colsc3 colsc4 l_ca125 asc sol irreg pap              [MODEL 1a]
9     187.2336  meno colsc3 colsc4 l_ca125 asc sol irreg pap shadows     [MODEL 1b]
```

2.4 Model Fitting, Diagnostics and Classification

In this section we will first fit MODEL1a based on all the data which contains no missing values. Then we will do logistic regression diagnostics and see the predictive accuracy of this model.

2.4.1 Model Fitting and Goodness-of-fit Test

Removing the observations with missing values, we are working on 425 observations with 134 malignant cases and 291 benign cases. Below is the output of the model fitting for MODEL 1a: `meno colsc3 colsc4 l_ca125 asc sol irreg pap`.

Response Profile

Ordered Value	path	Total Frequency
1	1	134
2	0	291

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	531.781	240.076
SC	535.833	276.544
-2 Log L	529.781	222.076

		[1]	
R-Square	0.5152	Max-rescaled R-Square	0.7231

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	[2] Pr > ChiSq
Likelihood Ratio	307.7050	8	<.0001
Score	249.9219	8	<.0001
Wald	99.3702	8	<.0001

[3] Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-6.3414	0.6883	84.8750	<.0001
meno	1	1.2427	0.3687	11.3588	0.0008
colsc3	1	1.3397	0.3954	11.4815	0.0007
colsc4	1	1.7286	0.4877	12.5625	0.0004
l_ca125	1	0.4383	0.1194	13.4826	0.0002
asc	1	1.8664	0.4329	18.5890	<.0001
sol	1	2.3962	0.5244	20.8793	<.0001
irreg	1	1.7510	0.4792	13.3503	0.0003
pap	1	1.7116	0.4233	16.3494	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
meno	3.465	1.682	7.138
colsc3	3.818	1.759	8.287
colsc4	5.632	2.166	14.650
l_ca125	1.550	1.227	1.959
asc	6.465	2.768	15.102
sol	10.981	3.929	30.692
irreg	5.760	2.252	14.735
pap	5.538	2.416	12.696

[4] Association of Predicted Probabilities and Observed Responses

Percent Concordant	94.8	Somers' D	0.896
Percent Discordant	5.1	Gamma	0.897
Percent Tied	0.1	Tau-a	0.388
Pairs	38994	c	0.948

Partition for the Hosmer and Lemeshow Test

		path = 1		path = 0	
Group	Total	Observed	Expected	Observed	Expected
1	42	0	0.21	42	41.79
2	46	0	0.51	46	45.49
3	43	0	0.84	43	42.16
4	44	2	1.45	42	42.55
5	43	7	3.26	36	39.74
6	43	7	7.48	36	35.52
7	43	14	15.88	29	27.12
8	43	29	29.59	14	13.41
9	43	40	40.02	3	2.98
10	35	35	34.76	0	0.24

[5] Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
7.1088	8	0.5249

All the variables are significant at p≤0.05 [3], The parameter estimates suggest that, the effects of the *l_ca125, the presence of menopausal, normal blood flow or strong blood flow, the presence of ascites, solid mass, irregular internal wall, and papillarites > 3mm*, on the log of odds of the tumors being malignant are *all positive*.

The max-rescaled Rsquare [1] is the adjusted coefficient with value 0.7231, which is fairly good. And the Hosmer and Lemeshow goodness-of-fit test [5] can't reject the null hypothesis that the model provides a good fit to the data (p-value = 0.5249).

From the c-value = 0.948 we see that this model gives a good association between the probabilities and observed responses. However this c-value (or area under ROC curve) is somehow bias, because it use the training set for classification. In section 2.4.3 we will divide the data set into two exclusive parts, one for model fitting, the other part for testing.

2.4.2 Influence Measures and Diagnostics

Influence measures and diagnostics help us to determine whether individual observations have undue impact on the fitted regression model or the coefficients of individual predictors. However, since the response in logistic regression is discrete, additional problems occur when similar diagnostics, which are normally used in ordinary least squares regression, are applied in logistic regression.

1) **Leverage** measures the potential impact of an individual case on the result, which is directly proportional to how far an individual case is from the centroid in the space of the predictors. Leverage is computed as the diagonal elements, h_{ii}, of the "Hat" matrix $\boldsymbol{h}$.

$$\boldsymbol{h}=\boldsymbol{x}^*(\boldsymbol{x}^{*\prime}\boldsymbol{x}^*)^{-1}\boldsymbol{x}^{*\prime} ,$$

where $\boldsymbol{x}^*=\boldsymbol{v}^{1/2}\boldsymbol{x}$ and $\boldsymbol{v}=\text{diag}\{\hat{P}(1-\hat{P})\}$. As in ordinary least squares regression, leverage values are between 0 and 1, and a leverage value, $h_{ii}>2k/n$ is considered "large"; k = number of predictors, n = number of cases.

2) **Residuals** Pearson and deviance residuals are useful in identifying observations that are not well explained by the model. They are the (signed) square roots of the contribution that each case makes to the overall Pearson and deviance goodness-of-fit statistics.

3) **Influence measures** measure the effect that deleting an observation has on the regression parameters or the goodness-of-fit statistics.

   C or **CBAR** analogs of the Cook's Distance statistic in ordinary least squares regression are two standardized measures of the approximate change in all regression coefficients when the ith case is deleted.
   DIFCHISQ is an approximate measure of the amount of Delta chi-square by which the Pearson chi-square would decrease if the *i*th case were deleted. Values > 4 indicate "significant" change (since these are 1 df chi-square or squared normal values).
   DIFDEV is an approximate measure of the amount of Delta chi-square by which the likelihood ratio deviance chi-square would decrease if the *i*th case were deleted. Values > 4 indicate "significant" change.

Plots of the change in chi-square(DIFCHISQ or DIFDEV) against leverage or predicted probabilities are useful for detecting unduly influential cases. We use the SAS macro INFLOGIS (Doc: http://www.math.yorku.ca/SCS/vcd/inflogis.html) to generate the plot, which shows a measure of badness of fit for a given case (DIFDEV or DIFCHISQ) vs. the fitted probability (PRED) or leverage (HAT), using an influence measure (C or CBAR) as the size of a bubble symbol.

There are 17 observations identified as influential cases in these 425 data, because they have high leverage (leverage>(2k)/n=0.04235) or high influence (DIFCHISQ>4).

Num	Path	meno	Coll3	Coll4	L_CA12	Asc	Sol	Irreg	Pap	pred	stud	hat	diffchisq	diffdev	c
76	1	0	0	0	1.60944	0	0	1	1	.102	2.155	.02	8.943	4.718	0.15798
173	1	0	0	0	3.25810	0	0	1	1	.190	1.842	.02	4.357	3.413	0.09202
192	1	0	0	0	3.29584	0	1	0	0	.076	2.294	.02	12.434	5.403	0.24882
203	1	1	0	0	4.71850	0	0	0	0	.046	2.496	.01	20.962	6.408	0.25561
205	1	0	0	0	2.56495	0	0	1	1	.147	1.975	.02	5.891	3.938	0.11173
261	0	0	1	0	9.49251	0	0	1	1	.932	-2.364	.04	14.280	5.908	0.54538
305	1	1	0	0	4.21951	0	0	1	0	.183	1.876	.03	4.633	3.559	0.16454
341	1	0	0	0	2.94444	0	0	1	0	.036	2.592	.01	27.294	6.857	0.18567
361	0	0	0	1	3.33220	1	0	1	1	.898	-2.174	.03	9.126	4.877	0.32003
367	1	1	1	0	3.58352	0	0	0	0	.101	2.169	.02	9.146	4.815	0.23115
377	1	0	1	0	2.94444	1	0	0	0	.136	2.038	.04	6.598	4.254	0.28262
403	1	0	0	0	4.53260	0	0	1	0	.069	2.329	.01	13.702	5.537	0.18922
409	1	0	0	0	4.31749	0	0	1	0	.063	2.365	.01	15.034	5.710	0.18537
411	0	1	1	0	4.18965	1	0	1	0	.845	-1.979	.05	5.724	4.006	0.29361
443	1	1	0	0	3.33220	0	0	0	0	.026	2.715	.01	38.253	7.559	0.23228
487	1	1	0	0	2.07944	0	0	1	0	.080	2.263	.02	11.613	5.223	0.18569
500	1	0	1	0	0.00000	1	0	1	0	.200	1.871	.08	4.347	3.571	0.38668

Note : Studres: Studentized deviance residual = resdev / sqrt(1-hat);
Hat: Leverage (Hat value);
Difchisq: Change in Pearson Chi Square;
Difdev: Change in Deviance.

Checking the case number, we find that only 6 influential observations come from the first 300 cases, while the remaining 11 ill-fitted cases come from the rest of the 250 cases. From this information, we might expect that if we fit the logistic regression model based on the first 300 observations, the predicting accuracy on the testing data which come from the last 250 patients will certainly be much lower than the classification accuracy on the first 300 observations itself. This can partially explain our testing result in the model validation section.

Also interesting to mention is that, all these ill-fitted cases are all misclassified by the linear discriminant function when setting the probability cutoff value to 0.5; and all these cases have very high predicted probability (>0.75) or very low predicted probability (<0.25) in that discriminant analysis.

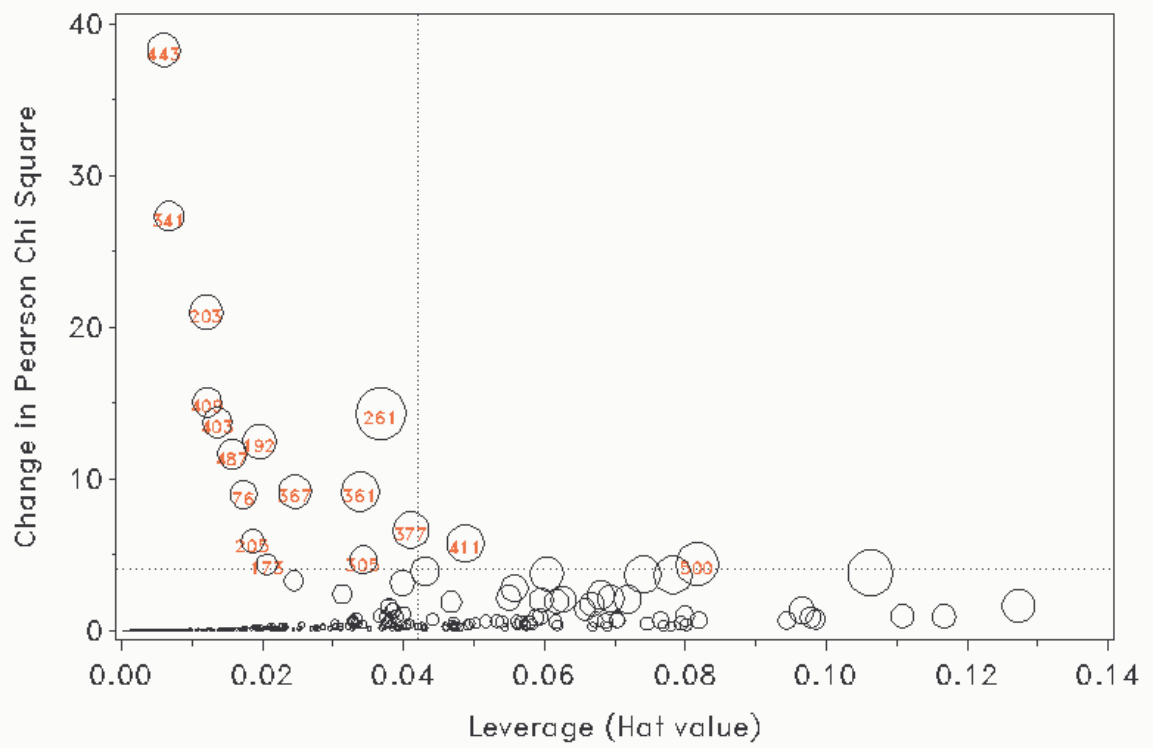


Fig.2.1.a Influence Plot. Cases with DIFCHISQ>4 or leverage>(2k)/n=0.04235 are influential, as indicated by the size of the bubble symbol.

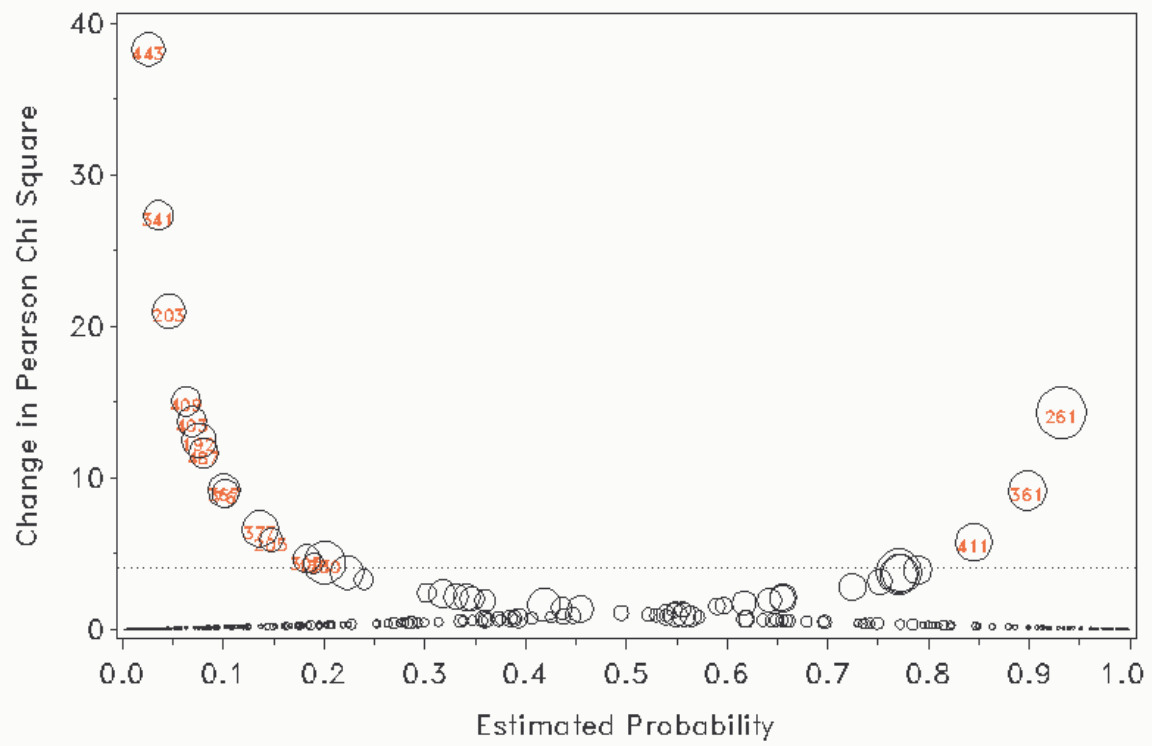


Fig2.1.b Changes in chi-square vs. predicted probability. The plot shows that most of the influential observations are those with very high or low predicted probabilities. The systematic pattern shown is inherent in the discrete nature of logistic regression.

2.4.3 Classification Table and ROC-Curve

A fitted model will be used to classify observations as events or nonevents. The model classifies an observation as an event if its estimated probability is greater than or equal to a given probability cutoff value (threshold). Otherwise the observation is classified as a nonevent. The classification table reports several measures of predictive accuracy for varying probability cutoff level. The higher the probability cutoff level, the more likely that an observation is classified as a nonevent.

Table2.1 shows the classification table output by SAS, based on the fitted MODEL1a.

Table 2.1

Classification Table

	[1] Correct		[2] Incorrect			[3] Percentages			
Prob Level	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.000	134	0	291	0	31.5	100.0	0.0	68.5	.
0.050	131	176	115	3	72.2	97.8	60.5	46.7	1.7
0.100	125	202	89	9	76.9	93.3	69.4	41.6	4.3
0.150	122	219	72	12	80.2	91.0	75.3	37.1	5.2
0.200	119	233	58	15	82.8	88.8	80.1	32.8	6.0
0.250	118	243	48	16	84.9	88.1	83.5	28.9	6.2
0.300	114	249	42	20	85.4	85.1	85.6	26.9	7.4
0.350	111	258	33	23	86.8	82.8	88.7	22.9	8.2
0.400	109	265	26	25	88.0	81.3	91.1	19.3	8.6
0.450	109	272	19	25	89.6	81.3	93.5	14.8	8.4
0.500	105	274	17	29	89.2	78.4	94.2	13.9	9.6
0.550	101	274	17	33	88.2	75.4	94.2	14.4	10.7
0.600	98	276	15	36	88.0	73.1	94.8	13.3	11.5
0.650	92	278	13	42	87.1	68.7	95.5	12.4	13.1
0.700	88	281	10	46	86.8	65.7	96.6	10.2	14.1
0.750	82	282	9	52	85.6	61.2	96.9	9.9	15.6
0.800	76	284	7	58	84.7	56.7	97.6	8.4	17.0
0.850	71	288	3	63	84.5	53.0	99.0	4.1	17.9
0.900	67	289	2	67	83.8	50.0	99.3	2.9	18.8
0.950	54	290	1	80	80.9	40.3	99.7	1.8	21.6
1.000	0	291	0	134	68.5	0.0	100.0	.	31.5

The column labeled correct [1] and Incorrect [2] give the frequency with which observations are respectively correctly and incorrectly classified as event or nonevent for each probability cutoff level. For instance, at cutoff level of 0.2 , the model correctly classifies 119 events (malignant cases) and 233 nonevents (benign cases). It incorrectly classifies 58 nonevents as events and 15 events as nonevents. The five percentages columns [3] measure the predictive accuracy of the model:

Correct gives the probability that the model correctly classifies the data at each probability cutoff level. In our example, at 0.2 cutoff level, 352 out of the 425

observations are correctly classified, which results in a correct percentage 82.8% (i.e. 352/425).

Sensitivity is a ratio of the number of correctly classified events over the total number of events (malignant cases). At the cutoff level 0.2, the sensitivity is 88.8% (i.e.119/134) as 119 out of the 134 events are correctly classified as events.

Specificity is a ratio of the number of correctly classified nonevents over the total number of nonevents (benign cases). At the cutoff level 0.2, the specificity is 80.1% (i.e. 233/291) since 233 out of 291 nonevents are correctly classified as nonevents.

False POS is the false positive rate, i.e. the ratio of the number of nonevents incorrectly classified as event over the number of observations classified as events.

False NEG is the false negative rate, i.e. the ratio of the number of events incorrectly classified as nonevents over the number of observations classified as nonevents.

The **Receiver operating characteristic (ROC)** curves provide visualization of the relationship between sensitivity and specificity of a test over all possible probability cutoff levels. A curve is constructed by plotting the sensitivity versus the false positive rate, or 1-specificity, for varying probability cutoff level.

The proportion of the whole area of the graph which lies below the ROC curve is a one-value measure of the performance of a test. The higher the proportion, the better the test. The area under the ROC curves have a nice statistical interpretation, i.e. it measures the probability of the classifier to correctly classify events and nonevents. Consider the situation in which patients are already correctly classified into two groups. If one randomly picks a sample X_m from the malignant group and one X_b from the benign group, and do the test, compute the model output $y(X_m)$ and $y(X_b)$ on both. The one with the more abnormal test result should be the one from the malignant group. The area under the curve is the percentage of randomly drawn pairs for which this is true (that is, the test correctly classifies the two patients in the random pair). The measurement is given by

$$\theta = \text{AUC} = \text{P}[y(X_b) < y(X_m)] = \frac{1}{N_m N_b} \sum_{k=1}^{N_b} \sum_{l=1}^{N_m} \left(y(X_b^{(k)}) < y(X_m^{(l)}) \right)$$

The uncertainty of the measurement is given by the Wilcoxon statistics:

$$SE(\theta) = \sqrt{\frac{\theta(1-\theta) + (N_m - 1)(Q_1 - \theta^2) + (N_b - 1)(Q_2 - \theta^2)}{N_b N_m}}$$

where
$$Q_1 = P\left[y(X_b) < y(X_m^{(1)}) \text{ I } y(X_b) < y(X_m^{(2)}) \right]$$
$$Q_2 = P\left[y(X_b^{(1)}) < y(X_m) \text{ I } y(X_b^{(2)}) < y(X_m) \right].$$

To compute the area under the ROC curves, two methods are commonly used. One is a parametric method using a maximum likelihood estimator to fit a smooth curve to the data points. What we use in this analysis is the other one: a non-parametric method based

on the Wilcoxon statistic, using the trapezoidal rule to approximate the area. This method also gives a standard error that can be used for comparing two different ROC curves (Hanley and McNeil).

Fig 2.1 shows the ROC curve for the fitted model: `meno colsc3 colsc4 l_ca125 asc sol irreg pap`. The curve rises quickly, indicating that the predictive accuracy of this logistic regression model is good. The area under the ROC curve (AUC) is 0.948.

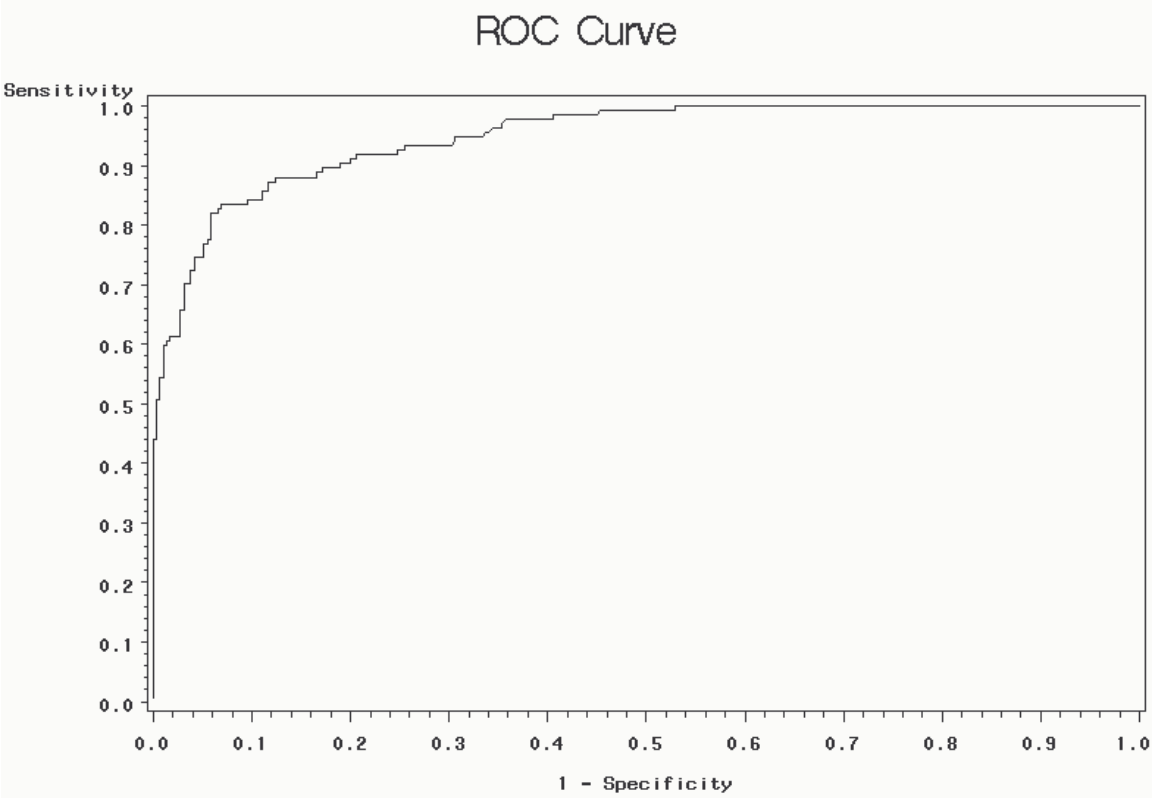


Fig 2.1 ROC curve

2.5 Model Validation

In order to test the generalization ability which means the ability to predict on the new data, we first take the first 300 observations out of the 525 as the training set, after eliminating the observations with missing value in the important predictor variables like l_ca125, we indeed have 265 observation in the training set. The remaining 225 observations will be taken as test set. However, after removing 90 observations with missing value, 160 are left for test.

Frequency table

	Benign (row %)	Malignant (row %)	Total (column %)
Training Set	185 (70%)	80 (30%)	265 (62%)
Test Set	106 (66%)	54 (34%)	160 (38%)
total	291 (68%)	134 (32%)	425 (100%)

The following is the partial output of model fitting on the training set:

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	326.601	119.587
SC	330.181	151.804
-2 Log L	324.601	101.587

R-Square 0.5690 Max-rescaled R-Square 0.8057

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	223.0140	8	<.0001
Score	171.6762	8	<.0001
Wald	44.7158	8	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-8.2509	1.3431	37.7364	<.0001
meno	1	0.7984	0.5495	2.1109	0.1462
colsc3	1	1.1884	0.5788	4.2163	0.0400
colsc4	1	2.4582	0.7922	9.6290	0.0019
l_ca125	1	0.4181	0.2054	4.1429	0.0418
asc	1	2.2811	0.7658	8.8735	0.0029
sol	1	4.7286	1.0907	18.7959	<.0001
irreg	1	2.1037	0.9147	5.2896	0.0215
pap	1	3.7761	0.8722	18.7414	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
meno	2.222	0.757	6.524

```
colsc3        3.282       1.056      10.205
colsc4       11.684       2.473      55.197
l_ca125       1.519       1.016       2.272
asc           9.787       2.182      43.900
sol         113.131      13.341     959.333
irreg         8.197       1.365      49.230
pap          43.645       7.897     241.208


Association of Predicted Probabilities and Observed Responses

Percent Concordant     97.2    Somers' D    0.945
Percent Discordant      2.7    Gamma        0.946
Percent Tied            0.1    Tau-a        0.400
Pairs                 14800    c            0.973

Hosmer and Lemeshow Goodness-of-Fit Test

   Chi-Square       DF     Pr > ChiSq
       5.8322        8         0.6660
```

We can see that the model still gives a good fit on the training set. c-value gives area under the ROC curve 0.973, which is rather high.

Testing this model on the complete new data, the resulting area under the ROC curve is 0.904.

Since the variable selection is not unique, we now try some other subset variables to build the model.

```
MODEL00: meno colsc3 colsc4 l_ca125 asc pap
MODEL0a: meno colsc3 colsc4 l_ca125 asc sol pap
MODEL0b: meno colsc3 colsc4 l_ca125 asc irreg pap
MODEL0c: meno colsc3 colsc4 l_ca125 asc smooth pap
MODEL1a: meno colsc3 colsc4 l_ca125 asc sol irreg pap
MODEL1b: meno colsc3 colsc4 l_ca125 asc sol irreg pap Shadows
MODEL2b: meno colsc3 colsc4 l_ca125 asc smooth pap bilat
```

	Training					Test
Model	#Var	AIC	SC	-2LogL	c	c
00	6	146.613	171.671	132.613	0.948	0.885
0a	7	124.031	152.669	108.031	0.969	0.879
0b	7	148.204	176.842	132.204	0.947	0.896
0c	7	126.545	155.183	110.545	0.966	0.907
1a	8	119.587	151.804	101.587	0.973	0.904
1b	9	117.100	152.897	97.100	0.973	0.904
2b	8	126.338	158.556	108.338	0.967	0.902

The important statistics in model fitting on the training set and classifying on the test set are shown in the table above.

We find that MODEL 0c has the highest c-value 0.907 on the test set, while its c-value 0.966 on the training set is slightly smaller than MODEL1a. This seems that it might also

be a good model for predicting the malignancy of the tumor. The partial output of the model fitting information for MODEL0c is listed below.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	326.601	126.545
SC	330.181	155.183
-2 Log L	324.601	110.545

R-Square 0.5541 Max-rescaled R-Square 0.7847

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	214.0555	7	<.0001
Score	166.0613	7	<.0001
Wald	48.8584	7	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-5.1125	0.9311	30.1464	<.0001
meno	1	1.2361	0.5055	5.9784	0.0145
colsc3	1	1.1917	0.5805	4.2142	0.0401
colsc4	1	1.9787	0.6521	9.2064	0.0024
l_ca125	1	0.5640	0.2023	7.7719	0.0053
asc	1	2.5546	0.7801	10.7246	0.0011
smooth	1	-3.7745	1.1348	11.0627	0.0009
pap	1	2.1222	0.5166	16.8730	<.0001

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
meno	3.442	1.278	9.272
colsc3	3.293	1.055	10.273
colsc4	7.233	2.015	25.968
l_ca125	1.758	1.182	2.613
asc	12.866	2.789	59.352
smooth	0.023	0.002	0.212
pap	8.349	3.033	22.984

Association of Predicted Probabilities and Observed Responses

Percent Concordant	96.6	Somers' D	0.932
Percent Discordant	3.4	Gamma	0.932

Percent Tied	0.0	Tau-a	0.394
Pairs	14800	c	0.966

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
1.5293	8	0.9922

2.6 Conclusion

So far we find out two good models for predicting the probability of tumors being malignant by logistic regression analysis.
The first logistic regression model **LR1**, containing 8 independent variables:

```
meno colsc3 colsc4 l_ca125 asc sol irreg pap,
```

can be written as:

$$\ln\frac{p}{1-p} = -8.251 + 0.798\times meno + 1.188\times colsc3 + 2.458\times colsc4 + 0.418\times l_ca125 + 2.281\times asc + 4.729\times sol + 2.104\times irreg + 3.776\times pap$$

The second model **LR2** has 7 independent variables:

```
meno colsc3 colsc4 l_ca125 asc smooth pap,
```

can be written as:

$$\ln\frac{p}{1-p} = -5.113 + 1.236\times meno + 1.192\times colsc3 + 1.979\times colsc4 + 0.564\times l_ca125 + 2.555\times asc - 3.775\times smooth + 2.122\times pap$$

Fig 2.2 (a) and (b) gives the ROC curve for the prediction on training set and test set, respectively, with logistic regression model LR1, LR2, and RMI. Table 2.2 gives the areas under the ROC curve and their standard deviation.

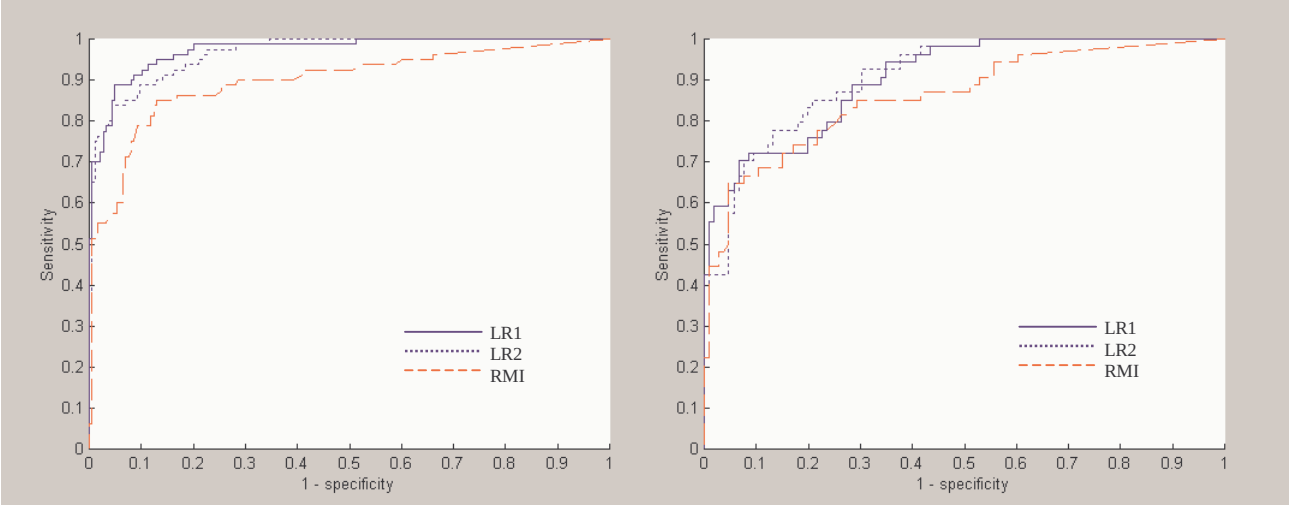


Fig2.2 (a) Training set

Fig2.2 (b) Test set

Utilizing the AUC value and its corresponding *SE*, we can conduct a pairwise *z*-test to see if the difference between models is significant or not. Table 2.3 report the *p*-value of the six *z*-tests to compare the performance of different models using both training set and test set. From these tests, we can confirm that the logistic regression model is significantly different from RMI when taking a retrospective test. However, when using the test set, the *p*-value doesn't give evidence to reject the null hypothesis that there is no significant

difference between LR model and RMI. The ROC curves shown in Fig2.2 (a) and (b) also prove this test result.

Table 2.2 Area Under the ROC curve (AUC) and its standard error

	Training		Test	
Model	AUC	SE	AUC	SE
RMI	0.898	0.0243	0.861	0.0343
LR 1	0.972	0.0130	0.904	0.0289
LR 2	0.966	0.0144	0.908	0.0285

Table 2.3 Resulting *p*-value of Pairwise Significant Difference *z*-test

Significance (p-value)	Training			Test		
Model	RMI	LR1	LR2	RMI	LR1	LR2
RMI	1	0.0014	0.0035	1	0.1446	0.1093
LR 1		1	0.2035		1	0.4194
LR 2			1			1

3 Artificial Neural Network Models

Artificial neural networks (ANNs) are networks of units called neurons that exchange information in the form of numerical values with each other via synaptic interconnections. Here we will use two types of feed-forward neural networks, which are known as universal approximators, to create a nonlinear mapping between a set of input variables and the output variables; they are multi-layer perceptrons and radial basis function neural networks. Many successful applications in pattern recognition using these two types of networks have been reported (Bishop).

- Multi-Layer Perceptron (MLP)

We begin by considering an example of a layered feed-forward neural network, whose architecture is shown in Fig. 3.1. In this network there are *d* inputs, *M* hidden units and *c* output units. The output of the jth hidden unit is obtained by first forming a weighted linear combination of the *d* input values and adding a bias, then transforming the weighted sum using an activation function g(.):

$$z_j = g\left(\sum_{i=0}^{d} w_{ji}^{(1)} x_i \right) \qquad (3.1)$$

Here w_{ji} denotes a weight in the first layer, going from input *i* to hidden unit *j*, and w_{j0}, denote the bias for hidden unit *j*, i.e. weight from extra variable $x_0 = 1$ to unit *j*.
The outputs of the network are obtained by transforming the activations of the hidden units using a second layer of processing elements:

$$y_k = \tilde{g}\left(\sum_{j=0}^{M} w_{kj}^{(2)} z_j \right) \qquad (3.2)$$

Combining (3.1) and (3.2), we obtain an explicit expression for the complete function represented by the network diagram in Fig. 3.1 in the form

$$y_k = \tilde{g}\left(\sum_{j=0}^{M} w_{kj}^{(2)} g\left(\sum_{i=0}^{d} w_{ji}^{(1)} x_i \right) \right) \qquad (3.3)$$

Note that, the activation functions can be linear or nonlinear, and can be different for different layers. The typical ones are the logistic sigmoidal function, tanh activation function, radial basis functions, and threshold functions etc.

The use of multi-layer perceptrons (MLP) can be seen as a generalization of the methodology of logistic regression analysis described in Section 2. Or, the logistic regression model can be seen as a special MLP that has no hidden layer, and takes a logistic sigmoidal function as its output activation function.

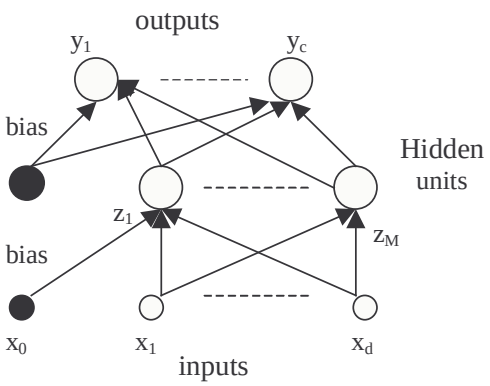


Fig 3.1 Architecture of MLP

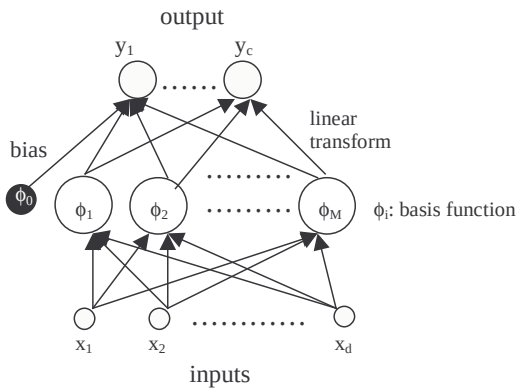


Fig 3.2 Architecture of RBF-NN

- Radial basis function network (RBF-NN)

Radial basis function networks are also feedforward, but have only *one* hidden layer. For the architecture shown in Fig 3.2, we can write the RBF-NN mapping from input vector $\boldsymbol{x}$ to kth output y_k in the following form:

$$y_k(x) = \sum_{j=0}^{M} w_{kj}\phi_j(x) \qquad (3.4)$$

where $\phi_j(x) = \exp\left(-\frac{\|x-\mu_j\|^2}{2\sigma_j^2}\right)$ for the case of Gaussian basis functions, and $\boldsymbol{x}$ is the d-dimensional input vector with elements x_i, and μ_i is the vector determining the center of basis function ϕ_j and has elements μ_{ji}. ϕ_0 is an extra basis function whose activation is set to 1. M is the number of hidden neurons. Its output activation function hence is a linear transform of the activations of the hidden neurons (or M basis functions).

The roles of the first and second layers of weights are different in radial basis function networks. This leads to a two-stage training procedure for training RBF-NNs. In the first stage the input data set $\{\boldsymbol{x}^N\}$ alone is used to determine the parameters of the basis function (e.g. μ_i and σ_i for the Gaussian basis function). In the second phase, the basis functions are kept fixed, while the second layer weights are found by supervised learning.

- Neural Network Training

The training of the feed-forward neural network (parameter estimation) is often done by an iterative backpropagation procedure, until the discrepancy between the target output t_k and actual response y_k is minimized. The commonly used error function which reflects this discrepancy within a set of N data, is the sum of squares error (*sse*) function

$$sse = \sum_{k=1}^{N}(t_k - y_k)^2 \qquad (3.5)$$

A properly trained neural network should have the capability to extract the unknown relationships from the training data and the generalization possibility towards unseen cases. Overfitting occurs when the error on the training set is driven to a very small value, but when new data is presented to the network the error is large. The more complex of the neural network, the higher the risk for overfitting. We will consider this generalization problem in both the architecture design and training for the following two types of neural networks for this classification problem.

3.1 Network Design and training

3.1.1 Multi-layer Perceptron

In order to avoid overfitting, we take only the variables selected by the logistic regression (8 variables in Model 1 and 7 variables in Model 2 as described in Section 2.6) as the candidate input variables. Only one hidden layer and three hidden neurons are used. The activation function for both hidden layer and output layer is the logistic sigmoidal function of the form:

$$g(a) = \frac{1}{1+\exp(-a)} \qquad (3.6)$$

The architecture of the MLP, we use in the simulation is illustrated in Fig. 3.3.

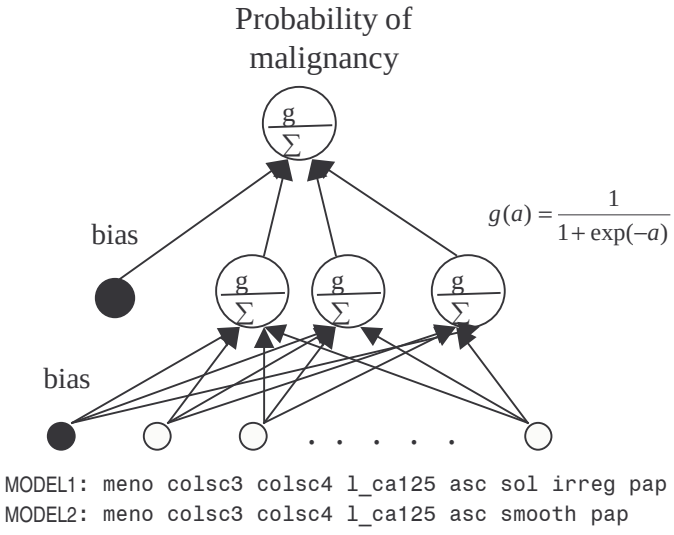


Fig 3.3 Architecture of MLPs for Predicting Malignancy of Ovarian Tumor

We use matlab function *trainbr* to train the MLP. This procedure updates the weight and bias values according to the Levenberg-Marquardt optimization. The error function it tends to minimize is a combination of the sum of squared error (*sse*) and the sum of

squared weights (ssw): $E_{reg} = \alpha sse + \beta ssw$, where $ssw = \sum_{j=1}^{N} w_j^2$, α and β are the regularization hyperparameters which are determined by a Bayesian approach. Under the Bayesian frame work, the weights and biases of the network are assumed to be random variables with specified distributions. The regularization parameters are related to the unknown variance with specified distributions.

For notational convenience, we call the MLPs with 8 input variables of Model1 as MLP1, for the MLPs with 7 input variables of Model2 as MLP2.

3.1.2 Generalized Regression Neural Networks

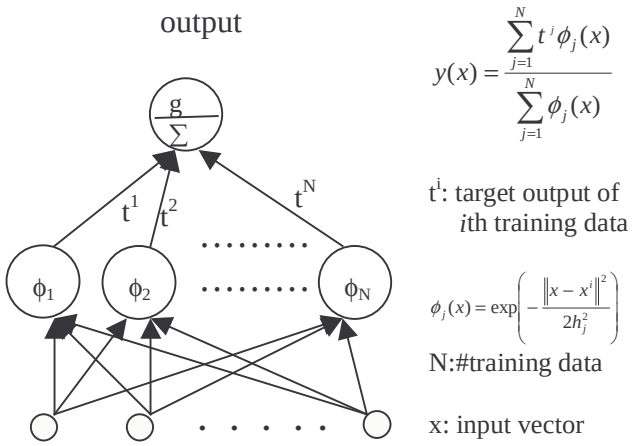
The generalized regression neural networks (GRNNs) are the paradigms of RBF-NNs, often used for function approximations. It's another term for Nadaraya-Watson kernel regression, and has the following form for the function mapping (Bishop).

$$y(x) = \frac{\sum_n t^n \exp\{-\|x - x^n\|^2 / 2h^2\}}{\sum_n \exp\{-\|x - x^n\|^2 / 2h^2\}} \qquad (3.7)$$

GRNNs share a special property, namely that they do not require iterative training; the hidden-to-output weights are just the target values, so the output is simply a weighted average of the target values of training cases close to the given input case. It can be viewed as a normalized RBF network in which there is a hidden unit centered at every training case. These RBF units are called "kernels" and are usually probability density functions such as the Gaussians considered in (3.7). The only weights that need to be learned are the widths of the RBF units *h*. These widths (often a single width is used) are called "smoothing parameters" or "bandwidths" and are usually chosen by cross-validation. GRNN is a universal approximator for smooth functions, so it should be able to solve any smooth function-approximation problem given enough data. The main drawback of GRNN is that, like kernel methods in general, it suffers seriously from the curse of dimensionality. GRNN cannot ignore irrelevant inputs without major modifications to the basic algorithm.

Here we use the function *grnn* in the Matlab6 neural network toolbox to create a radial basis function neural network. The input variables are again selected from the logistic regression process: Model1 and Model2, with 8 and 7 variables respectively.

We denote the GRNN with 8 input variables of Model1 as GRN1, while the GRNN with 7 input variables of Model2 is denoted by GRN2.




```
MODEL1: meno colsc3 colsc4 l_ca125 asc sol irreg pap
MODEL2: meno colsc3 colsc4 l_ca125 asc smooth pap
```

Fig 3.4 Architecture of GRNNs for Predicting Malignancy of Ovarian Tumors

3.2 Simulation Results

The training and simulation of the above 4 neural networks, which encompass 2 kinds of architectures and 2 sets of input variables, are done with the neural network toolbox of Matlab6.

The overall data both for input variables and output variable are first preprocessed: the continuous variable l_ca125 is standardized and the binary variables {0,1} are transferred to {-1,1}, since the two algorithms perform best on data within [-1, 1].

The whole dataset is first split into two parts, the first 300 data are used for training and the remaining 225 cases for testing. After removing the data with missing value in l_ca125, the final training set is composed of 265 observations, and the test set has 160 observations. This process is the same as the one for logistic regression model building.

For MLPs, the initial values of the weights and biases are randomly chosen from a normal distribution with mean zero and variance one. The training is repeated 100 times with different initializations; the parameters of the MLP which exhibit the best performance, i.e. the one with the highest AUC on the test set, will be taken as the final parameters of the MLP.

For GRNN, by searching in the interval [0.5, 5], the optimal value for the width of the radial basis function can be chosen. When the width is set to 3, GRN1 achieves the best performance on the test set with AUC = 0.9111; while GRN2 reaches its best performance with AUC = 0.9050 if the width equals 2.7. Note that, unlike logistic regression and MLPs, since GRNNs are doing function approximation, the output value

is within [-1, 1]. The ROC curve is created from this output, but the output can not be interpreted as a probability within the range of [0, 1].

Table 3.1 reports the results of the four neural networks, both on the training set and test set. The performance of Risk of Malignancy Index (RMI, computation see section 0.2.) and two logistic regression models (LR1 and LR2) are also shown for comparison.

Table 3.1 Area Under the ROC curve (AUC) and its standard error

	Training		Test	
Model	AUC	SE	AUC	SE
RMI	0.898	0.0243	0.861	0.0343
LR 1	0.972	0.0130	0.904	0.0289
LR 2	0.966	0.0144	0.908	0.0285
MLP1	0.975	0.0123	0.924	0.0261
MLP2	0.964	0.0149	0.917	0.0271
GRN1	0.966	0.0145	0.911	0.0280
GRN2	0.968	0.0141	0.905	0.0288

4 Performance Measure and K-Fold Cross-Validation

The most commonly used performance measure of a classifier or a model is the classification accuracy, that is the probability of correctly classifying a randomly selected instance. But it assumes equal misclassification costs (for false positive and false negative errors), and assumes that the class distribution in the target environment will be constant and relatively balanced. Both of these assumptions are not suitable in real-world problems. Unlike classification accuracy, ROC is independent of class distributions or error costs and has been widely used in the biomedical field. Furthermore, a one-value measure, namely the area under the ROC curves (AUC) which is independent of the selection of cutoff values, can be extracted from the ROC. Hence in this study, ROC analysis is conducted, and AUC is taken as measure for assessing the performance of different models.

So far, we have done the prospective study by taking the data of the 300 first treated patients as training set. The test set contains the data of patients who were treated more recently. But how much confidence we can get from these results? How representative is the test set we choose in this way? To answer this question, we first do some experiments to train and test different subsets of the data. The experimental results require an evaluation of the performance in a statistical way. Cross validation and bootstrap are commonly used in estimating the classification accuracy (Kohavi). Here we will apply *k*-fold cross-validation to obtain an estimate of the expected ROC.

4.1 ROC Analysis on Different Subsets

Practically, the data set is increasing with time. Very often we take the existed data set for training and test on the data set of newly treated patients. Here we will also follow this tendency and do the training and testing 'incrementally'. These experiments have three stages:

1) In the first stage, the accessible data set included only the first 191 observations, of which 173 data without missing l_ca125 (which is the logarithm of the serum marker ca125 level). Hence the 173 data are split up into a training set with the first 116 observations, and a test set with the later 57 ones.
2) In the second stage, the available data increased to the first 300 observations, of which 265 observations without missing important predictive values are used. This time, we can take the oldest 173 data for training and the new 92 ones for test.
3) In the last stage the whole dataset of 525 observations was available. The splitting of the training set and test set is performed in the same way as described in section 3 and section 2.

We do these experiments only for logistic regression, and the performance of Risk of Malignancy Index is also calculated in different stages as a baseline. Note that, the proportions of the malignant cases among different subsets are different, as here we only

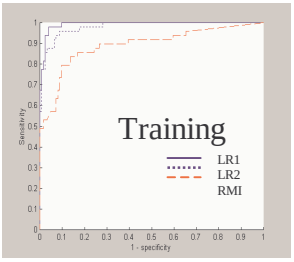
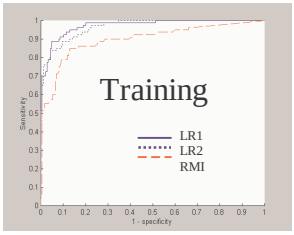
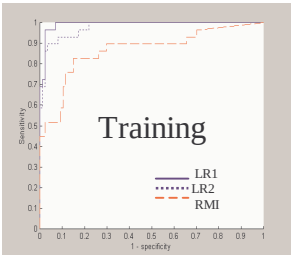
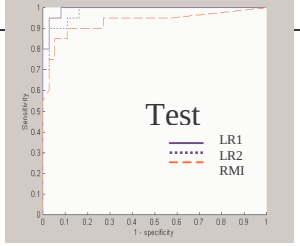
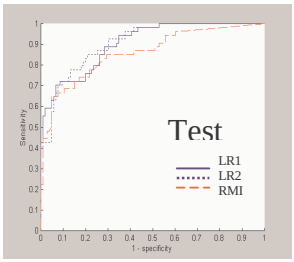
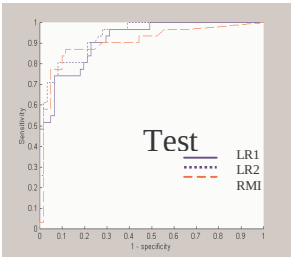
do the simplest split of the data set without randomizing and stratifying. The experimental results are shown in the table 4.1. For each result column, the white block reports the retrospective performance on the training set; the grey block reports the performance on the test set. The underlined number is the size of the training set or test set. The number of malignant cases and benign cases are reported below the underlined number. Also reported are the areas under the ROC curves (AUCs) computed from different model outputs and their corresponding standard errors *SE* according to the method proposed by Hanley and McNeil. The ROC curves on the training set and test set are also plotted in the table respectively.

From this table, we can see that logistic regression models outperform RMI consistently, in the sense that they have higher AUCs than RMI. The difference between the performance of RMI and logistic regression is significant on the training set while it's not always the case for performance on the test set.

One remarkable problem is that there exists a large variance between the accuracy estimation of the same models over different stage of experiments. Even the AUCs of RMI vary from 0.86 to 0.93. For the two logistic models, we got very high AUCs on the first stage prospective test (around 0.98), relatively low at the third stage prospective test (around 0.90).

To solve this problem, we will use k-fold cross validation to evaluate the expected performance of different models, and give the confidence interval of the estimation.

Table 4.1 ROC of Models on Different Subsets of Data

	Stage 1	Stage 2	Stage 3
1 116	<u>116</u> Malignant: 29 Benign: 87 RMI 0.87 ± 0.045 LR1 0.99 ± 0.012 LR2 0.98 ± 0.020	<u>173</u> Malignant: 49 Benign: 124 RMI 0.89 ± 0.032 LR1 0.99 ± 0.0091 LR2 0.98 ± 0.014 	 <u>256</u> Malignant: 80 Benign: 185 RMI 0.898 ± 0.024 LR1 0.972 ± 0.013 LR2 0.966 ± 0.014 MLP1 0.950 ± 0.017 MLP2 0.965 ± 0.014 GRN1 0.966 ± 0.014 GRN2 0.968 ± 0.014
173 265	<u>57</u> Malignant: 20 Benign: 37 RMI 0.93 ± 0.041 LR1 0.99 ± 0.014 LR2 0.98 ± 0.022  	<u>92</u> Malignant: 31 Benign: 61 RMI 0.90 ± 0.038 LR1 0.91 ± 0.036 LR2 0.93 ± 0.032	
425			<u>160</u> Malignant: 54 Benign: 106 RMI 0.861 ± 0.034 LR1 0.904 ± 0.029 LR2 0.908 ± 0.029 MLP1 0.924 ± 0.026 MLP2 0.917 ± 0.027 GRN1 0.911 ± 0.028 GRN2 0.905 ± 0.029

4.2 ROC Analysis with K-Fold Cross-Validation

Let's first introduce how to extract the area under the ROC curve from k-fold cross-validation, and how to construct the expected ROC curves with confidence intervals. Our design and results of the experiments are described afterwards.

4.2.1 AUC from Cross-Validation

Holdout method is one kind of often used cross-validation, it partitions the data into two mutually exclusive subsets namely training set and test set. This is what we have used in the previous experiments. One can also repeat the holdout method for *k* times, each time a different partition is chosen, then the estimated AUC is derived by averaging over all the runs. However, in medical practice, the holdout method makes inefficient use of the data set, which is usually smaller than desired. For example, one third of the data set is not used for training the classifier.

K-fold cross-validation is a variant of cross-validation. The data set is randomly divided into *k* ($k>1$) mutually exclusive subsets (*k* folds) of approximately equal size. The model is trained on all the subsets except for one, and the validation error is measured by testing it on the subset left out. This procedure is repeated k times, each time using a different subset for validation. The performance of the model is assessed by averaging the AUC under validation over the *k* estimates. Repeating the *k*-fold cross-validation for multiple runs can provide a better statistical estimate.

The cross-validation estimate is a random number that depends on the division of the data set. We hope that the estimates have low bias and low variance. Leave-one-out is a special k-fold cross-validation, in which the number of folds equals the number of available data. This method is almost unbiased, but has high variance, leading to unreliable estimates. When choosing the number of folds, we would like to tradeoff bias for low variance.

4.2.2 Combining ROC curves

There are mainly two ways to generate an expected ROC curve from k-fold cross-validation. One is known as pooling proposed by Swets and Pickets (Bradely 1997), in which the *i*th points making up each raw ROC curve are averaged.

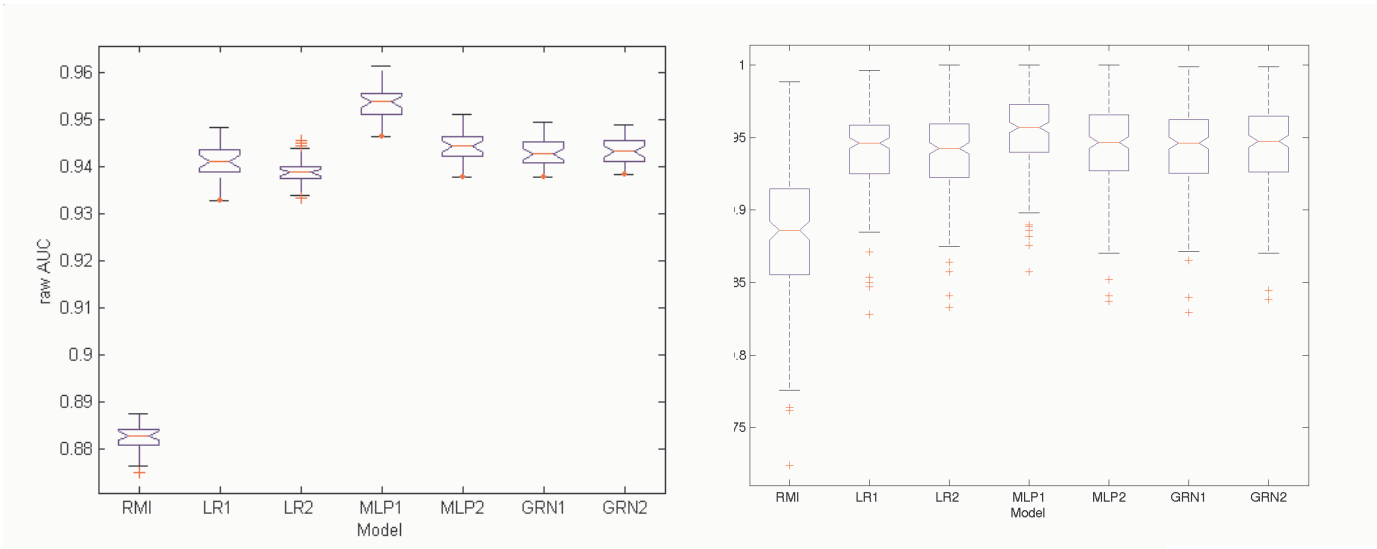
What we will use is another method, called averaging (Provost 1998). It averages the ROC curves in the following manner. For *k*-fold cross-validation, the ROC curve from each of the folds is treated as a function, R_i, such that $TP = R_i(FP)$, where TP is the true positive rate, FP, is the false positive rate, the points in ROC space correspond to its (FP, TP) pair. R_i is obtained by linear interpolations between points in the ROC space (if there are multiple points with the same FP, the one with the maximum TP is chosen). The

averaged ROC curve is the function R(FP) = mean(R_i(FP)). The confidence intervals of the mean of TP are computed under the assumption of binomial distribution.

4.2.3 Experimental Results

As we have 425 data and two classes, we developed a stratified cross-validation with *k*=7. Stratification forces an equal proportion of malignant cases (31.7%) in each fold. For each subset of the data, a model is developed with around 365 data samples in the training set and 60 in the test set. This procedure is repeated 30 times, by randomly dividing the dataset into seven stratified folds.

In this experiment, the architecture of the neural network and training method is kept the same as the one we used in section 3. To accelerate the training of MLPs, we limit the maximum training epochs to 60, since the previous experiments show that training of MLPs in this case normally converges within 60 times. The width of all the GRNNs is set to 3 for both models. The estimated AUC for each trial of 7-fold cross-validation is the average of AUCs over all the 7 validations, denoted to mAUC. Then the mean of the 30 averaged AUCs, i.e. mAUCs, and variance can be computed. The experimental results including mAUCs and raw AUCs for logistic regression models, multi-layer perceptrons and generalized regression networks with two subsets of variables are shown below.



(a) **(b)**

Fig. 4.1 Box plot of (a) mean AUC and (b) raw AUC from different models.
The lower and upper lines of the "box" are the 25th and 75th percentiles of the sample. The distance between the top and bottom of the box is the interquartile range. The line in the middle of the box is the sample median. The "whiskers" are lines extending above and below the box. They show the extent of the rest of the samples (unless there are outliers). Assuming no outliers, the maximum of the sample is the top of the upper whisker. The minimum of the sample is the bottom of the lower whisker. An outlier is a value that is more than 1.5 times the interquartile range away from the top or bottom of the box. The plus sign at the top of the plot is an indication of an outlier in the data. The notches in the box are a graphic confidence interval about the median of a sample.

In Fig 4.1, (a) shows the boxplot of 7 groups of mAUCs, (b) is the boxplot of 7 groups of raw AUCs. We can visually see that the best model among the six is MLP1, which has the highest mean AUC and a comparatively small variance. The predictive variables in Model1 (with 8 variables) seem to generally have slightly higher AUC than the variables in Model2 (with 7 variables).

We will next use Analysis of Variance (ANOVA) techniques to test the equality of means of mAUC over several model-building methods simultaneously (Neter, 1996). If the one-way ANOVA test on AUC provides evidence that all the means of AUCs are not equal, then we can further use a multiple comparison procedure to determine which pairs of means are significantly different.[1] Here the Tukey-Kramer procedure is chosen to perform this task. For the difference between two subsets of means to be significant it must exceed a certain value. This value is called the 'honest significant range' for the k means, R_k, and for equal sample size n, is given by

$$R_k = q(1-\alpha; k, nk-k)\sqrt{s^2/n}$$

where $q(1-\alpha;\ k,\ nk-k)$ is the $100(1-\alpha)$th percentile derived from the studentized range distribution, s^2 is the estimate of the common variance over all the nk samples.

To be consistent with the boxplots above, we will do two Tukey multiple comparisons. We first compare mAUC, each group or method has 30 mAUCs from 30 replications of the 7-fold cross-validation. The subsets of adjacent means that are not significantly different at 95% confidence level are shown in Table 4.2 (a), and are indicated by drawing a wavy line under the subsets. The second multiple comparison is done onto the raw AUCs, each group has 30x7=210 raw AUCs, results are shown in Table 4.2 (b).

Table 4.2(a) Rank ordered significant subgroups from multiple comparison on mAUC

Models	RMI	LR2	LR1	GRN1	GRN2	MLP2	MLP1
mAUC							
Mean	0.8824	0.9394	0.9413	0.9429	0.9436	0.9444	0.9536
SD	0.0030	0.0031	0.0035	0.0030	0.0027	0.0030	0.0031

Table 4.2(b) Rank ordered significant subgroups from multiple comparison on raw AUC

Models	RMI	LR2	LR1	GRN1	GRN2	MLP2	MLP1
AUC							
mean	0.8824	0.9394	0.9413	0.9429	0.9436	0.9444	0.9536
SD	0.0458	0.0285	0.0282	0.0285	0.0279	0.0294	0.0259

It's clear that from both tables, AUCs generated from LRs, MLPs, or GRNNs are all significantly different from RMI. AUCs from MLP1 are significantly different from all the others. If we check Table 4.2(a), we find that mAUCs from neural networks (MLPs

[1] One can also perform a series t-test, one for each pairs of means, i.e. for 7 groups of models, at least 14 *t*-tests need to be conducted. However, simultaneous inference using different intervals of level 1-α does not lead to a simultaneous confidence interval 1-α.

and GRNNs) are significantly different from those computed from the logistic regression models too. On the other hand, the raw AUCs of these models do not exhibit significant differences. However, the conclusion can still be drawn that neural networks have the potential to give a higher AUC for predicting the malignancy of the tumor than RMI and LR models.

Note that, the statistical tests we've done in this section are different from the *z*-test that we did based on one AUC value and its *SE* (standard error) in section 2.6.

We then construct the expected ROC curves by averaging the 210 ROC curves in the 30 replications of 7-fold cross-validation. Though the mean AUC from each replicate of 7-fold cross-validation has very small variance, but this variance results mainly from the way to split data. What we are more interested in is the variance resulting from the limited size of the available data set (134 malignant cases and 291 benign cases) and the method we use to generate the models. Hence the 95% confidence interval of each point at the expected ROC curve, i.e. for a certain FP, the confidence interval of corresponding TP is computed based on 134 malignant samples under the assumption of binomial distribution. Fig. 4.2 shows the expected ROC curves (bold solid) of different models and their 95% confidence intervals (dashed). The points distributed in the ROC space are the points from 30x7 raw ROC curves. Fig. 4.3 illustrates only the expected ROC curves, the expected ROC curve of RMI can be seen as a baseline for comparison.

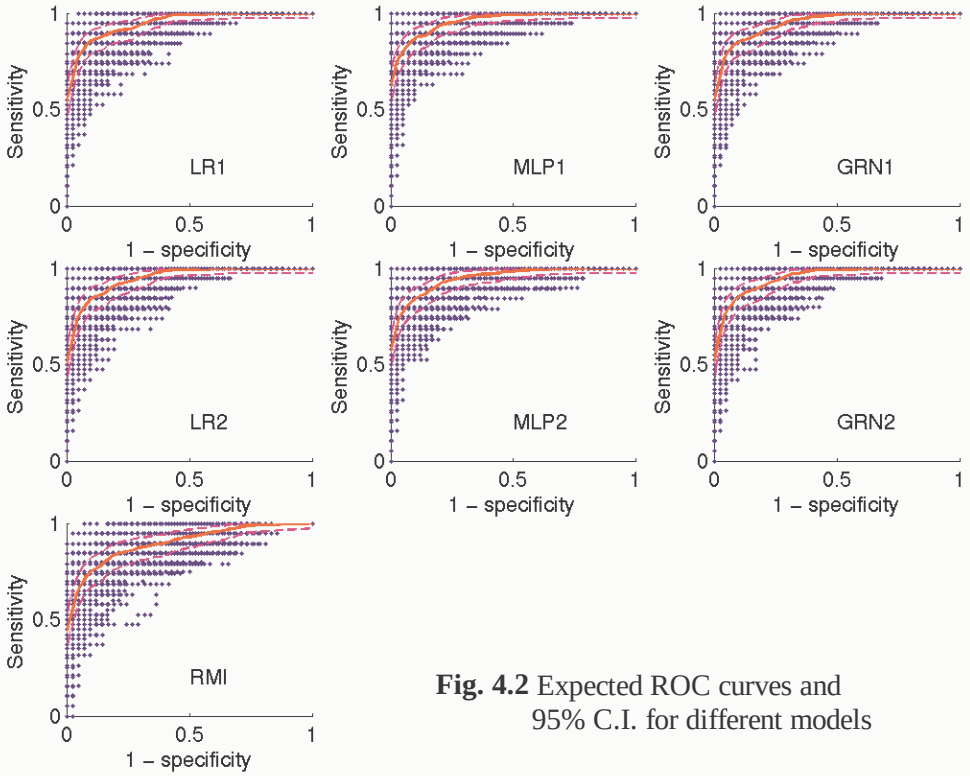


Fig. 4.2 Expected ROC curves and 95% C.I. for different models

Again, all these expected ROC curves can't be distinguished easily, they look close to each other. However, we can still see that the expected ROC curve of RMI is obviously below the other ROCs generated from the other methods. And the expected ROC curve of MLP1 is always slightly higher than the other curves in the most interesting region (where TP is considerably high while FP is low).

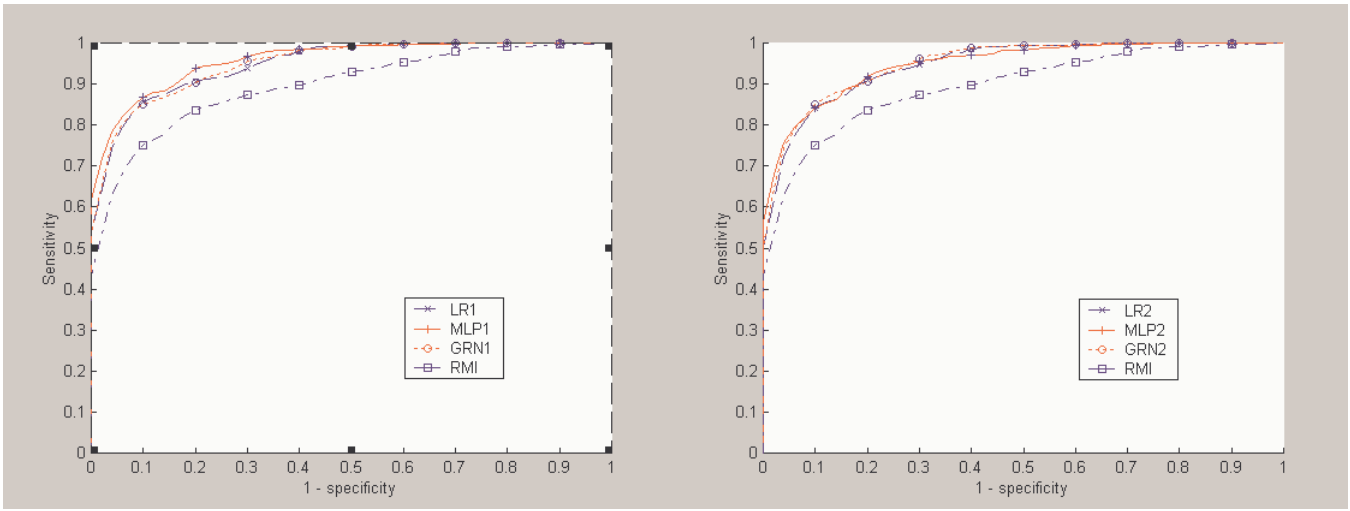


Fig. 4.3 (a) Expected ROC curves

Fig. 4.3 (b) Expected ROC curves

5 Conclusions

5.1 Conclusion from This Study

By doing logistic regression analysis, we first select two sets of input variables as the predictor candidates. Then we use three methods, i.e. logistic regression, multi-layer perceptrons and generalized regression neural networks, to build models with these two sets of variables. The performance of the models is then measured according to the ROC analysis. Both, a single-value metric AUC and the ROC curves have been shown. In order to make a statistical comparison between the methods, and see which methods have the potential for generating a better model in predicting the malignancy of tumor, multiple runs of 7-fold cross-validation were applied.

The experimental results indicate that the predictor candidate MODEL1 has slightly higher predictive power than MODEL2. It has eight input variables:

meno, colsc3, colsc4, l_ca125, asc, sol, irreg, pap.

By doing one-way ANOVA followed by multiple comparison, we conclude that: all the models including LRs, MLPs and GRNs, have higher expected AUCs than the risk of malignancy index (RMI); the multi-layer perceptrons have higher expected AUC and smaller variance than the models generated from the other methods.

5.2 Related Research and Publications

This research was initiated in 1997, and the size of dataset steadily increased. A lot of related research has been done and a number of papers have been published since then. Here I would like to summarize some issues about the model building that we have encountered in this study and previous research.

- Model building algorithms: logistic regression models [1][2], boosting (logistic regression models) [9], MLPs [1][3][4][5], support vector machines [8], and Bayesian network models [6] have been applied previously. In this study, besides LRs and MLPs, we have also tried GRNNs.

- MLP training: quasi-Newton optimization, Levenberg-Marquardt optimization and simulated annealing techniques have been used [1][3][4][5][9]; early stopping and Bayesian regularizer [5][9] were applied for regularization. In this report, Levenberg-Marquardt optimization combined with Bayesian regularizer were chosen, since this method converges faster than the other ones in this case and has no problem of choosing an additional validation set as required by early stopping technique. The cost functions that have been experimented previously include the weighted MSE in which weight for errors in malignant cases is 2 times as high as in the benign ones. In this report, simple MSE without special weighting for malignant cases was chosen, since the early experiments show no obvious improvement in the AUC by using weighted MSE. Another way of training MLP has also been tried in [4] by maximizing the AUC using Boltzmann simulated annealing technique directly, this training style gives comparable result as the one by minimizing MSE. Besides

maximum likelihood estimation as mentioned above, Bayesian learning techniques can also be used to find the parameters of the model, which has been reported in [6][10].

- Input variables: the input variables selected based on the larger data set (425 observations), are similar to the one obtained based on only the first 173 observations [1][2]. In the previous study, the input variables selected by logistic regression were used mostly for logistic regression. When building models for MLPs, variables were selected by exhaustive search (either by maximum likelihood estimation procedure or Bayesian learning procedure [6][10]). Some practical and medical consideration might also be taken into account. Bayesian network models are white box models, their variable selection would be much more related to the prior knowledge [7].

- Performance measure: the ROC curve analysis is done for evaluation of the model performance in most researches. Normally, 1/3 of data is selected randomly as test set, the performance of the trained model is given by the AUC on the test set. In [7], the performance of the Bayesian model is given by the averaged AUC over 1000 cross-validations (75% of the data set is used for training, 25% used for test). In this analysis, the expected AUC of each method for building models is extracted from multiple runs of 7-fold cross-validation.

- Comparison of AUC: for AUC obtained from a single test, a *z*-test proposed by Hanley and McNeil is usually conducted to compare the two AUCs. For comparing several groups of AUCs obtained from k-fold cross-validation, as the case in this study, an ANOVA technique is used followed by a Tukey multiple comparison procedure.

5.3 Future Work

The performance of LS-SVMs on this classification problem is still interesting to consider. Moreover, experimental results suggest that, more variables should be used and prior knowledge need to be embedded into the model for more accurate prediction. For example, a hybrid methodology might be more promising which combines the advantages of Bayesian network models (understandable knowledge representation), and the black-box models (efficiently learnable representation) [7]. Another direction might be committee machines, which use the strategy of divide and conquer, distributing the learning task among a number of experts, such as hierarchical mixture of experts.

References

Christopher M. Bishop (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Kohavi, Ron (1995). *A study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection*, in the International Joint Conference on Artificial Intelligence (IJCAI).

Andrew P. Bradley (1997). *The use of Area under the ROC Curve in the evaluation of Machine Learning Algorithms.* Pattern Recognition. Vol 30. No. 7 pp 1145-1159, 1997.

Foster Provost, Tom Fawcett, Ron Kohavi (1998). *The Case Against Accuracy Estimation for Comparing Induction Algorithms.* Proceedings of the Fifteenth International Conference on Machine Learning (IMLC-98), Madison, WI, 1998.

Hanley J.A. and McNeil B. (1982). *The Meaning and Use of the Area Under a Receiver Operating Characteristic Curve.* Diagnostic Radiology, vol. 143, no. 1, pp.29-36, 1982.

Hanley J.A. and McNeil B. (1983). *A method of comparing the area under the receiver operating characteristics curves derived from the same cases*, Radiology, Vol 148, pp. 839-843, 1983.

John Neter, Michael H. Kutner, Christopher J.Nachtsheim, William Wasserman (1996). *Applied Linear Statistical Models*, fourth edition. WCB/McGraw-Hill.

Related Publications:

[1] Jos De Brabanter (1997). *Logistic Regression and Artificial Neural Network Models for Predicting Malignancy in Ovarian Tumors: a Statistical Analysis.* Master thesis, supervisor: S. Van Huffel , K.U.Leuven 1997.

[2] Dirk Timmerman, Thomas H. Bourne, Anil Tailor, William P. Collins, Herman Verrelst, Kamiel Vandenberghe, and Ignace Vergote (1999). *A comparison of methods for preoperative discrimination between malignant and benign adnexal masses: The development of a new logistic regression model.* Am J Obstet Gynecol 1999.

[3] D. Timmerman, H. Verrelst, T.H.Bourne, B. De Moor, W.P.Collins, I. Vergote and J.Vandewalle (1999). *Artificial neural network models for the preoperative discrimination between malignant and benign adnexal masses.* Ultrasound Obstet Gynecol 1999; 13:17-25.

[4] Verrelst H., Moreau Y., Vandewalle J. Timmerman D. (1997) *Use of a Multi-Layer Perceptron to Predict Malignancy in Ovarian Tumors.* Proceedings of NIPS'97, Denver, MIT Press, Colorado, USA, Dec 1997, pp. 978-984.

[5] E.Lerouge, S. Van Huffel (1999). *Generalization Capacity of Neural Networks for the Classification of Ovarian Tumors.* Proceedings of the 20th Symposium on Information Theory in the Benelux, Haasrode, Belgium. May 27-28, 1999, pp 149-156

[6] Herman Verrelst, Joos Vandewalle, Bart De Moor. *Bayesian Input Selection for Neural Network Classifiers.* In Proceedings Of the Third International Conference on Neural Networks and Expert Systems in Medicine and Healthcare (NNESMED'98), Pisa, Italy, Sep. 1998, pp. 125-132.

[7] P.Antal, H. Verrelst, D.Timmerman, Y. Moreau, S.Van Huffel, B.De Moor, I. Vergote (2000). *Bayesian Networks in Ovarian Cancer Diagnosis: Potentials and Limitations.* Proceedings 13th IEEE Symposium on Computer-Based Medical Systems, CBMS 2000, 22-24 June 2000, Houston, Texas, USA.

[8] Sabine Van Huffel (2000). *Cancer Diagnosis: Preoperative classification of Ovarian Tumors Using Support Vector Machines*. Lecture Notes for Case Studies in Biomedical Data Processing. November 2000, K.U.Leuven.

[9] Tom Bellemans, Ellen Lerouge (1998). *Trainingsaloritmen voor neurale netwerken toegepast op ovariumkanker classificatie*. Master thesis, supervisor: Joos Vandewalle, Sabine Van Huffel. K.U.Leuven, 1997-1998.

[10] Pascal Geerts, Sabien Geeraerts (1998). *Parameter-selectie voor ovariumkanker classificatie d.m.v. Bayesiaanse neurale netwerken*. Master thesis, supervisor: Joos Vandewalle, Sabine Van Huffel. K.U.Leuven, 1997-1998.