Applied and Computational Mathematics
2020; 9(4): 130-145
http://www.sciencepublishinggroup.com/j/acm
doi: 10.11648/j.acm.20200904.15
ISSN: 2328-5605 (Print); ISSN: 2328-5613 (Online)

Explore the Characteristics of Age, BMI and Blood Composition of Breast Cancer Patients Based on Multivariate Statistical Analysis

Ruixuan Dong

Department of Statistic, East China Normal University, Shanghai, China

Email address:
dongruixuan@163.com

To cite this article:
Ruixuan Dong. Explore the Characteristics of Age, BMI and Blood Composition of Breast Cancer Patients Based on Multivariate Statistical Analysis. *Applied and Computational Mathematics*. Vol. 9, No. 4, 2020, pp. 130-145. doi: 10.11648/j.acm.20200904.15

Received: July 12, 2020; **Accepted**: August 18, 2020; **Published**: August 22, 2020

Abstract: In this paper, through a series of analysis and testing of breast cancer detection data, the statistical rules of multiple objects and multiple indicators are analyzed in the case of their correlation. First of all, univariate diagnosis and multivariate diagnosis were performed on the data. Among them, when studying the correlation between variables, it was found that HOMA had a clear linear positive correlation with insulin content in blood. It is worth noting that some patients with breast cancer show a high degree of insulin resistance and blood insulin content, which is a feature not found in samples without breast cancer. Then, through single factor analysis of variance, we believe that there were significant differences in blood test conditions, ages, and BMI indicators of samples of different health conditions. Next, the principal component analysis was used to reduce the dimension of the data. In this study, the differences in age, BMI, and blood component content between the two groups with different health conditions can be summarized by these two independent factors. Among them, the absolute value of the MCP-1 (monocyte chemoattractant protein 1) coefficient in the main component 1 is large, reflecting the characteristics of the blood component of the sample; the load values of glucose and leptin in the main component 2 are large, reflecting similar results. Then, assuming the use of m = 3 factor model and the use of maximum likelihood method and principal component method, the original data and factor rotation data are re-analyzed, so that the variables are reduced to 3 factors for analysis. Among them, the maximum likelihood method is used to estimate the factor rotation data. The first factor reflects the insulin resistance factor attributed to insulin and HOMA indicators, and the second factor reflects the body fat and thin factor attributed to BMI and leptin. The third factor reflects the glucose content in the blood. Finally, by setting different misjudgment costs for discriminant analysis, the obtained APER is 0.1638 and EAER is 0.1872. Among them, the probability of discriminating patients with breast cancer from not having breast cancer is 0.09375, which is a low rate of misjudgment and also means the model established in this paper is efficient.

Keywords: Data Diagnosis, One-Way MANOVA, Principal Component Analysis, Factor Analysis, Discriminant Analysis

1. Introduction

Breast cancer is becoming a leading cause of death among women in the whole world, meanwhile, it is confirmed that the early detection and accurate diagnosis of this disease can ensure a long survival of the patients [1]. The breast cancer incidence and mortality rates among Chinese women were increasing rapidly, especially in rural area during the recent 10 years, though they were still in low level worldwidely. The distribution of breast cancer incidence and mortality among Chinese women by age and district were showing significant characters [2].

Yang Ling et al estimated and predicted the incidence and mortality of breast cancer in China in recent years using a log-linear model, and concluded that due to the multiple effects of risk factors, population growth and aging, breast cancer will be one of the most growing malignant tumors in China [3].

Therefore, combined with relevant factors to accurately diagnose individuals who check whether they have breast

cancer, breast cancer patients can be screened as early as possible, so that patients can start treatment as soon as possible. M. Eskelinen et al. compared the testing power of 7 tumor markers CEA, AFP, CA15-3, TPS and NEU in the diagnosis of breast cancer [4]. Moreover, Na Liu et al. established a novel intelligent classification model for breast cancer diagnosis, which employed information gain directed simulated annealing genetic algorithm wrapper (IGSAGAW) for feature selection [5], while a rough set (RS) based supporting vector machine classifier (RS_SVM) is proposed for breast cancer diagnosis [1].

However, the several studies mentioned above are based on some difficult-to-obtain indicators for modeling, and cannot be used for preliminary screening of breast cancer under more general conditions. To solve this problem, we have to develop an easier way. Form the data collected by the University Hospital Centre of Coimbra [6], some of the more accessible indicators are fully displayed, and the samples are also divided into those with breast cancer and those without breast cancer.

Therefore, the main purpose of this article are

Explore the relationship between the content of seven blood components and age, BMI indicators, and test whether the sample comes from a multivariate normal population;

Check whether there is a significant difference between the values of variables of breast cancer patients and non-breast cancer patients by multivariate analysis method;

Reduce the data to obtain several principal components, and explore whether the differences in age, BMI, and blood component content between the two groups with different health conditions can be summarized by these several principal components;

Summarize 9 continuous variables into several types of indicators, so as to more intuitively reflect the relationship between variables;

Re-determine whether each examiner is a breast cancer patient according to the dependent variable, and calculate the misjudgment rate.

2. Data Description

2.1. Data Sources

The blood composition, age, illness, and BMI index data of 116 examiners selected in this paper are the test results from the University Hospital Centre of Coimbra [6].

2.2. Variable Introduction

The data selected in this paper contains 9 continuous variables and one binary variable. The continuous variables are all dependent variables and the binary variables are the corresponding variables. The names, dimensions and introduction of these variables are shown in *Table 1* below

Table 1. *Variable Introduction.*

Variable	Unit	Introduction
Age	years	The length of time a person survives from birth to calculation
BMI	kg/	At present, a standard commonly used in the world to measure the degree of body fat and thin and whether it is healthy
Glucose	mg/dL	The glucose in the blood is called blood glucose (Glu). Glucose is an important component of the human body and an important source of energy.
Insulin	U/mL	Insulin is the main hormone in the body that lowers blood glucose levels.
HOMA	-	Reflects the degree of resistance of cells in the body to insulin
Leptin	ng/mL	A hormone that controls weight and fat deposition by regulating metabolism and appetite
Adiponectin	g/mL	Adiponectin is a new type of protein specifically secreted by adipocytes [7].
Resistin	ng/mL	Fat cell production and secretion, used to reduce the body's sensitivity to insulin
MCP-1	pd/dL	A kind of small molecule proteins that play an important role in the physiological functions of the human body, mostly secreted by immune cells and glial cells, and have chemotactic activity [8].
Classification	-	Divided into two types with breast cancer and no breast cancer

3. Data Diagnosis

After understanding the basic situation of breast cancer, it is necessary to understand the basic data structure of breast cancer screening blood test. In actual work, the data we encounter is often the original data. These data are generally not clear and complete. For example, there may be missing data and the sample ID is not unique, which cannot be directly used for modeling. The data diagnosis work helps us understand the defects of the data in order to further clean up and integrate the data. The data diagnosis in this paper can be started from the aspects of completeness, accuracy, rationality, etc. The diagnostic methods include the following, the diagnosis of a single variable, and the diagnosis of the relationship between multiple variables.

3.1. Univariate Diagnosis

The first is the integrity problem of univariate. In some samples, some variable indicators are missing. This may be because the data is missing during the data sampling process, or it may be that the sample itself does not have the record of the indicator during the blood test. It is necessary to distinguish between these two situations. The meaning of the representative is different. If there are missing values in the modeling data, the missing values must be completed, or the records with missing values must be deleted before normal modeling can be performed. The results of univariate diagnosis are summarized in *Table 2* and *Table 3*

Table 2. *Univariate diagnosis of categorical variables.*

Variable	Category value	Total number of samples	Number of classification samples	Classification ratio
State of health	Healthy	116	52	44.83%
	Patients	116	64	55.17%

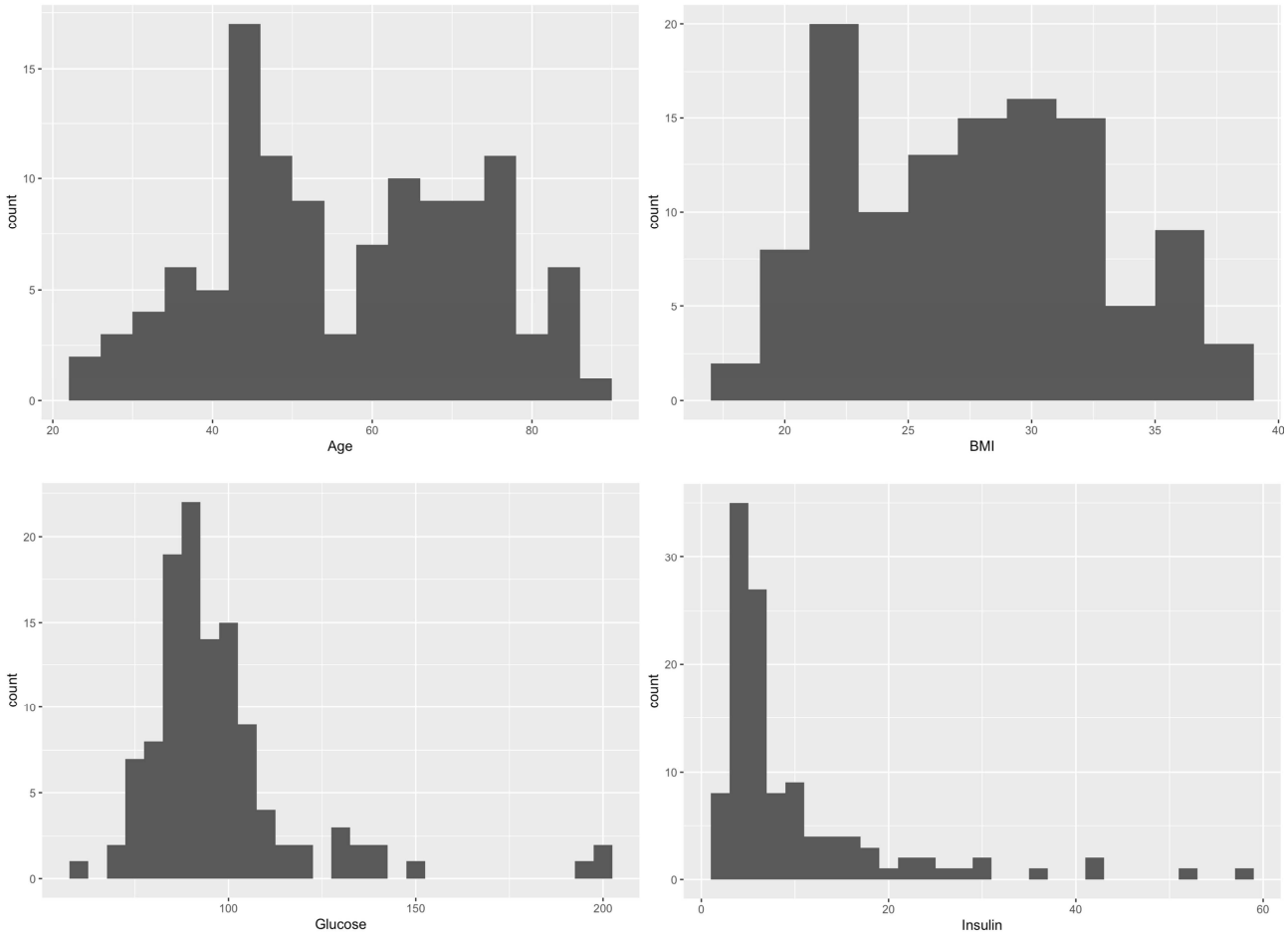
Remarks (1): HOMA variable is an evaluation index of islet β cells, the maximum value is 25.05, but the mean is only 2.69, and the standard deviation is very small. It can be inferred that there may be abnormal values or data accuracy problems that need further testing. Since HOMA is an evaluation index of islet β cells, it can also be verified whether it is related to the insulin index.

For each variable, the number of missing samples is 0, so the data performs perfect in terms of completeness, but whether the data has outliers needs further testing.

Then the data of the five variables are plotted as histograms respectively, and the results are shown in Figure 1 below. The results show that the blood glucose content is concentrated at 75-125 mg/dL, and there are also particularly large values, such as 201 mg/dL, but the number is very small; the blood insulin content, adiponectin content, resistin content and MCP-1 The content and HOMA indicators are also similar. Their distributions are right-biased, that is, there are large and particularly small values; the distribution of age and BMI is relatively uniform; the distribution of insulin content is similar to the HOMA distribution. The correlation coefficient is large, which shows that the two have a strong correlation.

Table 3. *Univariate diagnosis of continuous variables.*

Variable	Total number of sample	Number of missing samples	Mean	Standard deviation	Maximum	Minimum	Media	Remark
Age	116	0	57.30	16.11	89.00	24.00	56.00	
BMI	116	0	27.58	5.02	38.58	18.37	27.66	
Glucose	116	0	97.79	22.53	201.00	60.00	92.00	
Insulin	116	0	10.01	10.07	58.46	2.43	5.92	
HOMA	116	0	2.69	3.64	25.05	0.47	1.38	(1)
Leptin	116	0	26.62	19.18	90.28	4.31	20.27	
Adiponectin	116	0	10.18	6.84	38.04	1.66	8.35	
Resistin	116	0	14.73	12.39	82.10	3.21	10.83	
MCP-1	116	0	534.65	345.91	1698.44	45.84	471.32	



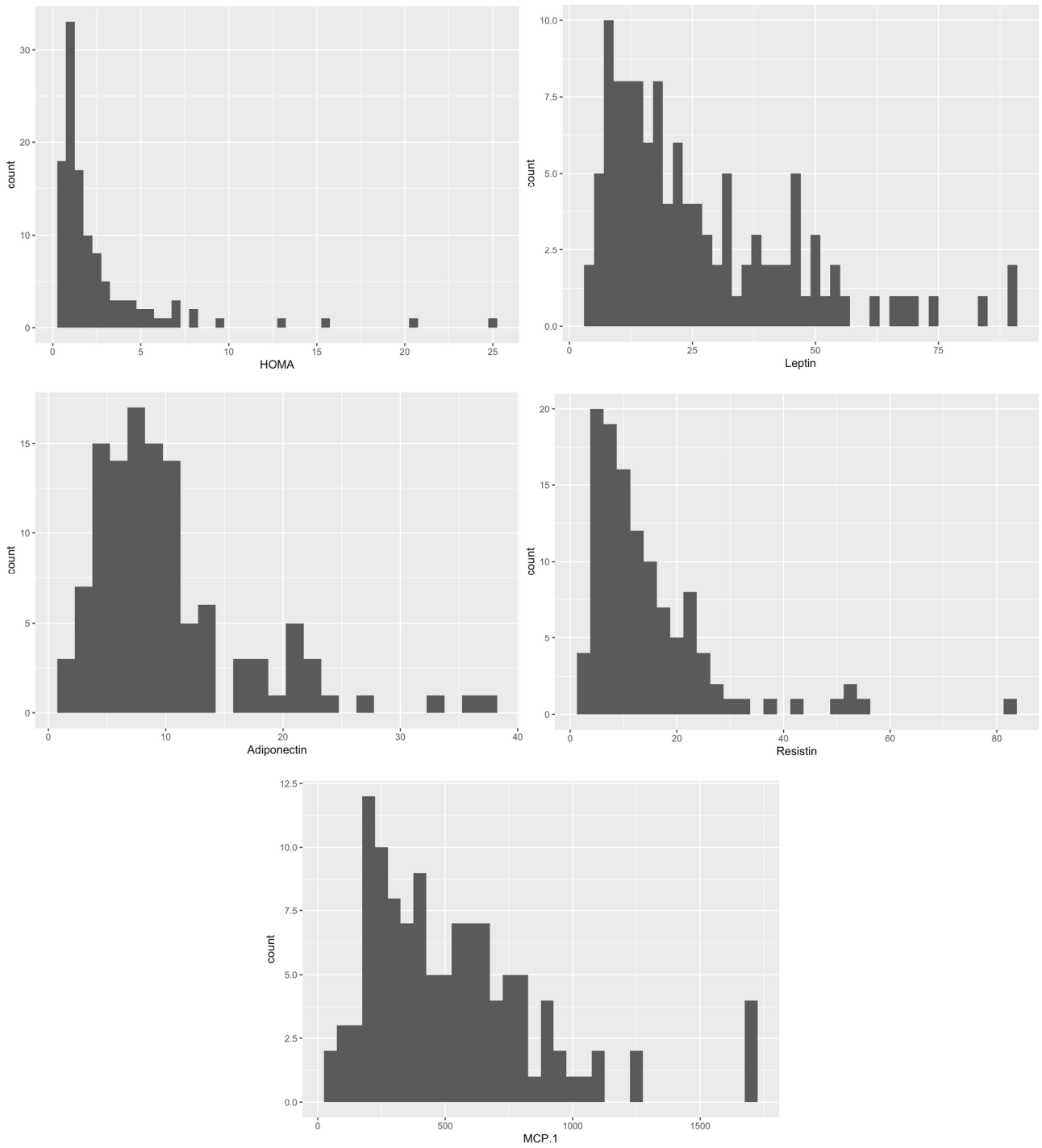


Figure 1. *Annotation of this figure.*

3.2. Multivariate Diagnosis

The correlation coefficients of nine groups of continuous related variables are calculated, and the results are shown in *Table 4*.

It can be seen from *Figure 1* that the distribution of insulin content is similar to the distribution of HOMA. Homeostatic model assessment (HOMA) is a method for assessing β-cell function and insulin resistance (IR) from basal (fasting) glucose and insulin or C-peptide concentrations [9].

Combining *Table 4* above shows that the correlation coefficient between the two is large, indicating that the two have a strong correlation. *Figure 2* shows that there is a correlation between the insulin content and the HOMA indicator. And, since *Figure 2* draws the health status as Healthy and the sample and the health status as Patients separately, we can see that those with higher HOMA index (≥8) and higher blood insulin content (≥30 μU/mL) The samples are all sick samples. At the same time, the correlation coefficient of the variables in *Table 4* also shows that the correlation coefficient between the HOMA index and the glucose content in the blood is 0.696, so it is considered

that the correlation between these two variables is relatively strong.

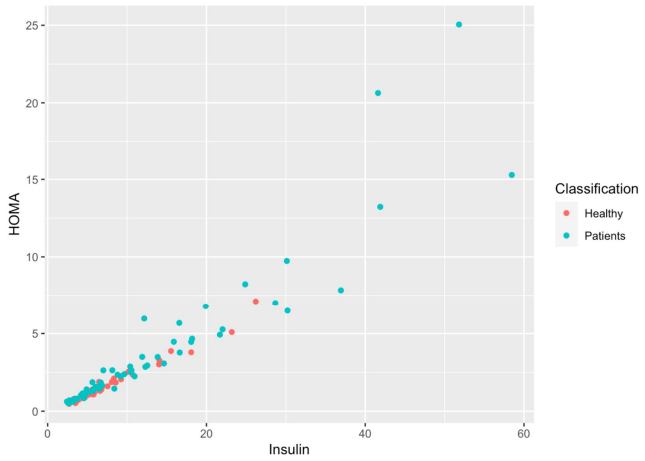


Figure 2. *Scatter plot of insulin content and HOMA indicator.*

3.3. Normality Test

For continuous variable data sets, first check whether the marginal distribution of each continuous variable is normal. The methods to test whether the variables follow the univariate normal distribution include drawing *Q-Q* diagrams and Shapiro-Wilk test methods. Among them, *Q-Q* plot is often used to intuitively assess whether a sample comes from a normal population. The construction steps are as follows:

1. Sort samples $x_1, \ldots, x_n$ from small to large to get $x_{(1)}, \ldots, x_{(n)}$, and their corresponding probability values are $\frac{1-\frac{1}{2}}{n}, \frac{2-\frac{1}{2}}{n}, \ldots, \frac{n-\frac{1}{2}}{n}$;
2. Calculate the lower quantile of the standard normal distribution $q_{(1)}, q_{(2)}, \ldots, q_{(n)}$;
3. Draw a graph of $(q_{(1)}, x_{(1)}), \ldots, (q_{(n)}, x_{(n)})$, and check whether these points are all approximately on a straight line.

For each continuous variable, draw a *Q-Q* graph, as shown in *Figure 3* below.

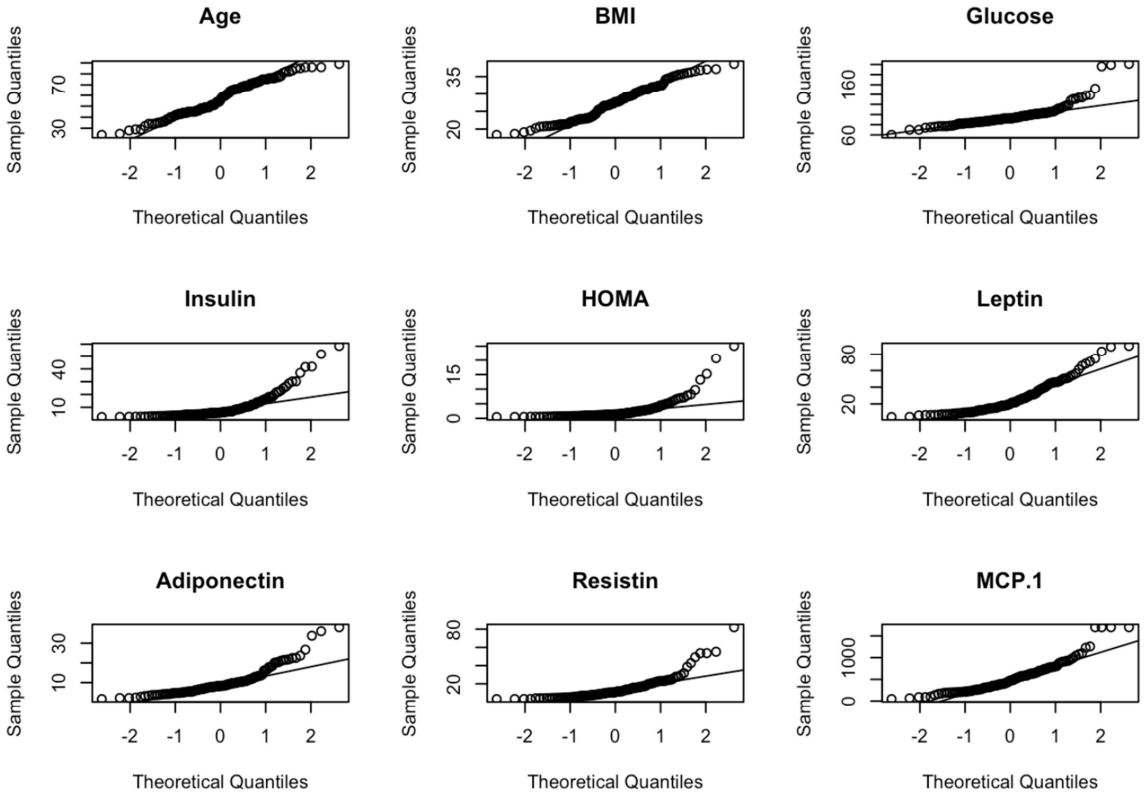


Figure 3. *Q-Q Plot of continuous variables.*

It can be seen that in addition to the age and BMI variables, the *Q-Q* Plot of the remaining seven variables are all upward at the right end, which indicates that the sample distribution is right-biased and the tail is thick. This conclusion is consistent with the variable distribution plotted in *Figure 1*. The Shapiro-Wilk test [10] results are listed below, as shown in *Table 4* below. It can be seen from the Shapiro-Wilk test that p-Value is less than $\alpha = 0.05$, so is rejected, and the nine continuous variables are considered not to follow the univariate normal distribution.

Table 4. *Shapiro-Wilk test results for continuous variables.*

Variable	Test Statistic	P-Value
Age	0.9692	0.0089
BMI	0.9684	0.0077
Glucose	0.7544	<0.001
Insulin	0.6804	<0.001
HOMA	0.5580	<0.001
Leptin	0.8702	<0.001
Adiponectin	0.8287	<0.001
Resistin	0.7453	<0.001
MCP-1	0.8844	<0.001

To test whether the sample comes from a multivariate normal population, you need to use a chi-square plot. For the data set, draw a Chi-Square *Q-Q* plot, as shown in *Figure 4* below.

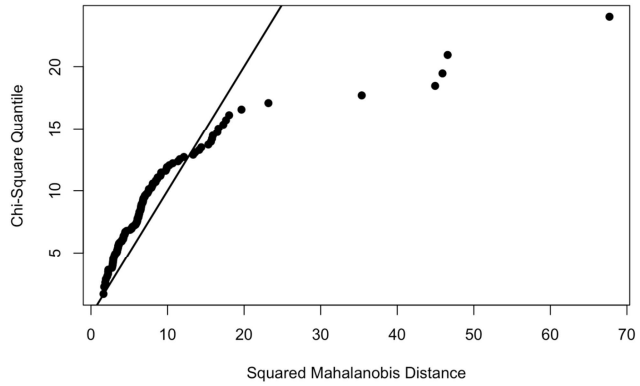


Figure 4. *Chi-Square Q-Q Plot.*

It can be seen from the figure that is $\left(q_{c,p}\left(\frac{j-\frac{1}{2}}{n}\right), d_{(j)}^2\right)$ obviously not on a straight line, so it is considered that the original data set does not follow a multivariate normal distribution. When the sample does not satisfy the normality assumption, some transformations can be performed on the sample to make the sample obtained after the transformation satisfy the normality assumption. The transformation method proposed by Box and Cox [11] is

$$x^{(\lambda)} = \begin{cases} \frac{x^{\lambda}-1}{\lambda}, \lambda \neq 0 \\ lnx, \lambda = 0 \end{cases}$$

Use the power Tranasform (object, family = "bcPower") function in the R package car package to calculate and calculate the λ value used for the transformation. Then use the BoxCox(x,lambda) function in the forecast package and the resulting lambda to perform Box-Cox transformation on the data. For the classification of different health conditions, the Chi Square *Q-Q* Plot of the transformed data is shown in *Figure 5.*

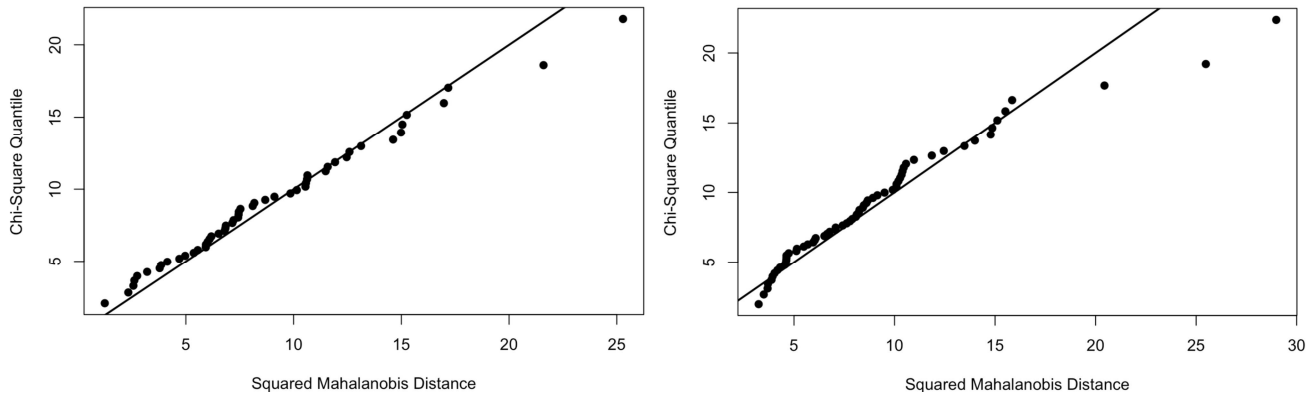


Figure 5. *Chi-Square Q-Q Plot.*

It can be seen from *Figure 5* that for different health conditions, the points $\left(q_{c,p}\left(\frac{j-\frac{1}{2}}{n}\right), d_{(j)}^2\right)$ are almost distributed on a straight line. Therefore, it can be considered that for the transformed data, the sample with Healthy status and the sample with Patients status, all of the nine continuous variables contained in it follow a multivariate normal distribution.

4. Data Analysis

4.1. One-Way MANOVA

To study whether there are differences in blood test status, age, and BMI indicators of samples of different health conditions, One-Way MANOVA method, that is, One-Way analysis of variance, is required. The theoretical derivation part refers to Dai's [12] paper, so the MANOVA model used is

$$X_{lj} = \mu + \tau_l + e_{lj}, \quad j = 1, \ldots, n_l \ and \ l = 1,2$$

where $e_{lj} \sim N_p(0, \Sigma)$, μ is the mean vector of the entire sample, τ_l is the effect of the lth group of treatments, and $\sum_{l=1}^{2} n_l \tau_l = 0$. The null hypothesis of the model is

$$H_0: \tau_1 = \tau_2 = 0 \ v.s. H_1: at \ least \ one \ l \ makes \ \tau_l \neq 0$$

Test statistic is

$$\Lambda^* = \frac{|W|}{|B+W|} = \frac{\left|\sum_{l=1}^{2}\sum_{j=1}^{n_l}(x_{lj}-\overline{x}_l)(x_{lj}-\overline{x}_l)'\right|}{\left|\sum_{l=1}^{2}\sum_{j=1}^{n_l}(x_{lj}-\overline{x})(x_{lj}-\overline{x})'\right|}$$

When the following inequality is true, reject the null hypothesis and think there is a difference

$$-\left(n-1-\frac{(p+g)}{2}\right) ln\,\Lambda^* > \chi^2_{p(g-1)}(\alpha)$$

where $\chi^2_{p(g-1)}$ is the -quantile of the chi-square distribution, its degree of freedom is $p(g\text{-}1)$, $g = 2$, $p = 9$. The results of MANOVA analysis are shown in *Table 5*.

Table 5 shows that p-Value = 1.811× <0.05 =, so the null hypothesis is rejected, and it is considered that when the sample's health status is different, the sample's age, BMI, and

blood test index values such as insulin in this study are significantly different. So it is necessary to carry out the next research.

Table 5. *Results of One-Way MANOVA.*

Source of Variation	Test Statistic	df	P-Value
Treatment	0.70709	1	1.811×
Residuals		114	

4.2. Principal Component Analysis

According to Guo's opinion, the PCA method is a dimensionality reduction method that maintains the maximum overall dispersion. Its advantage is that it uses a smaller dimensionality to reflect the structural relationship between samples [13]. In a word, principal component analysis can use a few principal components to reveal the internal structure of multiple variables. Firstly, from *Table 3* we can get

$$\overline{x'} = [57.30,\ 27.58,\ 97.79,\ 10.01,\ 2.69,\ 26.62,\ 10.18,\ 14.73,\ 534.65]$$

Principal component coefficient, eigenvalue and cumulative variance contribution rate are shown in *Table 6*. The selection of the number of principal components can refer to the screen plot of the cumulative variance contribution rate, as shown in *Figure 6* below.

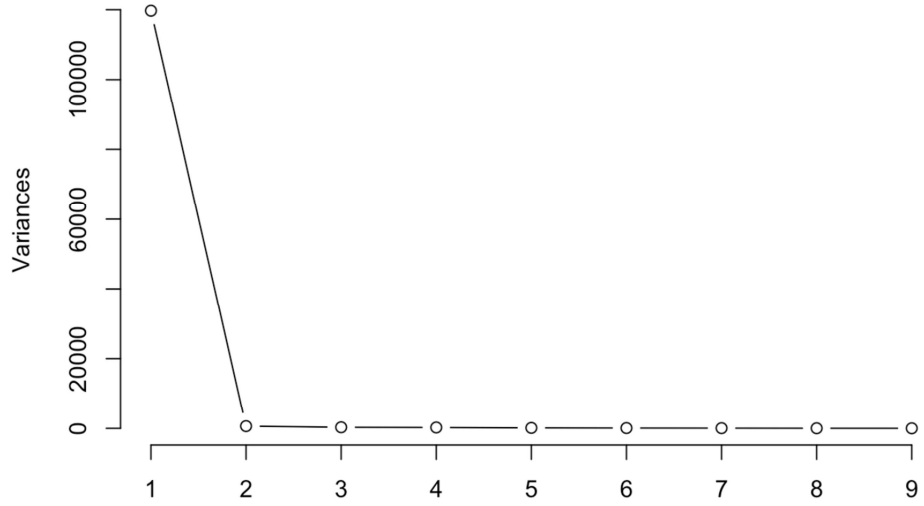


Figure 6. *Screen Plot.*

Table 6. *Principal component coefficient, eigenvalue and cumulative variance contribution rate.*

Variable	$\hat{e}_1$	$\hat{e}_2$	$\hat{e}_3$	$\hat{e}_4$	$\hat{e}_5$	$\hat{e}_6$	$\hat{e}_7$	$\hat{e}_8$	$\hat{e}_9$
Age	0.0006	0.2226	0.3493	-0.8965	-0.0390	-0.1039	-0.1040	0.0390	0.0079
BMI	0.0033	0.0625	-0.1454	-0.0369	0.0250	0.0454	0.2607	0.9485	-0.0568
Glucose	0.0173	0.7725	0.4601	0.3374	0.0281	0.2726	0.0173	0.0166	0.0414
Insulin	0.0051	0.2186	0.0129	0.1533	0.2068	-0.8990	0.0451	0.0351	0.2727
HOMA	0.0027	0.0989	0.0272	0.0519	0.0544	-0.2473	-0.0075	-0.0453	-0.9594
Leptin	0.0008	0.5165	-0.7902	-0.2004	0.1680	0.1125	-0.0543	-0.1573	0.0084
Adiponectin	-0.0040	-0.0386	-0.0009	0.1024	0.1687	0.0107	-0.9441	0.2611	0.0033
Resistin	0.0131	0.1535	-0.1407	0.0651	-0.9458	-0.1753	-0.1569	0.0476	0.0059
MCP-1	0.9997	-0.0177	-0.0054	-0.0064	0.0112	0.0027	-0.0029	-0.0029	0.0006
Variance($\hat{\lambda}_i$)	346.00	25.11	17.23	15.40	10.98	8.41	6.46	3.62	0.82
Percentage of total variance	98.83	0.52	0.25	0.20	0.10	0.06	0.03	0.01	0.00
Cumulative percentage of total variance	98.8	99.4	99.6	99.8	99.9	100.0	100.0	100.0	100.0

The first principal component explained 98.8% of the total sample variance. The first two main components together accounted for 99.4% of the total sample difference. To set the principal component variance interpretation rate to 99%, you need to select the first two principal components to make their cumulative variance contribution rate reach 99.4%. Therefore, the sample change can be well summarized into two main components, so it is reasonable to reduce the data from 116 observations of 9 variables to 116 observations of 2 main components. The two principal components are:

$$\hat{y}_1 = 0.0006(Ages) + 0.0033(BMI) + 0.0173(Glucose) + 0.0051(Insulin) + 0.0027(HOMA)$$

$$+ 0.0008(Leptin) - 0.0040(Adiponectin) + 0.0131(Resistin) + 0.9997(MCP - 1)$$

$$\hat{y}_2 = 0.2226(Ages) + 0.0625(BMI) + 0.7725(Glucose) + 0.2186(Insulin) + 0.0989(HOMA)$$

$$+ 0.5165(Leptin) - 0.0386(Adiponectin) + 0.1535(Resistin) - 0.0177(MCP - 1)$$

Given the results of the above principal component coefficients (*Table 6*), the first major component seems to basically represent MCP-1 (monocyte chemoattractant protein 1). The second main component is basically a weighted sum of glucose and leptin. Dataset for 9 variables perform principal component analysis to obtain the contribution rate of each component, and extract the two principal components that contribute the most. It can be seen from Table 6 that the absolute value of the MCP-1 (monocyte chemoattractant protein 1) coefficient in the main component 1 is large, reflecting the characteristics of the blood component of the sample; the load values of glucose and leptin in the main component 2 are large, It also reflects the characteristics of the blood components of the sample. In this study, the 9 components of the sample were combined into 2 principal components by principal component analysis, and the cumulative contribution rate of each principal component reached 99.4%, which met the requirement that the cumulative contribution rate was greater than or equal to 99%, indicating that the health status in this study was different The differences in age, BMI, and blood component levels between the two groups can be summarized by these two independent factors.

4.3. Factor Analysis

It is more common to use principal component analysis for the comprehensive evaluation of multiple indicators, but the evaluation results are unreasonable or even wrong due to the lack of consideration of application conditions [14]. Therefore, factor analysis is necessary. Assuming that the m = 3 factor model is used and the maximum likelihood method and principal component method are used, the data is re-analyzed. In *Table 7*, the estimated factor load, communalities, specific variances, and the proportion of the total sample variance explained by the principal component method and the maximum likelihood method for each factor obtained from the original data and the rotated data, respectively.

Table 7. *Factor analysis results using principal component analysis and maximum likelihood.*

Variable	MLE				PCA			
	Estimated factor loadings			Specific Variances	Estimated factor loadings			Specific Variances
	F_1	F_2	F_3		F_1	F_2	F_3	
Raw data								
Age	0.102	0.056	0.281	0.9071	0.218	-0.082	-0.223	0.8960
BMI	0.546	-0.835	0.000	0.0050	0.455	-0.616	0.460	0.2014
Glucose	0.635	0.249	0.506	0.2788	0.768	0.229	-0.141	0.3379
Insulin	0.862	0.390	-0.275	0.0293	0.776	0.477	0.101	0.1597
HOMA	0.891	0.446	0.037	0.0049	0.862	0.462	-0.013	0.0431
Leptin	0.531	-0.335	0.086	0.5990	0.580	-0.288	0.630	0.1837
Adiponectin	-0.181	0.244	-0.145	0.8867	-0.302	0.593	0.305	0.4644
Resistin	0.276	-0.054	0.272	0.8468	0.493	-0.375	-0.312	0.5194
MCP-1	0.312	-0.064	0.241	0.8409	0.445	-0.260	-0.537	0.4462
Cumulative proportion of total sample variance	0.304	0.511	0.869		0.340	0.639	0.872	
Rotated Data								
Age	0.011	0.013	0.304	0.9071	0.116	0.300	-0.022	0.8960
BMI	0.039	0.997	-0.014	0.0050	-0.010	0.223	0.865	0.2014
Glucose	0.454	0.131	0.705	0.2788	0.742	0.321	0.091	0.3379
Insulin	0.977	0.108	0.068	0.0293	0.907	0.018	0.129	0.1597
HOMA	0.919	0.084	0.380	0.0049	0.963	0.145	0.091	0.0431
Leptin	0.237	0.564	0.164	0.5990	0.294	-0.021	0.854	0.1837
Adiponectin	0.022	-0.306	-0.138	0.8867	0.135	-0.677	-0.243	0.4644
Resistin	0.101	0.197	0.322	0.8468	0.149	0.648	0.195	0.5194
MCP-1	0.136	0.223	0.301	0.8409	0.177	0.721	-0.050	0.4462
Cumulative proportion of total sample variance	0.232	0.487	0.869		0.275	0.460	0.872	

The factor scores obtained by principal component analysis (PCA) and maximum likelihood estimate (MLE) methods are shown in *Table 8*.

Table 8. *Factor scores using PCA and MLE methods.*

Factor	MLE		PCA	
	Weighted Least Squares Method	**Regression Method**	**Weighted Least Squares Method**	**Regression Method**
Raw Data				
F_1	-0.9069	-0.5475	-0.9743	-0.0912
F_2	0.3839	-0.4814	0.2901	-1.3516
F_3	-0.1353	0.7399	-0.3972	-2.1088
Rotated Data				
F_1	-0.4986	-0.9300	-0.5980	-0.9563
F_2	-0.8031	0.1457	-0.3223	2.1818
F_3	-0.3076	0.4390	-0.8542	-0.7793

The proportion of the total variance explained by the three-factor solution obtained by applying the principal component method to the original data is significantly larger than the proportion of the two-factor solution. However, for $m = 3$, the value produced by is usually greater than the sample correlation coefficient. This is especially true for r_{69}. Obviously, on the first factor F1, most variables have very high loads on the factor, and the loads are approximately equal, with the exception of adiponectin content, so F_1 can be regarded as reflecting the blood adiponectin content and other variables Factors of difference between. The second factor compares age, BMI, some blood indicators with the rest of the blood indicators. From this factor, the negative load of BMI is relatively large, while adiponectin has a large positive load. On the third factor F_3, it mainly reflects the relationship between leptin and MCP-1. Similar conclusions can be drawn from the solutions obtained from the original data using the maximum likelihood method.

After rotation, the two solving methods seem to give some different results. If we focus on the principal component method and the cumulative proportion of the total sample variance, we see that a three-factor solution is obviously necessary. The third factor explains the "large number" of additional sample changes. The first factor is roughly the pancreatic function factor determined by the blood glucose, insulin and HOMA indicators; the second factor mainly reflects the comparison of adiponectin and resistin content in the blood, which can be attributed to the factors that explain obesity and diabetes; The third factor mainly reflects the comparison of BMI and blood leptin content, which can be attributed to the factors that explain the body's fatness and thinness. The maximum likelihood factor load after rotation is similar to the load generated by the principal component factor method for the first factor, but is inconsistent with factors 2 and 3. For the maximum likelihood method, the second factor can also be attributed to explaining the body fat and thin. The third factor mainly reflects the effect of glucose content in the blood.

4.4. Discriminant Analysis

Discriminant analysis is to establish discriminant functions based on various variables of the research object, discriminate and classify various groups, and predict the attribution of new samples. Because the main research purpose of this article is to infer whether a sample has breast cancer through the information given in the data set

For this disease, the discriminant analysis method is suitable for classifying new samples. The expected cost of the misjudgment rate of discriminant analysis is:

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2$$

The total probability of misjudgment is

$$TPM = p_1 \int_{R_2} f_1(x)dx + p_2 \int_{R_1} f_2(x)dx$$

Data converted using the BoxCox method, so that both categories of data can be considered to be from a multivariate normal distribution population, whose population density function is

$$f_i(x) = \frac{1}{(2\pi)^{p/2}\left|\Sigma_i\right|^{-1}} exp\left[-\frac{1}{2}(x-\mu_i)'\,\Sigma^{-1}(x-\mu_i)\right],$$

$$i = 1,2, p = 9$$

For the converted data, first examine the scatter plot of two or two variables. The two-variable scatter plots drawn using the ggplot2 package of the R program are presented in Appendix. Most of these scatter plots are relatively disordered, scattered, uniform, and there is no obvious ellipse area, except for the scatter plots of insulin and HOMA indicators, as shown in *Figure 7* below

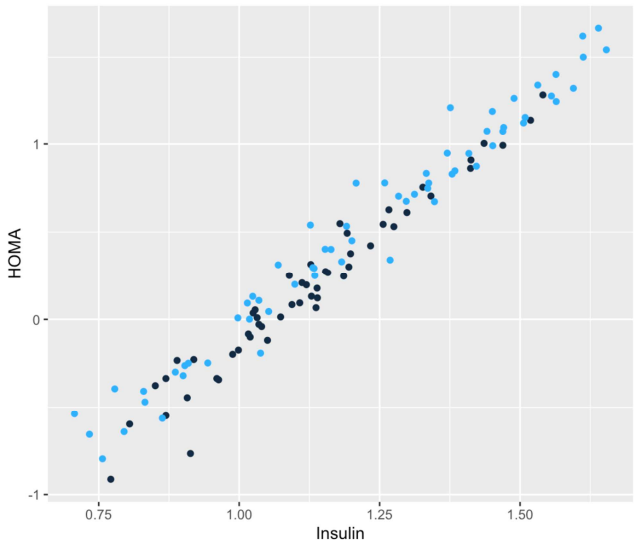


Figure 7. *Scatterplot of insulin and HOMA indicators (translated data).*

Among them, the dark dots represent Healthy status, that is, the group without breast cancer, and the light dots represent Patients, the group with breast cancer. The data in Figure 6 above seems to form a fairly elliptical shape, so for these two variables, the assumption of multiple normality does not seem appropriate. However, because there is no obvious correlation between them and variables, this problem is ignored for now, and the data after BoxCox transformation is considered to be from a multivariate normal population.

If the following inequality is true, x_0 is judged as category1 (π_1), and the sample's health status is considered as "Healthy", that is, the sample does not have breast cancer:

$$-\frac{1}{2}x_0'(\Sigma_1^{-1}-\Sigma_2^{-1})x_0+\left(\mu_1^{'}\Sigma_1^{-1}-\mu_2^{'}\Sigma_2^{-1}\right)x_0-k\geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

If the following inequality is true, x_0 is judged as category2 (π_2), and the sample's health status is "Patients", which means that the sample has breast cancer:

$$-\frac{1}{2}x_0^{'}\ (\Sigma_1^{-1}-\Sigma_2^{-1})x_0+\left(\mu_1^{'}\Sigma_1^{-1}-\mu_2^{'}\Sigma_2^{-1}\right)x_0-k< \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right]$$

where the formula of k is

$$k=\frac{1}{2}\ln\left(\frac{|\Sigma_1|}{|\Sigma_2|}\right)+\frac{1}{2}\left(\mu_1^{'}\Sigma_1^{-1}\mu_1-\mu_2^{'}\Sigma_2^{-1}\mu_2\right)$$

Among them, μ_1, μ_2, Σ_1 and Σ_2 are all unknown statistics, so in actual use, the sample statistics will be used instead. The apparent error rate (APER) and expected actual error rate, E(AER) can also be calculated based on the model obtained by fitting the sample data. The formulas are as follows:

$$APER=\frac{n_{1M}+n_{2M}}{n_1+n_2},\quad \hat{E}(AER)=\frac{n_{1M}^{(H)}+n_{2M}^{(H)}}{n_1+n_2}$$

Here n_{1M} represents the number of samples that belong to category1 are misjudged to category2, n_{2M} represents the number of samples that belong to category 1 is misjudged to category2; $n_1^{(H)}$ represents the number of the sample of category1 is misjudged to the number of category2 with Lachenbruch's "holdout" process [15], and $n_2^{(H)}$ is the opposite.

In this study, the determination of a sample with breast cancer as not having breast cancer would cause patients to fail to receive timely treatment, and the consequences are more serious. On the contrary, if a sample without breast cancer is identified as having Breast cancer, after a progressive screening will increase the probability of being correctly classified, so the consequences of this misjudgment are lighter. Therefore, the misjudgment costs $c(1|2)$ and $c(2|1)$ are assigned 0.75 and 0.25, respectively. The calculated linear discriminant function is more complicated, so it is no longer listed. The resulting confusion matrix form is shown in *Table 9*.

Table 9. *Confusion matrix.*

		Forecast category	
		π_1	π_2
True category	π_1	39	13
	π_2	6	58

After calculation, $APER$ = 0.1638 and $\hat{E}(AER)$ = 0.1872. Among them, the probability that a patient with breast cancer is judged as not suffering from breast cancer is 0.09375, which is a low rate of misjudgment, which is in line with the expected effect of adjusting the cost of misjudgment.

5. Conclusion and Outlook

5.1. Research Conclusion

In this paper, through a series of analysis and testing of breast cancer detection data, the main conclusions are:

1) HOMA is one of the evaluation indicators for testing insulin resistance in the blood, and it has a clear linear correlation with the insulin content in the blood, as shown in Figure 2, which shows that the stronger the insulin resistance, the less likely the insulin in the blood is Use, so its content in the blood will be correspondingly higher. It is worth noting that compared with the samples without breast cancer, some patients with breast cancer showed a high degree of insulin resistance and high insulin content in the blood, which is a feature that the samples without breast cancer do not have.
2) Through One-Way MANOVA, the blood test conditions, age, and BMI indicators of samples of different health conditions are obtained. There are significant differences in the blood test conditions, age, and BMI indicators of samples of different health conditions.
3) Principal component analysis can be used to reduce the dimension of the data to obtain two principal components. Therefore, the differences in age, BMI and blood component content between the two groups with different health conditions in this study can be summarized by these two independent factors.. Among them, the absolute value of the MCP-1 (monocyte chemoattractant protein 1) coefficient in the main component 1 is very large, reflecting the characteristics of the blood component of the sample; the load values of glucose and leptin in the main component 2 are large, reflecting the same It is the characteristic of the blood component of the sample.
4) Assuming the use of m = 3 factor model and the use of maximum likelihood method and principal component method, the original data data and factor rotation data are re-analyzed, and the variables are reduced to 3 factors for analysis. Among them, the maximum likelihood method is used to estimate the factor rotation data. The first factor reflects the insulin resistance factor attributed to the insulin and HOMA indicators, and the second factor reflects the body fat and thin factor attributed to BMI and leptin. The third factor reflects the glucose content in the blood.
5) By setting different misjudgment costs for discriminant analysis, the obtained APER is 0.1638 and EAER is

0.1872. Among them, the probability of identifying patients with breast cancer as not having breast cancer is 0.09375, which is a low rate of misjudgment.

5.2. Inadequacies in Research

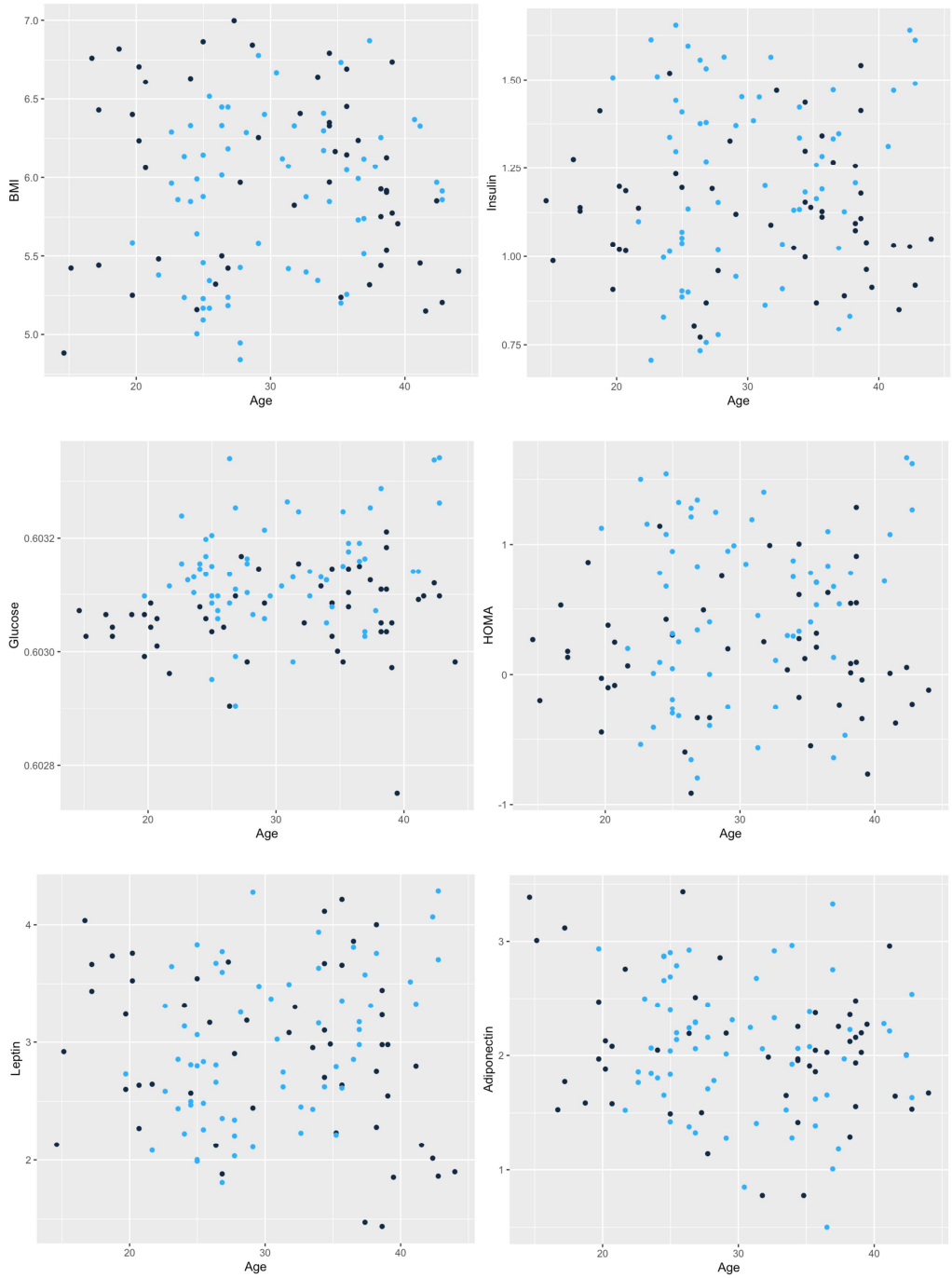
Due to the limitation of objective conditions, this study still has deficiencies in the following two aspects:

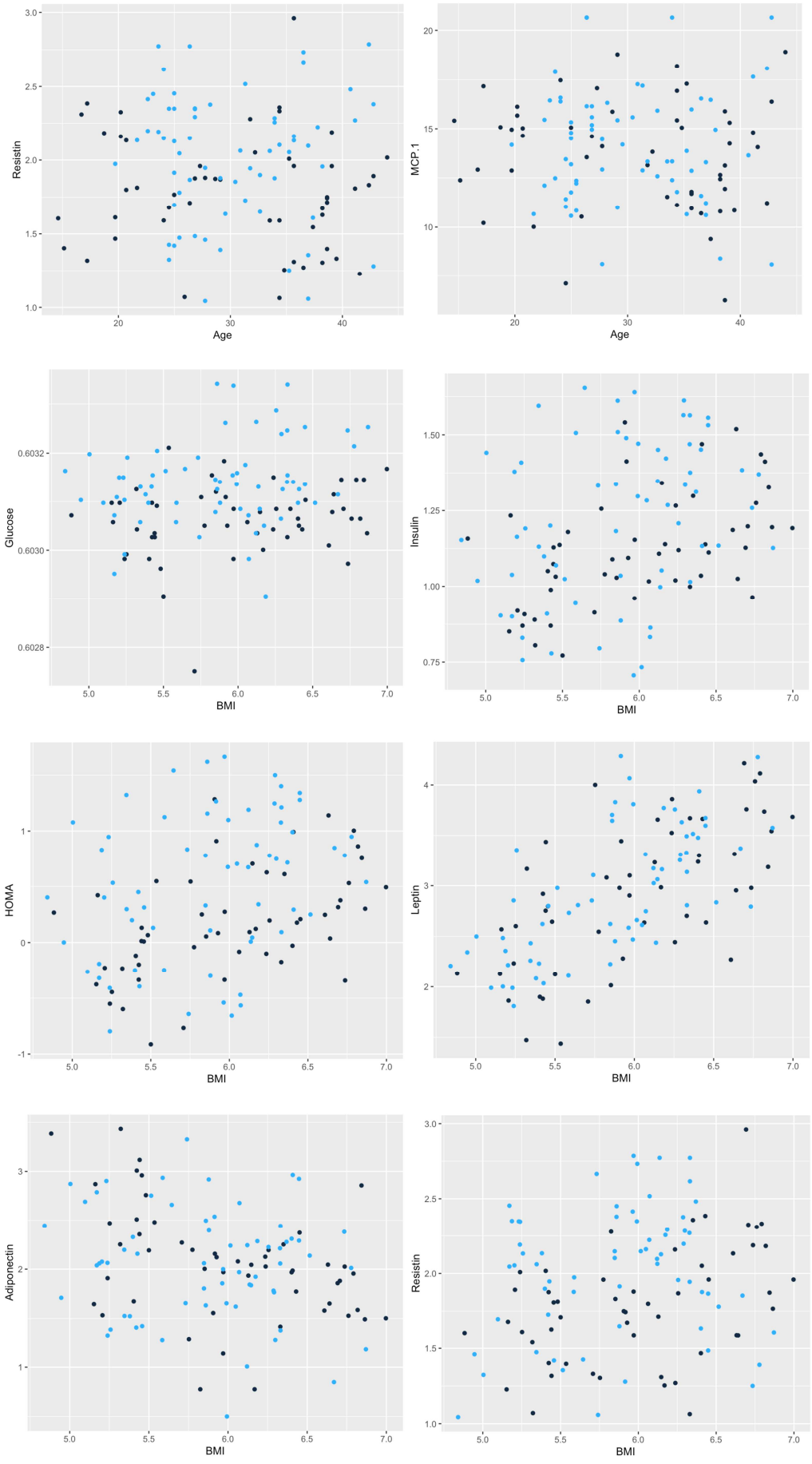
Since no tumor marker protein content data is collected in the blood, only analysis based on the content of conventional blood components, age, and BMI indicators, there is still room for a decline in the overall misjudgment rate.

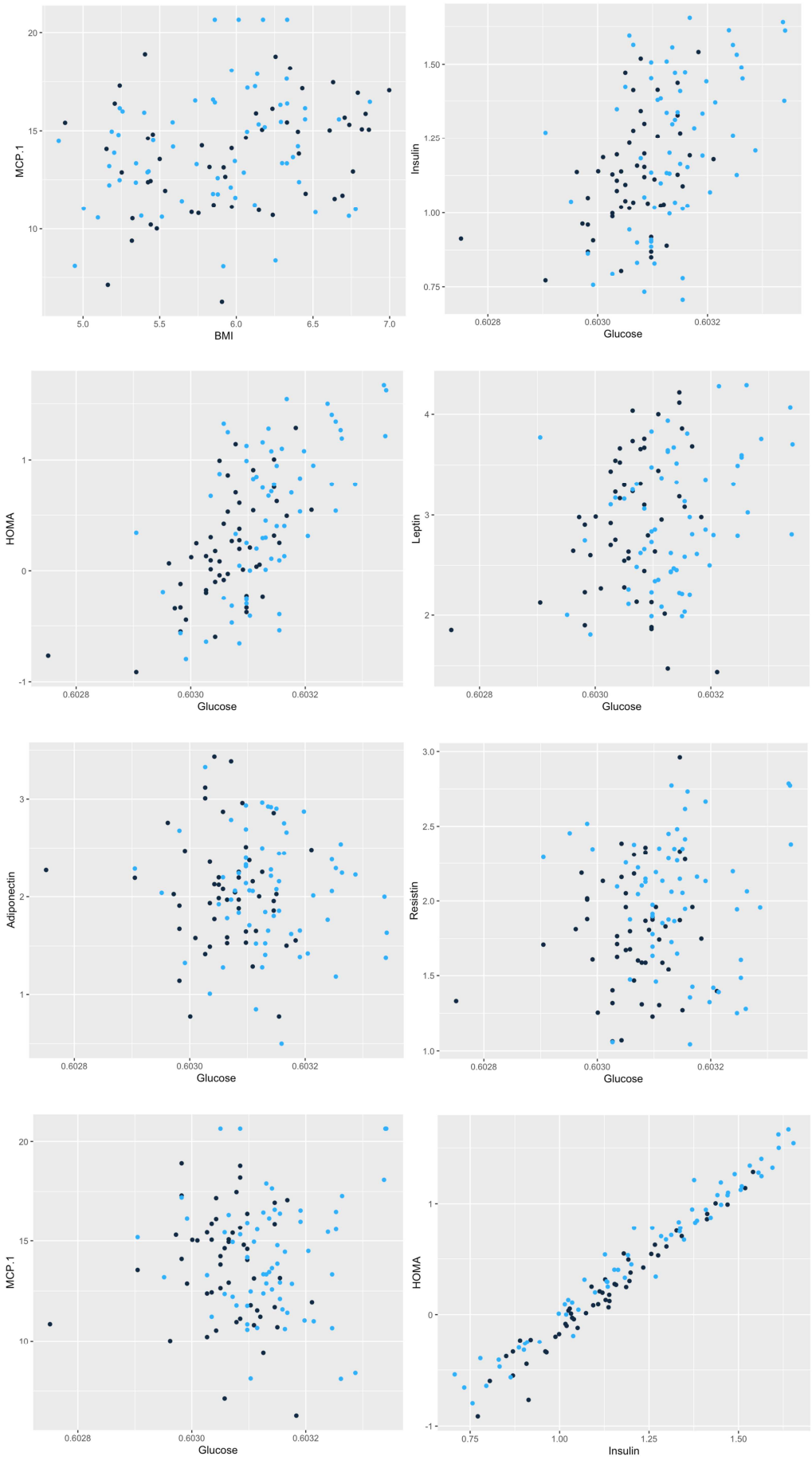
The samples in the data are all from Coimbra Hospital, so there are certain geographical limitations, and they cannot represent and reflect the general condition of the population.

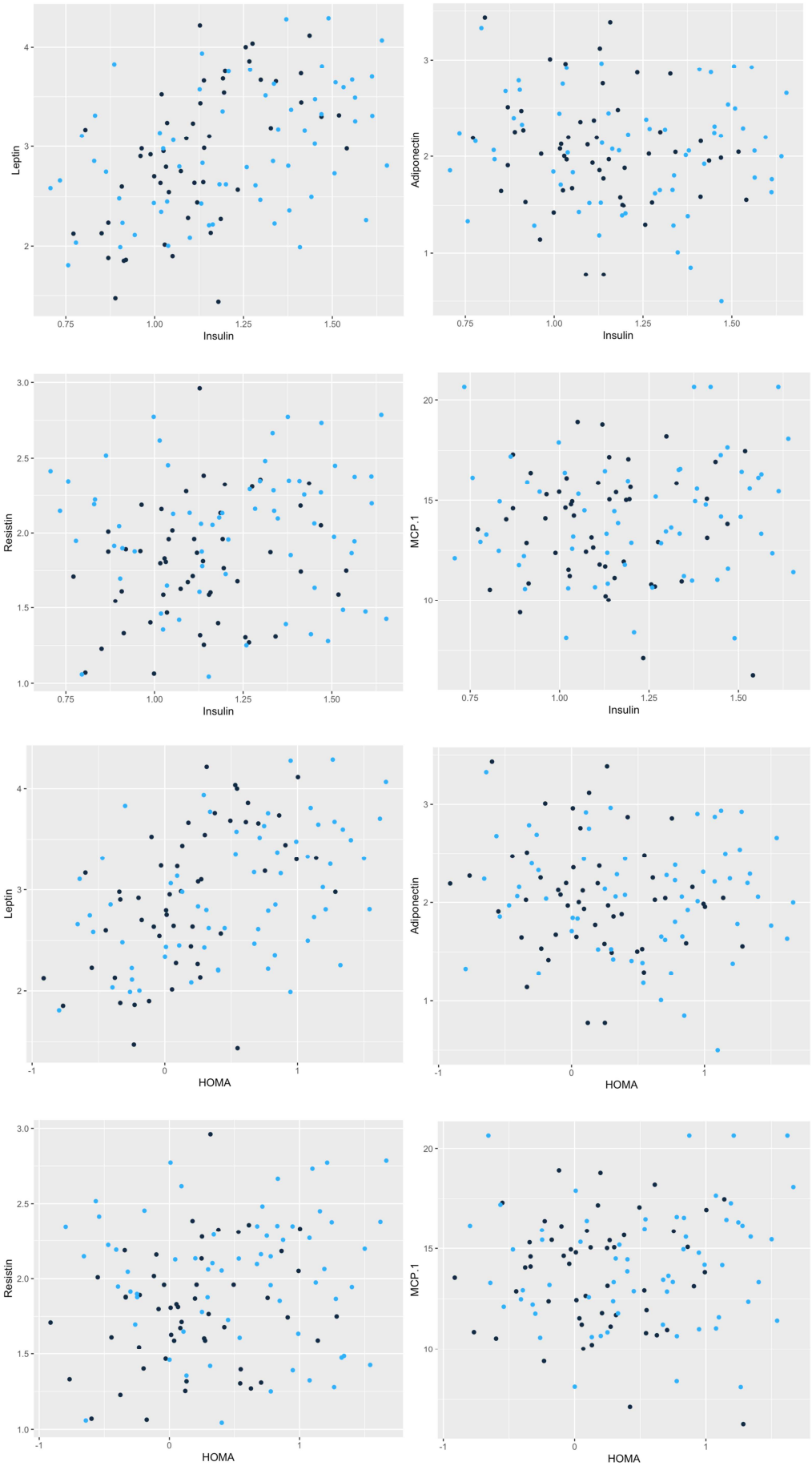
Appendix

For the converted data, use the `ggplot2` package of the R program to draw a two-variable scatter plot. In the following figure, the dark dots represent healthy groups, that is, those who do not suffer from breast cancer, and the light dots represent patients, which are groups with breast cancer.




Resistin
Age
MCP.1
Age
Glucose
BMI
Insulin
BMI
HOMA
BMI
Leptin
BMI
Adiponectin
BMI
Resistin
BMI


MCP.1
BMI
Insulin
Glucose
HOMA
Glucose
Leptin
Glucose
Adiponectin
Glucose
Resistin
Glucose
MCP.1
Glucose
HOMA
Insulin


Leptin
Insulin
Adiponectin
Insulin
Resistin
Insulin
MCP.1
Insulin
Leptin
HOMA
Adiponectin
HOMA
Resistin
HOMA
MCP.1
HOMA

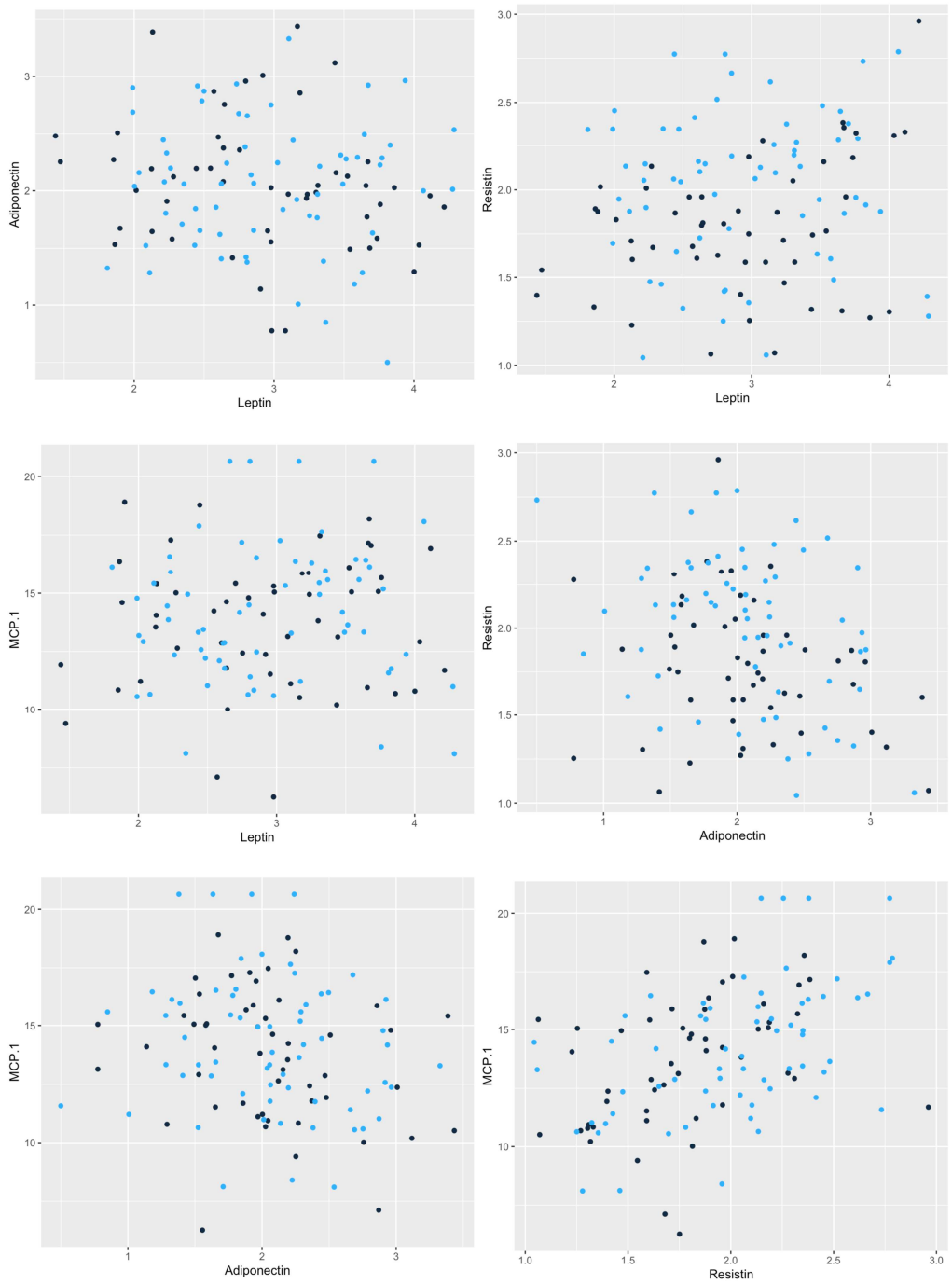


Figure 8. *Scatter plot for each pair of variables (translated data).*

References

[1] Hui-Ling Chen, Bo Yang, Jie Liu, Da-You Liu. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis [J]. Expert Systems With Applications, 2011, 38 (7).

[2] Zheng Ying, Wu Chunxiao, Zhang Minlu. The prevalence and disease characteristics of breast cancer in China [J]. Chinese Journal of Cancer, 2013, 23 (008): 561-569. (in Chinese).

[3] Yang Ling, Li Liandi, Chen Yude, et al. Estimation and prediction of the incidence and death trend of breast cancer in China [J]. Chinese Journal of Oncology, 2006, 28 (006): 438-440. (in Chinese).

[4] M. Eskelinen, E. Hämäläinen, V.-M. Kosmat, I. Penttilä, E. Alhava, K. Syrjänent. 7 Comparison of tumour markers CEA, AFP, CA15-3, TPS and NEU in breast cancer diagnosis [J]. The Breast, 1995, 4 (1).

[5] Na Liu, Er-Shi Qi, Man Xu, Bo Gao, Gui-Qiu Liu. A novel intelligent classification model for breast cancer diagnosis [J]. Information Processing and Management, 2019, 56 (3).

[6] M. Patrício, J. Pereira, J. Crisóstomo, P. Matafome, M. Gomes, R. Seic A, and F. Caramelo. Using resistin, glucose, age and bmi to predict the presence of breast cancer. Bmc Cancer, 18 (1): 29, 2018.

[7] Jiang Yina, Chen Naihong. Research on the mechanism of CCL2/MCP-1 in related diseases [J]. Chinese Pharmacological Bulletin, 2016, 32 (12): 1634-1638. (in Chinese).

[8] Yue Chen. Adiponectin-a new type of lipid-derived hormone [J]. Medical Journal of Chinese People's Liberation Army, 2003 (02): 183-185. (in Chinese).

[9] Wallace, Tara M., Levy, Jonathan C., Matthews, & David R. Use and Abuse of HOMA Modeling. [J]. Diabetes Care, 2004.

[10] Srivastava M. S, Hui T. K. On assessing multivariate normality based on shapiro-wilk W statistic. 1987, 5 (1): 15-18.

[11] Liu-Cang Wu, Deng-Ke Xu. Maximum Likelihood Estimation of Normal Distribution Parameters under Data Transformation [J]. Journal of Data Analysis, 2010, 5 (5): 15-24. (in Chinese).

[12] Dai Jinhui, Yuan Jing. Comparison of single-factor analysis of variance and multiple linear regression analysis methods [J]. Statistics and Decision, 2016 (09): 23-26. (in Chinese).

[13] Guo Zhibo, Liu Huajun, Zheng Yujie, et al. Enhanced linear discriminant analysis criteria based on the unification principle of PCA and LDA [J]. Journal of Image and Graphics, 2008, 13 (4): 702-708. (in Chinese).

[14] Lin Haiming, Du Zifang. Problems that should be paid attention to in the comprehensive evaluation of principal component analysis [J]. Statistical Research, 2013, 30 (08): 25-31. (in Chinese).

[15] P. A. Lachenbruch and M. R. Mickey. Estimation of error rates in discriminant analysis. Technomet- rics, 10 (1): 1–11, 1968.