Team Number :	85256

Problem Chosen :	A

2018 APMCM summary sheet

Common in the elderly population, falls have a negative impact on the lives of older people. The core objective of this paper is to establish a mathematical model to describe the relationship between the gait parameters, the data of base condition of subjects and the risk of falls, thus we can build a balance risk assessment system which can measure the risk of falls in the elderly.

In the first question, we calculate four types of gait parameters from the coordinates in the Annex 2 according to credit medical experiments, which include space-time parameters, kinematics characteristics, asymmetry indices and deviation degree of gravity center deviation. After that, we test these parameters for significance and preserve the 25 indicators as body balance features.

In the second question, we establish the classification based on Logistic regression analysis (CL model) to build the balance risk assessment. Before we doing the logistic regression, we pre-process the data with four steps. First, we adopt SMOTE(Synthetic Minority Over-sampling Technique) to implement the data processing. Besides, we standardize these variables to avoid the influence of dimension and range. After that we use the PCA(principal component analysis) method to find principal components to represent the original indicators. Finally, we build up the CL model to assess the balance risk.

In the third question, we first utilize CL model to make an analog computation with the original input data from the Annex 2, and the proportion of analog correct is 94.73% . Then according to the principle of case control analysis, we implement Logistic stepwise regression analysis. We conclude that body balance ability is important and the complaint1 (orthopedics illness) , sex and IBM also matter.

Furthermore, through comparative analysis, we can obtain a more comprehensive assessment of the fall risk of the elderly and provide effective measures from both physical and medical perspectives.

Key words: gait analysis; significance test; PCA; Logistic regression model; comparative analysis

Contents

1. Problem Description

Background:

Falls may cause many complications in elderly people due to their poor rehabilitation ability, and the fear from falls may put a limit on movement thus worsen the quality of life. Therefore to make a balance ability assessment for elderly people is benefit for preventing accidental falls. The basic data of all the elderly subjects contains age, sex, weight, height, etc. And the data from gait analysis by deploying 42 monitoring points on the body of the elderly subjects includes the coordinates of motion of a monitoring point (x, y, z).

The problem to be solved:

Problem 1: Extract 25 body balance features from the system consisting of the 42 monitoring points based on an analysis of steps, the center of gravity and motion, which can assess body balance of elderly people comprehensively.

Problem 2: Build a balance risk assessment system by using the 25 body balance features extracted in problem one to assess the balance ability of elderly people, and provide corresponding advice as well.

Problem 3: Based on the actual data provided, to make an analog computation and a comparative analysis of the body balance force. Provide elderly people weak in balance with effectual advice.

2. Problem Analysis

2.1 Analysis of Problem 1

Medical research shows that, due to the decline of physiological function and the combined effects of various nervous system, musculoskeletal diseases, drugs, psychology and so on, the elderly often have a variety of complex problems in gait. While the abnormal gait during walking increases the risk of falling in elderly people, so it is of great importance to clarify the gait characteristics of the elderly to prevent falls.

By consulting relevant literature at home and abroad and considering the physical meaning, we selected two main types of gait parameters calculated from the data in Annex 2.What's more, we calculated the relative difference between the left and right limbs as an asymmetric index. For the center of gravity, we compute the weighted value of 3D coordinates by using the weight ratio of each part of the body, thus the relationships between center of gravity projection and supporting center of gravity is obtained.

Then we will seek 25 features by eliminating variables that are not significant, which can assess body balance of elderly people comprehensively. Because there are discrepancies in data in Annexes 1 and 2, after screening, we remained 76 samples with the same parts.

2.2 Analysis of Problem 2

The problem 2 asked us to build a balance risk assessment system based on 25 indicators to assess the balance ability of elderly people.

First of all, as the original database always contains noisy information, ,we do data pre-processing at first in order to establish accurate risk assessment system. The data pre-processing of this problem can be divided into four phases, generate classification tag, over-sampling, standardization and PCA method to reduce dimension.

After that we can do the logistic regression to establish a balance risk assessment system and analyse the result of our assessment system.

2.3 Analysis of Problem 3

The problem 3 asks us to make an analog computation firstly, we can use the model set up in problem 2 for simulation testing and thus the interpretation ability of analytical model can be observed.

Further a comparative analysis of the body balance force is required. We noted that the original database contains incomplete, redundant, and noisy information as expected in any real-world data. There were several features that could not be treated directly since they had a high percentage of missing values. Then we We deal with continuous variables and discrete variables respectively and make regression analysis to find influential variables. Finally we made a comparative analysis of these variables.

3. Problem1:Feature Extraction Model

3.1.Assumptions

1) Assuming that the gait parameters we calculated from the coordinates is of great significance for the gait analysis.

2) Assuming that the height of the center of gravity calculated by the weighted method is close to the actual position.

3) Assuming that the original data of gait parameters have strong correlation with each other.

4) Assuming that the selected body balance features are highly representative.

3.2Establishment of Model

3.2.1Calculation of 29 Original Indicators

1)Space-time parameters.

Gait cycle refers to the time between the beginning of the contact between one heel and the ground and the second landing of the heel on the ground.

Step width refers to the distance between bilateral arches when two consecutive feet touch the ground.

Foot angle is the angle between the direction of movement of the body and the long axis of the foot.

Walking velocity is the walking distance per unit time.

Step height is the highest point of foot movement.

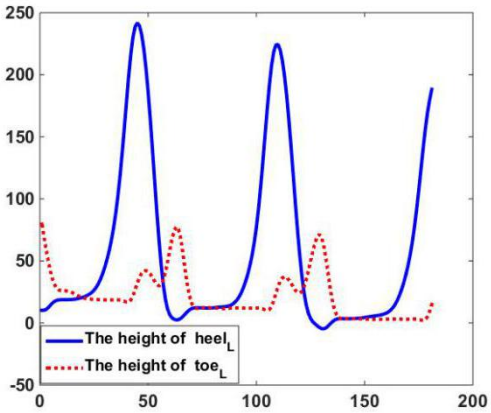

Figure1.1 Step height of left limb

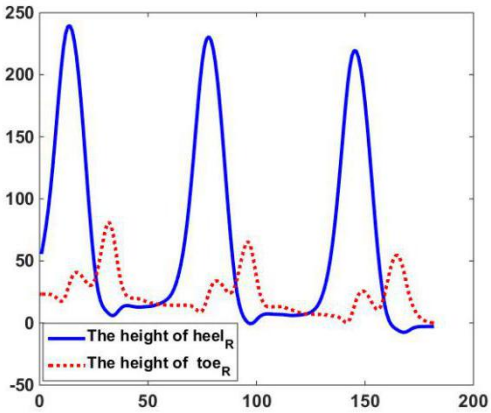

Figure1.2 Step height of right limb

2)Kinematics characteristics

We calculated series of angle values by using the coordinates of corresponding monitoring points, which include the angle of the knee, hip and ankle, the spinal curvature, landing angle and elevation angle.

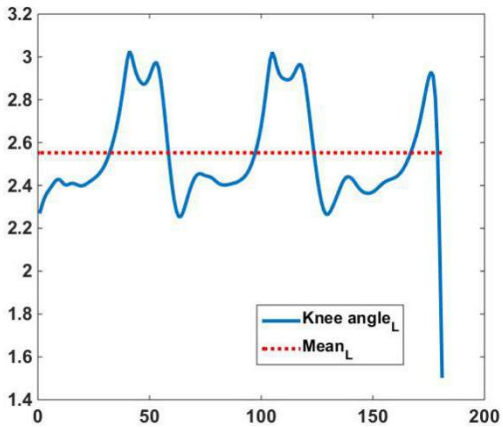

Figure2.1 Knee angle of left limb

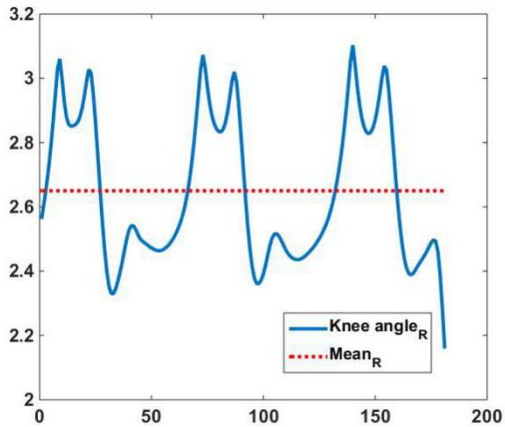

Figure2.2 Knee angle of right limb

3) What's more, we calculated the relative difference between the left and right limbs as an asymmetric index. Take Walking velocity for example, we use

|Walking velocity_L -Walking velocity_R| / Walking velocity_L as the

asymmetry indices of Walking velocity, and the same principle is applicable to other indicators.

4) For the center of gravity, we compute the weighted value of 3D coordinates by using the weight ratio of each part of the body, the relationships between center of gravity projection and supporting center of gravity is obtained.

The three-dimensional coordinate common formula of the center of gravity of the human body obtained by applying the kinematics model SKC (Segment Kinematic Centroid) is:

$$x_{\text{com}} = \frac{1}{M}\sum_{i=1}^{m} x_i m_i = \sum_{i=1}^{m} x_i \frac{m_i}{M}$$

$$y_{\text{com}} = \frac{1}{M}\sum_{i=1}^{m} y_i m_i = \sum_{i=1}^{m} y_i \frac{m_i}{M}$$

$$z_{\text{com}} = \frac{1}{M}\sum_{i=1}^{m} z_i m_i = \sum_{i=1}^{m} z_i \frac{m_i}{M}$$

(x_i, y_i, z_i) is the central coordinates of each part, $\mathrm{m_i} / M_i = W_i$.

Then we calculate the mean of the height of the center of gravity, maximum distance between Center of Gravity projection and supporting Center of Gravity, the mean of distance between gravity projection and supporting center of gravity and coefficient of variation in distance between barycenter projection and supporting center of gravity.

Table1.1 :the weight ratio of each part of the body

Part	Mass ratio for Man	Mass ratio for Woman	Wi（mean of man and woman)
Head	0.0440	0.0370	0.0405
Neck	0.0330	0.0260	0.0295
Body	0.4790	0.4870	0.4830
Upper arms	0.0530	0.0510	0.0520
Forearms	0.0300	0.0260	0.0280
Hands	0.0180	0.0120	0.0150
Thighs	0.2000	0.2230	0.2115
Shanks	0.1070	0.1070	0.1070
Foots	0.0380	0.0300	0.0340

3.2.2Test for Significance

By using column ‘Fall times in one year’ in Annex 1, we define logical variable with label =1 if times is more than zero, otherwise label =0.

Then, We tested 49 original indicators for significance, and the results are as follows.

Table 1.2: Result of Significance test

Number	Category	Index	p-value
1	Space-time parameters	Gait cycle	0.3694
2		Step width	0.3433
3		Foot angle_L	0.0068
4		Foot angle_R	5.7933E-05
5		Walking velocity	0.9235
6		Step height_L	0.8485
7		Step height_R	0.7876
8	Kinematics characteristics	Knee angle_L (for Knee angle of left limb)	0.3591
9		Knee angle_R (for Knee angle of right limb)	0.4929
10		Hip angle_L	0.0047
11		Hip angle_R	0.0376
12		Ankle angle_L	0.0898
13		Ankle angle_R	0.0580
14		Spinal curvature	0.0891
15		Landing angle_L	0.0532
16		Landing angle_R	0.0102
17		Elevation angle_L	0.0702
18		Elevation angle_R	0.4844
19	Asymmetry indices	Foot angle	1.7628E-05
20		Step height	0.6727
21		Knee angle	0.0005
22		Hip angle	0.0973
23		Ankle angle	0.0004
24		Landing angle	0.9432
25		Elevation angle	0.0325
26	centre of gravity	The mean of the height of the center of gravity	0.1607
27		Maximum distance between Center of Gravity projection and supporting Center of Gravity	0.0595
28		The mean of distance between gravity projection and supporting center of gravity	0.0048

29	Coefficient of variation in distance between barycenter projection and supporting center of gravity	0.1210

3.3.3 Screening of Data

As the test result above, most variables have relatively small p value, while for Walking velocity, Step height of left limb, Step height of right limb and asymmetry indices of Landing angle, larger values are observed. Therefore we deleted these indicators and 25 gait parameters are preserved as final body balance features.

Table 1.3: Result of Significance test for 25 selected indicators

Number	Category	Index	p-value
1	Space-time parameters	Gait cycle	0.0656
2		Step width	0.3158
3		Foot angle_L	0.0039
4		Foot angle_R	0.0000
5	Kinematics characteristics	Knee angle_L (for Knee angle of left limb)	0.2888
6		Knee angle_R (for Knee angle of right limb)	0.5323
7		Hip angle_L	0.0035
8		Hip angle_R	0.0250
9		Ankle angle_L	0.0892
10		Ankle angle_R	0.0475
11		Spinal curvature	0.0002
12		Landing angle_L	0.0105
13		Landing angle_R	0.0066
14		Elevation angle_L	0.0319
15		Elevation angle_R	0.5157
16	Asymmetry indices	Foot angle	0.0000
17		Step height	0.0007
18		Knee angle	0.0003
19		Hip angle	0.0725
20		Ankle angle	0.0002
21		Elevation angle	0.0142
22	centre of gravity	The mean of the height of the center of gravity	0.1419
23		Maximum distance between Center of Gravity projection and supporting Center of Gravity	0.0396
24		The mean of distance between gravity projection and supporting center of gravity	0.0032

25	Coefficient of variation in distance between barycenter projection and supporting center of gravity	0.0468

3.3 Strength and Weakness

Strength:

（1）We compute several types of gait parameters from the original data to describe the gait of elderly people, which have more more obvious physical implications. The selected 25 features are highly representative.

（2）MATLAB software is used to calculate gait parameters from the original data, which makes the calculation more simple and easy to implement.

（3）MATLAB software is used to implement test for significance. The results showed that the decision deleting variables that are not significant is scientific, which can lay a good foundation for the establishment of the model in second problems.

Weakness:

（1）Some information was lost in the process of calculating gait parameters from the original data.

（2）Because the selected 25 body balance features could not represent all the information of the data, the information contained in the characteristic value is fuzzy to a certain extent.

4. Problem2:The Classification Model Based on Logistic Regression Analysis(CL Model)

4.1Assumptions

1) Assuming that the selected principal components are highly representative.

2) Assuming that the balance capacity is closely related to the probability of falling down.

4.2Data Pre-processing

The original database of problem 2, which comes up with 25 indicators and the value of fell down, contains incomplete, redundant, and noisy information as expected in any real-world data. In order to establish a reliable balance risk assessment system, we decide to do meticulous data pre-processing at first.

4.2.1 Generate Classification Tag

From the data given in Annex 1, we can find that only 13 of the 76 test samples fell down in one year, which count for a low proportion. And the number of samples falling down more than 1 times in one year is much less than the whole.

As we decide to build up balance risk assessment system using classify model, we divide the elderly people into two groups, have or have not fell down in one year so that the value of fall times becomes boolean one.

4.2.2 Over-sampling Technique

However, the proportion of people who have fell down still count for a low proportion, this leads to an increase in the probability of misjudgment in our discriminant analysis and cause the problem of class-imbalance, thus we need to pre-process the data.

We adopt SMOTE(for Synthetic Minority Over-sampling Technique) to implement the data processing.

The SMOTE algorithm is put forward to counter the effectiveness of having a few instances of the minority class in data sets. The main idea is to balance the data sets. The basic principle of SMOTE algorithm is shown as follow:

Algorithm SMOTE:

For each point p in S:

1.Compute its k nearest neighbors in S.

2.Randomly choose r<k of the neighbors (with replacement)

3.Choose a random point along the lines joining p and each of the r selected neighbors.

4.Add these synthetic points to the data set with class S.

$$p_i = x + rand(0,1) \cdot (y_i - x), i = 1,2,\ldots,N$$

Where p is new minority sample, x is sample, and y is randomly selected k nearest neighbors.

4.2.3 Standardization

In order to limit the influence of dimension and range of the 25 indicators, we standardize these variables with the formulation:

$$x' = \frac{x - \mu}{\sigma}$$

Where μ is the average and σ is the standard deviation.

4.2.4 PCA(Principal component analysis)

Principal component analysis is used to reduce the dimension of raw data. The value of the principal component can replace the original amount of data.
F1 present the principal component index of the first linear combination of 29 gait parameters data:

$$F_1 = a_{11}X_1 + a_{21}X_2 + ... + a_{p1}X_p$$

If the first principal component F1 is not sufficient to represent the original data of the five groups of information, and then consider the selection of second principal component F2. The information contained in F1 and F2 remain independent, covariance as follows:

$$COV(F_1, F_2) = 0$$

And so on,constructing principal components F1 , F2 , , Fm were used as the first、second、...、the m-th component of the original data：

$$F_m = a_{m1}X_1 + a_{m2}X_2 + ... + a_{mp}X_p$$

According to the above analysis, we can get the principal components with Fi and Fj not related, which means and the covariance cov(Fi,Fj)=0.

4.3Establishment of Model:Logistic Regression

After pre-processing the value of fell times become binary one, we choose logistic regression to deal with the problem. The most commonly used ordinal logistic model was called the proportional odds(PO) model. The PO model is best stated as follows, for a response variable having levels : 0,1,2,…， k

$$\Pr[Y \geq j|X] = \frac{1}{1+\exp[-(\alpha_j + X\beta)]}, j = 1,2,\ldots,k$$

And we utilize logistic regression by MATLAB.

4.4Model Solving Process

The specific steps to establish a balance risk assessment system based on 25 indicators from question 1 using logistic regression are as follows:

4.4.1 Data Pre-processing

First, we generate classification tag and classify the value of fell times into two value, has fell down or not. Then we use SMOTE (for Synthetic Minority Over-sampling Technique) to deal with the problem of class-imbalance problem.

Table2.1 : The sample size before and after SMOTE

	Fell down	Never Fell
Before SMOTE	13	63
After SMOTE	63	63

After that we utilize the PCA to reduce the dimension of database from 25 to 20 and the percentage of total variance explained is 98.2531. The 20 selected variables can explain most of the information from 25 variables.

4.4.2 Logistic Regression

After the We treat the 25 body balance features from question1 as 25 independence and the value of fell times in one year as dependence. By using logistic regression, we can get the formulation of logistic regression and some accordingly parameters.

Table 2.2: The parameters of logistic regression

	Estimate	SE	T-Stat	P Value
1	-0.0244	0.0426	-0.5727	0.5680
2	0.1910	0.0484	3.9469	0.0001
3	-0.0188	0.0496	-0.3800	0.7047
4	0.0492	0.0587	0.8387	0.4036
5	-0.0036	0.0414	-0.0870	0.9308
6	-0.0757	0.0670	-1.1305	0.2609
7	-0.2782	0.0687	-4.0517	0.0001
8	-0.1285	0.0648	-1.9826	0.0500
9	-0.0210	0.0504	-0.4164	0.6780
10	0.2842	0.0566	5.0176	0.0000
11	0.0178	0.0607	0.2939	0.7694
12	0.0734	0.0720	1.0204	0.3099
13	-0.1022	0.0548	-1.8633	0.0652
14	-0.0849	0.0588	-1.4430	0.1520
15	-0.0839	0.0784	-1.0700	0.2871
16	0.1819	0.0623	2.9169	0.0043
17	0.2025	0.0561	3.6090	0.0005
18	0.1868	0.0650	2.8730	0.0049
19	0.1005	0.0460	2.1859	0.0310
20	-0.1961	0.0533	-3.6755	0.0004
(Intercept)	0.5000	0.0334	14.9892	0.0000

Putting the processed data into the model, we can get a excellent result of the balance risk assessment system and only 4 of 76 elderly people are classified incorrectly.

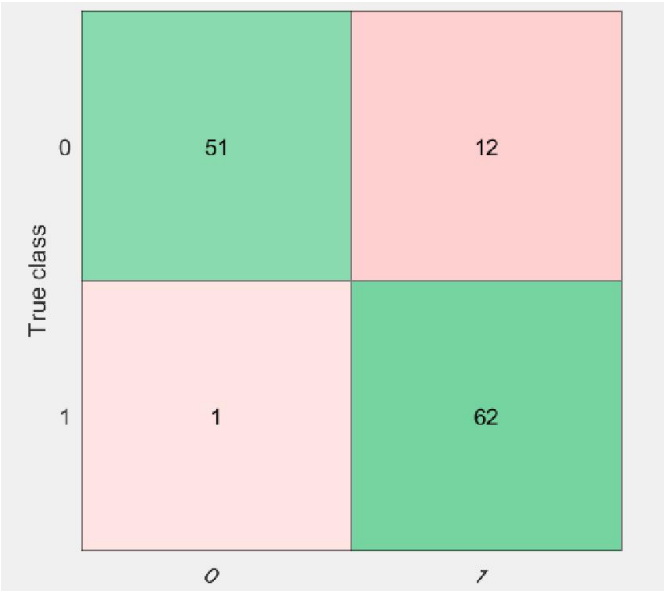

Figure3 : The result of balance risk assessment system (processed data)

4.5 Give Advice

According to the analysis above, we provide the elderly with some suggestions from four aspects as to help prevent falls in elderly people .

(1) The elderly should pay attention to controlling the walking speed while walking. At the same time, the elderly should also pay attention to the width of the step and foot height.

(2) From a kinematics point of view, the elderly should adjust their walking posture as much as possible. For the abnormal joint angle increases the risk of falling, it is of great importance for them to maintain a proper walking posture.

(3) As the larger the asymmetry indices, the more likely to fall in elderly people, it helps to keep the symmetry of the left and right limbs. This requires their conscious adjustment, if necessary, through medical machinery.

(4) Because the relationships between center of gravity projection and supporting center of gravity have impact on the result, the elderly are suggested to pay attention to the position of center of gravity, and sometimes crutches can help.

5. Problem3: Analog Computation and a Comparative Analysis model

5.1 Analog Computation

In this part, we utilize the CL model to make an analog computation with the original data from the Annex 2. It is because that according to the model assumption of problem 2, we assume that the balance risk is closely relative to the probability of falling down. In that case, this model can greatly explain the probability of falling down.

After input the ordinary data from Annex2 we discover that the model can describe the probability of falling down very well and the proportion of analog correct is 94.73%.

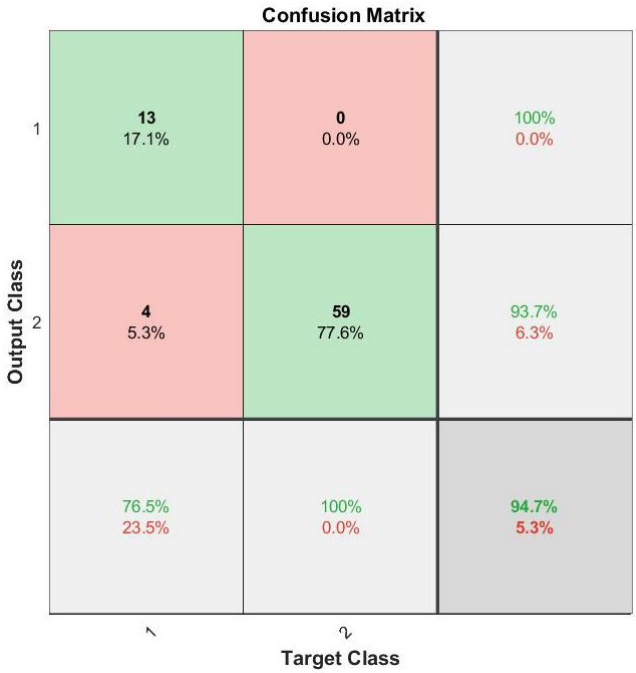

Figure4 : The result of analog computation

5.2 A Comparative Analysis of the Body Balance Force

Due to the original data always composed of incomplete, redundant, and noisy information, it is necessary for us to preprocess the data before doing the comparative analysis. Then we use the multivariate logistic regression to access the estimate of coefficients of each independent variables. Finally, we can compare the estimate of coefficients of each variables to analyse the influence of each element.

5.2.1 Data Pre-processing

As original database contains redundant information, there were several features that could not be treated directly since they had a high percentage of missing values. These values are shown as following.

Table3.2 : Missing value of some lists in Annex 1

list	W	X	Z	AA	AC	AD	AF	AG	AI	AJ
Missing ratio	41.25 %	40.00 %	41.25 %	41.25 %	38.75 %	38.75 %	38.75 %	38.75 %	38.75 %	38.75 %

Then there are some other variables which have useless actual meanings. Fall data was considered to be too sparse and it was not included in further analysis. Fall

times in year 2015 was removed since its value is closely related with fall times in one year.

There are 3 results of force platform tests' lead leg, so we calculate the final result of the three tests as a synthesis one. We convert left to 1 , right to 0 and calculate the Boolean calculation as the final result. The missing value are converted to 0.5. The final result of three staircase tests are handled with the same method.

Besides, as there are some char variables, we convert them to discrete variables according to the law from the table.

Table 3.3: Logistic Regression Assignment

Variables	Assignment
Sex	1 = male, 0 = female
BMI	1= less than 18.5, 2 = 18.5-23.9, 3 = 24-27, 4 = 28-32, 5 = more than 5
Fall times in one year	0=0, 1=1 or 2 or 3
Complains1（*Orthopedic diseases*）	1 = if complains contain bone fracture history or Osteoporosis or Sarcopenia, 0 = else
Complains2（*Cardiovascular diseases*）	1 = if complains contain anoxia or Head trauma or visual impairment unadjustable by lenses or hypertension or drug/alcohol withdrawal symptoms, 0 = else
Complains3（*Neurological diseases*）	1= if Neuropathic disorder and vestibular disorder, 0 = else

Finally, we do numerical standardization of numerical variables with the formulation as following to eliminate the influence of dimension and range.

$$x' = \frac{x - \mu}{\sigma}$$

Where μ is the average and σ is the standard deviation.

5.2.2 Comparative Analysis of Body Balance Force

We can use the CI model to evaluate the body balance force. After pre-processing the database of problem 3, we can utilize software SAS to do logistic regression. And we can assess the result as following.

Table3.4:Odds Ratio Estimates (result from logistic regression)

Effect	Point Estimate	95% Wald Confidence Limits	
Age	1.1300	0.4210	3.0340
Height	0.2170	0.0030	16.5900
Weight	57.1130	0.0320	>999.999
BMI	0.0290	<0.001	16.5430
High press	0.7210	0.2110	2.4690

Low press	1.3130	0.4390	3.9290
Heart rate	0.6850	0.2730	1.7180
sex	0.1250	0.0130	1.2210
com1	3.1230	0.6360	15.330
com2	0.5760	0.1060	3.1240
com3	0.7350	0.0420	12.9840
force_platform	>999.999	<0.001	>999.999
staircase	1.0350	0.2080	5.1540
aid	2.0300	0.1970	20.9180

Table3.5:Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Pr > ChiSq	Pr > ChiSq
Intercept	1	-13.4538	306.8000	0.0019	0.965
Age	1	0.1224	0.5038	0.0590	0.808
Height	1	-1.5268	2.2121	0.4764	0.4901
Weight	1	4.0450	3.816	1.1237	0.2891
BMI	1	-3.5325	3.234	1.1932	0.2747
High press	1	-0.3265	0.6277	0.2707	0.6029
Low press	1	0.2721	0.5593	0.2367	0.6266
Heart rate	1	-0.3785	0.4693	0.6503	0.4200
sex	1	-2.0834	1.1648	3.1996	0.0737
com1	1	1.1387	0.8118	1.9678	0.1607
com2	1	-0.5509	0.8623	0.4082	0.5229
com3	1	-0.3084	1.4654	0.0443	0.8333
force_platform	1	12.3323	306.8000	0.0016	0.9679
staircase	1	0.0341	0.8193	0.0017	0.9668
aid	1	0.7082	1.1900	0.3542	0.5518

By observing the table3.4 and 3.5, we come to the conclusion that complaint1 (orthopedics illness) , sex and IBM is important, and the female take greater risks than the male. Also , despite the outlier, most kinds of complaints generally play a key role in the assessment of falling risk, and the subjects not sick see a lower risk of falls(with OR<1).

Furthermore, as some of the independent variables are not necessary for explanation dependent variables, we consider use stepwise logistic regression to study the problem further.

Stepwise logistic regression is a systematic method for adding and removing terms from logistic model based on their statistical significance in explaining the response variable. We uses forward and back ward stepwise regression to determine a final model and access the final result in table3.6.

Table3.6:Analysis of Maximum Likelihood Estimates (result from stepwise logistic regression)

	Estimate	SE	T Stat	P Value
Intercept	0.1869	0.0693	2.6955	0.0087
Heart rate	0.0099	0.0530	0.1863	0.8527
Complain3	-1.0353	0.3309	-3.1286	0.0025
staircase	-0.0240	0.0979	-0.2448	0.8073
Heart rate*complain3	-0.8218	0.1407	-5.8429	1.2846E-07
Complain3*staircase	0.9015	0.3731	2.4157	0.0182

According to the table above, we can conclude that complain3(mental illness) has relatively more significant impact on the risk of falls.

To observe the influence of some variables more visible, we draw several figures as following.

5.2.3Advice to Elderly People with Weak Balance Ability

As we can see from the analysis above, the elderly with weak balance ability are suggested to keep a healthy lifestyle to maintain good physical condition and thus to reduce the risk of falls.

1) Multiple types of exercise contribute to the maintenance of bone mass. Also, good nutrition is important in preventing osteoporosis, including adequate calcium, vitamin D, vitamin C, and protein. Consequently the elderly should do regular exercise with proper intensity and pay attention to intake of nutrient elements like calcium which helps to relieve osteoporosis as well. Strong bones is of help to maintain body balance.

2) Older people are vulnerable to cardiovascular disease, so they should avoid intense movement and lead a regular life. In terms of diet, more fruit and vegetables and less salt are preferred. Control of weight is also important, obesity may limit the movement and reactive potency so that the risk of falls increases.

3) Nerve injury or head injury has an obvious influence on the balance ability of the elderly. So they are expected to discontinue the habit of smoking and drinking, and keep a good sleep condition. What' more, a positive attitude is necessary for the elderly population to the treatment of diseases.

5.3 Strength and Weakness

Strength:

1) By creating Boolean variables from the original data in Annex1, we make the data easier to analyze and the results are more readable.

2) We built Logistic regression model based on the case-control principle, the result shows the decision is scientific. And the use SAS software make the operations more

convenient and effective.

3) By observation of the p value and the odds ration of each parameter, we get a more reliable results of comparison.

Weakness:

1) During the pre-processing of raw data, some information get lost, the constructed variables may not able to include complete information.

2) The test result is not satisfying due to the small sample size.

6. Conclusion

6.1 Advantages

1) The model of problem1 firstly raise 29 balance related variables according to credit books.

2) The database of problem 2 and 3 are preprocessed carefully with the proper methods to deal with the missing value, char variables, data standardization and reducing dimension(PCA). It greatly enhance the credibility of our models.

3) The principal component analysis method can reduce the dimension of original data and generate new variables which can make the logistic regression more feasible. Besides , those new variables can still be explained as they are related with original data by coefficients.

6.2 Disadvantages

1) The establish of the classification model based on Logistic regression are dependent on the assumption that the balance risk are closely related with the probability of falling down. However, the balance risk may not have the absolute strong relativity with the probability of falling down.

2) Because of the possible presence of multiple linear in the multiple logistic regression, it is possible that the model spends more time than the superior one.

3) The data pre-processing inevitably lead to some loss of information of the original data, which may influence the result to some extent.

6.3 Improve Method of the Model

When we deal with problem 3, we may get more accurate results if we could divide the data base into different categories to analyse the feature of each category. And some of the variables losing most original information can be treated as extra group to analyse their property.

7. Reference

[1] Yu Weihua, Wang Li, Xu Zhongmei. A study on the correlation between gait characteristics and falls of elderly people in community [J]. Chinese Journal of Nursing, 2017 (01): 44-48.

[2] Li Zhongyi. Characteristics and gait analysis of human movement in knee joint rehabilitation robot design [J]. Henan Science and Technology, 2018 (07): 14-16.

[3] Yang Ninfeng. Coordination of Human Motion and its Parametric description [D]. Tsinghua University, 2001

[4] Sui Shao Kun. Research and system realization of human gravity detection method [D]. Hebei University, 2015.

[5] Xionggao Zou,Yueping Feng,Huiying Li,Shuyu Jiang. Improved over-sampling techniques based on sparse representation for imbalance problem[J]. Intelligent Data Analysis,2018,22(5).

[6] Shi Xiaoyu, Guo Jinchen, Liu Lanlin, Feng Ye, Cheng Yue, Li Kailin, Zhou Chao, Gong Yue, Qi Zhuocao, Gao Ting. Physical constitution investigation and Logistic regression analysis of 135patients with insomnia [J]. Journal of Shanxi Institute of traditional Chinese Medicine, 201819 (05): 5-8 11

[7] Kong Fancheng, Xia Xin Yan. Comparative study on Interprovincial Science and Technology Innovation ability based on Principal component Analysis [J]. Frontier: 1-7 [2018-11-26].

[8] Gao Wenlong, Liu Xiaoning, Yan Hong. SAS implementation of a logistic regression ratio estimation method [J]. China Health Statistics, 2014: 31 (03): 530-531.

[9] Tian Jia-min, Hu Lingyan. Stepwise regression and Neural Network influencing factors on Sleep quality [J]. Journal of Jiamusi University (Natural Science Edition), 2018N 36 (04): 625-628 644.

[10] Chen Bo. Prevention of cardiovascular disease based on scientific management [J]. Journal of electrocardiogram (EKG), 2018, 7 (03): 200-202.

```
function feature = generateFeature(data)
%% Features
feature = zeros(1,29);

[nrow, ncol] = size(data);

frame = data(:,1);
time = data(:,2);
variables = cell(1,42);
for i = 1:42
  variables{i} = data(:,3*i:3*i+2);
end

%% feature01
point1l = variables{33};
point2l = 0.5*(variables{10}+variables{28});
point3l = variables{8};
point1r = variables{33};
point2r = 0.5*(variables{9}+variables{27});
point3r = variables{7};
feature01Left = zeros(nrow,1);
feature01Right = zeros(nrow,1);
for k = 1:nrow
  pointsl = [point1l(k,:);point2l(k,:);point3l(k,:)];
  feature01Left(k) = getAngle(pointsl);
  pointsr = [point1r(k,:);point2r(k,:);point3r(k,:)];
  feature01Right(k) = getAngle(pointsr);
end

feature(1) = mean(feature01Left);
feature(2) = mean(feature01Right);

%% feature02
point1l = variables{28};
point2l = variables{33};
point3l = variables{33}+[0 0 -1];
point1r = variables{27};
point2r = variables{33};
point3r = variables{33}+[0 0 -1];
feature02Left = zeros(nrow,1);
feature02Right = zeros(nrow,1);
for k = 1:nrow
  pointsl = [point1l(k,:);point2l(k,:);point3l];
  feature02Left(k) = getAngle(pointsl);
  pointsr = [point1r(k,:);point2r(k,:);point3r];
  feature02Right(k) = getAngle(pointsr);
end

feature(3) = mean(feature02Left);
feature(4) = mean(feature02Right);
```

```
%% feature03
point1l = variables{35};
point2l = variables{26};
point3l = variables{8};
point1r = variables{34};
point2r = variables{25};
point3r = variables{7};
feature03Left = zeros(nrow,1);
feature03Right = zeros(nrow,1);
for k = 1:nrow
  pointsl = [point1l(k,:);point2l(k,:);point3l(k,:)];
  feature03Left(k) = getAngle(pointsl);
  pointsr = [point1r(k,:);point2r(k,:);point3r(k,:)];
  feature03Right(k) = getAngle(pointsr);
end

feature(5) = mean(feature03Left);
feature(6) = mean(feature03Right);

%% feature04
point1 = variables{32};
point2 = variables{15};
point3 = variables{15} + [0 0 1];
feature04 = zeros(nrow,1);
for k = 1:nrow
  points = [point1(k,:);point2(k,:);point3];
  feature04(k) = getAngle(points);
end

feature(7) = mean(feature04);

%% feature05
feature05Left = variables{37}(:,3);
feature05Right = variables{36}(:,3);

feature(8) = max(feature05Left);
feature(9) = max(feature05Right);

%% feature06
[~,indexL] = min(variables{37}(:,3));
point1l = variables{35}(indexL,:);
point2l = variables{37}(indexL,:);
point3l = [(point1l(:,1:2)-point2l(:,1:2)),0];
pointsl = [point1l;point2l;point3l];

[~,indexR] = min(variables{36}(:,3));
point1r = variables{34}(indexR,:);
point2r = variables{36}(indexR,:);
point3r = [(point1r(:,1:2)-point2r(:,1:2)),0];
pointsr = [point1r;point2r;point3r];

feature06Left = getAngle(pointsl);
```

```
featureO6Right = getAngle(pointsr);

feature(10) = featureO6Left;
feature(11) = featureO6Right;

%% featureO7
[~,indexL] = min(variables{35}(:,3));
point1l = variables{37}(indexL,:);
point2l = variables{35}(indexL,:);
point3l = [(point1l(:,1:2)-point2l(:,1:2)),0];
pointsl = [point1l;point2l;point3l];

[~,indexR] = min(variables{34}(:,3));
point1r = variables{36}(indexR,:);
point2r = variables{34}(indexR,:);
point3r = [(point1r(:,1:2)-point2r(:,1:2)),0];
pointsr = [point1r;point2r;point3r];

featureO7Left = getAngle(pointsl);
featureO7Right = getAngle(pointsr);

feature(12) = featureO7Left;
feature(13) = featureO7Right;

%% featureO8

featureO8 = 2*abs(feature(1)-feature(2))/(feature(1)+feature(2));
feature(14) = featureO8;

%% featureO9

featureO9 = 2*abs(feature(3)-feature(4))/(feature(3)+feature(4));
feature(15) = featureO9;

%% feature10

feature10 = 2*abs(feature(5)-feature(6))/(feature(5)+feature(6));
feature(16) = feature10;

%% feature11

feature11 = 2*abs(feature(7)-feature(8))/(feature(7)+feature(8));
feature(17) = feature11;

%% feature14

feature14 = abs(mean(variables{36}(:,2))-mean(variables{37}(:,2)));
feature(18) = feature14;

%% feature15
point1l = variables{35};
point2l = variables{37};
point3l = variables{37} + [1 0 0];
```

```
point1r = variables{34};
point2r = variables{36};
point3r = variables{36} + [1 0 0];
feature15Left = zeros(nrow,1);
feature15Right = zeros(nrow,1);
for k = 1:nrow
  pointsl = [point1l(k,:);point2l(k,:);point3l];
  feature15Left(k) = getAngle(pointsl);
  pointsr = [point1r(k,:);point2r(k,:);point3r];
  feature15Right(k) = getAngle(pointsr);
end

feature(19) = mean(feature15Left);
feature(20) = mean(feature15Right);

%% feature16
feature16 = 2*abs(feature(19)-feature(20))/(feature(19)+feature(20));
feature(21) = feature16;

%% feature17
len = max(variables{37}(:,1))-min(variables{37}(:,1));
feature17 = len/length(time);
feature(22) = feature17;

%% feature18
feature18 = length(time)/3;
feature(23) = feature18;

%% feature19
feature19 = 2*abs(feature(10)-feature(11))/(feature(10)+feature(11));
feature(24) = feature19;

%% feature20
feature20 = 2*abs(feature(12)-feature(13))/(feature(12)+feature(13));
feature(25) = feature20;

%% feature21
gravity1 = 0.0405*variables{41};
gravity2 = 0.0295*variables{32};
gravity3 = 0.483*(0.5*variables{32}+0.25*variables{16}+0.25*variables{17});
gravity4 = 0.052*0.25*(variables{16}+variables{18}+variables{17}+variables{19});
gravity5 = 0.028*0.25*(variables{20}+variables{18}+variables{21}+variables{19});
gravity6 = 0.015*0.25*(variables{20}+variables{38}+variables{21}+variables{39});
gravity7 = 0.2115*(variables{11}+variables{12});
gravity8 = 0.107*(variables{7}+variables{8});
gravity9 = 0.034*(variables{1}+variables{2});
gravity = gravity1;
for k = 2:9
  eval(['gravity','=','gravity','+','gravity',char(num2str(k)),';']);
end
feature21 = mean(gravity);
feature(26) = feature21(3);
```

```
center = 0.25*(variables{34}+variables{35}+variables{36}+variables{37});
dist = zeros(nrow,1);
for k = 1:nrow
  dist(k) = norm(gravity(k,1:2)-center(k,1:2));
end

feature(27) = max(dist);
feature(28) = mean(dist);
feature(29) = std(dist)/mean(dist);
```

第二问:

```
clear, clc
%% main
load('dataFile.mat');

nameList = [
    "liangzengli_g9" % "liangzhengli_g9"
    "wenyanfang_g9"
    "yangxijin_g9"
    "hanwenshan_g9"
    %"hujiawei_g9"
    "pengguizhu_g9"
    "wangfulan_g9"
    "yangwanlin_g9"
    "baiyulin_g9"
    "guodafa_g9"
    "jiashengpu_g10"
    "lidee_g9"
    "wangdeqin_g9"
    "yangzongfen_g10"
    "zhaojuan_g9"
    "gengxiulin_g9"
    %"hanjianshe_g9"
    "haoyubin_g9"
    "lixiaoru_g5"
    "machunxia_g9"
    "renweiyi_g9"
    "sunshaohua_g9"
    "xiangxu_g9"
    "lirenfan_g9"
    %"lishuhau_g9"
    "wangchunling_g9"
    "tianguilin_g9"
    "wujinzhan_g9"
    "xieshuli_g9"
    "xuxiuyun_g9"
    "fujianying_g9"
    "yanzuozhou_g9"
    "zhangzuoyin_g9"
    "baiqingquan_g9"
    "guanpeihua_g9"
    "liangyuxing_g9"
```

```
    "zhaojinhua_g9"
    "zhaoshurong_g9"
    "chenfue_g9"
    "cuizhenhua_g9"
   % "gaoyuling_g9"
    "hanyingchun_g9"
    "liuzaoheng_g9"
    "niuzhenhang_g9"
    "xingjunmiao_g9"
    "yangchengyu_g9"
    "zhujianguo_g9"
    "cangyongli_g9"
    "ruanshuyin_g9"
    "tongyuhua_g9"
    "wangjianmin_g9"
    "zhangzhundao_g9"
    "aizhenjiang_g10"
    "lijingjing_g8"
    "wangzhining_g9"
    "wangjiuhong_g9"
    "maguicheng_g9"
    "liwenlong_g9"
    "zhangzuowen_g9"
    "zongkeqin_g9"
    "pengruiying_g9"
    "yangbaoling_g9"
    "qijianming_g9"
    "zhaojie_g9"
    "songfang_g9"
    "lijianke_g9"
    "zhaoshufen_g9"
    "wangxiong_g9"
    "hanyongchang_g9"
    "wanghancai_g9"
    "jinyan_g9"
    "sunxingjian_g9"
    "cuixiulan_g9"
    "jianglimin_g9"
    "weixiurong_g9"
    "lvjun_g9"
    "litongsheng_g9"
    "maochengai_g9"
    "zonglanfang_g9"
    "liuyuexian_g9"];

n = length(nameList);
Features = zeros(n,29);
for k = 1:n
  eval(['data','=',char(nameList(k)),';']);
  Features(k,:) = generateFeature(data);
end

csvwrite('features.csv',Features);
```

```
%% add labels

load('label.mat')

Featureslabel = zeros(76,1);
for k = 1:76
  name = char(nameList(k));
  name = name(1:end-3);
  for index = 1:80
    NameEle = char(Name(index));
    if strncmp(NameEle,name,9)
      Featureslabel(k) = label(index);
      break
    end
  end
end

result = [Features, Featureslabel];

mdl0 = fitglm(Features,Featureslabel);

csvwrite('result.csv',result);

%% PYTHON!!!


% clear, clc
%% SMOTE
X_smo = load('smote_X.csv');
y_smo = load('smote_y.csv');
X_norm = (X_smo-mean(X_smo))./std(X_smo);
colIndex = [1 2 3 4 5 6 7 10 11 12 13 14 15 16 17 18 19 20 21 23 25 26 27 28 29];
X_norm = X_norm(:,colIndex);
data_smo = [X_norm,y_smo];

index1 = y_smo==1;
index0 = y_smo==0;
class1 = X_norm(index1,:);
class0 = X_norm(index0,:);
csvwrite('class1.csv',class1);
csvwrite('class0.csv',class0);



%%
mdl = fitglm(X_norm,y_smo);

result_table = mdl.Coefficients;
writetable(result_table,'table.csv');

%% pca
[coeff,score,latent,tsquared,explained] = pca(X_norm);
```

```
X_new = X_norm*coeff';
mdl_new = fitglm(X_new(:,1:20),y_smo);
data_smo2 = [X_new(:,1:20),y_smo];

result_table2 = mdl_new.Coefficients;
writetable(result_table2,'table2.csv');


%% NewData

% rand

index1 = y_smo==1;
index0 = y_smo==0;
class1 = X_smo(index1,:);
class0 = X_smo(index0,:);

mean0 = mean(class0);
std0 = std(class0);
mean1 = mean(class1);
std1 = std(class1);

colIndex = [1 2 3 4 5 6 7 10 11 12 13 14 15 16 17 18 19 20 21 23 25 26 27 28 29];

X0 = std0.*randn(1e+3,29)+mean0;
X0_norm = (X0-mean(X_smo))./std(X_smo);
X0 = X0_norm(:,colIndex);
X1 = std1.*randn(1e+3,29)+mean1;
X1_norm = (X1-mean(X_smo))./std(X_smo);
X1 = X1_norm(:,colIndex);
% use toolbox
load('model.mat')
y0 = trainedModel1.predictFcn(X0);
y1 = trainedModel1.predictFcn(X1);
error = (sum(y0)+sum(y1))/2000;



X_data = Features;
y_data = Featureslabel;
X_norm_data = (X_data-mean(X_data))./std(X_data);
colIndex = [1 2 3 4 5 6 7 10 11 12 13 14 15 16 17 18 19 20 21 23 25 26 27 28 29];
X_norm_data = X_norm_data(:,colIndex);
data_data = [X_norm_data,y_data];
% use toolbox
load('model.mat')
yfit = trainedModel1.predictFcn(X_norm_data);

%% test
res = zeros(1,4);
for k = 1:length(yfit)
  if y_data(k)==0
    if yfit(k)==0
```

```
      res(1) = res(1)+1;
    else
      res(2) = res(2)+1;
    end
  else
    if yfit(k)==0
      res(3) = res(3)+1;
    else
      res(4) = res(4)+1;
    end
  end
end

bar(res)

figure
plot(yfit,'bo')
hold on
plot(y_data,'r*')
```

Python

```
import pandas as pd
from imblearn.over_sampling import SMOTE

# data before oversampling
data = pd.read_csv('result.csv',header=None)
X = data.iloc[:,range(29)]
y = data.iloc[:,29]
# smote
smo = SMOTE(random_state=666)
X_smo, y_smo = smo.fit_sample(X, y)
X_smo = pd.DataFrame(X_smo)
y_smo = pd.DataFrame(y_smo)
# save
X_smo.to_csv("smote_X.csv",index=False,sep=',',header=None)
y_smo.to_csv("smote_y.csv",index=False,sep=',',header=None)
```