IDENTIFYING SIGNIFICANT FACTORS INFLUENCING METABOLIC SYNDROME IN CHINA

A Paper
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Xiaoxue Gu

In Partial Fulfillment of the Requirements
for the Degree of
MASTER OF SCIENCE

Major Department:
Statistics

April 2015

Fargo, North Dakota

North Dakota State University
Graduate School

Title

IDENTIFYING SIGNIFICANT FACTORS INFLUENCING

METABOLIC SYNDROME IN CHINA

By

Xiaoxue Gu

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

MASTER OF SCIENCE

SUPERVISORY COMMITTEE:

Seung Won Hyun
Chair

Yarong Yang

Mark A. Strand

Approved:

04/16/2015
Date

Rhonda Magel
Department Chair

ABSTRACT

Metabolic Syndrome occurs when a person's body does not properly use and store energy. The disease has five criteria: abdominal obesity, insulin resistance, hypertension, dyslipidemia, and impaired glucose regulation. The purpose of this paper was to analysis a longitudinal data obtained from China. The data was collected using surveys in 2008 and 2012. For finding the factors that contributed significantly to the development of Metabolic Syndrome, a marginal model was applied. To fit the marginal model, the Generalized Estimating Equation method was used. The developed model did not have high accuracy of presenting the proportion of true results (Metabolic Syndrome observed and no Metabolic Syndrome observed).

ACKNOWLEDGEMENTS

I would like to express my sincerest gratitude to the following individuals:

Dr. Seung Won Hyun and Dr. Rhonda Magel, my major advisers, thank you for your help, instruction, and advice during the courses of my graduation education and the preparation for this paper.

Dr. Mark A. Strand, thank you for your help and encouragement during my whole graduation education, and providing the data which is used in this paper. I appreciate that you gave me an assistantship. Thank you for your trust.

Dr. Yarong Yang, thank you for serving as my graduate committee, giving me valuable comments.

Curt Doetkott, thank you for your help on the SAS program. Thanks to your effort I now have a strong understanding of SAS and how to work with the output.

Matthew Warner, thank you for your guidance on my writing skill. My meetings with you helped me to express my knowledge. Thank you for your suggestions.

Hanzhe Li, my husband, thank you for you being in my life, giving me infinite love, understanding, support, and encouragement. You have been there for me and I know you will always be there for me.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1. INTRODUCTION

In recent decades, tremendous economic development in China had multifarious effects on the Chinese population. These effects occur in many areas of life including education level, work status and lifestyle habits, such as amount of physical activity, dietary choices, smoking, and alcohol consumption. These areas may have effects on people's health status and the disease of Metabolic Syndrome. There is a high prevalence of Metabolic Syndrome in United State, nearly 23.7%, based on the data which collected from the Third National Health and Nutrition Examination Survey (1988-1994)[1]. For the Chinese population, not much investigation has been done to establish trends between these factors and diseases including Metabolic Syndrome.

The Metabolic Syndrome (MetS) is a disorder disease that occurs when a person's body does not properly use and store energy. Complications attributed to MetS include increased risk of diabetes and cardiovascular disease[2]. The diagnosis of MetS involves assessing a person with the following criteria: abdominal obesity, insulin resistance, hypertension, dyslipidemia, and impaired glucose regulation. Categorization into three or more of these criteria indicates a person has MetS. Gu's study [3] investigated MetS prevalence among Chinese adults and found 9.8% of men and 17.8% of women of the participants had MetS but among overweight individuals the rates were 26.9% for men and 31.1% for women. The incidence of MetS has increased quickly and is very prevalent in the Chinese population.

In this paper, I study the associations between the risks linked with the occurrence of Metabolic Syndrome and various factors, such as gender, dietary habits, amount of physical activity, smoking status, and alcohol consumptions. An existing longitudinal data which was collected in Yuci District, Jinzhong Prefecture, Shanxi Province, China is used. The data contains 637 individuals and each individual was observed twice, in 2008 and 2012, respectively. In a longitudinal data, the time-dependent covariate problem should be taken into consideration. For this longitudinal data, a marginal model will be developed, which can explain the relationships between the factors (explainary variables) and MetS (the response variable) by addressing the time-dependent covariate problem. To fit the marginal model, the Generalized Estimating Equation (GEE) method will be used. Based on the marginal model, this study identifies influential factors on the MetS and their effects.

In chapter 2, the background is explained. In chapter 3, data description is presented. In chapter 4, the research method is explained, and all results shows in chapter 5. In the final chapter, chapter 6, the conclusion is made.

CHAPTER 2. BACKGROUND

2.1. Collinearity

For any actual estimation of the model, a consideration of collinearity among explainary variables is necessary. If a collinearity exists, the coefficients of regression is fluctuate and the estimated variances become large. Therefore, a check for collinearity among the variables is necessary. One way to measure the level of collinearity is using the condition index [4]. The function of the condition index is defined as $\sqrt{\frac{\lambda_{\max}}{\lambda_j}}$, where λ_{max} is the largest eigenvalue, and λ_j (j=1,2, ... , p, and p is the number of explainary variables) is the corresponding eigenvalue. The range of condition index is from 1 to infinite. If the condition index for j^{th} in explainary variable is not greater than 10, there is no indication of collinearity for the j^{th} explainary variable. Then the variable can be used for further analysis.

The collinearity diagnostics table was used in this paper. In collinearity diagnostics table, the value of the condition index for each explainary variable is obtained. The collinearity diagnostics table is obtained by using the PROC REG procedure of SAS. The resulting collinearity table is only based on the relationships among the explainary variables, so the consumptions about the response variable does not matter. For example, a sample of individuals are observed in several time periods repeatedly, this may results that the response in one time period is correlated with the response in others. Even though the response variables are correlated within each other, still we can use the collinearity table to check if there is a collinearity among explainary variables, from the regression model.

2.2. Logistic Regression

Logistic regression is a very prevalent statistical method. It is extensively used in the bio-medical field to determine the relationship between a dichotomous response variable (occurrence/non-occurrence, true/false, female/male,etc.) and a group of explainary variables. Logistic regression permits us to achieve two goals. One goal is that we can check if the probability of getting a specific value of the response variables is linked with the explainary variables. The other goal is that we try to find the best fitting model to predict the probability of getting a specific value of the response variables based on the explainary variables [5].

Let the response variable (Y) be a binary variable, where the value of Y is 1 for occurrence and 0 for non-occurrence. The Y will be either 0 or 1, and p be the probability of Y=1, equivalently p=P[Y=1]. Now we can define the odds ratio (*OR*) in favor of Y=1:

$$OR = \frac{p}{1-p} = \frac{p[Y=1]}{p[Y=0]}$$

By the definition of the logistic regression model, the odds ratio is modeled simply by the form of $\frac{p}{1-p} = e^{\beta_0+\beta_1 x}$, and the natural logarithm of the odds ratio becomes

$$logit(p) = \ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X$$

Where X is an independent variable, β_1 represents the change in *logit(p)* by each 1-unit increased in X and the β_0 is *logit(p)* when X=0. Modeling the *logit(p)* based on a linear combination of independent variables is known as logistic regression.

2.3. Marginal Model

A longitudinal study tracks the same subjects over at least two different time periods. In general, this type of study design typically has time-dependent covariate on account of the repeated measurements from the same sample of individuals. Marginal model is often applied to the longitudinal data for addressing the time-dependent covariate problem. The marginal model can also be called as marginal mean model, because it stands for the population-averaged responses over individuals at the same time point. The advantage of population-averaged is that the conclusion about the comparisons of the population groups at different time points is easier to make. Specifically, it can balance the study individuals to get the mean estimates of overall responses by using time factor for dividing each level of independent variables. The advantage of marginal model is that if a sample can be balanced by time, then the effects of the factors will be unbiased. For a binary response variable, the marginal model is used to predict the logit of marginal probabilities logit[$P(Y_t=1)$], where Y_t means that the response of variable Y at t time period, to express the relationship between response variable Y and X in longitudinal study.

$$\text{logit}[P(Y_t = 1)] = \beta_0 + \beta_1 X_1 + \beta_2 t$$

where X_1 is the explainary variable, t is the number of time periods, $P(Y_t=1)$ stands for the probability of positive responses at time point t, and β_0 and β_1 are the coefficients that explain the effects of the variables.

2.4. Generalized Estimating Equation

Generalized Estimating Equation (GEE) method can provide poluation-averaged effects for longitudinal data [6]. It is known as an estimation for longitudinal marginal models and correlation structure. The method is based on quasi-likelihood estimation, and needs no consumption about the distribution of response subjects [7].

For binary repeated measurements, it is very hard to define joint distribution and this leads us to not using maximum-likelihood method but using quasi-likelihood method, and also we need to consider correlation among response variables. Working correlation structure can specify the correlation of the responses. Working correlation structure can has several different structure, such as “Independent” structure, “exchangeable” structure, “AR-1”structure, “Toeplitz” structure, etc. “The goal of selecting a working correlation structure is to estimate β more efficiently” [8]. For auto correlated data over time periods, the working correlation structure is assumed to have “exchangeable” structure [6].The correlation between each pair responses can be presented as:

$$\mathrm{Corr}(Y_i, Y_j) = \begin{cases} 1, i = j \\ \alpha, i \neq j \end{cases} \quad (1 \leq i,j \leq t)$$

For the assumed working correlation structure, the GEE method can estimate the correlations based on the given data.

The GEE also presents the value of QIC and QIC_u. QIC stands for the criterion for quasi-likelihood model selection. In likelihood-based model selection, AIC (Akaike's Information Criterion) is used as a criterion. It assumes that response variables are independent. For the correlated response variables, Pan [8] developed a refined version of AIC, named the QIC (Quasi-likelihood under the Independence model Criterion), for model selection in the

Generalized Estimating Equations (GEE). When several different models are compared, the one with smallest QIC value is preferred.

2.5. The Definition of Metabolic Syndrome

Several definitions for Metabolic Syndrome exist. For this research project, we will use the five criteria that the NIH identified in report by the National Cholestrol Education Program [9]. Research suggests for specific populations additional criteria are nessesary, so we will use the suggested cut-off points of waist circumference for Asians[10]. Though there are five criteria, a person needs to meet only three or more criteria to be classified into the Metabolic Syndrome group.

1. Increased waist circumference (≧ 90 cm for men, ≧ 80 cm for women)
2. Elevated Triglyceride ≧ 1.7 mmol/L
3. Low HDL cholesterol (HDL cholesterol ≤ 1.03 mmol/L for men and ≤ 1.29 mmol/L for women)
4. Elevated blood pressure (systolic blood pressure ≥ 130 mmHg and/or diastolic blood pressure ≥ 85 mmHg or current use of antihypertensive drugs)
5. Impaired fasting glucose (fasting plasma glucose ≥ 5.6 mmol/L)

CHAPTER 3. DATA DESCRIPTION

This study uses an existing longitudinal data for the research on a population in Yuci District, Jinzhong Prefecture, Shanxi Province, China. The data was collected in 2008 and 2012 from the same group people who are from three age cohorts, born in 1956, 1960-1961 and 1964. In 2008, a total of 793 subjects completed the study. When it came to 2012, a total of 643 completed the subsequent study. Since there is 6 observations have some missing independent variables, a total of 637 subjects will be used in this study. 637 individuals were observed in twice, in 2008 and 2012, respectively. The data used and the methods of data collection are described in detail Strand's paper [11]. In this paper, a total of 13 variables from the data will be used to build a model. The binary response variables is the status of MetS, and there are 12 independent variables. All variables are ordinal and description of each variable is as followed:

$\underline{Y}$: Status of MetS

$\underline{X_1}$: Gender;

$\underline{X_2}$: Is peopel's physical activity greater than 150 mins per week?

$\underline{X_3}$: Status of people's alchohol consumption;

$\underline{X_4}$: Smoking status;

$\underline{X_5}$: Frequency of eating bed-time snacks;

$\underline{X_6}$: Frequency of eating fruit;

$\underline{X_7}$: Frequency of eating meat;

$\underline{X_8}$: Frequency of eating tofu;

$\underline{X_9}$: Frequency of eating fry food;

$\underline{X_{10}}$: Frequency of eating preserved food;

$\underline{X_{11}}$: Frequency of drinking milk;

t: time period.

Table 1 is the frequency distribution of all variables in the two time period. According to Table 1, there is 41.92% people have MetS in 2008 and the rate is increased to 47.72% in 2012.

Table 1. A Frequency Table of Time Period by Specified Component

Characteristics	Status	Time Period (t)		
		No. (% in column)		
		2008	2012	Total
MetS (y)	No MetS	370 (58.08)	333(52.28)	703
	Have MetS	267 (41.92)	304 (47.72)	571
Gender (X_1)	Men	209 (32.81)	211 (33.12)	420
	Women	428 (67.19)	426 (66.88)	854
Physical Activity (X_2)	< 150 min/week	614 (96.39)	603 (94.66)	1217
	≧ 150 min/week	23 (3.61)	34 (5.34)	57
Regular Alcohol Consumption (X_3)	No	436 (68.45)	374 (58.71)	810
	Occasionally	95 (14.91)	142 (22.29)	237
	Quit (>1 year)	20 (3.14)	39 (6.12)	59
	2-3 times/week	86 (13.50)	82 (12.87)	168
Smoking (X_4)	Never	479 (75.20)	469 (73.63)	948
	Quit (>10 year)	24 (3.77)	33 (5.18)	57
	Yes	134 (21.04)	135 (21.19)	269
Bedtime Snacks (X_5)	Rarely	582 (91.37)	594 (93.25)	1176
	Occasionally	45 (7.06)	29 (4.55)	74
	4 times/week	10 (1.57)	14 (2.20)	24

Table 1. A Frequency Table of Time Period by Specified Component (continued)

Characteristics	Status	Time Period (t)		
		No. (% in column)		
		2008	2012	Total
Fruit (X_6)	Rarely	71 (11.15)	59 (9.26)	130
	Occasionally	155 (24.33)	205 (32.18)	360
	4 times/week	411 (64.52)	373 (58.56)	784
Meat (X_7)	Rarely	88 (13.81)	96 (15.07)	184
	Occasionally	257 (40.35)	299 (46.94)	556
	4 times/week	292 (45.84)	242 (37.99)	534
Tofu (X_8)	Rarely	22 (3.45)	52 (8.16)	74
	Occasionally	267 (41.92)	346 (54.32)	613
	4 times/week	348 (54.63)	239 (37.52)	587
Fry Food (X_9)	Rarely	428 (67.19)	508 (79.75)	936
	Occasionally	191 (29.98)	121 (19.00)	312
	4 times/week	18 (2.83)	8 (1.26)	26
Preserved Food (X_{10})	Rarely	298 (46.78)	373 (58.56)	671
	Occasionally	235 (36.89)	170 (26.69)	405
	4 times/week	104 (16.33)	94 (14.76)	198
Milk (X_{11})	Rarely	287 (45.05)	311 (48.82)	598
	Occasionally	129 (20.25)	152 (23.86)	281
	4 times/week	221 (34.69)	174 (27.32)	395

CHAPTER 4. RESEARCH METHODS

In this paper, first the collinearity table is obtained to check if there is a collinearity among all independent variables. After that, the marginal logistic model and GEE method would be applied to the longitudinal data for finding significant factors on MetS. Finally, the accuracy for predicting MetS by the proposed model is obtained to check the goodness of fit of the model.

4.1. Influential Factors on MetS

The longitudinal data analysis is necessary to uncover the sequential appearance of the explainary variables in the progression of MetS. In the longitudinal study, same subjects are observed in two different time period. In each time point, the responses for every subject are independent, but for the same subject, the responses among the different time periods are correlated to each other, this correlation must be taken into account. For a data with a single binary response for each subject, logistic regression could be used because the relationship between responses and time-dependent covariates could be ignored . For the time-dependent covariate problem, however, the method for addressing the time correlation is necessary. Hence, the marginal logistic regression model is used to analyze the data.

In this study, the longitudinal data is constituted by an binary response variable Y, twelve explanatory variables, X_1, ..., X_{12}. Among these explanatory variables, there is a time covariate variable, t=1, 2 (1 for the year of 2008, 2 for the year of 2012). Then the marginal model can be written as:

$$\text{logit}[P(Y_t = 1)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5$$

$$+ \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \beta_{10} X_{10} + \beta_{11} X_{11} + \beta_{12} t$$

where $\beta_0, \beta_1, \ldots, \beta_{12}$ are the coefficients for the independent variables and corr(Y_1, Y_2)=α, α≠0.

In order to identity the influential factors on MetS, Wald test statistic is used for hypothesis test. The null hypothesis H_0: $\beta_j=0$ versus the alternative hypothesis H_1: $\beta_j\neq0$, where $j=1,...,12$. In this paper, Type I Errors set to be 0.10. If the p-value for β_j is less than the type I errors, 0.10, then we reject the null hypothesis and conclude that this variable makes a significant contribution to MetS. If p-value > 0.10, then we fail to reject the null hypothesis, meaning there is no strong evidence to show that this variable contributes significantly to the MetS.

To fit the marginal model, the Generalized Estimating Equation (GEE) method will be used. We fit the GEE models using PROC GENMOD procedure of SAS. GEE allows us to estimate the correlation, Corr(Y_1,Y_2), based on the data and estimate the model parameters taking into account the correlation.

4.2. Prediction Accuracy

In order to check the prediction ability of the model, the probability of the accuracy for predicting MetS is obtainded. The range of probability is continuous between 0 and 1. Hence, we created a new variable, called prediction, which is a binary outcome. Let the cutoff point for the probability of predicting that a person has MetS is 0.5. If the probability is less than 0.5, then we treat the prediction value as 0, it means that no MetS observed. In contrast, when the probability is greater or equal to 0.5, then we treat the prediction value as 1, and it means Mets observed. The predicted MetS and the observed MetS for each subjects are obtained in a 2x2 frequency table, Table 2.

Table 2. The Accuracy of Predicting MetS

Table of MetS by prediction			
MetS	**prediction**		
	0	**1**	**Total**
0	n_{11}	n_{12}	n_{11+} n_{12}
1	n_{21}	n_{22}	n_{21+} n_{22}
Total	n_{11+} n_{21}	n_{12+} n_{22}	N

Based on the frequency table, we can calculate the accuracy rate of predicting true results [12]. True results means that if a person do not have MetS, the prediction from the model shows no MetS observed, also if a person has MetS, the prediction from the model shows MetS observed. The higher the accuracy we get, the better the prediction of the model is. The function for accuracy calculation is:

$$Accuracy = \frac{n_{11} + n_{22}}{N}$$

Where $N = n_{11+}$ n_{12+} n_{21+} n_{22} in the 2x2 frequency table.

CHAPTER 5. RESULTS

In Table 3 that shows the condition index, the largest Condition Index is 3.4825, which is smaller than 10. It indicates that there is no indication of collinearity problems among all independent variables, so all the independent variables can be used in the marginal model.

Table 4 shows the information of each parameter and this is used to interpret the result of the analysis of GEE parameter estimates in Table 5. For each independent variable, it uses a reference cell. One level of the variable will be set as the reference. In this model, the reference is the last level of the variable, and it would not be showed in Table 4. The parameter estimates are calculated by comparing each level with the reference level. As long as there is one *p*-value that is significant, we can conclude that the corresponding explainary variable contributes significantly to the MetS. For example, in Table 1, it shows that the variable X_4, smoking status, has three levels, 0, 1 and 2. In Table 4, the variable X_4 presented just two levels, 0 and 1. Level 2 would be treated as the reference cell. The parameter estimater for level 0 is calculated by comparing level 0 to level 2, and the parameter estimater for level 1 is calculated by comparing level 1 to level 2. If the *p*-value of level 0 is insignificant and the *p*-value of level 1 is significant, it tells us that level 1 is different from level 2, and level 0 is indifferent from level 2, then we can conclude that the variable X_4 has a significant effect on the MetS because at least one level in the variable is different from others.

Table 3. The Collinearity among All Independent Variables

Collinearity Diagnostics													
Number	**Eigen-value**	**Condition Index**	**Proportion of Variation**										
			Time	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**X_6**	**X_7**	**X_8**	**X_9**	**X_{10}**	**X_{11}**
1.0000	2.7055	1.0000	0.0000	0.0366	0.0021	0.0487	0.0384	0.0008	0.0278	0.0132	0.0009	0.0069	0.0075	0.0101
2.0000	1.4992	1.3434	0.1152	0.0002	0.0001	0.0001	0.0009	0.0366	0.0447	0.1043	0.1136	0.1130	0.0107	0.0674
3.0000	1.1085	1.5622	0.1798	0.0020	0.0715	0.0061	0.0022	0.1255	0.0284	0.0311	0.0393	0.0001	0.2005	0.1487
4.0000	1.0575	1.5995	0.0002	0.0102	0.3199	0.0061	0.0058	0.2152	0.0147	0.0244	0.0201	0.0774	0.0872	0.0924
5.0000	0.9460	1.6912	0.0032	0.0009	0.5206	0.0004	0.0004	0.1542	0.0193	0.0013	0.1954	0.1156	0.0004	0.0032
6.0000	0.9276	1.7078	0.0822	0.0004	0.0076	0.0013	0.0000	0.1826	0.1107	0.0705	0.1793	0.0254	0.3234	0.0188
7.0000	0.8681	1.7654	0.0008	0.0015	0.0274	0.0152	0.0047	0.2086	0.0360	0.1075	0.0126	0.2581	0.2825	0.0968
8.0000	0.8337	1.8015	0.4161	0.0050	0.0414	0.0000	0.0093	0.0165	0.0063	0.1043	0.2933	0.0044	0.0111	0.1819
9.0000	0.7088	1.9537	0.0419	0.0076	0.0077	0.0287	0.0032	0.0187	0.2273	0.5028	0.0499	0.3077	0.0195	0.0046
10.0000	0.6811	1.9930	0.1309	0.0050	0.0001	0.0004	0.0086	0.0362	0.4821	0.0385	0.0797	0.0845	0.0403	0.3743
11.0000	0.4410	2.4769	0.0298	0.0674	0.0004	0.8633	0.1858	0.0015	0.0029	0.0019	0.0094	0.0053	0.0095	0.0019
12.0000	0.2231	3.4825	0.0001	0.8633	0.0010	0.0297	0.7407	0.0036	0.0000	0.0001	0.0065	0.0016	0.0072	0.0000

Table 4. Parameter Information

Parameter Information													
Parameter	**Effect**	**Time**	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**X_6**	**X_7**	**X_8**	**X_9**	**X_{10}**	**X_{11}**
Prm1	Intercept												
Prm2	Time	2008											
Prm3	**X_1**		0										
Prm4	**X_2**			0									
Prm5	**X_3**				0								
Prm6	**X_3**				1								
Prm7	**X_3**				2								
Prm8	**X_4**					0							
Prm9	**X_4**					1							
Prm10	**X_5**						1						
Prm11	**X_5**						2						
Prm12	**X_6**							1					

Table 4. Parameter Information (continued)

Parameter Information													
Parameter	**Effect**	**Time**	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**X_6**	**X_7**	**X_8**	**X_9**	**X_{10}**	**X_{11}**
Prm13	**X_6**							2					
Prm14	**X_7**								1				
Prm15	**X_7**								2				
Prm16	**X_8**									1			
Prm17	**X_8**									2			
Prm18	**X_9**										1		
Prm19	**X_9**										2		
Prm20	**X_{10}**											1	
Prm21	**X_{10}**											2	
Prm22	**X_{11}**												1
Prm23	**X_{11}**												2

In Table 5, the parameter estimate, standard error, 95% confidence limits, Z-value and p-value of all independent variables are reported. According to the table, only three variables have *p*-value less than 0.10, indicating significant effect. The three variables are Time, alchohol consumption (X_3) , and bed-time snacks (X_5). The parameter for the time variable represents the effect of time in the year of 2008 compared to the reference level (2012). The parameter estimate means from 2012 to 2008, the time has a negative significant effect to the development of MetS. In other words, the risk of developing the MetS is increased from 2008 to 2012.

The variable X_3 represents the status of alcohol consumption, level 0 indicates "no alcohol consumption", level 1 indicates "occasionally alcohol consumption", level 2 indicates "person quit consuming alcohol more than 1 year ago", and level 3 indicates "person consumes alcohol 2-3 times per week". It shows that the parameter for X_3 is significant when X_3 takes the level of 0. It means that people who drink alcohol 2-3 times per week are more likely to have MetS than people who do not drink alcohol. Hence, alcohol consumption is a positive significant variable to the development of MetS.

The variable X_5 represent the frequency of eating bed-time snacks, level 0 indicates "person rarely eats bed-time snacks", level 1 indicates "person occasionally eats bed-time snacks", and level 3 indicates "person eats bed-time snacks more than 4 times per week". It shows that the parameter for X_5 is significant when X_5 takes the level of 1. It means that people who rarely eat bed-time snacks are more likely to have MetS than people who eat bed-time snacks 4 more times per week. Hence, bed-time snacks is a negative significant variable to the development of MetS.

Table 5. Analysis of GEE Parameter Estimates

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		**Estimate**	**Standard Error**	**95% Confidence Limits**		**Z**	**Pr>|Z|**
Intercept		-0.7158	0.5961	-1.8841	0.4524	-1.20	0.2298
Time	2008	-0.2105	0.0854	-0.3778	-0.0432	-2.47	0.0137
$\mathbf{X_1}$	0	0.1951	0.2450	-0.2852	0.6753	0.80	0.4259
$\mathbf{X_2}$	0	0.3043	0.2340	-0.1544	0.7630	1.30	0.1935
$\mathbf{X_3}$	0	-0.3935	0.2193	-0.8233	0.0363	-1.79	0.0727
$\mathbf{X_3}$	1	-0.0955	0.1960	-0.4797	0.2887	-0.49	0.6263
$\mathbf{X_3}$	2	0.2549	0.3417	-0.4148	0.9245	0.75	0.4557
$\mathbf{X_4}$	0	0.0883	0.2504	-0.4025	0.5791	0.35	0.7243
$\mathbf{X_4}$	1	-0.2372	0.3066	-0.8382	0.3637	-0.77	0.4391
$\mathbf{X_5}$	1	0.7261	0.4137	-0.0847	1.5370	1.76	0.0792
$\mathbf{X_5}$	2	0.7289	0.4456	-0.1444	1.6022	1.64	0.1019
$\mathbf{X_6}$	1	0.2592	0.1930	-0.1190	0.6374	1.34	0.1792
$\mathbf{X_6}$	2	-0.0487	0.1229	-0.2896	0.1923	-0.40	0.6922
$\mathbf{X_7}$	1	0.0469	0.1862	-0.3180	0.4119	0.25	0.8010
$\mathbf{X_7}$	2	0.0763	0.1169	-0.1528	0.3054	0.65	0.5140
$\mathbf{X_8}$	1	-0.1548	0.2301	-0.6058	0.2963	-0.67	0.5013
$\mathbf{X_8}$	2	0.0531	0.1049	-0.1524	0.2587	0.51	0.6125

Table 5. Analysis of GEE Parameter Estimates (continued)

Analysis Of GEE Parameter Estimates							
Empirical Standard Error Estimates							
Parameter		**Estimate**	**Standard Error**	**95% Confidence Limits**		**Z**	**Pr>|Z|**
$\mathbf{X_9}$	1	-0.2285	0.2883	-0.7934	0.3365	-0.79	0.4281
$\mathbf{X_9}$	2	-0.4217	0.2957	-1.0011	0.1578	-1.43	0.1538
$\mathbf{X_{10}}$	1	-0.2045	0.1655	-0.5289	0.1199	-1.24	0.2167
$\mathbf{X_{10}}$	2	-0.1086	0.1621	-0.4264	0.2092	-0.67	0.5029
$\mathbf{X_{11}}$	1	0.1898	0.1335	-0.0719	0.4515	1.42	0.1552
$\mathbf{X_{11}}$	2	-0.0142	0.1438	-0.2960	0.2676	-0.10	0.9215

In order to check model goodness of fit, QIC and the accuracy are obtained in Table 6, and Table 7. In this GEE model, the value of QIC is 1755.0257 and the accuracy is 57.85% (557 for no MetS and 180 for having MetS, giving the total number is 1274). Then, we delete all the insignificant independent variables from the original marginal model, and obtain the value of QIC and accuracy again in Table 6 and Table 8. The value of new QIC is 1740.8040, which smaller than 1755.0257 and the accuracy is 57.69% (543 for no MetS and 192 for having MetS, giving the total number is 1274). We can see that the new model with influencing factors only shows small improvement from the original marginal model. However, we don't see any difference in terms of the prediction accuracy.

Table 6. GEE Fit Criteria

GEE Fit Criteria		
	With 11 independent variables	With 3 Selected Varaibles
QIC	1755.0257	1740.8040

Table 7. The Accuracy of Predicting MetS by using 11 Dependent Variables

Table of MetS by prediction			
MetS	**prediction**		
	0	**1**	**Total**
0	557	146	703
1	391	180	571
Total	948	326	1274

Table 8. The Accuracy of Predicting MetS by using 3 Dependent Variables

Table of MetS by prediction			
MetS	**prediction**		
	0	**1**	**Total**
0	543	160	703
1	379	192	571
Total	922	352	1274

As we can see in both QIC value and the calculated accuracy of the two models, the improvement is actually trivial. This means that both models are not good for predicting the MetS based on the data used for this paper. Based on the results that we have, we may need to consider why the models does not provide good prediction. It is possible to see that there is a big change between data collected in 2008 and in 2012. The PROC LOGISTIC procedure of SAS is used to perform stepwise selection using the variables from X_1 to X_{11} and the significant level 0.10. Wald test is used for the examination of each individual parameter. If the *p*-value of any parameter estimate is less than 0.10, then the variable will stay. Otherwise, it will be removed from the model. In 2008, there are five significant variables (fruit, alcohol consumptions, physical activity, milk, and preserved food) which contribute to MetS, but in 2012, these variables all turn to insignificant (see Table 8)..

Table 10 shows the accuracy of the selected model (i.e. the model with five significant explainary variables) for 2008. The accuracy of the 2008 model is 64.36%, shows in 2008 model, there is 64.36% can successfully predict people's MetS status. It is higher than the accuracies obtained from two marginal models (57.85% and 57.69%). Since there is no significant variables existing in 2012, we can not get the accuracy for 2012.

Table 9. Summary of Stepwise Selection in Each Time Period

Summary of Stepwise Selection				
Step	**Effect in year 2008**		**Effect in year 2012**	
	Entered	**Pr > ChiSq**	**Entered**	**Pr > ChiSq**
1	**Fruit**	<.0001		
2	**Alchohol Consumptions**	0.0065		
3	**Physical Activity**	0.0072		
4	**Milk**	0.0213		
5	**Preserved Food**	0.0912		

Table 10. The Accuracy of Predicting MetS in Year 2008

Table of MetS by prediction			
MetS	**prediction**		
	0	**1**	**Total**
0	310	60	370
1	167	100	267
Total	477	160	637

CHAPTER 6. CONCLUSIONS

As we can see from the results, there are three factors (time, alcohol consumption, and bed-time snacks) that contribute significantly to the development of Metabolic Syndrome (MetS). However, the QIC values of the two models with time-dependent variable show that the improvement is trivial. Also, the first marginal model with all independent variables had an accuracy of 57.85%, and the second marginal model with three selected independent variables had an accuracy of 57.69%. The accuracies are both a mediocre amount of predictive power. But when we calculate the accuracy for 2008, the value is 64.36%, which is a higher accuracy than the two marginal models.

The results shows that the design requires attention, wether it is the variables measured or the way we record the observations for each individual. There are several possible explanations for no significant factors in the 2012 data. The variables collected reflect behavioral information collected through self-report survey method. Since this is the second time completing the survey, the participants may have not been careful in their responses, or perhaps attemped to anticipate the desired answers, resulting in interviewer bias. Recall bias is also possible. The 24-hour recall method would have been more reliable, and with a previously validated instrument, if possible. It is also possible that researcher fatigue was such that data collection was done with less rigor, and thus lower quality, in 2012. Loss to follow up reduced the sample size which may have resulted in a less representative sample. The laboratory equipment used for the blood marker analysis may have not been calibrated on schedule. All of these are merely speculative, but they do indicate the importance of consistency in data collection, particularly in a longitudinal study.

There is another possible explanation for limited variability in responses. China is a collective society, with significant homogeneity in behavior and values. The power of statistics is

invariability within the sample. Because the subjects demonstrated limited variability in dietary intake and physical activity levels, it was impossible to detect any significant impact of these variables on the presence or absence of MetS if this effect existed.

The longitudinal study is still in progress, and new data will arrive in 2016. When we have the third time period of data, we should try to fit the marginal model to this extended data so that we can cehck how the model provides good fit to the data. Another possiblity to make the model beter is adding two-way interactions between some of the explanatory variables and the time variable. In our paper, for simplicity, we didn't include the interactions in the model. Adding the interations could help to explain the relationships between the MetS and the covariates is changed by the time.

REFERENCES

[1]. Ford ES, Giles WH, Dietz WH. "Prevalence of the metabolic syndrome among US adults findings from the third national health and nutrition examination survey." The Journal of the American Medical Association 2002; 287; 3–356.

[2]. Lorenzo C, Williams K, Hunt KJ, and Haffner SM. "The National Cholesterol Education Program–Adult Treatment Panel III, International Diabetes Federation, and World Health Organization Definitions of the Metabolic Syndrome as Predictors of Incident Cardiovascular Disease and Diabetes." *Diabetes Care* 2007; 30; 8-13

[3]. Gu DF, Reynolds K, Wu XG, Chen J, Duan XF, Reynolds RF, Whelton PK, and He J. "Prevalence of the metabolic syndrome and overweight among adults in China." *The Lancet* 2005; 365; 1398-1405.

[4]. Belsley DA, Kuh E, and Welsch RE (1980). *Regression Diagnostics: Identifying Infuential Data and Sources of Collinearity*. Wiley Series in Probability and Statistics, New York: Wiley Interscience.

[5]. McDonald JH (2014). *Handbook of Biological Statistics.* 3rd ed. Sparky House Publishing, Baltimore, Maryland.

[6]. Liang KY and Zeger SL. "Longitudinal data analysis using generalized linear models." *Biometrika* 1986; 73; 13-22.

[7]. Wedderburn RWM. "Quasi-likelihood functions, generalized liner models, and the Gauss-newton method." *Biometrika* 1974;61:439 – 447.

[8]. Pan W. "Akaike's information criterion in generalized estimating equations," *Biometrics* 2001; 57; 120-125.

[9]. National Institutes of Health. *Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult Treatment Panel III).* 2002.

[10]. Moy FM and Bulgiba A. "The modified NCEP ATP III criteria maybe better than the IDF criteria in diagnosing Metabolic Syndrome among Malays in Kuala Lumpur." *BMC Public Health* 2010; 10; 678.

[11]. Strand MA, Will T, Gu XX, and Perry J. "A descriptive study of the progression of the metabolic syndrome in middle aged Chinese persons." *The International Quarterly of Community Health Education* 2015; 35(2);165-178.

[12]. Anooj PK. "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules." *Journal of King Saud University – Computer and Information Sciences* 2012; 24; 27–40.

APPENDIX. SAS PROGRAM FOR IDENTIFYING SIGNIFICANT FACTORS INFLUENCING METABOLIC SYNDROME

```
proc genmod data=mark.mets1274 descending;
class tag_no time sex_1b  PA wine_36 SM_35 bed_38 fruit_39
      meat_42 tofu_43   fry_37 pre_46 milk41/ param=ref;
model atp_stat= time sex_1b  PA wine_36 SM_35 bed_38 fruit_39
      meat_42 tofu_43   fry_37 pre_46 milk41 / dist=bin link=logit;
repeated subject=tag_no/type=CS;
output       out       = Residuals
             pred      = Pred
             resraw    = Resraw
             reschi    = Reschi
             resdev    = Resdev
             stdreschi = Stdreschi
             stdresdev = Stdresdev
             reslik    = Reslik;;
run;


data few;
  set residuals;
  if pred >=0.5 then prediction=1;
   if pred <0.5 then prediction=0;
proc freq data=few;
  table atp_stat*prediction/nocol norow nopercent;
  run;
```