Serious Genetic Disease Screening

I-Huei Ho
ihuei.ho25@uga.edu

University of Georgia, Department of Statistics

Contents

List of Figures

List of Tables

1 Executive Summary

The purpose of this study is to discover the association of the probability whether carrying a serious genetic disease (SGD) and various blood markers obtained by the blood tests which is possible to screen elevated levels in SGD. Then establish a predictive model based on effectiveness and low cost. This information would give further insight on age of the subjects took the blood test and the time that the blood tests were taken. In the future, the results from this study could help with predicting the probability of having serious genetic disease based on own blood test results and personal background.

The researchers are interested in screening women for the genetic predisposition for SGD by a predictive model based on effective and inexpensive blood tests and they have collected 209 samples from 209 women of four blood tests results. The response variable of this experiment is binary dependent variable *SGD* which indicates 1 for known carrier of SGD and 0 for unknown carrier of SGD. There are a few predictor variables that were collected as following: *SubjID* (Unique ID for subjects), *SampleNo* (Blood sample number), *Age* (Age in years), *Month* (1=January,...,12=December), *Year* (in years), and *BM1, BM2, BM3, BM4* (concentration of blood markers). The goal of this experiment is to determine which of these variables can effectively predict the probability whether carrying SGD or not. The possibility of influence on blood tests' measurements by water supply change is also of interest.

However, before conducting any analyses on this case, there are several obstacles that must be addressed. First, there are total 15 missing values rather in blood marker 3 or blood marker 4, and the researcher did not explain the reasons of missing values occurred. The samples contained missing values represent part of important information of specific years and it will be necessary to fill in the missing values. The second issue should be investigated before the analysis is characteristic of each samples. The rough ages interval and seasons are interests of this study, then the specific ages and months will be categorized into new levels for better discussion. Lastly, the contradiction between researchers explanation and dataset needs to be clarified before starting the analyses.

The analyses conducted during this study includes logistic regressions, which are used to select significant variables to predict the probability whether carrying SGD or not. A few diagnostics are performed to strengthen the selected model and interpret the reason of dropping particular three observations. After clarifying the significance of each variables included in the final model, the model contains blood marker 1, 3 and 4, and also the variable *Age*. Several one-way analysis of variances are also implemented in order to test the relationship among each blood markers and time variables, which may indicates the effect from water supply change on blood markers' measurements. The blood marker 2 are the only marker tested to react to time change but it is not included in the selected model.

On the other hand, there are several potential problems might bias the results of this study such as the particular years included missing values in blood marker 3 and 4 will probably impact the final model if it contains special trend.

2 Introduction

The binary results whether an asymptomatic carrier of serious genetic disease (SGD) have provided with 209 women in the sample. Each of these known and unknown carriers has seven different factors that are associated with it. These variables are: *SubjID, SampleNo, Age, Month, Year, BM1, BM2, BM3, and BM4*. Each of variables of blood markers *BM1,BM2,BM3,BM4* are continuous measured in concentration and variables *SubjID* and *SampleNo* are categorical. Variables *Age, Month, Years* are numerical but will be reassigned in appropriate categories for better comprehension in subsequent section and explained the reasons. The probability of a SGD carriers is important since it is the expected prediction of interest and the researchers want to find influential various factors on it. They also desire the effectiveness and low cost of the experiment of screening genetic susceptibility for SGD; hence, the less blood tests included the more preferable they are. Otherwise, they expected to determine the probable influence on blood markers' measurements by the water supply change of the laboratory.

In the next Data Summary section 3 of this paper, an exploratory data analysis is conducted. The data will be described in more detail, as well as further explanation of the meaning for each variable. In order to decide on which statistical techniques will be best to analyze the data, it is important to conduct some preliminary phases of exploratory analysis on each variable. Once there is a better comprehension of each variable, a decision can be made on the best statistical techniques to continue the analysis. In the Analysis section 4, the data is analyzed by logistic regression. The relevant assumption will be incorporated alongside visuals, such as S-shaped curves graphs and residual plots. Then the predictive model will be selected and validated. Additionally, several tests and graphs will be present for clarifying the association among concentration of blood markers and time, which attempting to specify the time of water supply change. Next, a conclusion will be drawn from the analysis and it will contain an overview of predictive model and overall work and probable extended concerns in this project. Lastly, there are appendixes for various figures and tables that were not included in the earlier sections and for R code that are used for whole project.

3 Data Summary

The dataset is composed of 209 sampled women with 75 SGD carriers and 134 non-SGD carriers. In this study, the researchers are interested in predicting SGD carrier according to deficient and inexpensive blood tests. The variables and their descriptions from the dataset are:

- *SGD* - Binary variable for 209 target women whether carrying SGD (1) or not carrying SGD (0)
- *SubjID* - Unique ID for each subject from which one or more blood samples were obtained
- *SampleNo* - Blood sample number of the subject is contribeted
- *Age* - Age of subject in years, range from 20 to 60 years old
- *Month* - Month of the year in which blood sample was taken, 1 for January,..., 12 for December
- *Year* - Year in which blood sample was taken, range from 1988 to 1991
- *BM1* - Concentration of blood marker 1 from frozen blood samples
- *BM2* - Concentration of blood marker 2 from frozen blood samples
- *BM3* - Concentration of blood marker 3 from fresh samples, recorded missing value as -99
- *BM4* - Concentration of blood marker 4 from fresh samples, recorded missing value as -99

According to Table 3.1 below, the results showed that all the samples consist of missing values in BM3 were taken either in 1990 or 1991 and in BM4 were taken in 1988. Since there are only 7 samples from 1988 and 3 results from 1991, which all contains missing values, it is not tenable to eliminate all incomplete samples due to instability for reduction of sample size and the following analyses may be questioned by possible influence of years on blood markers results. In this research, using classification and regression tree (CART) to predict the missing data based on the present values, which is the method regards each variable BM3 and BM4 as response and other three blood markers variables as explanatory variables to predict missing values and fill in.

Table 3.1: Missing values in BM3 and BM4 (15 duplicates)

SGD	subjID	sampleNo	age	month	year	BM1	BM2	BM3	BM4
0	1	1	27	10	1988	22	99	10.8	-99
0	11	1	31	11	1988	29	94	11.8	-99
0	13	1	22	12	1988	22	85.5	15	-99
0	15	4	25	10	1988	41	87.3	15	-99
0	17	1	26	12	1988	28	93.5	7	-99
0	19	1	38	12	1988	45	108	13.7	-99
0	26	2	24	10	1988	26	94.2	11.7	-99
1	597	2	32	5	1990	79	9	-99	137
1	831	4	36	11	1990	144	24.4	-99	329
1	837	3	40	12	1990	123	25.4	-99	275
1	840	4	32	12	1990	610	111.7	-99	593
1	857	2	30	12	1990	510	60.2	-99	272
1	875	3	36	1	1991	55	20.7	-99	262
1	880	5	31	1	1991	45	13.8	-99	217
1	882	3	59	1	1991	25	9.2	-99	316

Otherwise the lucid relationship between blood markers and SGD carriers, the researchers are also interested in the effects from ages of samples, seasons and years they were taken. The particular ages and months will be too specific in this case. For better classification, establish new categories of variable *Age* and *Month* as following rules:

- *Age* - Divided into five categories in years as younger than 25, 26 to 30, 31 to 35, 36 to 40, and older than 40 with level 1, 2, 3, 4, or 5
- *Season* - Divided into four categories as month 9 to 11, month 12 to 2, month 3 to 5, and month 6 to 8 with level Fall, Winter, Spring, or Summer
- *Year* - Divided into four categories in year with level 1988, 1989, 1990, or 1991

Besides the missing values in blood markers and categorizing three other variables, the inconsistency of variable *SubjID* and *Age* is needed to develop some assumptions. According to the samplers and dataset, there are two illogical statements: "The data are composed of 209 women and some of them took blood marker tests twice", and "The subject ID is unique for each subject, but the dataset showed different subjects for the same subject ID", which are presented by Table 6.3. However, if the variable *SubjID* and *SampleNo* are neglected, the dataset seems reasonable to regard as 209 different individuals since those

samples with the same ID have all different values of other variables. Consequently, this project will work under the assumption that regarding 209 samples as all different sampled women.

4 Analysis

4.1 Assumption

To begin the analysis, certain assumptions must be checked. First of all, the explanatory variables Blood Markers 1 to 4 are examined in order to see how the values of each individual blood marker varies in the probability whether carrying SGD. A Logistics Regression is performed in order to predict the probability of carrying SGD based on the results of blood markers, sample's age and the time that the sample was taken. Points of interest for the study are in both effectiveness and economical cost of detecting carriers. Before conducting the analyses, the following assumptions must be met:

- The dependent variable should be binary
- The factor level 1 of dependent variable should represent desired outcome, which is SGD carrier
- The independent variables should be linearly related to the log odds
- Each observation should be independent

To verify that these assumptions are met, the method of data collection are considered, and then several graphs composed of S-shaped growth curve are produced. Since mentioned in previous section, the dependent variable *SGD* is binary whose level 1 is for SGD carriers and level 0 is for non-SGD carriers. According to Figure 6.1 in Appendix, set $\pi(x)$ as the probability whether carrying SGD or not, the approximate S-shapes in each plots of $\pi(x)$ against blood marker are shown and the linearities between log odds and each blood marker are shown after taking logistic function of $\pi(x)$.

In order to see the independence among each observation, it is showed that the residuals are all approximately distributed around horizontal line 0 depending on the residual plot of residuals against fitted values with loess smooth curve in Figure 6.2 in Appendix. Though there are slightly ascending and descending trends of the blue loess smooth curve, they are very subtle and allowed to ignore. Therefore, all the assumptions are satisfied and the model building process can continue.

4.2 Model Selection

In order to select the significant model for predicting probability whether a SGD carrier or not, apply the most general approach that splitting dataset into two pieces, training set and validation set, establish the model based on training set, and test the selected model by validation set. In this case, randomly sample 85% of observations, 177 observations, to training dataset; remained 15% of observations, 32 observations, for validating.

Employ 85% dataset and proceed the full stepwise procedure as the output showed in Table 7.1 in Appendix, it is known that the model included blood marker *BM1*, *BM3*, and *BM4*, and the categorical variable *Age* is selected and performed as below:

$$log\frac{\pi(x)}{1-\pi(x)} = \beta_0 + \beta_1(BM_1) + \beta_2(BM_3) + \beta_3(BM_4) + \beta(Age) + \epsilon \quad (1)$$

where $\pi(x)$ = probability whether carrying SGD or not

However, the model only consider first-order independent variables, the interaction and quadratic terms are all ignored in this stage. The higher order terms will be discussed in following Diagnostics section and adjusted if it is inadequate.

4.3 Model Validation

Exploit the rest 15% dataset and perform the cross validation approach. Compared following results of each dataset based on the selected model concluded in previous section, the comparison is showed as Table 4.2 below.

For the coefficient estimates of both dataset, obvious differences are detected among each level of *Age* and the variable *BM1*. However, for the root mean square error (RMSE) and Pseudo R^2 values above, both datasets have similar results. Roughly, it can be concluded that even the coefficient estimates are diverse, the variability that these variables can explain are still the same, which attributes the difference to other possible reasons.

There are several problems may cause these discrepancies: First, There are only 32 observation remained in the validation dataset, which is close to the minimum sample size of defined large sample, thus the results may not be robust because of the small sample size. Moreover, the possible effect by influential observation, outliers, or multi-collinearity are not considered particularly in this stage, which probably and seriously increase or reduce the prediction of $\pi(x)$. Among these reasons, succeeding section is going to diagnose the conceivable problems that may not be noticed previously.

Table 4.2: Cross Validation

	Training dataset	**Validation dataset**
Coefficients		
(Intercept)	0.333	-20.181
$BM1$	7.774	1.869
$BM3$	1.572	2.347
$BM4$	1.586	1.182
Age_2	0.625	20.688
Age_3	1.550	20.657
Age_4	-0.145	3.256
Age_5	20.707	37.777
RMSE	0.062	0.097
Pseudo R^2		
McFadden R^2	0.693	0.557
CoxSnell R^2	0.596	0.512
Nagelkerke R^2	0.817	0.707

4.4 Diagnostics

Based on the residual plot of chosen model (1) in Table 6.3 in Appendix, the loess smooth curve is more close to the line of residuals equal to zero, which means the fit of new model is more appropriate than the full model. However, as the problems mentioned in prior sections, the further discussions of marginal effects, outliers, and collinearity will be performed to strengthen the model or the necessity of adjustment.

4.4.1 Marginal Effects

In order to diagnose the effects of individual variable on probability $\pi(x)$, the marginal effects are utilized. Remove the effect of the other variable from response $logit(\pi(x))$ and independent variable that is interested in, and obtain the residuals from each model and residual plots of $logit(\pi(x))$ against curious variables. In the Figure 4.1, all plots behaved positively increasing trends and the linear terms in each variables may be useful, which strengthen the decision of model selection.

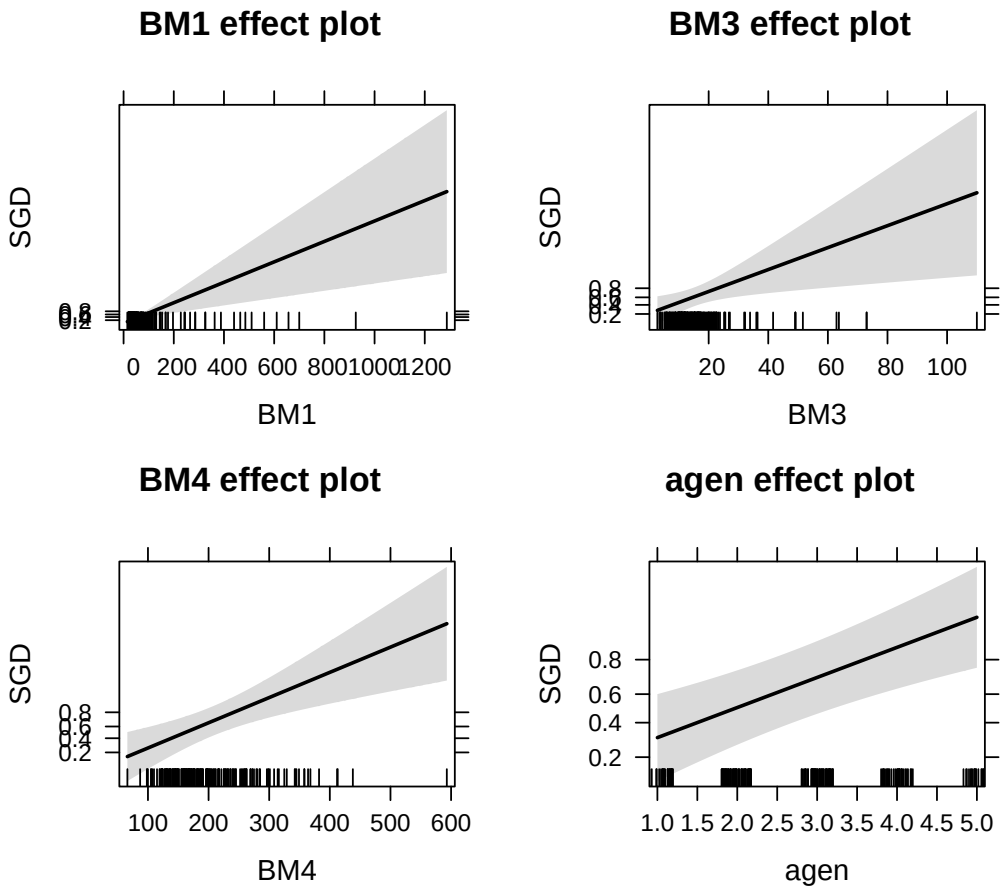

Figure 4.1: Marginal Effect

4.4.2 Outliers

As reported by the Cook's Distance Figure 4.2 below, there are several observations are considered as potential outliers. For instance, the cook's distance marked by observations' numbers in the Figure 4.2 are those possessing extremely larger cook's distance than most of other observations: Observation 52, 171, and 191 have cook's distance values larger than 0.08 and observation 92, 195, and 207 have have cook's distance values larger than 0.06. Since other possible problems are successively proved by either graphs or tests, the outliers could be the reason influence the coefficient estimates in previous parts. On account of not reducing the sample size as far as possible, only regard three observations as outliers: Eliminate observation 52, 171, and 191 from the dataset in this stage.

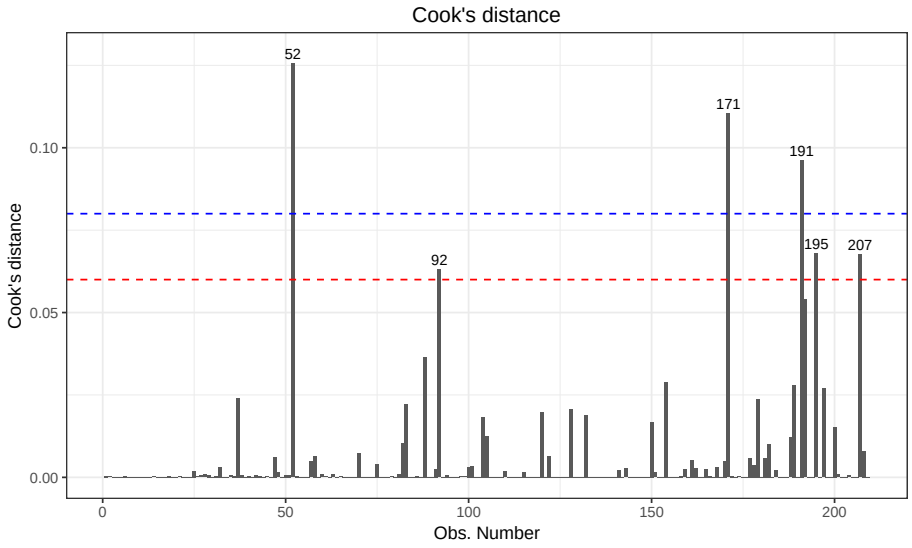

Figure 4.2: Cook's Distance

4.4.3 Multi-collinearity

In the interest of not including higher-orders or interaction terms in the model, measured the inflation of standard errors compared to a lack of linear relationship and obtained the variance inflation factors (VIFs). As stated in Table 6.4 in Appendix, all the VIF values are larger than 1 but not extremely large. Since practical cases are always not perfect to get VIFs equal to 0 to state that there are no linear relationship among the variable and other variables, the VIFs acquired here are tenable enough to conclude that there are no multi-collinearity among the variables. Consequently, the probable cause of failure in model validation section is attributed to outliers. After comparing each elements of the selected model (1) based on the whole dataset except for three assumed outliers and the selected model (1) based on the training dataset in Table 6.5 in Appendix, all of the coefficients are similar and the chosen model is clarified.

4.5 Underlying Effect

As the researchers have mentioned, the water supply of the laboratory was changed during the course of this study at unknown exact timing and it was an underlying effect that might involve in the measurements of blood markers.

In Figure 6.4, there are intense ebbs and flows in blood marker 1 and 2 during summer time in 1990. At the same time, blood marker 3 and 4 do not have exceptional displays. After four individual one-way analysis of variance (ANOVA) tests as outputs in Table 7.2 for testing the association among each blood markers and time variables–variable *Season* and *Year*, the results showed that blood marker 2 has significant negative relationship with variable *Year*. Therefore, there are two conjectures of the reason: First, the water supply change might happen during summer time in 1990 but only influence measurement of blood marker 2; Second, the water supply change might happen during some other timing and it did not directly affect measurement of blood markers. There might be other circumstance happened at that time to make the highs and lows occur in blood marker 1 and 2. Therefore, since the most concerned blood marker 2 here is not contained in the selected model, the model (1) is still available after this discussion.

5 Conclusion

After conducting the analysis, it is able to conclude a model that were effective predictors for probability of a SGD carrier. The selected model includes *BM1*, *BM3*, *BM4* (Blood Marker 1, 3, and 4), and *Age*. Throughout the analysis performed formerly, the logistic regression was used under complete assumption checks and discarded insignificant variables as *BM2*, *Season*, and *Year* from the full model. Given that the dataset did not involve collinearity among variables and the marginal effects all showed obvious linearity between each variable and probability, the outliers was founded that is the possible factor caused some biases. Then, those outliers were considered to eliminate from the dataset at the end and the model became more appropriate and reliable. The model prediction would be more accurate if a larger sample size in terms of the number of women.

As mentioned in the previous sections, there are still several possible factors that would unintentionally affect predicting probability whether a SGD carrier. First of all, back to the most beginning assumption supposed, the analysis was proceeded by regarding all samples are from different individual women and none of them took more than once blood test. If the typo in the dataset means opposite way that those samples with the same subject ID should be considered as being tested by the same women and the contradicted ages of both are typos, then how to weight the duplicates in the analysis will be another concern needed to deeply discuss. Secondly, the analysis were conducted by filling those missing values based on available data. Equally as the reason they are filled at the first section in this case, most of the missing values represented the data of particular whole year. If analyzed without datapoint with missing values, the results would be considerably biased. The statement also indicates that if there were specific circumstance in those particular years related to blood marker 3 and 4, the results might be significantly different.

Lastly, the researchers are interested in cost down but the final model selected contains both blood markers from fresh samples. Due to the potential problem of missing values, the researchers may expect to only contain one of these two blood markers and it would need more suitable analysis to extend this demand.

6 Appendix: Figures and Outputs

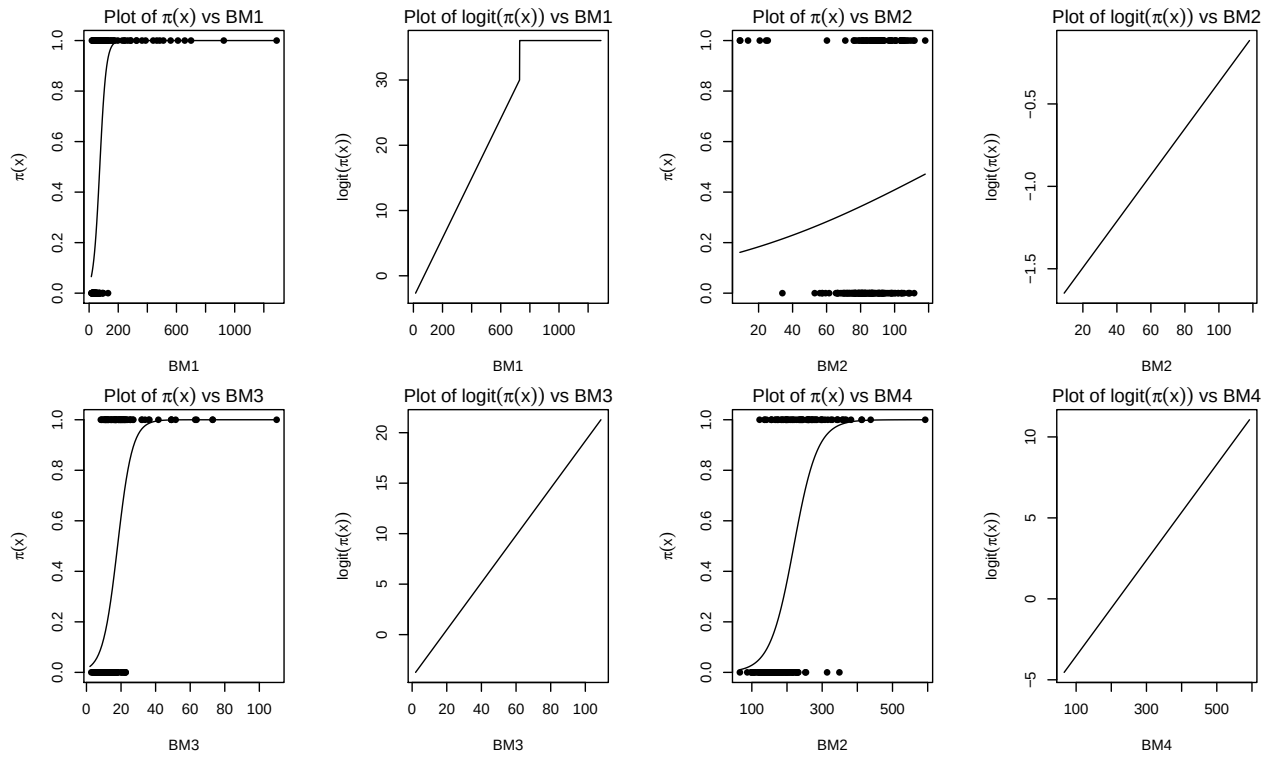

Figure 6.1: Scatter Matrix

Table 6.3: Duplicated Subjects (17 duplicates)

SGD	subjID	sampleNo	age	month	year	BM1	BM2	BM3	BM4
0	133	1	27	8	1989	15	87	13.5	232
0	133	1	32	7	1989	28	82.5	17.4	144
0	255	1	29	2	1990	74	80.4	8.9	207
1	255	1	35	6	1989	48	98	16.4	233
0	257	1	36	2	1990	40	72.7	7	131
1	257	1	34	6	1989	73	105.5	17	285
0	258	1	30	2	1990	69	66.7	8.7	119
1	258	1	38	6	1989	286	109.5	31.9	260
0	273	1	27	3	1990	27	87.2	12.5	99
1	273	1	53	6	1989	59	93	22.2	240
1	291	1	29	8	1989	53	76	14	174
0	291	6	39	4	1990	25	98.7	10	174
0	293	1	31	5	1990	35	90.3	15.3	124
1	293	1	42	8	1989	78	118	15.5	212
0	353	2	25	6	1990	59	72.5	10.7	314
1	353	3	35	9	1989	42	100.1	17.1	184
1	593	1	33	5	1990	57	88	8.9	190
0	593	2	34	7	1990	87	76.3	6	87
1	597	2	32	5	1990	79	9	11	137
0	597	3	27	7	1990	24	57.5	5.6	130
1	633	1	53	6	1990	101	77.5	11.7	280
0	633	3	28	9	1990	72	66.3	16.4	156
0	634	1	24	9	1990	25	92	14	166
1	634	2	36	6	1990	104	87.5	16.7	256
0	640	1	25	10	1990	42	65.5	13.3	216
1	640	2	45	6	1990	35	86.3	14.4	184
1	647	1	59	6	1990	560	106	21	345
0	647	2	34	10	1990	48	83	13.7	228
0	649	1	36	10	1990	55	78.2	21.8	188
1	649	1	48	6	1990	115	79	14.2	258
0	650	1	22	11	1990	30	104	22.6	230
1	650	2	39	6	1990	228	104	10.2	236
0	651	1	21	11	1990	26	79.3	16.4	123
1	651	2	26	6	1990	700	90	49.1	343

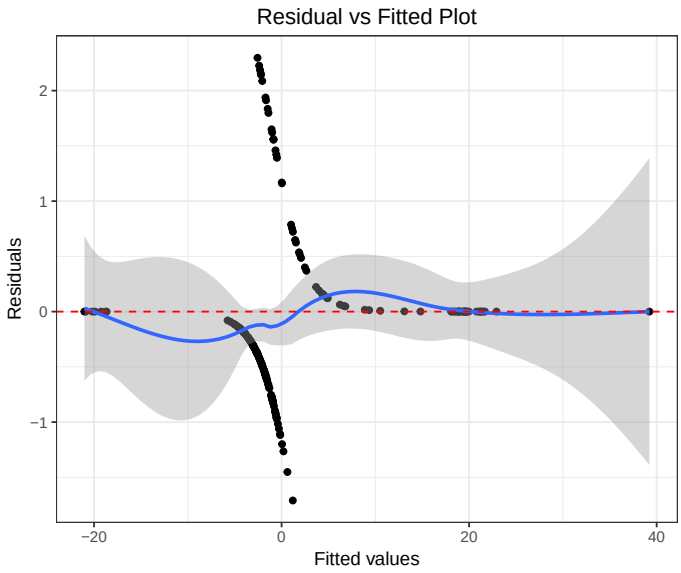

Figure 6.2: Residual Plot with loess smooth curve of Full Model

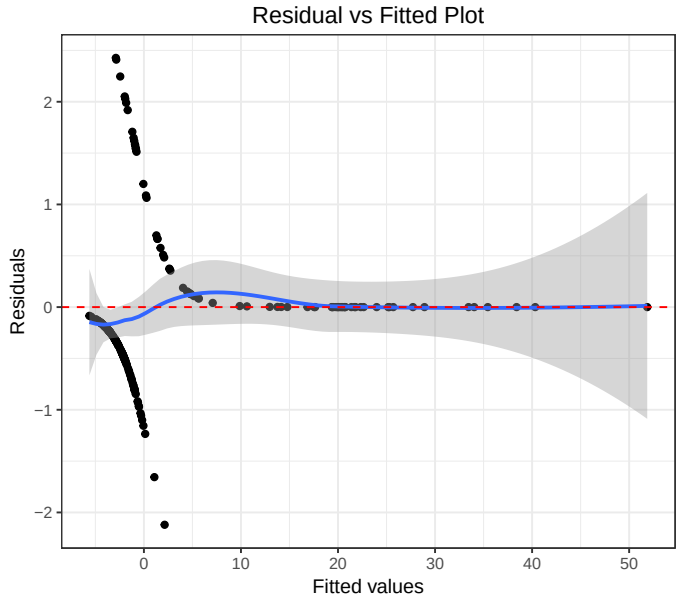

Figure 6.3: Residual Plot with loess smooth curve of Selected Model (1)

Table 6.4: Variance Inflation Factor (VIF)

Variables	VIF
SGD	1.627
BM1	2.503
BM2	1.020
BM3	2.628
BM4	2.057

Table 6.5: Cross Comparison

	Training dataset	**Fulldataset w/o outliers**
Coefficients		
(Intercept)	0.333	-0.1305276
$BM1$	7.774	8.3367543
$BM3$	1.572	2.1336028
$BM4$	1.586	1.4086971
Age_2	0.625	1.6500766
Age_3	1.550	2.3435243
Age_4	-0.145	0.4195481
Age_5	20.707	21.4429961
RMSE	0.062	0.06151851
Pseudo R^2		
McFadden R^2	0.693	0.6949477
CoxSnell R^2	0.596	0.5964924
Nagelkerke R^2	0.817	0.8181415

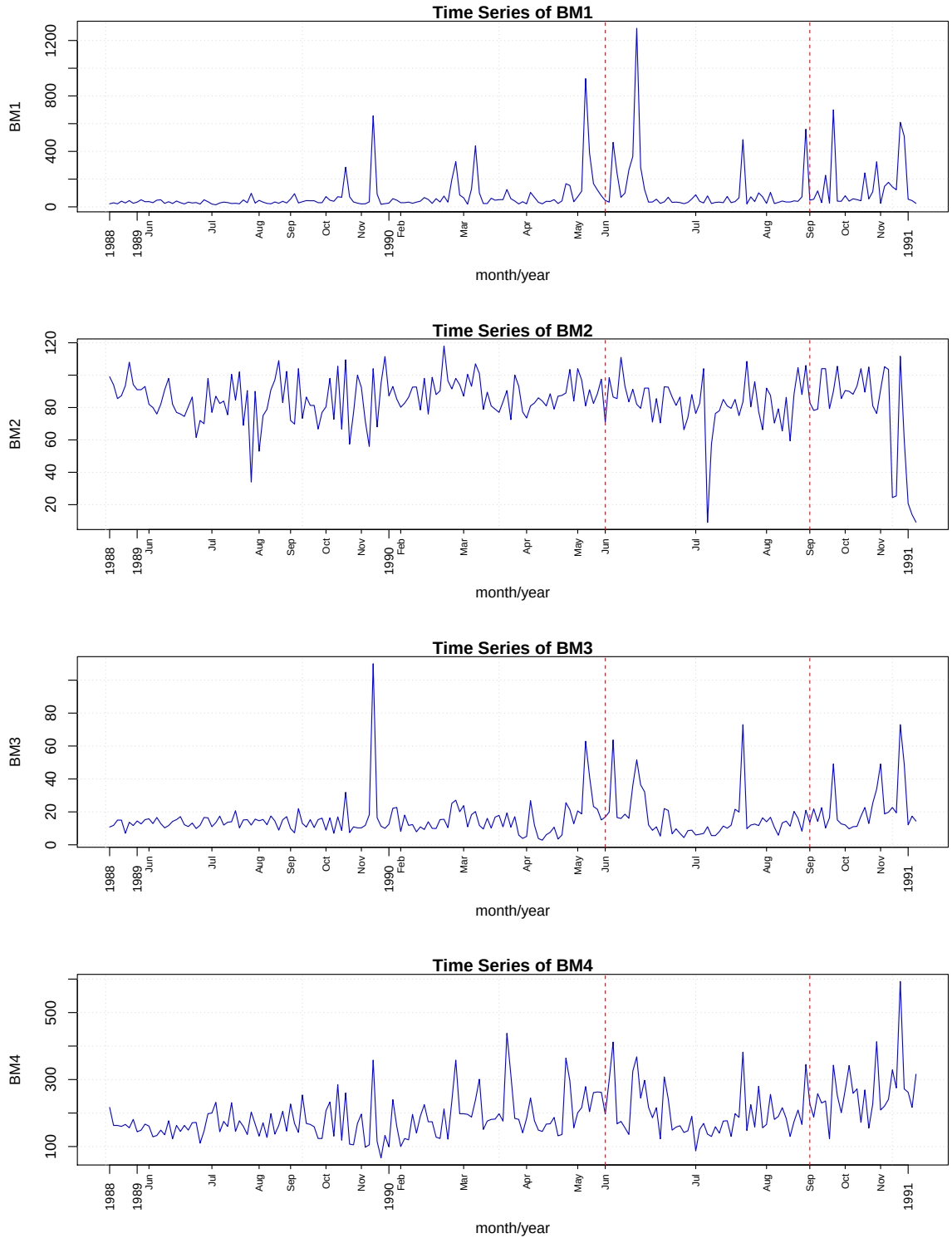

Figure 6.4: Time Series

7 Appendix: R code and outputs

```
#########################################################
####### [MS Qualifying Examination], May, 2017 ######
############### by examinee 9590 ####################
################ 050517--050717 #####################
#########################################################

sgd=
  read.csv("~/Documents/1- UGA/[QEM]/QEM-050517/sgdData.csv", header = T)
#suppressWarnings( = read.csv("~/Documents/1- UGA/[QEM]/QEM-050517/.csv", header = T) )

#par(mfrow=c(2,2),mar=c(4.5,4.5,1.5,2), oma = c(0,0,2.5,0))
#col="cadetblue","darkolivegreen3","indianred1","tan1"
#capuchins.n2[capuchins.n2[,"group2"]=="C",]$Crack

#-----------------------------------------------------------------------#
### Data Summary ###
#-----------------------------------------------------------------------#
#1 missing values created
nan = sgd[(sgd[,"BM3"]==-99 | sgd[,"BM4"]==-99),]
na = sgd[!(sgd[,"BM3"]==-99 | sgd[,"BM4"]==-99),]
# created missing values in BM3
coef3 = as.vector(summary(lm(BM3~BM1+BM2+BM4,data=na))$coef[,1])
BM3n=rep(NA,nrow(sgd))
for(i in 1:nrow(sgd)){
  if(sgd[i,"BM3"]==-99){
    BM3n[i] = as.matrix(sgd[i,c("BM1","BM2","BM4")])%*%coef3[2:4]+coef3[1]}else{
      BM3n[i] = sgd[i,"BM3"]}}
# created missing values in BM4
coef4 = as.vector(summary(lm(BM4~BM1+BM2+BM3,data=na))$coef[,1])
BM4n=rep(NA,nrow(sgd))
for(i in 1:nrow(sgd)){
  if(sgd[i,"BM4"]==-99){
    BM4n[i] = as.matrix(sgd[i,c("BM1","BM2","BM3")])%*%coef4[2:4]+coef4[1]}else{
      BM4n[i] = sgd[i,"BM4"]}}
# new data without missing values
sgd=cbind(sgd,BM3n,BM4n)

# use Cart method to fill in the missing values
install.packages("mice")
library("mice")
sgd[sgd[,"BM4"]==-99,][,"BM4"]=NA
sgd[sgd[,"BM3"]==-99,][,"BM3"]=NA
micedata = mice(sgd, m=3, maxit=50, method="cart",seed=500)
sgd = complete(micedata,2)
sgd[sgd[,"year"]==1991,]
```

```
#2 age group
# <=25; (25,30]; (30,35]; (35,40]; >40
agen=rep(NA,nrow(sgd))
for(i in 1:nrow(sgd)){
  if(sgd[i,"age"]<=25){agen[i]=1}
  else if(sgd[i,"age"]<=30){agen[i]=2}
  else if(sgd[i,"age"]<=35){agen[i]=3}
  else if(sgd[i,"age"]<=40){agen[i]=4}
  else {agen[i]=5}
}
# new data with group age
sgd=cbind(sgd,agen)

#3 change monthes into season
# 9/10/11:Fall, 12/1/2:Winter, 3/4/5:Spring, 6/7/8:Summer
season=rep(NA,nrow(sgd))
for(i in 1:nrow(sgd)){
  if(sgd[i,"month"]==9|sgd[i,"month"]==10|sgd[i,"month"]==11){season[i]="Fall"}
  else if(sgd[i,"month"]==12|sgd[i,"month"]==1|sgd[i,"month"]==2){season[i]="Winter"}
  else if(sgd[i,"month"]==3|sgd[i,"month"]==4|sgd[i,"month"]==5){season[i]="Spring"}
  else {season[i]="Summer"}
}
# new data with group season
sgd=cbind(sgd,season)

#4 duplicated ID -> decided to ignore
dup= sgd[duplicated(sgd$subjID)|duplicated(sgd$subjID,fromLast=T),]
nrow(sgd[duplicated(sgd$subjID),])  #17
nrow(sgd[!duplicated(sgd$subjID),]) #192

set.seed(500)
dup.sa1 = sample(seq_len(nrow(dup[dup[,"SGD"]==1,])),size = 0.5*nrow(dup[dup[,"SGD"]==1,]))
dup.sa0 = sample(seq_len(nrow(dup[dup[,"SGD"]==0,])),size = 0.5*nrow(dup[dup[,"SGD"]==0,]))

sgd = sgd[-as.numeric(row.names(dup[dup[,"SGD"]==1,][-dup.sa1,])),]
sgd = sgd[-as.numeric(row.names(dup[dup[,"SGD"]==0,][-dup.sa0,])),]
nrow(sgd)


#-----------------------------------------------------------------------#
### Analysis ###
#-----------------------------------------------------------------------#

### 1 Assumption check ------------------------------------------#
#-----------------------------------------------------------------------#
summary(sgd)
# S-shaped check
```

```
l1 = glm(SGD~BM1,family='binomial',data = sgd)
l2 = glm(SGD~BM2,family='binomial',data = sgd)
l3 = glm(SGD~BM3,family='binomial',data = sgd)
l4 = glm(SGD~BM4,family='binomial',data = sgd)

# Draw Curves
par(mfrow=c(2,4),mar=c(4.5,4.5,1.5,2.5), oma = c(0,0,1.5,1.5))
# BM1
r = range(sgd$BM1)
x_range = seq(r[1],r[2],1); x_range = as.integer(x_range)
y = predict(l1,data.frame(BM1=x_range),type="response")
plot(sgd$BM1, sgd$SGD, pch = 16,
     xlab = "BM1", ylab = expression(pi(x)),
     main = expression(paste("Plot of ",pi(x)," vs BM1")))
lines(x_range,y)
plot(x_range, log(y/(1-y)), pch = 16,
     xlab = "BM1", ylab = expression(logit(pi(x))), type="n",
     main = expression(paste("Plot of ",logit(pi(x))," vs BM1")))
lines(x_range,log(y/(1-y)))
# BM2
r = range(sgd$BM2)
x_range = seq(r[1],r[2],1); x_range = as.integer(x_range)
y = predict(l2,data.frame(BM2=x_range),type="response")
plot(sgd$BM2, sgd$SGD, pch = 16,
     xlab = "BM2", ylab = expression(pi(x)),
     main = expression(paste("Plot of ",pi(x)," vs BM2")))
lines(x_range,y)
plot(x_range, log(y/(1-y)), pch = 16,
     xlab = "BM2", ylab = expression(logit(pi(x))), type="n",
     main = expression(paste("Plot of ",logit(pi(x))," vs BM2")))
lines(x_range,log(y/(1-y)))
# BM3
r = range(sgd$BM3)
x_range = seq(r[1],r[2],1); x_range = as.integer(x_range)
y = predict(l3,data.frame(BM3=x_range),type="response")
plot(sgd$BM3, sgd$SGD, pch = 16,
     xlab = "BM3", ylab = expression(pi(x)),
     main = expression(paste("Plot of ",pi(x)," vs BM3")))
lines(x_range,y)
plot(x_range, log(y/(1-y)), pch = 16,
     xlab = "BM3", ylab = expression(logit(pi(x))), type="n",
     main = expression(paste("Plot of ",logit(pi(x))," vs BM3")))
lines(x_range,log(y/(1-y)))
# BM4
r = range(sgd$BM4)
x_range = seq(r[1],r[2],1); x_range = as.integer(x_range)
y = predict(l4,data.frame(BM4=x_range),type="response")
```

```
plot(sgd$BM4, sgd$SGD, pch = 16,
     xlab = "BM2", ylab = expression(pi(x)),
     main = expression(paste("Plot of ",pi(x)," vs BM4")))
lines(x_range,y)
plot(x_range, log(y/(1-y)), pch = 16,
     xlab = "BM4", ylab = expression(logit(pi(x))), type="n",
     main = expression(paste("Plot of ",logit(pi(x))," vs BM4")))
lines(x_range,log(y/(1-y)))

par(mfrow=c(1,1),mar=c(8,4.5,1.5,2.5), oma = c(0,0,1.5,1.5))
### Independence check
## full model
lm.f = glm(SGD~scale(BM1)+scale(BM2)+scale(BM3)+scale(BM4)
           +factor(agen)+factor(season)+factor(year)
           ,data = sgd, family="binomial")
summary(lm.f)

par(mfrow=c(1,2))
plot(lm.f,which=1)
fitted.values(lm.f)

install.packages("ggplot2")
library("ggplot2")
res =
  ggplot(lm.f, aes(.fitted, .resid)) + geom_point()+
  stat_smooth(method="loess")+
  geom_hline(yintercept=0, col="red", linetype="dashed")+
  xlab("Fitted values")+ ylab("Residuals")+
  ggtitle("Residual vs Fitted Plot")+ theme_bw()+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(plot.margin = unit(c(1,0.8,0.8,1), "cm"))
res

### 2 Model Selection ----------------------------------------#
#-------------------------------------------------------------------#
# sample into two data
set.seed(500)
smp.size = floor(0.85*nrow(sgd))
sgd.sp = sample(seq_len(nrow(sgd)),size = smp.size)
sgd.t = sgd[sgd.sp,]
sgd.v = sgd[-sgd.sp,]

# model building data
lm.t = glm(SGD~scale(BM1)+scale(BM2)+scale(BM3)+scale(BM4)
           +factor(agen)+factor(season)+factor(year)
           ,data = sgd.t, family="binomial")
summary(lm.t)
```

```
plot(lm.t) #independence
step(lm.t, direction = "both")
```

Table 7.1: Output: Stepwise Procedure

```
Start:  AIC=91.95
SGD ~ scale(BM1) + scale(BM2) + scale(BM3) + scale(BM4) + factor(agen) +
    factor(season) + factor(year)

                 Df Deviance     AIC
- factor(season)  3   64.095  88.095
- scale(BM2)      1   63.227  91.227
<none>                61.947  91.947
- factor(year)    3   68.877  92.877
- scale(BM3)      1   70.327  98.327
- scale(BM4)      1   71.591  99.591
- factor(agen)    4   83.920 105.920
- scale(BM1)      1   82.183 110.183

Step:  AIC=88.09
SGD ~ scale(BM1) + scale(BM2) + scale(BM3) + scale(BM4) + factor(agen) +
    factor(year)

                 Df Deviance     AIC
- scale(BM2)      1   65.409  87.409
<none>                64.095  88.095
- factor(year)    3   70.893  88.893
+ factor(season)  3   61.947  91.947
- scale(BM3)      1   71.121  93.121
- scale(BM4)      1   72.699  94.699
- factor(agen)    4   87.583 103.583
- scale(BM1)      1   83.811 105.811

Step:  AIC=87.41
SGD ~ scale(BM1) + scale(BM3) + scale(BM4) + factor(agen) + factor(year)

                 Df Deviance     AIC
- factor(year)    3   71.120  87.120
<none>                65.409  87.409
+ scale(BM2)      1   64.095  88.095
+ factor(season)  3   63.227  91.227
- scale(BM3)      1   73.109  93.109
- scale(BM4)      1   75.547  95.547
- scale(BM1)      1   84.095 104.095
- factor(agen)    4   93.493 107.493
```

```
Step:  AIC=87.12
SGD ~ scale(BM1) + scale(BM3) + scale(BM4) + factor(agen)

                 Df Deviance     AIC
<none>                71.120  87.120
+ factor(year)    3   65.409  87.409
+ scale(BM2)      1   70.893  88.893
+ factor(season)  3   68.954  90.954
- scale(BM3)      1   78.769  92.769
- scale(BM4)      1   83.271  97.271
- scale(BM1)      1   87.777 101.777
- factor(agen)    4   99.109 107.109

Call:  glm(formula = SGD ~ scale(BM1) + scale(BM3) + scale(BM4) + factor(agen),
    family = "binomial", data = sgd.t)

Coefficients:
  (Intercept)     scale(BM1)     scale(BM3)     scale(BM4)
       0.3332         7.7735         1.5718         1.5857
factor(agen)2  factor(agen)3  factor(agen)4  factor(agen)5
       0.6248         1.5496        -0.1455        20.7075

Degrees of Freedom: 176 Total (i.e. Null);  169 Residual
Null Deviance:     231.6
Residual Deviance: 71.12  AIC: 87.12


lm.c = glm(SGD~scale(BM1)+scale(BM3)+scale(BM4)+factor(agen),
           family="binomial",data = sgd.t)
summary(lm.c)
plot(lm.c)

### 3 Model Validation ----------------------------------------#
#-------------------------------------------------------------------#
lm.v = glm(SGD~scale(BM1)+scale(BM3)+scale(BM4)+factor(agen),
           family="binomial",data = sgd.v)
# coefficient
summary(lm.c)$coefficients[,1]
summary(lm.v)$coefficients[,1]
# estimated standard deviation
#summary(lm.c)$coefficients[,2]
#summary(lm.v)$coefficients[,2]
# root mean square error
pred.c = predict(lm.c, data = sgd.t, type = "response")
RMSE.c <- mean((sgd.t$SGD - pred.c)^2)
pred.v = predict(lm.v, data = sgd.v, type = "response")
```

```
RMSE.v <- mean((sgd.v$SGD - pred.v)^2)
#R2
install.packages("pscl")
library("pscl")
pR2(lm.c)[4:6]
pR2(lm.v)[4:6]


### 4 Diagonostics ------------------------------------------#
#-----------------------------------------------------------------------#
lm = glm(SGD~scale(BM1)+scale(BM3)+scale(BM4)+factor(agen),
         family="binomial",data = sgd)
summary(lm)
res.c =
  ggplot(lm, aes(.fitted, .resid)) + geom_point()+
  stat_smooth(method="loess")+
  geom_hline(yintercept=0, col="red", linetype="dashed")+
  xlab("Fitted values")+ ylab("Residuals")+
  ggtitle("Residual vs Fitted Plot")+ theme_bw()+
  theme(plot.title = element_text(hjust = 0.5))+
  theme(plot.margin = unit(c(1,0.8,0.8,1), "cm"))
res.c

# marginal effects
install.packages("effects")
library("effects")
m = glm(SGD~scale(BM1)+scale(BM3)+scale(BM4)+agen,data = sgd,family = "binomial")
plot(allEffects(m, default.levels=50,multiline=T,factor.names=T))

# outliers
install.packages("car")
library("car")
cook = cooks.distance(lm)
plot(cook, type = "h", main = "Cook's Distance Plot",col = "tan1", pch = 19)
text(row.names(as.matrix(cook[cook>=0.06])),
cook[cook>=0.06],labels= row.names(as.matrix(cook[cook>=0.06])))

x = as.matrix(cook[cook>=0.06])
y = cook[cook>=0.06]
la= c(52,92,171,191,195,207)

ck =
  ggplot(lm, aes(seq_along(.cooksd),.cooksd))+
  geom_bar(stat="identity", position="identity")+
  xlab("Obs. Number")+ ylab("Cook's distance")+
  ggtitle("Cook's distance")+ theme_bw()+
  theme(plot.title = element_text(hjust = 0.5))+
```

```
  theme(plot.margin = unit(c(1,0.8,0.8,1), "cm"))+
  geom_hline(yintercept=0.06, col="red", linetype="dashed")+
  geom_hline(yintercept=0.08, col="blue", linetype="dashed")+
  annotate("text", x = la, y = y+0.003, label = la,size=3.25)
ck

new = sgd[-c(52,171,191),]
ob = glm(SGD~scale(BM1)+scale(BM3)+scale(BM4)+factor(agen),
           family="binomial",data = new)
plot(ob)

# coefficient
summary(lm.c)$coefficients[,1]
summary(ob)$coefficients[,1]
# estimated standard deviation
#summary(lm.c)$coefficients[,2]
#summary(lm.v)$coefficients[,2]
# root mean square error
pred.o= predict(lm.c, data = sgd.t, type = "response")
RMSE.o <- mean((sgd.t$SGD - pred.o)^2)
pred.n = predict(ob, data = new, type = "response")
RMSE.n <- mean((new$SGD - pred.n)^2)
#R2
install.packages("pscl")
library("pscl")
pR2(lm.c)[4:6]
pR2(ob)[4:6]

# multi-collinearity
install.packages("usdm")
install.packages("sp")
install.packages("raster")
library("usdm")
vif(sgd[,c(1,7,8,9,10)])


### 5 Underlying Effect ----------------------------------------#
#-------------------------------------------------------------------------#
# Time Series Plot
# sort the data
time = new[order(new$year , new$month),]
which(time[,"year"]==1988) #1,7
which(time[,"year"]==1989) #8,71
which(time[,"year"]==1990) #72, 203
which(time[,"year"]==1991) #207, 209
mon.1989=matrix(NA,12,2)
for (i in 1:12){
```

```
  mon.1989[i,1] = min(which(time[,"year"]==1989 & time[,"month"]==i))
  mon.1989[i,2] = i}
mon.1989 = mon.1989[is.finite(mon.1989[,1]),]
mon.1989[,2] = c("Jan","Feb","Mar","Jun","Jul",
                 "Aug","Sep","Oct","Nov","Dec")
mon.1990=matrix(NA,12,2)
for (i in 1:12){
  mon.1990[i,1] = min(which(time[,"year"]==1990 & time[,"month"]==i))
  mon.1990[i,2] = i}
mon.1990 = mon.1990[is.finite(mon.1990[,1]),]
mon.1990[,2] = c("Jan","Feb","Mar","Apr","May","Jun",
                 "Jul","Aug","Sep","Oct","Nov","Dec")


lab1 = c("1988","1989","1990","1991")
lab2 = c(mon.1989[-c(1,2,3,10),2],mon.1990[-c(1,12),2])
mon = c(mon.1989[-c(1,2,3,10),1],mon.1990[-c(1,12),1])
yea = c(1,8,72,204)

# BM1
par(mfrow=c(1,1),mar=c(4.5,4.5,1,1.5), oma = c(0,0,1.5,1.5))
plot(new$BM1,type="n",main = "Time Series of BM1",
     xlab = "month/year", ylab = "BM1", xaxt="n")
grid(lty=9, col=gray(0.85))
lines(new$BM1, type='l', col=4)
axis(1,labels=lab1,at=yea, cex.axis=0.85,
     las=2, mgp=c(3, 1, 0), tck=-0.045)
axis(1,labels=lab2,at=mon, cex.axis=0.65,
     las=2,mgp=c(3,0.5,0),tck=-0.02)
abline(v=mon.1990[c(6,9),1],col="red",lty=2)

# BM2
par(mfrow=c(1,1),mar=c(4.5,4.5,1,1.5), oma = c(0,0,1.5,1.5))
plot(new$BM2,type="n",main = "Time Series of BM2",
     xlab = "month/year", ylab = "BM2", xaxt="n")
grid(lty=9, col=gray(0.85))
lines(new$BM2, type='l', col=4)
axis(1,labels=lab1,at=yea, cex.axis=0.85,
     las=2, mgp=c(3, 1, 0), tck=-0.045)
axis(1,labels=lab2,at=mon, cex.axis=0.65,
     las=2,mgp=c(3,0.5,0),tck=-0.02)
abline(v=mon.1990[c(6,9),1],col="red",lty=2)

# BM3
par(mfrow=c(1,1),mar=c(4.5,4.5,1,1.5), oma = c(0,0,1.5,1.5))
plot(new$BM3,type="n",main = "Time Series of BM3",
     xlab = "month/year", ylab = "BM3", xaxt="n")
```

```
grid(lty=9, col=gray(0.85))
lines(new$BM3, type='l', col=4)
axis(1,labels=lab1,at=yea, cex.axis=0.85,
     las=2, mgp=c(3, 1, 0), tck=-0.045)
axis(1,labels=lab2,at=mon, cex.axis=0.65,
     las=2,mgp=c(3,0.5,0),tck=-0.02)
abline(v=mon.1990[c(6,9),1],col="red",lty=2)

# BM4
par(mfrow=c(1,1),mar=c(4.5,4.5,1,1.5), oma = c(0,0,1.5,1.5))
plot(new$BM4,type="n",main = "Time Series of BM4",
     xlab = "month/year", ylab = "BM4", xaxt="n")
grid(lty=9, col=gray(0.85))
lines(new$BM4, type='l', col=4)
axis(1,labels=lab1,at=yea, cex.axis=0.85,
     las=2, mgp=c(3, 1, 0), tck=-0.045)
axis(1,labels=lab2,at=mon, cex.axis=0.65,
     las=2,mgp=c(3,0.5,0),tck=-0.02)
abline(v=mon.1990[c(6,9),1],col="red",lty=2)
```

Table 7.2: Output: One-way ANOVA tests

```
> anova(lm(BM1~factor(season)+factor(year),data=new))
Analysis of Variance Table

Response: BM1
                Df  Sum Sq Mean Sq F value Pr(>F)
factor(season)   3   73117   24372  1.0293 0.3807
factor(year)     3   70566   23522  0.9934 0.3970
Residuals      199 4711881   23678
> anova(lm(BM2~factor(season)+factor(year),data=new))
Analysis of Variance Table

Response: BM2
                Df Sum Sq Mean Sq F value    Pr(>F)
factor(season)   3   2071   690.3  3.1518   0.02602 *
factor(year)     3  14250  4750.0 21.6891 3.393e-12 ***
Residuals      199  43582   219.0
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
> anova(lm(BM3~factor(season)+factor(year),data=new))
Analysis of Variance Table

Response: BM3
                Df  Sum Sq Mean Sq F value  Pr(>F)
```

```
factor(season)   3  1142.6  380.87  2.4606 0.06388 .
factor(year)     3   594.7  198.23  1.2807 0.28215
Residuals      199 30802.6  154.79
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
> anova(lm(BM4~factor(season)+factor(year),data=new))
Analysis of Variance Table

Response: BM4
                Df Sum Sq Mean Sq F value  Pr(>F)
factor(season)   3  55763 18587.8  3.7277 0.01224 *
factor(year)     3  26751  8917.0  1.7883 0.15065
Residuals      199 992283  4986.3
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1
```