Malaysian Journal of Mathematical Sciences 4(1): 107-133 (2010)

# Model-Building Approach in Multiple Binary Logit Model for Coronary Heart Disease

**[1]H. J. Zainodin and [2]G. Khuneswari**
*School of Science and Technology*
*Universiti Malaysia Sabah, Locked Bag No. 2073,*
*88999 Kota Kinabalu, Sabah, Malaysia*
*E-mail: [1]zainodin@gmail.com, [2]khuneswari@gmail.com*

## ABSTRACT

This paper develops a procedure to find the best model. An illustration of the model-building approach in "Multiple Binary Logit" analysis has been introduced. The dependent variable of Multiple Binary Logit model is a qualitative nature (taking a value of 0 or 1). Besides introducing multiple single independent variables into the model, all possible combinations of generated interaction variables are included in the model. In order to obtain a set of selected models, a progressive elimination (one by one, least significant first) of the insignificant variables is employed. It was also proposed to use the modified Eight Selection Criteria (8SC) by replacing SSE (sum square of error) Deviance Statistic ($G^2$) to finally single out the best model. A numerical illustration (case study) on coronary heart disease (CHD) was included in order to get a clear picture of the procedure of getting the best Multiple Binary Logit model. In addition, there are three quantitative (age, cholesterol level and body-mass index (BMI)) and five qualitative independent variables (4 blood pressure categories and smoking habit). Detailed procedure is exposed, illustrated and explained using this case study. The model building approach through the Multiple Binary Logit model was established. It was found that interaction variables as age and BMI, BMI and cholesterol level, age and blood pressure category, SBP 100-129 and DBP 60-79, age and blood pressure category, SBP 130-139 and DBP 80-89, BMI and blood pressure category, SBP 130-139 and DBP 80-89, BMI and blood pressure category, SBP >140 and DBP >90 and cholesterol level and blood pressure category, SBP 100-129 and DBP 60-79 are significant in the best model obtained. The dummy variables blood pressure category, SBP >140 and DBP >90 and smoking/non smoking are significant in the best model obtained.

Keywords: model-building approach, Multiple Binary Logit, interaction variables, elimination procedure, M8SC, Deviance Statistics

## INTRODUCTION

Multiple Binary Logit (MBL) is the extension of Logit model. MBL is known as qualitative choice model. The difference between Multiple Regressions (MR) and MBL is the characteristic of the dependent variable and the procedures involved in estimating the parameters. Due to the nature of the dependent variable, Binary Logit model cannot be estimated using ordinary least squares (OLS). A more suitable estimation method to estimate

parameters in the Binary Logit model is Maximum Likelihood (ML). Studenmund (2006) stated that ML estimation is inherently different from OLS in that it chooses coefficient estimates that maximise the likelihood of the sample data set being observed. OLS and ML estimates are not necessarily different from a linear equation that meets classical assumptions. It has a number of desirable large sample properties. ML is consistent and asymptotically efficient. ML has the added advantage of producing normally distributed coefficient estimates with very large samples, allowing the use of typical hypothesis testing techniques.

The main objective of this work is to illustrate the model-building approach. The whole structure starting from identifying the dependent and independent variable to getting the best MBL model will be illustrated using a numerical illustration. The best MBL model will represent the whole structure of the collected data so that further analysis can be carried out.

## METHODOLOGY

The Logit model is based on the cumulative logistic regression. It will give probability estimates that are bounded by 0 and 1. The dependent variable set up differently in Logit model (Halcoussis, 2005). The Logit model is similar in form to the linear regression model. The general Multiple Binary Logit model is

$$\mathrm{Y} = \Omega_0 + \Omega_1 \mathrm{W}_1 + \Omega_2 \mathrm{W}_2 + \ldots + \Omega_k \mathrm{W}_k + \mathrm{u} \tag{1}$$

and

$$Y_i = \ln\left(\frac{p_i}{1-p_i}\right) \tag{2}$$

where Y denotes the binary dependent variable, $W_j$ denotes the *j-th* variable (which can be single independent variable, or interaction variable (first-order interaction, second-order interaction, third-order interaction, ...), generated variable (polynomial and dummy variable) and transformed variables (Ladder transformation and Box-Cox transformation). The $\Omega_0$ denotes the constant term of the model and the $\Omega_j$ denotes the *j-th* coefficient of independent variable $W_j$. The k denotes the number of independent variables, (k+1) denotes the number of parameters and $p_i$ is the *i-th* probability of an event occurring for i=1, 2, …, n and j=1, 2, ..., k. The $p_i$ is the *i-th* probability of an event occurs whereas $1-p_i$ is the compliment event of $p_i$. Fitted value for the dependent variable is now representing the

logarithm of the odds that Y equals to 1. According to Halcoussis (2005) a change in an independent variable affects the logarithm of the odds that Y equals to 1. A coefficient estimate from a Binary Logit (BL) model tells us the change in the logarithm of the odds for one unit change in an independent variable. Other independent variables are kept constant. In BL the final estimated value is in the form of probability where $p_i$ is to be estimated.

In the development of the mathematical model, there are four phases involved. The phases are possible models, selected models, best model and goodness-of-fit test. The four phases are as follows:

**Phase 1**: All Possible Models
- Single independent variables and all possible product of related single independent variables (interaction variables)

**Phase 2**: Selected Models
- Eliminate source variable(s) of multicollinearity phenomenon
- Elimination, discard a variable with $|t_{cal}|$ less than critical value and nearest to zero

**Phase 3**: Best Model
- Using 8SC: minimise for each criterion and mark the chosen model. The most preferred model by the criteria is the best model.

**Phase 4**: Goodness-of-Fit
- Deviance and Pearson Chi-Square test

These four phases will be discussed accordingly in the following section. These are the main phases in model-buliding approach. Each phase has its specific tests and justification.

### *All Possible Models*

Before the analysis is carried out, the entire possible models up to complete higher-order interaction variables must be listed out and considered. This is to help in determining the significant variables that contribute to the dependent variable. Eventually, only the contributing variables stay in the best model. The number of possible models can be calculated as follows (Zainodin and Khuneswari, 2007; 2009):

$$N = \sum_{j=1}^{q} j({}^{q}C_{j}) \tag{3}$$

where ${}^{q}C_{j}$ is defined as $\frac{q!}{j!(q-j)!}$ number of possible combinations and q is number of single independent variables (for j=1, 2, 3, ...,q).

### *Selected Models*

Consider each possible model to be written in terms of model (1). Observe correlation coefficient matrix of all the variables involve in the model. If multicollinerity phenomenon exists then remove the source variable(s). Next, let's consider a model with $k$ variables (with $k$+1 number of parameters which include the constant term) as one of the possible models. In the process of getting the selected model from possible models, Global test, Coefficient test (eliminating insignificant variables) and Wald test should be carried out to get significant variable that contributed to the dependent variable.

### *Global Test*

The Global test is carried out to investigate whether it is possible for all the independent variables in the model to have zero net regression coefficients (Zainodin and Khuneswari, 2007). The global test is carried out for all the possible models. The hypothesis for global test is as follows:

$$\mathrm{H}_0 : \Omega_1 = \ldots = \Omega_{k-1} = \Omega_k = 0$$
$$\mathrm{H}_1 : \text{at least one } \Omega\text{'s is nonzero}$$

The $\mathrm{F}_{\mathrm{cal}}$ is $\left(SSR/k\right)/\left[SSE/(n-k-1)\right]$ (where SSR is the sum of square regression due to the considered model and SSE is sum of square error) and $\mathrm{F}_{\mathrm{critical}}$ is F (k, (n-k-1), $\alpha$). The decision is to reject the null hypothesis where all the regression coefficients are zero if $\mathrm{F}_{\mathrm{cal}}$ is greater than $\mathrm{F}_{\mathrm{critical}}$. In most cases, the null hypothesis is rejected. Thus, indicating that at least one of the coefficients is nonzero. Then, the next step is to search for a nonzero coefficient. So the process of elimination is applied to remove non-contributing (insignificant) variables from the model.

### *Coefficient Test*

The next step is to perform the Coefficient test for all the coefficients in the model. The objective of the Coefficient test is to identify the most

insignificant coefficient and eliminate the corresponding insignificant variable. According to Zainodin and Khuneswari (2007), the Coefficient test is carried out by testing the coefficient of the corresponding variable with the value of zero. The insignificant variable will be eliminated one at a time. This will lead to the elimination procedure. Consider the model as defined in equation (1) where the dependent variable, Y is a binary variable. The algorithm of the elimination procedure will be shown in detail is as follows.

**Step 1**: Estimate the parameters, Ω using Maximum Likelihood estimation.

**Step 2**: Find the Residuals, SSE and MSE

$$\text{Residual} = Y_i - \hat{Y}_i$$

$$\text{SSE} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \quad \text{and degrees of freedom, df =n-k-1}$$

$$\text{MSE} = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{n-k-1}, \qquad \hat{\sigma}^2 = \text{MSE}$$

**Step 3**: Find var($\hat{\Omega}_j$) where var($\hat{\Omega}_j$) = $\hat{\sigma}^2 C_{\text{jj}}$

$$(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1} = \begin{bmatrix} C_{00} & C_{01} & \ldots & C_{0k} \\ C_{10} & C_{11} & \ldots & C_{1k} \\ . & . & & . \\ . & . & C_{jj} & . \\ . & . & & . \\ C_{k0} & C_{k1} & \ldots & C_{kk} \end{bmatrix}$$

where $C_{\text{ij}}$ denotes the value of *i-th* row and *j-th* column of matrix $(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}$

(for i and j = 0,1, 2, ..., k)

$C_{\text{jj}}$ denotes the *j-th* element of the diagonal of matrix $(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{W})^{-1}$

**Step 4**: The hypothesis of Coefficient test for *j-th* coefficient:

$$\mathrm{H}_0 : \Omega_j = 0$$

$$\mathrm{H}_1 : \Omega_j \neq 0$$

**Step 5**: Calculate the $t_{cal} = t_j$ for each Ω's.

$$t_j = \frac{\hat{\Omega}_j - \Omega_j(H_0)}{se(\hat{\Omega}_j)} \text{ for j = 1, 2, ..., k (j =0 is not tested)}$$

where

$\hat{\Omega}_j$ denotes the *j-th* estimated coefficient value

$\Omega_j(H_0)$ denotes the *j-th* coefficient value under the null hypothesis

se( $\hat{\Omega}_j$ ) denotes the standard error of $\hat{\Omega}_j$ and se( $\hat{\Omega}_j$ ) = $\sqrt{\text{var}(\hat{\Omega}_j)}$

**Step 6**: The $t_{critical} = t_{\alpha/2,(n-k-1)}$

where the level of significance is α (usually of 5%), n is the sample size and k+1 is the number of parameters in the final model.

**Step 7**: Let t* be the minimum $\{t_1, t_2, ..., t_k\}$. If $t^* < |t_{critical}|$ and $t^* \to 0$, eliminate the corresponding independent variable for all possible coefficients where j=1, 2, ..., k.

**Step 8**: Repeat the Step 1 to Step 7 until there is no more independent variable with $t^* < |t_{critical}|$. Otherwise, the selected model is achieved.

These are the eight steps (Step 1 to Step 8) involved in the elimination procedure. The selected model is achieved once all the independent variables in the corresponding model are significant.

### *Wald Test*

After the coefficient test (the omission of the most insignificant variable) takes place one by one, the Wald test is carried out to justify the removal of the insignificant variables (Zainodin and Khuneswari, 2009). In this situation the restricted model is the selected model whereas the unrestricted model is the initial possible model. Consider the following situation where:
Unrestricted model (possible model) as defined in equation (1):

$$\mathbf{U}: \quad Y = \Omega_0 + \Omega_1 W_1 + \Omega_2 W_2 + ... + \Omega_m W_m + \Omega_{m+1} W_{m+1} + ... + \Omega_k W_k + u$$

Restricted model (selected model) equation:

$$\mathbf{R}: \quad Y = \Omega_0 + \Omega_1 W_1 + \Omega_2 W_2 + ... + \Omega_m W_m + r$$

where variables $W_{m+1}$, $W_{m+2}$, ..., $W_k$ (not necessarily in that order) were removed from unrestricted model. The hypothesis of Wald test for removing variables $W_{m+1}$, $W_{m+2}$, ..., $W_k$ from equation (1) is as follows:

$$H_0 : \Omega_{m+1} = \Omega_{m+2} = ... = \Omega_k = 0$$
$$H_1 : \text{At least one } \Omega' s \text{ in } H_0 \text{ is nonzero}$$

TABLE 1: ANOVA for Wald Test

| Source of variations | Sum of Squares | *df* | Mean Sum of Squares | F |
|---|---|---|---|---|
| **Differences (R-U)** | SSE(R) - SSE(U) | k-m | $\frac{[SSE(R)-SSE(U)]}{k-m}$ | $F=\frac{[SSE(R)-SSE(U)]/(k-m)}{SSE(U)/(n-k-1)}$ |
| **Unrestricted (U)** | SSE(U) | n-k-1 | $\frac{SSE(U)}{n-k-1}$ | |
| **Restricted (R)** | SSE(R) | n-m-1 | | |

The Table 1 shows the $F_{cal}$ and the critical value is F [df(R )–df(U), df(U), $\alpha$] or can be written as $F_{(k\text{-}m,\ n\text{-}k\text{-}1,\ \alpha)}$. The decision is to reject the null hypothesis if $F_{cal}$ is greater than $F_{critical}$. If the null hypothesis is accepted, it justified the elimination of insignificant variables as stated in earlier section (Coefficients Test). Similar procedure is carried out for all the remaining selected models.

### *Best Model*

The Best Model will be determined from selected models that were obtained from previous test conducted in the previous section. The best model will eventually emerge when selection criteria is used. In obtaining the best model, Zainodin and Khuneswari (2007; 2009) have explained in detail the use of Eight Selection Criteria (8SC). Each of the selected models is subjected to each of the Model Selection Criterion.

The model selection criteria are Akaike information criterion (AIC), finite prediction error (FPE), generalised cross validation (GCV), Hannan and Quinn criterion (HQ), RICE, SCHWARZ, SGMASQ and SHIBATA. Finite prediction error (FPE) and Akaike information criterion (AIC) was

developed by Akaike (1969, 1974). HQ criterion was suggested by Hannan and Quinn in 1979. Golub *et al.* (1979) developed generalised cross validation (GCV). Other criteria are included SCHWARZ (Schwarz, 1978), SHIBATA (Shibata, 1981) and RICE (Rice, 1984). These criteria take the form of the sum square of error (SSE) multiplied by a penalty factor that depends on the model complexity as measured by the number of parameters (k+1) to be estimated. SGMASQ (Ramanathan, 2002) is the estimated residual variance ( $\hat{\sigma}^2$ ). Detail discussion on each criterion is discussed in Zainodin and Khuneswari (2007; 2009).

The coefficients in Multiple Binary Logit model are estimated using Maximum Likelihood estimation. For logistic models, the Deviance statistics was used as model selection criteria. As defined by Kutner *et al.* (2008), the Deviance statistic (sum squares of the deviance residuals) is

$$\mathbf{G}^2 = -2\sum_{i=1}^{n}[Y_i \ln(\hat{p}_i) + (1-Y_i)\ln(1-\hat{p}_i)] \ . \tag{4}$$

Vogelvang (2005) stated that maximising likelihood (or minimising Deviance) is identical to minimising the sum square of error (SSE). Therefore, a modification was suggested on eight selection criteria where SSE was replaced by Deviance statistics where $G^2$ is -2 times the log-likelihood.

TABLE 2: Modified Eight Selection Criteria

| **MODIFIED EIGHT SELECTION CRITERIA (M8SC)** | |
|---|---|
| **AIC:** $\left(\frac{G^2}{n}\right)(e)^{2(k+1)/n}$ | **RICE:** $\left(\frac{G^2}{n}\right)\left[1-\frac{2(k+1)}{n}\right]^{-1}$ |
| **FPE:** $\left(\frac{G^2}{n}\right)\frac{n+k+1}{n-(k+1)}$ | **SCHWARZ:** $\left(\frac{G^2}{n}\right)(n)^{2(k+1)/n}$ |
| **GCV:** $\left(\frac{G^2}{n}\right)\left[1-\frac{k+1}{n}\right]^{-2}$ | **SGMASQ:** $\left(\frac{G^2}{n}\right)\left[1-\frac{k+1}{n}\right]^{-1}$ |
| **HQ:** $\left(\frac{G^2}{n}\right)(\ln n)^{2(k+1)/n}$ | **SHIBATA:** $\left(\frac{G^2}{n}\right)\frac{n+2(k+1)}{n}$ |

These criteria take the form of the sum of square of deviance residual (that is, Deviance statistics or $G^2$ as defined in equation (4)) multiplied by a penalty factor that depends on the model complexity as measured by the number of parameters *(k+1)* to be estimated. The summary of modified eight selection criteria (M8SC) are shown in Table 2. $G^2$ is the Deviance statistics (as defined in equation (4)), *k+1* is the number of estimated parameters and *n* stands for sample size. There is a condition to be fulfilled when using these model selection criteria, that is, $2(k+1) < n$. After all this criteria is computed, that the best model can be obtained by choosing the model which has lowest values for most of the criteria. Finally, the best Multiple Binary Logit model is obtained.

### *Goodness-of-Fit Test*

Once the best model had been obtained, the residual analysis was carried out to examine the appropriateness of the best model. The residual analysis for Multiple Binary Logit regression is more difficult than Multiple Regression because the dependent variable, $Y_i$ takes on only the values 0 and 1. Therefore, $u_i$ will take either one of the values corresponding to the $Y_i$:

$$u_i = \begin{cases} 1-\hat{p}_i & if \quad Y_i = 1 \\ -\hat{p}_i & if \quad Y_i = 0 \end{cases} \quad \text{for i= 1, 2, ..., n} \tag{5}$$

According to Kutner *et al.* (2008), the residual will not be normally distributed. Plots of ordinary residuals against predicted values or independent variables will generally be uninformative. There are two tests suggested by Kutner *et al.* (2008): Pearson Chi-Square Goodness-of-fit test and Deviance Goodness-of-fit test. Rosado *et al.* (2006) also suggested using Pearson Chi-Square Goodness-of-fit test to evaluate the adjustment of the model. The Pearson residuals (as defined by Kutner *et al.,* 2008) is

$$r_{pi} = \frac{Y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}} \tag{6}$$

where $\sqrt{\hat{p}_i(1-\hat{p}_i)}$ is the estimated standard error of $Y_i$ for i=1, 2, ..., n. The sum square of the Pearson residuals is numerically equal to the Pearson chi-square test statistics. The Pearson chi-square test statistics is

$$\chi_{r_{pi}}^2 = \sum_{i=1}^{n} \frac{(Y_i - \hat{p}_i)^2}{\hat{p}_i(1-\hat{p}_i)} \tag{7}$$

The deviance residuals (as defined by Kutner *et al.,* 2008) is

$$dev_i = \text{sign}\ (Y_i - \hat{p}_i)\ \sqrt{-2[Y_i \ln(\hat{p}_i) + (1 - Y_i)\ln(1 - \hat{p}_i)]} \qquad (8)$$

where $\text{sign} = \begin{cases} + & \text{when} \quad Y_i \geq \hat{p}_i \\ - & \text{when} \quad Y_i < \hat{p}_i \end{cases}$ for i=1, 2, ..., n.

The Deviance statistics (sum squares of the deviance residuals) is numerically equal to the deviance statistics (as defined in equation (4)).

Besides these two tests, scatter plot of these residual can be used as the supporting evidence to check the appropriateness of the model. According to Kutner *et al.* (2008), there are three common residual plots used in Binary Logit regression analysis: ordinary residual against estimated probability, Pearson residual against estimated probability and Deviance residuals against estimated probability. The plots suggest that if the model is correct, the plot of the residual against the probability $(p_i)$ should result approximately in a regression line with zero intercept.

## CASE STUDY: CORONARY HEART DISEASE

Coronary Heart Disease is now the leading cause of death worldwide. According to World Health Organization (WHO, 2003), 3.8 million men and 3.4 million women worldwide die each year from coronary heart disease. The coronary heart disease risk factors are divided into two categories: controllable and uncontrollable. The controllable risk factors are high blood pressure, high blood cholesterol, smoking, obesity, physical inactivity, diabetes and stress. Whereas the uncontrollable risk factors are gender, heredity (family history of CHD) and age.

A study was conducted by Western Collaborative Group Study (WCGS) in California. The study began in July 1985 and completed in June 1990 (Vittinghoff *et al.*, 2004). There were 3154 male volunteers. The description of variables is shown in Table 3. Three quantitative and five qualitative independent variables associated with Coronary Heart Disease (CHD) were used to investigate the occurrence of coronary heart disease. The blood pressure was divided into four categories based on Systolic blood pressure (SBP) and Diastolic blood pressure (DBP) according to World Health Organization (WHO, 2003).

TABLE 3: Data Description on Variables for Coronary Heart Disease

| Variable | Description | Type of Variable |
|---|---|---|
| **Y** | The Coronary Heart Disease:<br>1 if a coronary incident occurred<br>0 if otherwise | Qualitative |
| **$X_1$** | Age (in years) | Quantitative |
| **$X_2$** | Body mass index (BMI) | Quantitative |
| **$X_3$** | Cholesterol level (milligrams per DL) | Quantitative |
| **$D_1$** | 1 if SBP<100 and DBP<60<br>0 if otherwise | Qualitative |
| **$D_2$** | 1 if SBP 100-129 and DBP 60-79<br>0 if otherwise | Qualitative |
| **$D_3$** | 1 if SBP 130-139 and DBP 80-89<br>0 if otherwise | Qualitative |
| **$D_4$** | 1 if SBP>140 and DBP>90<br>0 if otherwise | Qualitative |
| **$D_5$** | 1 if a person smokes<br>0 if non smoker | Qualitative |

The Table 4 shows the descriptive statistics for quantitative variables. The dependent variable, Y, and independent variables, $D_1$, $D_2$, $D_3$, $D_4$ and $D_5$ each is a binary variable. Therefore, the descriptive analysis could not be carried out to explore the data. There are 257 men (8.18%) of 3142 with coronary heart disease.

TABLE 4: The descriptive qualitative independent variables

| | **Variables** | | | | |
|---|---|---|---|---|---|
| | **$D_1$** | **$D_2$** | **$D_3$** | **$D_4$** | **$D_5$** |
| 1's | 4 | 990 | 406 | 451 | 1495 |
| 0's | 3138 | 2152 | 2736 | 2691 | 1647 |

As in Table 5, could see that the variables $X_1$, $X_2$ and $X_3$ have positive values of skewness. This means that each of these variables are skewed to the right and each of the variable has mode < median < mean. As stated by Crawley (2006), the distribution of a variable is said to be normal if the value of skewness fall within [-0.0437, 0.0437] and kurtosis fall within [-0.0874, 0.0874].

TABLE 5: The descriptive Statistics for quantitative independent variables

| Statistics | Variables | | |
|---|---|---|---|
| | $X_1$ | $X_2$ | $X_3$ |
| Mean | 46.2747 | 24.5145 | 226.3720 |
| Standard Error | 0.0985 | 0.0457 | 0.7746 |
| Median | 45 | 24.3898 | 223 |
| Mode | 40 | 24.3898 | 212 |
| Standard Deviation | 5.5185 | 2.5638 | 43.4204 |
| Sample Variance | 30.4533 | 6.5731 | 1885.3300 |
| Kurtosis (Se(kutosis) = 0.0874) | -0.7699 | 2.0059 | 3.0388 |
| Skewness (Se(skewness) = 0.0437) | 0.5262 | 0.5365 | 0.6768 |
| Minimum | 39 | 11.1906 | 103 |
| Maximum | 59 | 38.9474 | 645 |

For all three variables, the value of skewness only varies between -0.5 to 0.68 (as in Table 5). Therefore, it shows that all these two variables can be assumed approximately normal according to the skewness value because the value is not too high. But only the variable $X_3$ has value of kurtosis nearer to 3 which means that only the data for variable $X_3$ is normal. However, the variables $X_1$ and $X_2$ are not normal since their values of kurtosis are below 3.

TABLE 6: The Pearson correlation for Coronary Heart Disease

| | Y | $X_1$ | $X_2$ | $X_3$ | $D_1$ | $D_2$ | $D_3$ | $D_4$ | $D_5$ |
|---|---|---|---|---|---|---|---|---|---|
| **Y** | 1 | .120(**) | .063(**) | .163(**) | -.011 | -.095(**) | .027 | .083(**) | .085(**) |
| | | .000 | .000 | .000 | .550 | .000 | .131 | .000 | .000 |
| **$X_1$** | .120(**) | 1 | .026 | .089(**) | -.005 | -.113(**) | .023 | .125(**) | .003 |
| | .000 | | .152 | .000 | .779 | .000 | .195 | .000 | .849 |
| **$X_2$** | .063(**) | .026 | 1 | .071(**) | -.022 | -.239(**) | .031 | .243(**) | -.143(**) |
| | .000 | .152 | | .000 | .217 | .000 | .081 | .000 | .000 |
| **$X_3$** | .163(**) | .089(**) | .071(**) | 1 | -.014 | -.118(**) | .041(*) | .092(**) | .097(**) |
| | .000 | .000 | .000 | | .417 | .000 | .022 | .000 | .000 |
| **$D_1$** | -.011 | -.005 | -.022 | -.014 | 1 | -.024 | -.014 | -.015 | .020 |
| | .550 | .779 | .217 | .417 | | .175 | .441 | .413 | .272 |
| **$D_2$** | -.095(**) | -.113(**) | -.239(**) | -.118(**) | -.024 | 1 | -.261(**) | -.278(**) | .014 |
| | .000 | .000 | .000 | .000 | .175 | | .000 | .000 | .445 |
| **$D_3$** | .027 | .023 | .031 | .041(*) | -.014 | -.261(**) | 1 | -.158(**) | .017 |
| | .131 | .195 | .081 | .022 | .441 | .000 | | .000 | .348 |

TABLE 6 (continued): The Pearson correlation for Coronary Heart Disease

| | **Y** | **$X_1$** | $X_2$ | **$X_3$** | **$D_1$** | **$D_2$** | **$D_3$** | **$D_4$** | **$D_5$** |
|---|---|---|---|---|---|---|---|---|---|
| **$D_4$** | .083(**) | .125(**) | .243(**) | .092(**) | -.015 | -.278(**) | -.158(**) | 1 | -.045(*) |
| | .000 | .000 | .000 | .000 | .413 | .000 | .000 | | .012 |
| **$D_5$** | .085(**) | .003 | -.143(**) | .097(**) | .020 | .014 | .017 | -.045(*) | 1 |
| | .000 | .849 | .000 | .000 | .272 | .445 | .348 | .012 | |

** Correlation is significant at the 0.01 level (2-tailed) and Correlation is significant at the 0.05 level (2-tailed).

Based on Table 6, there is no strong relationship (significantly) between coronary incident with respect to age ($X_1$), BMI ($X_2$) and cholesterol level ($X_3$). Correlation is significant at the 0.01 level (2-tailed). As can be seen from the highlighted triangle in Table 6, there is no existence of multicollinearity (|correlation coefficient| > 0.95) between the independent variables. Thus, no further treatment or modification is required on the given data set and the data is ready for further analysis.

*All Possible Models*

The number of all possible models for coronary heart diseases (CHD) data had been listed as in Table 7. Since there are three independent variables, the total number of possible model is 12 (using equation (1) with q=3) with all possible combination of variables.

TABLE 7: All Possible Models for three single Independent Variables

| **Number of Variables** | **Individuals** | **Interactions** | | **Total** |
|---|---|---|---|---|
| | | **First-Order** | **Second-Order** | |
| 1 | 3 | - | - | 3 |
| 2 | 3 | 3 | - | 6 |
| 3 | 1 | 1 | 1 | 3 |
| **Total** | 7 | 4 | 1 | **12** |
| **Model Number (as in Appendix A)** | M1-M7 | M8-M11 | M12 | |

Since there are three quantitative single independent variables, 12 possible models up to third-order interaction variables had been considered in the analysis. The corresponding five qualitative independent variables are to be added in the 12 possible models up to first-order interaction variables. The 12 possible models are listed in Appendix A.

Each possible model will be estimated systematically in the following step to obtain selected models with significant variables. The variable $D_1$ was eliminated earlier from each possible models because an error occurred in obtaining covariance matrix leading to error in estimating the parameters in the model.

*Selected Models*

After clearing from multicollinearity phenomenon (if exist one), the next step was to estimate the coefficients for the 12 possible models and carry out tests to obtain selected models. For illustration purpose, consider model M7 with the following detail:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + (\beta_{D_1} D_1 + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{D_4} D_4 + \beta_{D_5} D_5) + u$$

This model can be written in the form of equation (1) where

$\Omega_0 = \beta_0$
$\Omega_1 = \beta_1$ and $W_1 = X_1$
$\Omega_2 = \beta_2$ and $W_2 = X_2$
$\Omega_3 = \beta_3$ and $W_3 = X_3$
$\Omega_4 = \beta_{D1}$ and $W_4 = D_1$
$\Omega_5 = \beta_{D2}$ and $W_5 = D_2$
$\Omega_6 = \beta_{D3}$ and $W_6 = D_3$
$\Omega_7 = \beta_{D4}$ and $W_7 = D_4$
$\Omega_8 = \beta_{D5}$ and $W_8 = D_5$

The number of independent variable is 8 (=k) and number of parameters is 9 (=k+1). The variable $D_1$ was eliminated earlier because an error occurred in obtaining covariance matrix. Therefore, only seven variables had been estimated. The model after variable $D_1$ eliminated is model M7.1. Table 8 represents the ANOVA for Global test. The hypothesis of Global test for model M7.1 is as follows:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_{D_2} = \beta_{D_3} = \beta_{D_4} = \beta_{D_5} = 0$$
$$H_1 : \quad \text{at least one } \beta\text{'s in } H_0 \text{ is nonzero}$$

As can be seen from Table 8, the $F_{cal}$ is 28.5105 and the $F_{critical}$ is $F_{0.05,7,3134}$=2.0100. Since $F_{cal}$ is greater than $F_{critical}$, the decision is to reject the null hypothesis where the entire regression coefficients in model M7.1 are zero. The next step is to search for insignificant variables by performing the Coefficient test for the entire coefficient in the model M7.1. The hypothesis of Coefficient test for the first coefficient $\beta_1$ is as follows:

$$H_0 : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

The $t_{cal}$ is 5.6901 and the $t_{critical}$ is $t_{0.025,\ 3133}$=1.960. The decision is to reject the null hypothesis where the $\beta_1$ is nonzero since $|t_{cal}|$ is greater than $|t_{critical}|$. This means that $\beta_1$ is significant in the model. From Table 9, it is shown that dummy $D_1$ was eliminated much earlier due to the fact that the variance of this $D_1$ is far too small. Similar procedure is carried out for other coefficients in the model where the test leads to the elimination of variables $D_3$ and $D_4$ respectively.

TABLE 8: The ANOVA for Global test of model M7.1

| Source of variations | Sum of Squares | *df* | Mean Sum of Squares | F |
|---|---|---|---|---|
| **Regression** | 14.2293 | 7 | 2.0328 | 28.5105 |
| **Residuals** | 223.5807 | 3134 | 0.0713 | |
| **Total** | 237.8100 | 3141 | | |

The Table 9 shows the estimated coefficient value and $t_{cal}$ value in parentheses. As seen from Table 9 (model M7.1), there are two variables, each has $|t_{cal}|$ less than $|t_{critical}|$=1.96. So, the dummy variable $D_3$ ($t_{cal}$= 0.6080) was eliminated and the new model M7.2 was rerun. The resulting $t_{cal}$ after eliminating variable $D_3$ is shown in Table 9 (model M7.2). There are insignificant variables in the new regression model. The insignificant variable $D_4$ ($t_{cal}$=1.6107) was eliminated and the new model M7.3 was obtained. All the variables in model M7.3 are significant ($|t_{cal}|$ are greater $|t_{critical}|$).

TABLE 9: Illustration of Elimination Procedure in getting Selected Model (model M7.3)

| Variables | Models | | |
|---|---|---|---|
| | **M7.1** | **M7.2** | **M7.3** |
| Constant | -9.9985 | -9.9764 | -10.2550 |
| $X_1$ | 0.0673<br>(5.6901) | 0.0674<br>(5.7009) | 0.0693<br>(5.9013) |
| $X_2$ | 0.0628<br>(2.3516) | 0.0628<br>(2.3520) | 0.0720<br>(2.7512) |
| $X_3$ | 0.0108<br>(7.2639) | 0.0108<br>(7.2700) | 0.0110<br>(7.3497) |
| $D_2$ | -0.5439<br>(-2.8599) | -0.5748<br>(-3.1456) | -0.6315<br>(-3.5376) |
| $D_5$ | 0.6431<br>(4.6014) | 0.6450<br>(4.6159) | 0.6386<br>(4.5771) |
| $D_4$ | 0.3067<br>(1.7127) | 0.2757<br>**(1.6107)** | - |

TABLE 9 (continued): Illustration of Elimination Procedure in getting Selected Model (model M7.3)

| Variables | Models | | |
|---|---|---|---|
| | **M7.1** | **M7.2** | **M7.3** |
| $D_3$ | 0.1199 **(0.6080)** | - | - |
| $D_1$ | - | - | - |
| Elimination Step | | Step 1 | Step 2 |
| | | | |
| **SSE** | 223.5807 | 223.4979 | 223.8264 |
| $\mathbf{G^2}$ | 1618.9652 | 1619.3298 | 1621.8572 |
| $^*\mathbf{t_{critical}}$ = 1.960 and value in parentheses is the $t_{cal}$ | | | |

Table 9, shows that the sum square of deviance residuals (Deviance statistics) was increased from 1618.9652 (for model M7) to 1621.85724 (for model M7.2) through the two elimination steps. This is because the eliminated insignificant variables at each step from the previous model were absorbed into Deviance of resulting model. The next step is to justify the elimination of variables $D_3$ and $D_4$. Therefore, the Wald Test was carried out to justify the elimination of these two variables.

The Wald test was carried out to the final model where the restricted model (model M7.1) is the selected model and the unrestricted model is the initial possible model (M7.3).
The unrestricted model (Possible model): M7.1

$$\mathrm{U}: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + (\beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{D_4} D_4 + \beta_{D_5} D_5) + u$$

The restricted model (Selected model): M7.3

$$\mathrm{R}: Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + (\beta_{D_2} D_2 + \beta_{D_5} D_5) + r$$

The hypothesis of Wald test for removing variables $D_3$ and $D_4$ from model M7 is as follows:

$$\mathrm{H}_0 : \beta_{D_3} + \beta_{D_4} = 0$$
$$\mathrm{H}_1 : \text{at least one } \beta\text{'s in } \mathrm{H}_0 \text{ is nonzero}$$

TABLE 10: Wald test for model M7.1 and model M7.3

| Source of variations | Sum of Squares | *df* | Mean Sum of Squares | F |
|---|---|---|---|---|
| **Differences (U - R)** | 0.2457 | 2 | 0.1229 | 1.7213 |
| **Unrestricted (U)** | 223.5807 | 3133 | 0.0714 | |
| **Restricted (R)** | 223.8264 | 3135 | | |

Based on Table 10, the decision was to accept the null hypothesis where all regression coefficients of variables $D_3$ and $D_4$ are zero since the $F_{cal}$ (=1.7213) is less than $F_{critical}$ ($F_{0.05,\ 2,\ 3133}$=3.0000). Thus, this justified the removal of the insignificant variables $D_3$ and $D_4$ in the elimination procedure. Therefore, the selected model is M7.3 that is $\hat{Y} = -10.2550 + 0.0693X_1 + 0.0720X_2 + 0.0109X_3 - 0.6315D_2 + 0.6386D_5$.

A similar procedure and tests (Global test, Coefficient test and Wald test) were carried out for the remaining possible models. The summaries of selected models are shown in Appendix B. There are 12 selected models obtained in this Phase 2 (the selected models obtained, each with reduced number of parameters).

*Best Model*

The modified eight model selection criterion (M8SC) values for each selected model were obtained and the corresponding values are shown in Table 11.

TABLE 11: The corresponding selection criteria value for the selected models

| MODEL | k+1 | AIC | FPE | GCV | HQ | RICE | SCHWARZ | SGMASQ | SHIBATA |
|---|---|---|---|---|---|---|---|---|---|
| M1.1 | 5 | 0.5365 | 0.5365 | 0.5365 | 0.5383 | 0.5365 | 0.5417 | 0.5356 | 0.5365 |
| M2.1 | 5 | 0.5453 | 0.5453 | 0.5453 | 0.5472 | 0.5453 | 0.5506 | 0.5445 | 0.5453 |
| M3.1 | 5 | 0.5287 | 0.5287 | 0.5287 | 0.5305 | 0.5287 | 0.5338 | 0.5278 | 0.5287 |
| M4.1 | 6 | 0.5346 | 0.5346 | 0.5346 | 0.5368 | 0.5346 | 0.5408 | 0.5336 | 0.5346 |
| M5.1 | 6 | 0.5191 | 0.5191 | 0.5191 | 0.5213 | 0.5191 | 0.5251 | 0.5181 | 0.5191 |
| M6.1 | 6 | 0.5277 | 0.5277 | 0.5277 | 0.5299 | 0.5277 | 0.5338 | 0.5267 | 0.5277 |
| M7.3 | 6 | 0.5182 | 0.5182 | 0.5182 | 0.5203 | 0.5182 | 0.5242 | 0.5172 | 0.5182 |
| M8.7 | 9 | 0.5298 | 0.5298 | 0.5298 | 0.5331 | 0.5298 | 0.5390 | 0.5282 | 0.5298 |
| M9.9 | 7 | 0.5170 | 0.5170 | 0.5170 | 0.5195 | 0.5170 | 0.5241 | 0.5159 | 0.5170 |
| M10.8 | 8 | 0.5240 | 0.5240 | 0.5240 | 0.5269 | 0.5240 | 0.5322 | 0.5227 | 0.5240 |
| M11.16 | 11 | 0.5111 | 0.5111 | 0.5112 | 0.5151 | 0.5112 | 0.5221 | 0.5094 | 0.5111 |
| M12.16 | 12 | 0.5118 | 0.5118 | 0.5118 | 0.5161 | 0.5119 | 0.5238 | 0.5099 | 0.5118 |

All the criteria shown in Table 11 indicate that the model M11.16 as the Best Model. The result of coefficient test showed that the entire coefficients in the best model M11.16 are significant (all the $t_{cal}$ value for each variable in the model M11.16 is greater than 1.96). Thus the best model M11.16 is

$$\hat{Y} = -9.2030 + 0.0190X_3 + 0.0036X_{12} - 3.84x10^{-4}X_{23} + 3.5378D_4 + 0.6633D_5 - 0.0539X_1D_2 - 0.0634X_1D_3 + 0.1259X_2D_3 - 0.1263X_2D_4 + 0.0082X_3D_2. \quad (9)$$

The $|t_{cal}|$ for variables $X_1$, $X_2$, $X_{13}$, $D_1$, $D_2$, $D_3$, $X_1D_1$, $X_1D_4$, $X_1D_5$, $X_2D_1$, $X_2D_2$, $X_2D_5$, $X_3D_1$, $X_3D_3$, $X_3D_4$, $X_3D_5$ increased through the elimination steps but the $|t_{cal}|$ value is still less than 1.960. This shows that these variables are insignificant variables and were eliminated from the model. For the issue in Deviance statistics, in general the value increases as the elimination of variables is carried out. This shows that the insignificant variables are absorbed into Deviance statistics of the later model. The Table 12 shows the illustration of elimination procedure in getting a Selected Model (model M11.16) and because of space constraint, just a part of the illustration of elimination procedure is shown in Table 12.

TABLE 12: Illustration of Elimination Procedure in getting Best Model (model M11.16)

| **Variables** | **Models** | | | | | |
|---|---|---|---|---|---|---|
| | **M11** | **M11.1** | **...** | **M11.14** | **M11.15** | **M11.16** |
| Constant | -12.7159 | -12.7159 | ... | -13.4737 | -12.9191 | -9.2030 |
| $X_3$ | 0.0482<br>(2.3948) | 0.0482<br>(2.3948) | ... | 0.0488<br>(2.8466) | 0.0495<br>(2.8933) | 0.0190<br>(4.1810) |
| $X_{12}$ | 0.0073<br>(1.5192) | 0.0073<br>(1.5192) | ... | 0.0071<br>(4.0891) | 0.0066<br>(3.8706) | 0.0036<br>(7.2019) |
| $X_{23}$ | -8.89 x$10^{-4}$<br>(-1.6294) | -8.89 x$10^{-4}$<br>(-1.6294) | ... | -8.41 x$10^{-4}$<br>(-2.3326) | -9.58 x$10^{-4}$<br>(-2.7230) | -3.84 x$10^{-4}$<br>(-2.2985) |
| $D_4$ | 4.4541<br>(1.7624) | 4.4541<br>(1.7624) | ... | 4.0771<br>(2.7071) | 3.8953<br>(2.5890) | 3.5378<br>(2.3433) |
| $D_5$ | 2.9009<br>(1.4502) | 2.9009<br>(1.4502) | ... | 2.7642<br>(2.0554) | 0.6558<br>(4.6728) | 0.6633<br>(4.7267) |
| $X_1D_2$ | -0.0506<br>(-1.4784) | -0.0506<br>(-1.4784) | ... | -0.0446<br>(-2.5803) | -0.0445<br>(-2.5798) | -0.0539<br>(-3.1741) |
| $X_1D_3$ | -0.0539<br>(-1.5431) | -0.0539<br>(-1.5431) | ... | -0.0603<br>(-2.4236) | -0.0604<br>(-2.4210) | -0.0634<br>(-2.5916) |
| $X_2D_3$ | 0.1682<br>(1.0680) | 0.1682<br>(1.9919) | ... | 0.1197<br>(2.5689) | 0.1197<br>(2.5605) | 0.1259<br>(2.7461) |
| $X_2D_4$ | -0.1213<br>(-1.8701) | -0.1213<br>(-1.8701) | ... | -0.1473<br>(-2.5067) | -0.1403<br>(-2.3885) | -0.1263<br>(-2.1412) |
| $X_3D_2$ | 0.0049<br>(1.1883) | 0.0049<br>(1.1883) | ... | 0.0064<br>(1.9876) | 0.0064<br>(1.9866) | 0.0082<br>(2.6109) |
| $X_{13}$ | -3.19x$10^{-4}$<br>(-1.2467) | -3.19 x$10^{-4}$<br>(-1.2467) | ... | -3.77 x$10^{-4}$<br>(-2.0835) | -3.31 x$10^{-4}$<br>**(-1.8462)** | - |
| $X_2D_5$ | -0.0856<br>(-1.5707) | -0.0856<br>(-1.5707) | ... | -0.0837<br>**(-1.5782)** | - | - |
| $X_1D_5$ | -0.0066<br>(-0.2640) | -0.0066<br>(-0.2640) | ... | - | - | - |
| $X_2D_2$ | 0.0828<br>(1.0680) | 0.0828<br>(1.0680) | ... | - | - | - |

TABLE 12 (continued) : Illustration of Elimination Procedure in getting Best Model (model M11.16)

| **Variables** | **Models** | | | | | |
|---|---|---|---|---|---|---|
| | **M11** | **M11.1** | **...** | **M11.14** | **M11.15** | **M11.16** |
| $D_2$ | -1.3642 (-0.5139) | -1.3642 (-0.5139) | **...** | - | - | - |
| $X_3D_3$ | -0.0045 (-1.0017) | -0.0045 (-1.0017) | **...** | - | - | - |
| $X_3D_4$ | -0.0028 (-0.7211) | -0.0028 (-0.7211) | **...** | - | - | - |
| $X_3D_5$ | 0.0010 (0.3067) | 0.0010 (0.3067) | **...** | - | - | - |
| $X_1D_4$ | -0.0067 (-0.2118) | -0.0067 (-0.2118) | **...** | - | - | - |
| $D_3$ | -0.4268 (-0.1354) | -0.4268 (-0.1354) | **...** | - | - | - |
| $X_1$ | -0.0134 (-0.0954) | -0.0134 (-0.0954) | **...** | - | - | - |
| $X_2$ | -0.0220 (-0.0789) | -0.0220 (-0.0789) | **...** | - | - | - |
| $X_3D_1$ | -0.0399 (0.0000) | -0.0413 (0.0000) | **...** | - | - | - |
| $X_2D_1$ | -0.0382 (0.0000) | -0.0886 (0.0000) | **...** | - | - | - |
| $X_1D_1$ | -0.1062 (0.0000) | -0.3961 **(0.0000)** | **...** | - | - | - |
| $D_1$ | -13.6057 **(0.0000)** | - | **...** | - | - | - |
| ELIMINATION STEPS | | STEP1 | **...** | STEP14 | STEP15 | STEP16 |
| | | | | | | |
| **SSE** | 219.4469 | 219.4469 | **...** | 219.8696 | 220.2922 | 220.5815 |
| $\mathbf{G^2}$ | 1585.3740 | 1585.3740 | **...** | 1589.0096 | 1591.5108 | 1594.8262 |
| $^*\mathbf{t_{critical}}$ = 1.96 and value in parentheses is the $t_{cal}$ | | | | | | |

*Goodness-of-fit Test*

The following phase is to check the validity of the best model. There are two tests on the goodness-of-fit. The tests are carried out based on the residuals obtained from the best model M11.16. The hypothesis for Pearson Chi-Square test for best model M11.16 is as follows:

$$H_0 : E[Y] = [1 + \exp(-W^T \Omega)]^{-1}$$
$$H_1 : E[Y] \neq [1 + \exp(-W^T \Omega)]^{-1}$$

The sum of squares of the Pearson residual is $\chi_{r_{pi}}{}^2 = 3061.3122$ and the $\chi^2_{critical}$ is $\chi^2_{0.95,3131} = 3174.1663$. Since $\chi_{r_{pi}}{}^2$ is less than $\chi^2_{critical}$, the decision is to accept the null hypothesis where the best model M11.16 is an appropriate or correct model. The hypothesis for Deviance goodness-of-fit test is

$$H_0 : E[Y] = [1 + \exp(-W^T\Omega)]^{-1}$$
$$H_1 : E[Y] \neq [1 + \exp(-W^T\Omega)]^{-1}$$

The sum of squares of the Deviance residual (Deviance statistics) is $G^2 = 1594.8261$ and the $\chi^2_{critical}$ is $\chi^2_{0.95,3131} = 3174.1663$. The decision is to accept the null hypothesis since $G^2$ is less than $\chi^2_{critical}$. Therefore, the best model M11.16 is an appropriate model. The plot Figures 1, 2 and 3 show the distribution of ordinary residuals, Pearson residuals and Deviance residuals for best model M11.16 respectively.

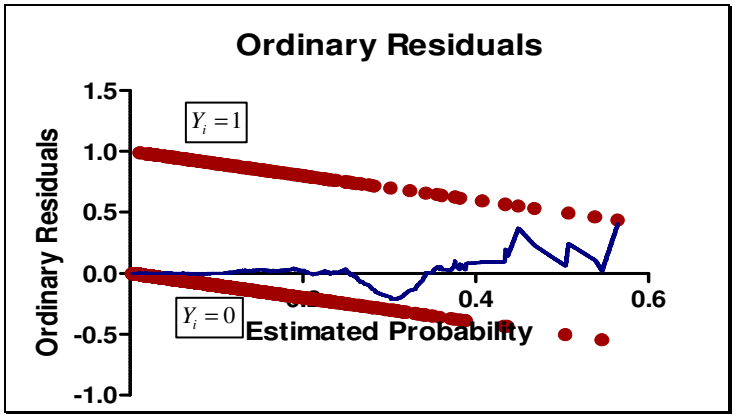


Figure 1: Ordinary Residuals for best model M11.16

In Figure 1, the ordinary residuals are plotted against the estimated probability. Here two trends of decreasing residuals with slope (for both lines) equal to -1 and form a parallel line.

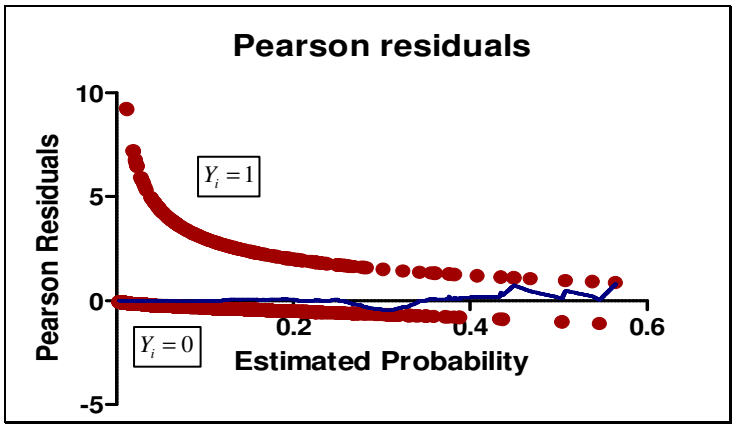


Figure 2: Pearson Residuals for best model M11.16

The Figure 2 shows that the trend of the Pearson residuals is plotted against the estimated probability with slope equal to -1.

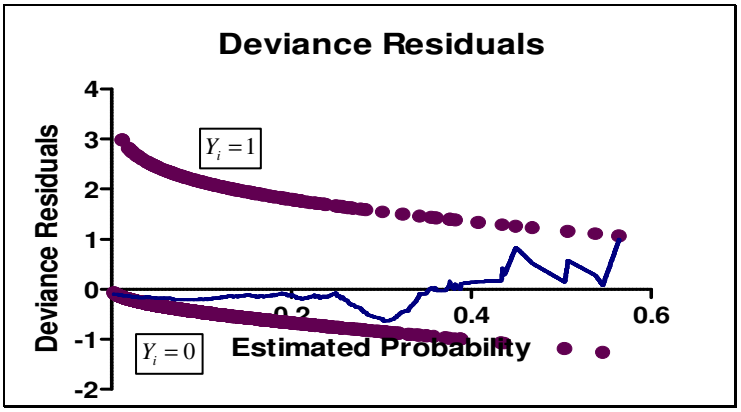


Figure 3: Deviance Residuals for best model M11.16

The Deviance residuals are plotted against the estimated probability. Here two trends of decreasing residuals with slope (for both lines) equal to -1 and form a parallel line. These three plots are used as supporting evidence to support the goodness-of-fit test. These three plots suggest that the model is correct where the trend of the plot of the residuals against the estimated probability ($= \hat{p}_i$) appear approximately in a horizontal line with zero intercept.

## DISCUSSION AND CONCLUSION

This numerical illustration example is an experiment of case study (WCGS data) where the variables weight and height were modified to be body mass index (BMI). The second modification was on the systolic and diastolic blood pressures. These two variables was joined together to form a four category dummy variable. This enables to identify which category has contributed significantly in determining the occurrence of Coronary Heart Disease. The best model obtained is model M11.16 as listed in equation (9). Thus, this best model M11.16 is ready for further analysis.

TABLE 13: Description on Best model 11.16

| **Variables** | **Coefficients** | **Comments** |
|---|---|---|
| $X_3$ | 0.0190 | -Cholesterol level<br>-Main factor |
| $D_4$ | 3.5378 | -Blood pressure category (1 if SBP > 140 and DBP >90, or 0 if otherwise)<br>-Main factor |
| $D_5$ | 0.6633 | -Smoking habit<br>-Main factor |
| $X_{12}$ | 0.0036 | -Age ($X_1$) and BMI ($X_2$)<br>-First-order interaction factor |
| $X_{23}$ | $-3.84\text{x}10^{-4}$ | -BMI ($X_2$) and Cholesterol level ($X_3$)<br>-First-order interaction factor |
| $X_1D_2$ | -0.0539 | -Age ($X_1$) and Blood pressure category (1 if SBP 100-129 and DBP 60-79, 0 if otherwise)<br>-First-order interaction factor |
| $X_1D_3$ | -0.0634 | -Age ($X_1$) and Blood pressure category (1 if SBP 130-139 and DBP 80-89, 0 if otherwise)<br>-First-order interaction factor |
| $X_2D_3$ | 0.1259 | -BMI ($X_2$) and Blood pressure category (1 if SBP 130-139 and DBP 80-89, 0 if otherwise)<br>-First-order interaction factor |
| $X_2D_4$ | -0.1263 | -BMI ($X_2$) and Blood pressure category (1 if SBP > 140 and DBP >90, 0 if otherwise)<br>-First-order interaction factor |
| $X_3D_2$ | 0.0082 | -Cholesterol level ($X_3$) and Blood pressure category (1 if SBP 100-129 and DBP 60-79, 0 if otherwise)<br>-First-order interaction factor |

As can be seen from the best model (in Table 13), the cholesterol level ($X_3$), blood pressure category, $D_4$ (1 if SBP > 140 and DBP > 90, or 0 if otherwise) and smoking/non smoking category ($D_5$) are the main factors that contribute to the occurrence of Coronary Heart Disease. The physical factors age ($X_1$) and BMI ($X_2$) interact together to indicate the strength of contribution in determining the occurrence of Coronary Heart Disease. The other blood pressure categories interact with factor age ($X_1$), BMI ($X_2$) and cholesterol level ($X_3$) to show the effect in the occurrence of Coronary Heart Disease.

## REFERENCES

Akaike, H. 1969. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, **21:** 243-247.

Akaike, H. 1974. A New Look at Statistical Model Identification. *IEEE Transactions Automatic Control AC* **19**: 716-723.

Crawley, M.J. 2005. *Statistics: An Introduction using R.* New Jersey: John Wiley & Sons.

Golub, G.H., Heath, M. and Wahba, G. 1979. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics,* **21**: 215-223.

Halcoussis, D. 2005. *Understanding Econometrics.* New York: South-Western.

Hannan, E.J. & Quinn, B. 1979. The Determination of the Order of an Autoregression. *J. Royal Stat. Society*, **41**(Series B): 190-195.

Kutner, M.H, Nachtsheim, C.J and Neter, J. 2008. *Applied Linear Regression Models.* (4th edition). Singapore: McGraw-Hill, Inc.

Ramanathan, R. 2002. *Introductory Econometrics with Application*. 5th edition. Ohio: South Western, Thomson Learning Ohio.

Rice, J. 1984. Bandwidth Choice for Nonparametric Kernel Regression. *Annals of Statistics,* **12**: 1215-1230.

Rosado, J.F.C, Solis, C.E.M., Sanchez, A.A.V., Rosado, A.J.C., Prado, B.H. and Burgos, L.A. 2006. Prevalence and associated factors for temporomandibular disorders in a group of Mexican adolescents and youth adults. *Clin Oral Invest*, **10**: 42-49.

Schwarz, G. 1978. Estimating the Dimension of a Model. *Annals of Statistics,* **6**: 461-464.

Shibata, R. 1981. An Optimal Selection of Regression Variables. *Biometrika,* **68**: 45-54.

Studenmund, A.H. 2006. *Using Economics: A Practical Guide*. 5th edition. New York: Addison-Wesley.

Vittinghooff, E., Glidden, D.V., Shiboski, S.C. and McCulloch, C.E. 2004. *Regression Methods in Biostatistics Linear, Logistic, Survival and Repeated Measures Models.* New York: Springer-Verlag.

Vogelvang, B. 2005. *Econometrics: Theory and Applications with EViews*. New York: Addison-Wesley.

WHO. 2003. World Health Organization (WHO) / International Society of Hypertension (ISH) statement on management of hypertension. *Journal of Hypertension,* **21**(11): 1983-1992.

Zainodin, H.J. and Khuneswari, G. 2007. Justification of Omitting Independent Variables in Selecting Best Model. *Proceedings of International Applied Science and Mathematics Conference,* 343-356.

Zainodin, H.J. and Khuneswari, G. 2009. A Case Study on determination of House Selling Price Model using Multiple Regression. *Malaysian Journal of Mathematical Sciences,* **3**(1): 27-44.

## APPENDIX A (All Possible Models)

M1: $Y = \beta_0 + \beta_1 X_1 + (\beta_{D_1} D_1 + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{D_4} D_4 + \beta_{D_5} D_5) + u$

M2: $Y = \beta_0 + \beta_2 X_2 + (\beta_{D_1} D_1 + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{D_4} D_4 + \beta_{D_5} D_5) + u$

M3: $Y = \beta_0 + \beta_3 X_3 + (\beta_{D_1} D_1 + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{D_4} D_4 + \beta_{D_5} D_5) + u$

M4: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + (\beta_{D_1} D_1 + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{D_4} D_4 + \beta_{D_5} D_5) + u$

M5: $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + (\beta_{D_1} D_1 + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{D_4} D_4 + \beta_{D_5} D_5) + u$

M6: $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + (\beta_{D_1} D_1 + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{D_4} D_4 + \beta_{D_5} D_5) + u$

M7: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$
$+(\beta_{D_1} D_1 + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{D_4} D_4 + \beta_{D_5} D_5) + u$

M8: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_{12}$
$+(\beta_{D_1} D_1 + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{D_4} D_4 + \beta_{D_5} D_5)$
$+(\beta_{1D_1} X_1 D_1 + \beta_{1D_2} X_1 D_2 + \beta_{1D_3} X_1 D_3 + \beta_{1D_4} X_1 D_4 + \beta_{1D_5} X_1 D_5)$
$+(\beta_{2D_1} X_2 D_1 + \beta_{2D_2} X_2 D_2 + \beta_{2D_3} X_2 D_3 + \beta_{2D_4} X_2 D_4 + \beta_{2D_5} X_2 D_5 + u$

M9: $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_{13} X_{13}$
$+(\beta_{D_1} D_1 + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{D_4} D_4 + \beta_{D_5} D_5)$
$+(\beta_{1D_1} X_1 D_1 + \beta_{1D_2} X_1 D_2 + \beta_{1D_3} X_1 D_3 + \beta_{1D_4} X_1 D_4 + \beta_{1D_5} X_1 D_5)$
$+(\beta_{3D_1} X_3 D_1 + \beta_{3D_2} X_3 D_2 + \beta_{3D_3} X_3 D_3 + \beta_{3D_4} X_3 D_4 + \beta_{3D_5} X_3 D_5) + u$

M10: $Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 + \beta_{23} X_{23}$
$+(\beta_{D_1} D_1 + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{D_4} D_4 + \beta_{D_5} D_5)$
$+(\beta_{2D_1} X_2 D_1 + \beta_{2D_2} X_2 D_2 + \beta_{2D_3} X_2 D_3 + \beta_{2D_4} X_2 D_4 + \beta_{2D_5} X_2 D_5)$
$+(\beta_{3D_1} X_3 D_1 + \beta_{3D_2} X_3 D_2 + \beta_{3D_3} X_3 D_3 + \beta_{3D_4} X_3 D_4 + \beta_{3D_5} X_3 D_5) + u$

M11: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_{12} + \beta_{13} X_{13} + \beta_{23} X_{23}$
$+(\beta_{D_1} D_1 + \beta_{D_2} D_2 + \beta_{D_3} D_3 + \beta_{D_4} D_4 + \beta_{D_5} D_5)$
$+(\beta_{1D_1} X_1 D_1 + \beta_{1D_2} X_1 D_2 + \beta_{1D_3} X_1 D_3 + \beta_{1D_4} X_1 D_4 + \beta_{1D_5} X_1 D_5)$

$$+(\beta_{2D_1}X_2D_1+\beta_{2D_2}X_2D_2+\beta_{2D_3}X_2D_3+\beta_{2D_4}X_2D_4+\beta_{2D_5}X_2D_5)$$
$$+(\beta_{3D_1}X_3D_1+\beta_{3D_2}X_3D_2+\beta_{3D_3}X_3D_3+\beta_{3D_4}X_3D_4+\beta_{3D_5}X_3D_5)+u$$

M12: $Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_{12}X_{12}+\beta_{13}X_{13}+\beta_{23}X_{23}+\beta_{123}X_{123}$
$$+(\beta_{D_1}D_1+\beta_{D_2}D_2+\beta_{D_3}D_3+\beta_{D_4}D_4+\beta_{D_5}D_5)$$
$$+(\beta_{1D_1}X_1D_1+\beta_{1D_2}X_1D_2+\beta_{1D_3}X_1D_3+\beta_{1D_4}X_1D_4+\beta_{1D_5}X_1D_5)$$
$$+(\beta_{2D_1}X_2D_1+\beta_{2D_2}X_2D_2+\beta_{2D_3}X_2D_3+\beta_{2D_4}X_2D_4+\beta_{2D_5}X_2D_5)$$
$$+(\beta_{3D_1}X_3D_1+\beta_{3D_2}X_3D_2+\beta_{3D_3}X_3D_3+\beta_{3D_4}X_3D_4+\beta_{3D_5}X_3D_5)+u$$

## APPENDIX B (Selected Models)

| Selected Models | Summary | k+1 | SSE | $G^2$ |
|---|---|---|---|---|
| **M1.2** | $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_{D_2}D_2+\hat{\beta}_{D_4}D_4+\hat{\beta}_{D_5}D_5$ | 5 | 227.7373 | 1680.2742 |
| **M2.2** | $\hat{Y}=\hat{\beta}_0+\hat{\beta}_2X_2+\hat{\hat{\beta}}_{D_2}D_2+\hat{\beta}_{D_4}D_4+\hat{\beta}_{D_5}D_5$ | 5 | 230.7942 | 1707.9998 |
| **M3.2** | $\hat{Y}=\hat{\beta}_0+\hat{\beta}_3X_3+\hat{\beta}_{D_2}D_2+\hat{\beta}_{D_4}D_4+\hat{\beta}_{D_5}D_5$ | 5 | 226.2680 | 1655.7950 |
| **M4.2** | $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_2+\hat{\beta}_{D_2}D_2+\hat{\beta}_{D_4}D_4+\hat{\beta}_{D_5}D_5$ | 6 | 227.4519 | 1673.2856 |
| **M5.2** | $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_3X_3+\hat{\beta}_{D_2}D_2+\hat{\beta}_{D_4}D_4+\hat{\beta}_{D_5}D_5$ | 6 | 223.5564 | 1624.7742 |
| **M6.2** | $\hat{Y}=\hat{\beta}_0+\hat{\beta}_2X_2+\hat{\beta}_3X_3+\hat{\beta}_{D_2}D_2+\hat{\beta}_{D_4}D_4+\hat{\beta}_{D_5}D_5$ | 6 | 226.5131 | 1651.6374 |
| **M7.3** | $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_2X_2+\hat{\beta}_3X_3+\hat{\beta}_{D_2}D_2+\hat{\beta}_{D_5}D_5$ | 6 | 223.8264 | 1621.8572 |
| **M8.10** | $\hat{Y}=\hat{\beta}_0+\hat{\beta}_{12}X_{12}+\hat{\beta}_{D_4}D_4+\hat{\beta}_{D_5}D_5+\hat{\beta}_{1D_2}X_1D_2+\hat{\beta}_{1D_3}X_1D_3+\hat{\beta}_{2D_3}X_2D_3+\hat{\beta}_{2D_4}X_2D_4+\hat{\beta}_{2D_5}X_2D_5$ | 9 | 225.5515 | 1655.0016 |
| **M9.12** | $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X_1+\hat{\beta}_3X_3+\hat{\beta}_{D_4}D_4+\hat{\beta}_{1D_2}X_1D_2+\hat{\beta}_{3D_2}X_3D_2+\hat{\beta}_{3D_5}X_3D_5$ | 7 | 222.6976 | 1617.3000 |
| **M10.11** | $\hat{Y}=\hat{\beta}_0+\hat{\beta}_2X_2+\hat{\beta}_3X_3+\hat{\beta}_{D_2}D_2+\hat{\beta}_{D_4}D_4+\hat{\beta}_{D_5}D_5+\hat{\beta}_{2D_2}X_2D_2+\hat{\beta}_{2D_4}X_2D_4$ | 8 | 224.9670 | 1638.1478 |

## APPENDIX B (Selected Models) (continued)

| Selected Models | Summary | k+1 | SSE | $G^2$ |
|---|---|---|---|---|
| **M11.16** | $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_3 X_3 + \hat{\beta}_{12} X_{12} + \hat{\beta}_{23} X_{23} + \hat{\beta}_{D_4} D_4 + \hat{\beta}_{D_5} D_5 + \hat{\beta}_{1D_2} X_1 D_2 + \hat{\beta}_{1D_3} X_1 D_3 + \hat{\beta}_{2D_3} X_2 D_3 + \hat{\beta}_{2D_4} X_2 D_4 + \hat{\beta}_{3D_2} X_3 D_2$ | 11 | 220.5815 | 1594.8262 |
| **M12.16** | $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_3 X_3 + \hat{\beta}_{12} X_{12} + \hat{\beta}_{123} X_{123} + \hat{\beta}_{D_4} D_4 + \hat{\beta}_{D_5} D_5 + \hat{\beta}_{1D_2} X_1 D_2 + \hat{\beta}_{1D_3} X_1 D_3 + \hat{\beta}_{2D_3} X_2 D_3 + \hat{\beta}_{2D_4} X_2 D_4 + \hat{\beta}_{2D_5} X_2 D_5 + \hat{\beta}_{3D_2} X_3 D_2$ | 12 | 219.9174 | 1589.8590 |