# CHAPTER 6

## Regression Analysis

### 6.1    Introduction

In the preceding Chapters, the factors that have significant association associated with the academic performance of students in the three subjects at the SPM examination were analyzed. This chapter first presents, the results of the Multiple Regression and then followed by the results of the Logistic regression. Multiple Regressions models is used to determine the important factors that significantly affect the performance of students in each of the three subjects, with the subject grades as dependent variables, while Logistic regression technique is used to estimate the effects of the independent variables on the likelihood of being high achievers in each of the three subjects. The selected independent variables that are used for these analyses are shown in Table 6.1.

### 6.2    Determinants Of Academic Performance in Selected Subjects

Stepwise regression was undertaken on demographic, home environment and educational background variables and the contribution of each significant variable towards the total variance in the specific subject performance was obtained. As most explanatory variable were not measured in interval scale, they were recoded into dummy variables for inclusion in the regression model. The relevant reference categories are shown in Table 6.1.

The estimated multiple regression model is as follows:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k + \alpha_1 D_1 + ... + \alpha_n D_n$$

where $\hat{Y}$ = estimated grade for a particular subject.

$X_1$ = represent family size

$X_2$ = represent subject grades for SRP examination

$D_1...D_7$ represents the dummy variables shown below (gender...ethnicity)

**Table 6.1: Independent Variables and the Corresponding Reference Categories**

| Independent Variables | Category |
|---|---|
| Gender | Female<br>Male (R) |
| Father's Education level | No Primary schooling<br>Primary<br>Lower secondary<br>Upper secondary<br>College/University (R) |
| Childhood Residence | Rural<br>Small Town<br>Large town (R) |
| Medium of instruction | Malay<br>Chinese<br>Others (R) |
| Ethnicity | Bumiputera<br>Chinese<br>Indians(R) |
| Stream | Science<br>Arts (R) |
| Tuition | Yes<br>No (R) |
| SRP grades | - |
| Family size | - |

(R) Indicate Reference Category

Detailed results of regression analysis are presented in Table 6.2, 6.3 and 6.4 for Bahasa Malaysia, English and Mathematics grades respectively. It is observed that the percentage of variation explained by the significant explanatory variables for Bahasa Malaysia is relatively low as compared to the other two subjects.

## 6.3    Results and Discussion

The following sections discuss the results displayed in Tables 6.2, 6.3 and 6.4. These results are for the models where the dependent variable is grades of Bahasa Malaysia, English and Mathematics obtained as the SPM examination respectively. Only the significant (p-values <0.01) independent variables are shown in each of the three tables.

### 6.3 .1   Regression Analysis of Performance in Bahasa Malaysia

Table 6.2 contains the regression coefficients and the associated R- square value, F statistics and the p value. The dependent variable is the grades obtained in Bahasa Malaysia at the SPM examination.

The variables that significantly affect the performance of student in this subject are student's ethnicity, gender, and prior achievement in SRP Bahasa Malaysia. These predictor variables only explained some 16.5 percent of the variance in Bahasa Malaysia grades. The estimated regression model is highly significant with F-statistics of 99.0 (p- value = 0.000).

**Table 6.2: Regression Results of Bahasa Malaysia grades.**

| Dependent variable | : | SPM Bahasa Malaysia | | | |
|---|---|---|---|---|---|
| R | : | 0.409 | | | |
| $R^2$ | : | 0.167 | | | |
| Adjusted $R^2$ | : | 0.165 | | | |
| Standard Error of the Estimate | : | 1.38 | | | |

**ANOVA**

| | Sum of squares | DF | Mean square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 947.074 | 5 | 189.415 | 99.056 | 0.000 |
| Residual | 4726.016 | 2471 | 1.912 | | |
| Total | 5673.090 | 2476 | | | |

**Coefficients**

| Variables | B | Std. error | t | Sig. |
|---|---|---|---|---|
| (Constant) | 3.030 | 0.129 | 23.493 | 0.000 |
| Performance in SRP - Bahasa Malaysia | 0.287 | 0.018 | 16.372 | 0.000 |
| Medium-Chinese | 0.439 | 0.100 | 4.399 | 0.000 |
| Gender-Female | -0.196 | 0.58 | -3.412 | 0.001 |
| Ethnicity-Malay | -0.417 | 0.106 | -3.917 | 0.000 |
| Ethnicity- Chinese | -0.358 | 0.126 | -2.850 | 0.004 |

The estimated model indicates that by holding all other variables constant, the SPM Bahasa Malaysia grades of female students will be approximately 0.2 points lower than male students. In other words, performance of female students was better than the male students in Bahasa Malaysia.

The model indicates that the performance of Bumiputera students in Bahasa Malaysia is much better than the Chinese and Indians students and this was observed in Chapter 5 as well. The mean grade and Figure 5.1 in Chapter 5 shows that most Bumiputera students achieved a strong credit (C3) in Bahasa Malaysia compared to a mean grade of approximately 4 for the two ethnic groups. Performance of students in SRP is directly related to the performance in SPM examination for Bahasa Malaysia. Students who did well in the former examination are most likely do well in the SPM examinations. Father's education level, childhood residence, and family size do not have significant effect on student performance in Bahasa Malaysia.

## 6.3.2 Regression Analysis of Performance in English

The relatively high adjusted R- square value of 0.840 in Table 6.3, indicates the explanatory power of the significant variables included in the regression model with the English grades as the dependent variable after adjusting for the number of independent variables and the sample size. The student's ethnicity, childhood residence, father's education level, family size, stream of study, tuition and the prior achievement in SRP English explain up to 84.0 percent of the variance in the English results at the SPM examinations.

The estimated regression model presented in Table 6.4 is highly significant as the F-statistics is large (1423.915) with a small p-value (p<0.001). This implies that there is a significant relationship between SPM English grades and the set of independent variables.

The coefficient of 1.299, indicates that holding all other variables constant, the performance of Bumiputera students in SPM English will be much lower than Indian students. Chinese students too, did not perform as well as the Indian students. It was observed in Chapter 5 that the Indian students had the most outstanding performance in English.  Mean grade of Indian students indicates a distinction, while the mean grade of the Chinese and the Bumiputera students were in the C4 and above P7 categories.

Student form rural areas performed poorly as compared to the students from small and large towns. Students' whose fathers' have upper secondary education have performed slightly better than the students whose fathers' have tertiary education. This could due to the high percentage Bumiputera  students, who indicated that their fathers' obtained tertiary education. The negative value of 0.233 in the equation indicates that generally, the science students have better command of English Language than the Arts students.

## 6.3: Regression Results of SPM English Grades.

| ndent variable | : | SPM English | | | |
|---|---|---|---|---|---|
| | : | 0.917 | | | |
| | : | 0.841 | | | |
| sted $R^2$ | : | 0.840 | | | |
| lard Error of the Estimate | : | 1.14 | | | |

### )VA

| | Sum of squares | DF | Mean square | F | Sig. |
|---|---|---|---|---|---|
| ession | 16608.596 | 9 | 1845.400 | 142.3.915 | 0.000 |
| dual | 3145.620 | 2427 | 1.296 | | |
| l | 19754.216 | 2436 | | | |

### fficients

| iables | B | Std. error | t | Sig. |
|---|---|---|---|---|
| nstant) | 1.092 | 0.89 | 12.244 | 0.000 |
| formance in SRP English | 0.673 | 0.12 | 57.016 | 0.000 |
| nicity-Malay | 1.299 | 0.081 | 16.068 | 0.000 |
| eam-Science | -0.233 | 0.060 | -3.896 | 0.000 |
| mily size | 0.0257 | 0.010 | 2.532 | 0.011 |
| inese Medium | 0.979 | 0.077 | 12.711 | 0.000 |
| ither's education level: upper condary | -0.400 | 0.078 | -5.148 | 0.000 |
| hildhood Residence | | | | |
| ural | 0.751 | 0.070 | 10.801 | 0.000 |
| mall Town | 0.545 | 0.070 | 7.827 | 0.000 |

### 6.3.3 Regression Analysis of Performance in Mathematics

It can be deduced from Table 6.4 that four variables, namely the student's, ethnicity, father's education level and the stream of study significantly affect the students performance in SPM Mathematics. These four variables explain about 83 percent of the total variation in the Mathematics grades. The estimated model is highly significant, as indicated by rather large F-value and the small p- value. The Chinese students have been identified as the top performers in Mathematic from preceding chapters. The value of −0.824 further confirms the excellent performance of Chinese students in mathematics as compared to Indian students. However, the performance of Indian students is much better than the Bumiputera students. Students whose fathers' have upper secondary education have done well in Mathematics as compared to students whose fathers' have tertiary education. Similar observation was made earlier for the regression model of English grades and therefore the same reasons as mentioned earlier could be applied.

The value of −1.204 indicates a pronounced difference in the Mathematics grades of science and arts stream students, where the performance of sciences students is much better than the arts. Percentage of arts student who obtained the fail grade was extremely high (36.6 percent) as compared to only 1.2 percent of science stream students in the same category (See Table 5.9).

## Table 6.4: Regression Results of Mathematics Grades

| Dependent variable | : | SPM Mathematics | | | |
|---|---|---|---|---|---|
| R | : | 0.912 | | | |
| $R^2$ | : | 0.832 | | | |
| Adjusted $R^2$ | : | 0.832 | | | |
| Standard Error of the Estimate | : | 1.2 | | | |

**ANOVA**

| | Sum of squares | DF | Mean square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 17154.642 | 5 | 3430.928 | 2392.164 | 0.000 |
| Residual | 3452.052 | 2407 | 1.434 | | |
| Total | 20606.694 | 2412 | | | |

**Coefficients**

| Variables | B | Std. error | t | Sig. |
|---|---|---|---|---|
| (Constant) | 2.776 | 0.102 | 27.339 | 0.000 |
| Ethnicity: | | | | |
| Bumiputera | 0.574 | 0.097 | 5.911 | 0.000 |
| Chinese | -0.824 | 0.097 | -8.511 | 0.000 |
| Father's education level: | | | | |
| Upper secondary | -0.173 | 0.079 | -2.199 | 0.028 |
| Stream- science | -1.204 | 0.064 | -18.727 | 0.000 |
| Performance in SRP Mathematics | 0.739 | 0.013 | -58.016 | 0.000 |

It is evident from Tables 6.2, 6.3 and 6.4 that the number of factors that effect student performance in Bahasa Malaysia is not as may as the number of factors affecting performance of students in English or Mathematics. Although the model is significant, the Adjusted R-square of regression analysis of Bahasa Malaysia grades is only 16 per cent . This indicates that the independent variables used in the model may not sufficiently explain the grade differentials in Bahasa Malaysia. There may be other important variables that may have greater influence on performance of students in Bahasa Malaysia, which were not included in this analysis.

However the Adjusted R-square values for English and Mathematic indicates that the independent variables in the regression model sufficiently explain the grade differentials in the two subjects. Table 6.5 below displays the number of significant variables for each of the three subjects.

Table 6.5: Significant Variables in the Regression Analysis for Respective Subjects

| Bahasa Malaysia | English | Mathematics |
| --- | --- | --- |
| Prior Achievement in SRP | Prior Achievement in SRP | Prior Achievement in SRP |
| Ethnicity | Ethnicity | Ethnicity |
| Medium of Instruction | Medium of Instruction | Father's Education Level |
| Gender | Father's Education Level | Stream of Study |
| | Stream of Study | |
| | Family Size | |
| | Childhood Residence | |

There are more common variables, which, explains the difference in performance of students in English and Mathematics. This leads to a conclusion that the variables that affect the performance of student will most likely also affect the performance of students in Mathematics. Interestingly, father's income level and occupation sector were not significantly affecting the performance of students in any of the three subjects.

**Figure 6.1: Histogram and Normal Probability Plots for Distributions Of Subject Grades**

### Histogram
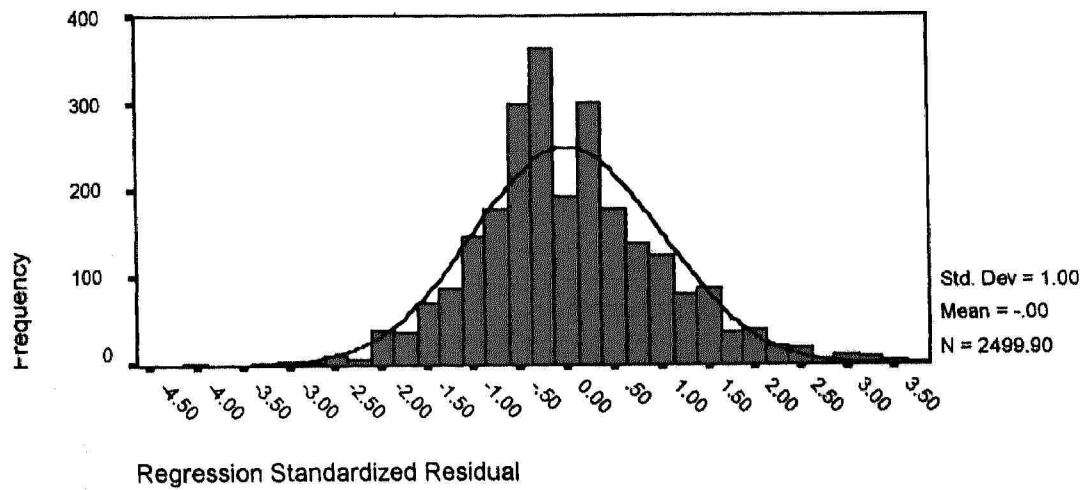
Dependent Variable: SPM BAHASA MALAYSIA
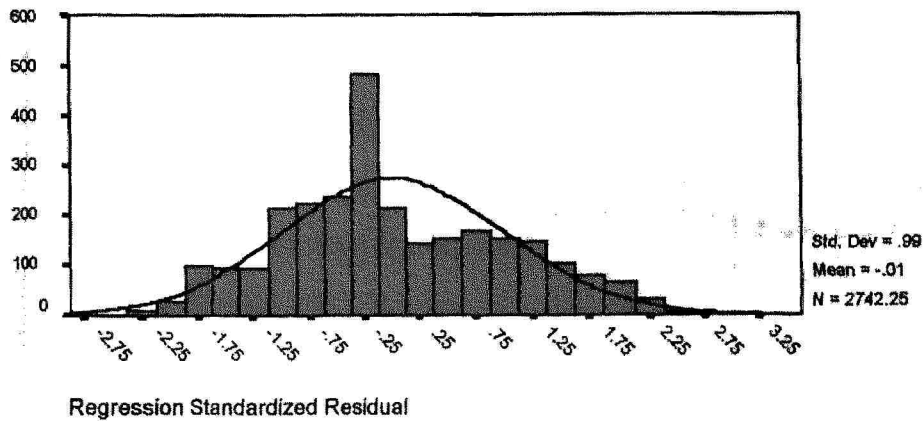


Std. Dev = .99
Mean = -.01
N = 2742.25

Regression Standardized Residual

### Normal P-P Plot of Regression Standardized Residual

Dependent Variable: SPM BAHASA MALAYSIA



Observed Cum Prob

# Histogram

## Dependent Variable: SPM ENGLISH



Std. Dev = 1.00
Mean = -.00
N = 2499.90
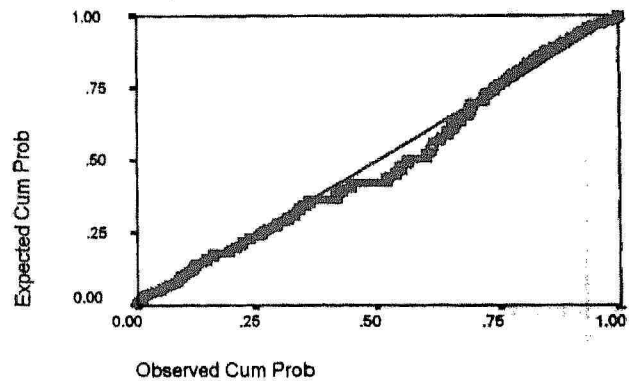
Regression Standardized Residual

# Normal P-P Plot of Regression Standardized

## Dependent Variable: SPM ENGLISH
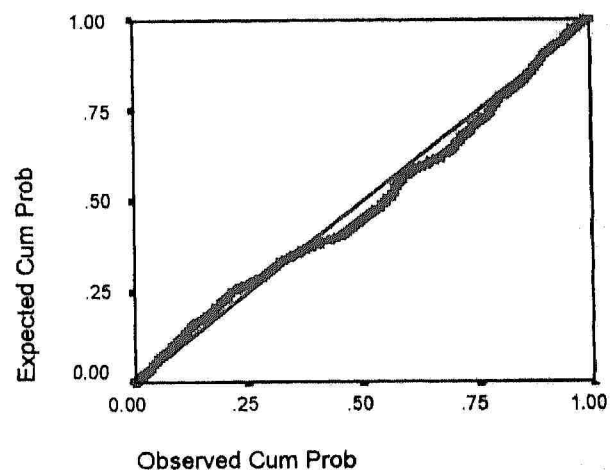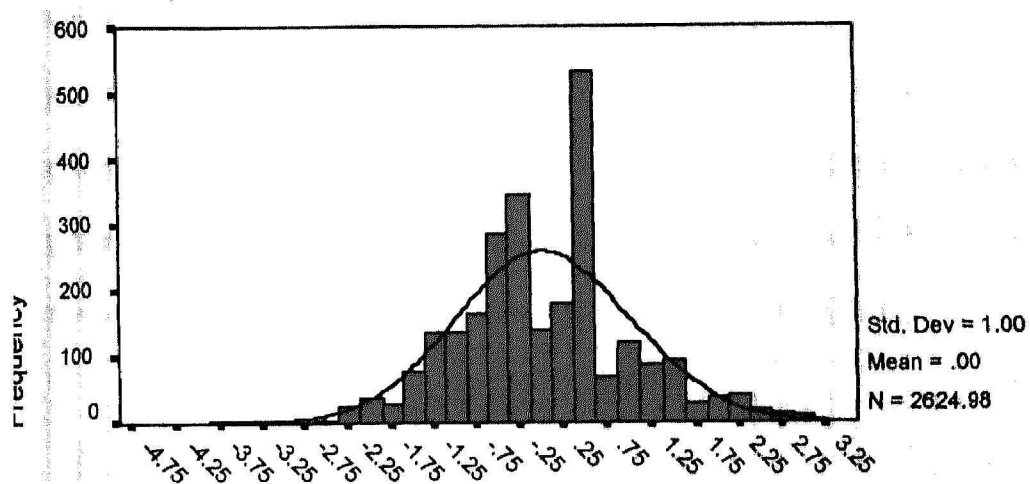


Observed Cum Prob

# Histogram

## Dependent Variable: SPM MATHEMATICS



Std. Dev = 1.00
Mean = .00
N = 2624.98

Regression Standardized Residual

## Normal P-P Plot of Regression Standardized Residual

## Dependent Variable: SPM MATHEMATICS



Observed Cum Prob

106
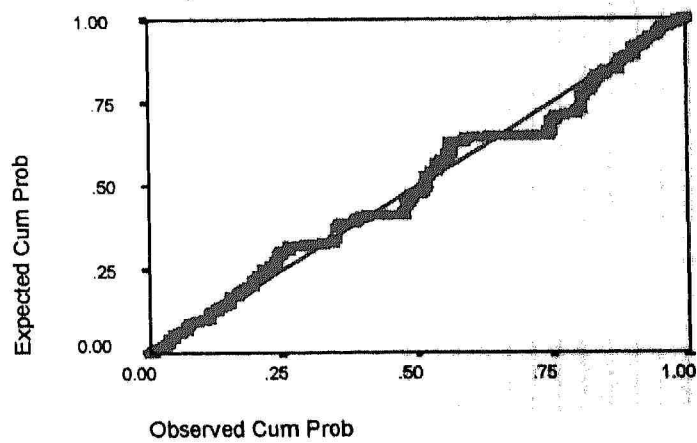
Students who have indicated that they have additional tuition did not outperform the students without tuition for any of the three subjects. The type of reading material ( newspaper, textbook, magazines) and the frequency of reading these materials were also found to be insignificant in explaining great differential in the three subjects.

The histograms and normal probability plots for the three regression models are shown in Figure 6.1. The normal distribution makes a straight diagonal line, and the plotted residuals are compared with diagonal. If a distribution is normal, the residual line closely follows the diagonal  (refer to Hair et., 1998). All the histograms and normal probability plots shown in Figure 6.1 seem to portray a distribution that is normal and therefore there is no serious violation of the normality assumptions.

## 6.4 Logistic Regression Analysis

In the previous section, the multiple regression analysis was used to identify the factors affecting students' performance in Bahasa Malaysia, English and Mathematics. This section presents the results of the logistic regression, which is carried out to determine the likelihood of students being high achievers in each of the three subjects. The dependent variable is the group of students who have performed exceptionally well in the three subjects. In this study, a student who has obtained a grade of A1, A2 and C3 in each of the subject is categorized as high achiever in that particular subject. The independent variables used in the this analysis are listed in Table 6.1. The parameters of the model are estimated using the maximum likelihood method, where the coefficients that will make the observed results most 'likely' are selected. Logistic regression estimates the odds ratio for each of the independent variables in the model.

Tables 6.6, 6.7 and 6.8 contain estimated coefficients (under column heading B) and related statistics from the logistic regression model that predicts high achievement in Bahasa Malaysia, English and Mathematics respectively. The estimates coefficient shows the effects of the independent variables on the likelihood of being a high achieving student in a specified subject. For instance, the odds of being a high achieving student in Bahasa Malaysia is the ratio of being a high achieving student in Bahasa Malaysia to those who are not high achievers (without a grade of A1, A2 or C3) in Bahasa Malaysia.

A positive coefficient indicates an increase in the log odds and consequently implies an increase in the probability of being a high achiever in the specified subject like wise, a negative coefficient indicates a decreases in the log odds of being a high achiever in the respective subject. By using indicator variables for coding the coefficients for the new variables represent the influence of each category in a variable compared to reference category.

**Table 6.6: Logistic Regression Analysis on the Likelihood of Students Being High Achievers in Bahasa Malaysia**

| Variables | Coefficients (B) | Standard Error | Significance Level | Exp (B) |
|---|---|---|---|---|
| **Ethnicity** Bumiputera Chinese | 0.9992 0.1831 | 0.1658 0.1691 | 0.0000 0.2686 | 2.7160 1.2010 |
| **Stream** | -0.1440 | 0.1064 | 0.1759 | 0.8655 |
| **Gender** | 0.4438 | 0.0856 | 0.0000 | 1.5587 |
| **Childhood Residence** Rural Small Town | -0.2465 0.0756 | 0.1184 0.1209 | 0.0374 0.5316 | 0.7815 0.9272 |
| **Father's Education level** No formal education Primary Lower Secondary Upper Secondary | -0.6217 -0.5265 -0.4238 -0.2758 | 0.2328 0.1897 0.1949 0.2094 | 0.0076 0.0055 0.0296 0.1877 | 0.5370 0.5906 0.6545 0.7589 |

Note: Refer to Table 6.1 for Reference Categories

It is observed from Table 6.6 that Bumiputra students have high chances of scoring A1, A2 and C3 in Bahasa Malaysia compare to other ethnic groups as the coefficient is 0.9992. The odds are increased by a factor of 2.7160, as shown in column in the Exp (B) column in Table 6.6. The negative coefficient of 0.1440 implies that science stream students are less likely to be high achievers in Bahasa Malaysia compared to the arts stream students.

The probability of being a high achiever in Bahasa Malaysia decreases with father's education level. Students whose fathers' are less educated are not likely to be high achievers compared to students whose fathers' have tertiary education. Female students appear to have better chances of being high achievers compared to the male students. With regards to childhood residence, as expected the rural students have the lowest probability of achieving a grade of A1or A2 or C3 in Bahasa Malysia.

**Table 6.7:Logistic Regression Analysis on the Likelihood of Students Being High Achievers in English**

| Variables | Coefficients (B) | Standard Error | Significance Level | Exp (B) |
|---|---|---|---|---|
| **Ethnicity**<br>Bumiputera<br>Chinese | -2.8072<br>-0.4566 | 0.2292<br>0.2236 | 0.0000<br>0.0412 | -0.2133<br>-0.0258 |
| **Stream** | 0.8040 | 0.1397 | 0.0000 | 0.0978 |
| **Gender** | 0.4908 | 0.1313 | 0.0002 | 0.0607 |
| **Childhood Residence**<br>Rural<br>Small Town | -1.9208<br>-1.2520 | 0.1557<br>0.1520 | 0.0000<br>0.0000 | -0.2149<br>-0.1423 |
| **Father's Education level**<br>No formal education<br>Primary<br>Lower Secondary<br>Upper Secondary | -1.8540<br>-1.7476<br>-1.1309<br>-0.4707 | 0.3576<br>0.2586<br>0.2661<br>0.2913 | 0.0000<br>0.0000<br>0.0000<br>0.0000 | -0.0875<br>-0.1159<br>-0.0703<br>-0.0137 |

Note: Refer to Table 6.1 for Reference Categories

It is observed from Table 6.7 that the likelihood of being high achiever in English is much lower among Bumiputra students as compared to the Indian students. The odds are decreased by a factor of 0.2133. Students from the science stream have highly likely to be high achievers in English compared to arts students.

The likelihood of being a high achiever in English is lowest among students whose fathers' have no formal education, however, the probability increases with fathers' education level. The more highly educated the father, the more likely the students will be high achievers in English. Rural students are less likely to achieve a grade of A1 or A2 or C3 in the SPM English.

**Table 6.8: Logistic Regression Analysis on the Likelihood of Students Being High Achievers in Mathematics**

| Variables | Coefficients (B) | Standard Error | Significance Level | Exp (B) |
|---|---|---|---|---|
| **Ethnicity** Bumiputera Chinese | -1.7172 1.8019 | 0.2180 0.2155 | 0.0000 0.0000 | 0.1796 6.6061 |
| **Stream** | 2.8723 | 0.1640 | 0.0000 | 17.6773 |
| **Gender** | -0.1819 | 0.1393 | 0.1916 | 0.8337 |
| **ChildHood Residence** Rural Small Town | -0.5634 -0.4182 | 0.1779 0.1777 | 0.0015 0.0186 | 0.5693 0.6582 |
| **Father's Education level** No formal education Primary Lower Secondary Upper Secondary | 0.0286 -0.5648 -0.2786 0.3242 | 0.3477 0.2700 0.2755 0.2926 | 0.9345 0.0364 0.3125 0.3676 | 1.029 0.5685 0.7569 1.3829 |

Note: Refer to Table 6.1 for Reference Categories

110

Table 6.8 presents the logistic regression estimates of the effects of selected independent variables on the likelihood of being high achiever in Mathematics. It is evident from Table 6.8 that the probabilities of achieving good grades (A1, A2 or C3) in Mathematics are the highest for the Chinese students as compared with the Indian students.

The probability of Bumiputera students achieving good grades in Mathematics is the lowest (-1.7172). Science students are most likely to be high achievers in Mathematics than arts students. The odds are increased by a factor of 17.6773. The effect of fathers' education and gender on the likelihood of being high achiever is insignificant as the significant level is more than 0.05.

Although these two variables were found to have significant association with Mathematics grades in Chapter 5, but now becomes insignificant when combined with other variables in the logistics regression model. Both, the rural and small town students are less likely to be high achievers in Mathematics as compared to students from large towns.

111