Chapter 2

Methodology

In this chapter we include a description of the methods used in the study, namely:

1. Computer programs
2. Data Management
3. Statistical Methods

2.1 Computer Programs

The following computer programs were used for data analysis and thesis preparation.

WebStat is a suite of functions written in HTML and VBScript for graphing and analysing statistical data stored in an SQL database. These programs use a web server. It is mainly used to perform preliminary data analysis and regression modeling.

Microsoft Excel is used initially to manage the data used for this research. Some functions are helpful for plotting graphs.

Microsoft Word is used to write and print the report of this research.

2.2 Data Management

The data stored in the EIS (Education Information System) program were imported to the Microsoft Excel spreadsheet file. Next, the data were checked for errors and these errors were corrected. After that, the categorical variables were coded and labeled. The data were imported to Microsoft SQL Server for analysis using *WebStat*.

Data Transformation

If the statistical assumptions of variance homogeneity and normality are not satisfied, the data might need to be transformed. Making a transformation of the data changes their skewness and kurtosis.

If the data are *right-skewed* we can transform by using logarithms, square roots, or cube roots. For logarithm transformations we can use base 2 or base 10 or natural (base e) and if the data have zero values we can transform by using $x' = \log(x+1)$. The base for the logarithmic transformation does not affect the shape of the resulting distribution, but it just affects the scale.

In our data BMI and parents' salary are right-skewed so we use a logarithmic transformation of the form $\ln(\text{BMI}-a)$ where a is constant chosen to induce symmetry for BMI, and we use $x' = \log(x+1)$ for parents' combined salary because these data have zero values.

Imputing missing data

This study had a substantial proportion of missing data for parents' combined salaries, coded as zeroes. To impute values by using a linear regression model, the association with age group and religion group is used, by fitting a multiple linear regression model of this outcome (parents' combined salary) with age group and religion group as predictors. The formula for imputing the missing parents' combined salaries is as follows.

$$\text{E}[\ln(\text{ParentSalary})] = \beta_0 + \beta_1(\text{age group}) + \beta_2(\text{religion group}), \qquad (2.1)$$

where β_0 is a constant, β_1 is a set of coefficients for age group (1, 2, …, 6) and β_2 is a set of coefficients of religion group (1, 2).

2.3 Statistical Methods

Two-sample t-test

The two-sample t-test is used to test the null hypothesis that the population means are the same, and this t-statistic is defined as follows.

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{2.2}$$

where if S_1 and S_2 denote the standard deviations of the two samples, respectively, it may be shown that the pooled sample standard deviation is given by the formula

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}. \tag{2.3}$$

A p-value is now obtainable from the table of the two-tailed t distribution with $n_1 + n_2 - 2$ degrees of freedom (McNeil, 2000).

Chi-squared decomposition and Odds ratio

Pearson' chi-squared test and 95% confidence intervals for odds ratios are used to assess the association between the determinant variables and the outcome of this study.

The odds ratio is a measure of the strength of the association between two binary variables (i.e., in which both the outcome and the determinant are dichotomous) (McNeil, 1996, 1998a, 1998b). The formulas for a two-by-two contingency table (McNeil, 1998b) have the counts a, b, c, and d, the estimated odds ratio is given as follows. Suppose that x is the determinant of interest and y is the outcome. Each variable is binary that the values are labeled 0 and 1. To illustrate the definition of the odds ratio, a two-by-two table is constructed as follows.

	$y = 0$	$y = 1$
$x = 0$	a	b
$x = 1$	c	d

The ratio of these odds is referred to as the odds ratio (McNeil, 1996). Thus the estimate of the odds ratio is

$$\text{OR} = \frac{ad}{bc}. \tag{2.4}$$

The standard error is given by

$$SE\left(\ln OR\right) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}. \tag{2.5}$$

A 95% confidence interval for the population odds ratio is thus

$$\text{OR} \times \exp\left(\pm\, 1.96\ \text{SE}\left[\ln \text{OR}\right]\right). \tag{2.6}$$

If $n = a + b + c + d$, Pearson's chi-square statistic is defined as

$$\chi^2 = \frac{(ad - bc)^2 n}{(a+b)(c+d)(a+c)(b+d)}. \tag{2.7}$$

The p-value is the probability that a chi-squared distribution with 1 degree of freedom exceeds this statistic.

In this study, some of variables are multicategorical. We use non-stratified $r \times c$ tables to compare them. These take the form

	$y = 1$	$y = 2$	$\cdots$	$y = c$
$x = 1$	a_{11}	a_{12}	$\cdots$	a_{1c}
$x = 2$	a_{21}	a_{22}	$\cdots$	a_{2c}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x = r$	a_{r1}	a_{r2}	$\cdots$	a_{rc}

An odds ratio associated with categories i and j of the table can be defined by

$$\mathrm{OR}_{ij} = \frac{a_{ij}d_{ij}}{b_{ij}c_{ij}}, \tag{2.8}$$

where

$$b_{ij} = \sum_{i=1}^{r} a_{ij} - a_{ij},\ c_{ij} = \sum_{j=1}^{c} a_{ij} - a_{ij},\ d_{ij} = n - a_{ij} - b_{ij} - c_{ij}, n = \sum_{i=1}^{r}\sum_{j=1}^{c} a_{ij}\ .$$

The standard error of the natural logarithm of the odds ratio is given by the same formula as for the two-by-two table. In general, the association is composed of $r \times c$ odds ratios, but only $(r-1)(c-1)$ of them are independent.

The standard error is given by

$$SE\left(\ln OR_{ij}\right) = \sqrt{\frac{1}{a_{ij}} + \frac{1}{b_{ij}} + \frac{1}{c_{ij}} + \frac{1}{d_{ij}}}\ . \tag{2.9}$$

A 95% confidence interval for the population odds ratio is thus

$$\mathrm{OR}_{ij} \times \exp\left(\pm\ 1.96\ \mathrm{SE}\left[\ln \mathrm{OR}_{ij}\right]\right). \tag{2.10}$$

Pearson's chi-squared statistic for independence (i.e., no association) in an $r \times c$ table is defined as

$$\chi^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(a_{ij} - \hat{a}_{ij})^2}{\hat{a}_{ij}}, \tag{2.11}$$

where a_{ij} is the count in cell (i, j), that is,

$$\hat{a}_{ij} = \frac{(a_{ij} + b_{ij})(a_{ij} + c_{ij})}{n}.$$

When the null hypothesis of independence is true, this has a chi-squared distribution with $(r-1)(c-1)$ degrees of freedom (McNeil, 1998b).

One-way Analysis of Variance

One-way analysis of variance (ANOVA) is a method for the analysis of data in which the outcome is continuous and the determinant is categorical. This null hypothesis may be tested by computing a statistic called the F-statistic and comparing it with the appropriate distribution to get a p-value. Suppose that there are n_j observations in sample j, denoted by y_{ij} for $i = 1, 2, \ldots, n_j$. The F-statistic is defined as

$$F = \frac{(S_0 - S_1)/(c-1)}{S_1/(n-c)}, \tag{2.12}$$

where

$$S_0 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \bar{y})^2, \qquad S_1 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}(y_{ij} - \bar{y}_j)^2,$$

and

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}, \qquad \bar{y} = \frac{1}{n}\sum_{j=1}^{c}\sum_{i=1}^{n_j} y_{ij}, \qquad n = \sum_{j=1}^{c} n_j.$$

Note that S_0 is the sum of squares of the data after subtracting their overall mean, while S_1 is the sum of squares of the residuals obtained by subtracting each sample mean. If the population means are the same, the numerator and the denominator in the F-statistics are independent estimates of the square of the population standard deviation (assumed the same for each population). Then the p-value is given by the area in the tail of the F-distribution with $c-1$ and $n-c$ degrees of freedom. The F-test also requires the further assumption that the adjusted data (that is, the data adjusted by subtracting the

population means from their respective samples) should have arisen from a normal distribution. Graphing the residuals against normal scores may check this assumption. If the normal scores plot shows a rough linear trend, the normality assumption might be reasonable for the data.

The standard errors used to compute confidence intervals for the mean are based on an estimate of the common standard deviation given by the formula

$$s = \sqrt{\frac{S_i}{n - c}} \tag{2.13}$$

(McNeil, 1996).

Multiple linear regression

Regression used to analyse data in which both the determinants and the outcome are continuous variables. It can summarise the data in the scatter plot by fitting a straight line. In conventional statistical analysis the line fitted is the *least squares line*, which minimises the distances of the points to the line, measured in the vertical direction. If there is more than one determinant, the method generalises to multiple linear regression, in which the *regression* line extends to the multiple linear relation represented as

$$\mathrm{E}[Y] = \beta_0 + \sum \beta_i \chi_i, \tag{2.14}$$

where Y is the outcome variable, β_0 is a constant, $\{\beta_i\}$ is a set of parameters (i = 1 to p), and $\{\chi_i\}$ is a set of determinants (i = 1 to p) (McNeil, 1998).

The model is fitted to data using least squares, which minimises the sum of squares of the residuals.

Linear regression analysis resets on three assumptions as follows.

(1) The association is linear.

(2) The variability of the error (in the outcome variable) is uniform.

(3) These errors are normally distributed.

In our case the multiple regression model is fitted to the data with ln(BMI-a) as an outcome. Thus,

$$Y = \ln(\text{BMI}-a) \tag{2.15}$$

Formula for Adjusted BMI

The linear regression model to ln(BMI–a) using the variables (1) (age, sex, religion) group, (2) birthplace, and (3) parents' combined salary with 0s imputed. The formula for predicting ln(BMI–a) is as follows. The expected value of ln(BMI–a) for a student is given by the formula

$$E[Y] = \beta_0 + \beta_1(\text{age, sex, religion})\text{group} + \beta_2\,\text{birth place} + \beta_3\ln(\text{parentSalary}) \tag{2.16}$$

where β_0 is a constant, β_1 is a set of coefficients for (age, sex, religion) group, β_2 is a set of coefficients for birth place, and β_3 is a coefficient for ln(parentSalary).

The expected value of the BMI is then calculated by assuming that ln(BMI–a) is normally distributed with mean m and standard deviation s, say, and consequently BMI has a lognormal distribution with mean $\exp(m+s^2/2)$. For example, if $E[Y] = m$ and $SD[Y] = s$,

$$Y = \ln(\text{BMI}-a)$$

$$e^Y = \text{BMI} - a$$

$$\text{BMI} = a + \exp(Y)$$

$$\text{E[BMI]} = a + \text{E}[\exp(Y)]$$

$$= a + \exp(m + s^2/2)$$

Thus, we obtain

$$\text{E[BMI]} = a + \exp(\beta_0 + \beta_1(\text{age, sex, religion})\text{group} + \beta_2 \text{birth place}$$

$$+ \beta_3 \ln(\text{parentSalary}) + s^2/2) \,. \qquad (2.17)$$