

|  |
| --- |
| A Study of Risk Factors Associated with Complications of Type 2  Diabetes |
|  |
| **March 2012** |

# Contents

Contents

[Contents 1](#_Toc65548045)

[Acknowledgements 3](#_Toc65548046)

[Chapter 1: Introduction 4](#_Toc65548047)

[1.1 Background 4](#_Toc65548048)

[1.2 The Edinburgh Type 2 Diabetes Study 5](#_Toc65548049)

[1.3 Aims 5](#_Toc65548050)

[1.4 The Variables 6](#_Toc65548051)

[Chapter 2: Methods 8](#_Toc65548052)

[2.1 A Continuous Response 8](#_Toc65548053)

[2.1.1 Correlation 8](#_Toc65548054)

[2.1.2 Linear Regression 9](#_Toc65548055)

[2.1.3 Variable Selection 9](#_Toc65548056)

[2.1.4 The Assumptions of a Linear Regression 10](#_Toc65548057)

[2.1.5 Coefficient of Determination 11](#_Toc65548058)

[2.1.6 Transformations 11](#_Toc65548059)

[2.2 A Categorical Response 11](#_Toc65548060)

[2.2.1 Logistic Regression 11](#_Toc65548061)

[2.2.2 Variable Selection 12](#_Toc65548062)

[2.2.3 The Assumptions of a Logistic Regression 12](#_Toc65548063)

[2.2.4 Deviance Explained 13](#_Toc65548064)

[2.3 Reducing Dimensionality 13](#_Toc65548065)

[2.3.1 The Method 13](#_Toc65548066)

[2.3.2 Correlation or Covariance Matrix? 14](#_Toc65548067)

[2.3.3 Scree Plot 14](#_Toc65548068)

[2.3.4 Application 14](#_Toc65548069)

[Chapter 3: Analysis 14](#_Toc65548070)

[3.1 Ankle Brachial Pressure Index 14](#_Toc65548071)

[3.2 Myocardial Infarction 21](#_Toc65548072)

[3.3 Principal Component Analysis 26](#_Toc65548073)

[Chapter 4: Conclusion 30](#_Toc65548074)

[4.1 Conclusions 30](#_Toc65548075)

[4.2 Discussion and Future Work 31](#_Toc65548076)

[References 33](#_Toc65548077)

[Appendix 36](#_Toc65548078)

# Acknowledgements

I would like to express thanks to the following people for all their support throughout this project;

# Chapter 1: Introduction

## Background

Diabetes (*diabetes mellitus*) is a chronic medical condition characterized by an excess of glucose in the blood due to problems with insulin production (*NHS Choices, 2010*). Insulin is a vital hormone produced by the pancreas which allows glucose to enter the body’s cells. There are two main types of diabetes, Type 1 and Type 2. Type 1 diabetes occurs when the body is unable to produce any insulin at all and usually presents in childhood. Type 2 diabetes occurs when the body is unable to produce enough insulin to compensate for an individual’s poor response to insulin, also referred to as insulin resistance. Type 2 diabetes is particularly common in elderly people over the age of 65. The global increase in diabetes is a serious cause for concern. In 2011 there were 366 million people with diabetes worldwide and it is expected that this figure will rise to 522 million by 2030. This equates to an increase of 50.7%, with the largest increases in older age groups (*Whiting, et al., 2011*). This global increase in numbers is generally linked to economic development and a change in lifestyles which has led to lower levels physical activity and rising levels of obesity. In the UK alone, there are 2.9 million people diagnosed with diabetes and an estimated 850,000 people who have the condition but are unaware of it. This is in comparison to 1.8 million people who were diagnosed with diabetes in the UK in 2004 (*Diabetes UK, 2004*). It is estimated that by 2025 five million people in the UK will have diabetes (*Diabetes UK, 2011*). The prevalence – the total number of cases in the population – of diabetes is 4.45% in the UK (*Diabetes UK, 2011*) and 4.6% in Scotland (*NHS Scotland, 2010*).

The more common type of diabetes is Type 2 and accounts for between 85% and 95% of diabetes cases in the UK (*Diabetes UK, 2010*). This condition can be controlled by diet during the early stages, but due to the progressive nature of the condition, medication may eventually be required. In elderly people with type 2 diabetes, the incidence of and mortality due to disease of the large blood vessels (macrovascular disease) is high, especially from coronary artery disease. Other complications associated with type 2 diabetes include microvascular disease (disease of the small blood vessels affecting the eyes, kidneys and peripheral nervous system) and increased prevalences of cognitive impairment and non- alcoholic fatty liver disease (*Diabetes UK, 2012*). With such a rapid increase in cases over the past few years, and a continuing increase expected over the coming decades, the study of risks associated with complications of diabetes is vital. Examples of these risks include a number of underlying physical markers such as levels of blood pressure, hormones and fats which could be considered warning signals for complications in diabetes. In-depth understanding of these potential risks will ideally lead to improved diagnoses and more effective, streamlined treatments.

## The Edinburgh Type 2 Diabetes Study

The Edinburgh Type 2 Diabetes Study is a prospective cohort study which aims to investigate the potential risk factors for complications of type 2 diabetes (*Price, et al., 2008*). The study began as a cross-sectional survey in 2006/7 when participants were recruited at baseline. Subjects attended a clinic a year later in 2007/8 and a follow-up clinic after four years in 2010/11. Initially, 5454 people aged between 60 and 74 years were selected by sex and 5- year age bands from the Lothian Diabetes Register (LDR). The LDR is a computerised database established in 2001 which includes details on over 20,000 patients with type 2 diabetes living in the Lothian Region of Scotland. From this sampling frame of 5454 potential participants, 1066 subjects were selected at random and hence recruited into the study. Data collected at baseline were based on a basic physical examination and questionnaires on demographic characteristics, cardiovascular risk factors and angina, chest pain and myocardial infarction diagnoses. Answers from questionnaires were combined with data on discharges from Scottish hospitals between 1981 and 2007 to validate reported cardiovascular events. After one year 940 subjects were still living and agreed to re-attend for further physical examination and questionnaires relating to alcohol intake and liver disorders. In 2010/11 the subjects underwent follow-up examinations of cognitive, vascular and liver assessments, the results of which are currently being collated. Analysis of the data from the Edinburgh Type 2 Diabetes Study has already resulted in a number of published articles over the last few years. In 2009 a paper was published which demonstrated significant relationships between obesity and depression, and cardiovascular disease and depression, in elderly patients with type 2 diabetes (*Labad, et al., 2009*). Reynolds, et al., 2010, found a strong association between ischemic heart disease and elevated levels of the steroid hormone cortisol. In 2011 the prevalence of non-alcoholic fatty liver disease within the Edinburgh Type 2 Diabetes sample was confirmed to be high (*Williamson, et al., 2011*). This type of analysis is currently on-going, with particular focus on the risk factors associated with cognitive impairment and dementia. Some less exhausted areas of research which will be explored here are unusual lipid measurements, thyroid hormone levels and sex hormone levels.

## Aims

In this project data from the baseline and Year-1 visits from the Edinburgh Type 2 Diabetes Study will be used to determine the cross-sectional association between a range of potential risk factors and vascular diseases in elderly patients with type 2 diabetes. In order to investigate this, both symptomatic vascular events and markers of underlying vascular disease will be considered. The main risk factors under study will be levels of thyroid hormones, steroid sex hormones (e.g. testosterone) and fats (lipids). The three key questions of interest are:

* + 1. Is there a relationship between plasma levels of thyroid hormones, sex hormones and/or lipids and Ankle Brachial Index (ABI: a marker of underlying vascular disease) after correcting for known confounding factors?
    2. Are plasma thyroid hormones, sex hormones and lipid levels associated with prevalence of the cardiovascular event of a myocardial infarction (MI), after correcting for known confounding factors?
    3. What are the relationships between potential risk factors for cardiovascular disease and can the dimensionality be reduced?

## The Variables

The data available for analysis consist of the following variables (further information on a selection of key variables can be found in Table A1, in the Appendix):

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable** | **Abbrev-**  **iation** | **Description** | **Type** | **Units/**  **Categories** |
| **Responses** (*recorded at baseline*) | | | | |
| Myocardial Infarction | MI | The interruption of the blood supply to a  part of the heart, causing heart cells to die | Binary | 0 = No  1 = Yes |
| Ankle Brachial Pressure Index | ABI | The ratio of the systolic blood pressure in the ankles to the systolic blood pressure in  the arms | Continuous |  |
| **Known Confounding Factors** | | | | |
| *Recorded at baseline* | | | | |
| Sex |  |  | Binary | 1 = Male  2 = Female |
| Body Mass Index | BMI |  | Continuous | kg/m² |
| Smoking Status |  | Self-reported levels of cigarette smoking | Categorical | 1 = Current  2 = Ex  3 = Never |
| Scottish Index of Multiple Deprivation | SIMD | Identifies small area concentrations of multiple deprivation (income, employment, health, education, etc.) across all of Scotland | Categorical | 1 = most deprived  …  10 = least  deprived |
| *Recorded at Year-1* | | | | |
| Age |  |  | Continuous | Years |
| Duration of Diabetes |  | Time since doctor diagnosis of diabetes | Continuous | Years |
| Treatment by diet |  | Self-reported use of diet alone to control  diabetes (blood sugar levels) | Binary | 0 = No  1 = Yes |
| Treatment by insulin |  | Self-reported use of insulin injections to  control diabetes (blood sugar levels) | Binary | 0 = No  1 = Yes |
| Treatment by tablets |  | Self-reported use of oral glucose-lowering medicines to control diabetes (blood sugar  levels) | Binary | 0 = No  1 = Yes |
| Cholesterol |  | A fatty substance found in the blood | Continuous | mmol/L |
| Systolic blood  pressure | sBP | The maximum pressure in the arteries  when the heart contracts | Continuous | mmHg |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Diastolic blood pressure | dBP | The minimum pressure in the arteries between heart beats when the heart  relaxes | Continuous | mmHg |
| Glycated hemoglobin | HBA1c | Identifies the average plasma glucose  concentration | Continuous | % |
| High-density  lipoprotein | HDL | Carries cholesterol away from cells and  back to the liver where it is broken down | Continuous | mmol/L |
| **Explanatory Variables** (*recorded at Year-1*) | | | | |
| **Lipids:** | | | | |
| Apolipoprotein A1 | Apo. A1 | A major protein component of HDL in  plasma | Continuous | g/L |
| Apolipoprotein B | Apo. B | The primary protein component of LDLs  (low-density lipoproteins) | Continuous | g/L |
| Free fatty acids | FFA | Fatty acids released into the bloodstream  and circulating throughout the body | Continuous | mmol/L |
| **Thyroid Hormones:** | | | | |
| Free T3 | FT3 | The available triiodothyronine | Continuous | pmol/L |
| Total T3 | TT3 | The total triiodothyronine | Continuous | nmol/L |
| Free T4 | FT4 | The available thyroxine | Continuous | pmol/L |
| Total T4 | TT4 | The total thyroxine | Continuous | nmol/L |
| Thyrotrophin-  stimulating hormone | TSH | Signals the thyroid to produce more  thyroid hormone | Continuous | mIU/L |
| **Sex Hormones:** | | | | |
| Testosterone |  | A steroid hormone, primarily secreted in the testes of males and the ovaries of  females | Continuous | nmol/l |
| Sex Hormone Binding  Globulin | SHBG | A protein produced mostly by the liver  that binds to sex hormones | Continuous | nmol/l |

**Table 1.1: Table of all available variables**

The two response variables, ABI and MI, are of particular interest in patients with type 2 diabetes. As previously mentioned, the incidence of and mortality due to vascular disease in these patients is high. ABI is a widely-used and valuable measurement as it allows diagnosis of peripheral arterial disease, or hardened arteries, which in many cases is asymptomatic (*Al-Qaisi, et al., 2009*). A lower ABI value is associated with cardiovascular disease as a low blood pressure in the ankle, in comparison to that in the arm, indicates hardening of the arteries (*Wild, et al., 2006*). A further advantage of ABI is that it is simple and inexpensive to implement in primary care services – the services provided by GP practitioners, dentists and community pharmacists. MI is commonly referred to as a “heart attack”, although it is not the only cause of a heart attack. Globally it is a major health burden, particularly in more economically developed countries such as the UK where it is the leading cause of death (*Lopez and Murray, 1998*). The classic symptoms of a myocardial infarction include chest pain, shortness of breath and overwhelming anxiety and require immediate emergency attention. The mortality rate of diabetic patients after their first myocardial infarction is extremely high (*Miettinen, et al., 1998*) and hence further understanding of the risks associated with MI is clearly beneficial.

Expert advice recommended a number of known, potential confounding variables which would have to be considered in a statistical model along with the particular variables of interest. These are split into a group of continuous variables (age, duration of diabetes, BMI, cholesterol, sBP, dBP, HDL and HbA1C) and a group of categorical variables (sex, SIMD, smoking status, treatment by diet, treatment by insulin and treatment by tablets). Further information on the categorization of smoking status stated that patients who reported being non-smokers but had quit within the previous 6 months were recorded as current smokers. Some less researched risk factors were suggested, such as unusual lipid measurements (Apolipoproteins A1 and B and FFA), thyroid hormones (triiodothyronine, thyroxine and TSH) and sex hormones (testosterone and SHBG). Although there are well- known fats associated with cardiovascular disease, such as cholesterol, it would be of interest to investigate whether alternative measurements could provide additional information. The link between thyroid hormones and cardiovascular disease in diabetic patients is documented (*Johnson and Duick, 2002*), but not in as great a detail as, for example, smoking and cardiovascular disease. There is thought to be a link between testosterone and cardiovascular disease, particularly in older males (Zmuda, et al., 1997).

One issue with the data which will be discussed in greater detail in Chapter 3 is the difference in time points at which measurements were made- baseline and the Year-1 visit. Again expert advice informed us that, in relation to cardiovascular diseases, one year is not regarded as a long period of time and that clinicians would not expect measurements to change drastically. Therefore, initially, measurements will be considered as being taken at essentially the same time point.

# Chapter 2: Methods

## A Continuous Response

### Correlation

The relationship between two continuous variables, X and Y, can be quantified using the sample correlation coefficient, assuming the relationship is linear. This measures the strength of a linear relationship using the formula:

√

where ∑( ̅) ( ̿) ; ∑( ̅) and ∑( ̅).

The range of is between -1 and 1, where a value closer to -1 indicates a stronger negative linear relationship, a value closer to 0 indicates a weak or no linear relationship and a value

closer to 1 indicates a stronger positive linear relationship. A formal test can be carried out to compare the hypotheses:

H₀: the true population correlation coefficient is zero H₁: the true population correlation coefficient is not zero

A test statistic based on the above formula is calculated and compared to a t-distribution with appropriate degrees of freedom. A corresponding p-value can be calculated and H₀ is rejected in favour of H₁ if this is less than 0.05.

### Linear Regression

The first response under investigation is a continuous variable. Therefore, a natural starting point for analysis is a multiple linear regression, which aims to discover a subset of the potential explanatory variables that linearly describes a continuous response. Given a set of data of n observations, multiple linear regression takes the general form:

where is the response variable, are the explanatory variables, are the estimated coefficients and is a random error term. The errors are assumed to be independently and identically distributed, following a normal distribution with mean zero and constant variance. The coefficients of the parameters are estimated using the method of ordinary least squares which minimises the sum of squares. This gives a unique and unbiased estimate of the parameters. The significance of the effect of each covariate on the response is given by a p-value which is calculated using a T-test. The hypotheses being compared are

H₀:

H₁:

and a p-value gives the probability of observing a test statistic as or more extreme if H₀ is true. Confidence intervals corresponding to each estimate can also be calculated using the t distribution.

### Variable Selection

In order to identify important terms in a model the Akaike Information Criterion (AIC) is commonly used. AIC is a measure of the goodness-of-fit of a model and imposes a penalty for increasing numbers of parameters. It is calculated as:

where  is the maximised log-likelihood of the model and is the number of parameters in the model. Low values of AIC indicate the preferred model i.e. the model that includes the fewest parameters while still providing an adequate fit to the data.

There are a number of potential methods which can be used to compare models. The two discussed and used in this report are an F-test and a Likelihood Ratio Test (LRT). An F-test compares a full model and a reduced model nested within the full model using the following hypotheses:

H₀: 2.1.3.1

H₁:

An observed F statistic is compared to the F distribution with appropriate degrees of freedom and H₀ is rejected in favour of H₁ if F is large in comparison. This can also be represented in a corresponding p-value.

For the LRT tests the same hypotheses are compared. The ratio, λ, between the maximum log-likelihood for the simpler model ( ) and the maximum log-likelihood for the full model () is computed and used to calculate the observed test statistic:

(  )

This is compared to a Chi-square distribution with appropriate degrees of freedom and H₀ is rejected in favour of H₁ if it is comparatively large. Again a corresponding p-value can be calculated.

### The Assumptions of a Linear Regression

The key assumptions that underlie a linear regression model are that the deterministic part of the model is appropriate, normality of the errors, constant variance in the errors and independence of the errors. The last of these is assumed to hold from the study design. The first three can be checked using diagnostic plots of residuals.

A residual, ̂, is defined to be the difference between the sample value and the value estimated by a model. A problem with residuals is that they may not be independent or have constant variance and therefore the standardized residuals are most commonly used:

̂

√ ̂ ( )

where ̂ is the original residual value, ̂ is the variance of the residuals and is the th element of the so-called Hat Matrix:

where A is the design matrix.

( )

Since the parameter estimates and the residuals are independent, the residuals and the fitted values are also independent. Therefore, a plot of the residuals against the fitted values should show no correlation if the assumption of linearity holds. There should also be no ‘fanning out’ in the pattern for the assumption of constant variance to hold. Finally, if the ordered residuals against the normal theoretical quantiles (a Q-Q plot) show a straight line then the assumption of normality holds.

### Coefficient of Determination

The coefficient of determination assesses what fraction of the variability in the data is explained by the model. It is calculated as:

where ,

and

are as above. The coefficient of determination has a range from 0 to

1, with larger values indicating a better fit. It is usually multiplied by 100 to give the percentage of variability in the data explained by the model.

### Transformations

If the Q-Q plot shows extreme deviations from the normal line or the response does not appear to be normally distributed then a transformation of the response may be required. One technique for finding a suitable transformation is the Box-Cox method. This suggests an appropriate transformation from a family of transformations in the following way:

{

The appropriate value of λ is estimated by maximum likelihood. If a 95% confidence interval for λ includes 1 then this suggests that there is no need to transform the response. Note that this method is only appropriate when the response is strictly positive.

## A Categorical Response

### Logistic Regression

The second response of interest is a binary response which suggests that a logistic regression is appropriate. The key aim is to model how a binary outcome variable, X, depends on one or more explanatory variables. The logistic regression model assumes that a non-response (X=0) happens with probability 1-θ and that a response (X=1) happens with probability θ where

()

(     )

However, instead of θ, a logistic transformation is used in the regression analysis:

where is the log-odds of a response. The log-odds have a linear relationship with the explanatory variables. Maximum likelihood methods are used to estimate the coefficients of the log-odds model:

In order to obtain the odds of response the exponential is taken and in order to revert back to probabilities the above formula for θ is used. Corresponding p-values are calculated for each covariate using a Wald test which indicates the significance of the effect of each covariate on the response. The test statistic for a Wald test is the regression coefficient divided by its estimated standard error. This is compared to a standard normal distribution to determine significance.

### Variable Selection

In order to reduce a model the AIC is compared, in the same way as described in Section

2.1.3. When comparing two logistic models two are again used: a Chi-square test (similar to the F-test in linear regression) and a Likelihood Ratio Test (See Section 2.1.3).

### The Assumptions of a Logistic Regression

It should first be noted that the residuals and goodness-of-fit tests are far more tentative than their linear regression equivalents and more often than not provide less than satisfactory results. The assumptions of a logistic regression are independence of observations, absence of outliers and normality in the residuals. Again the first assumption is assumed to hold by the study design. Residuals are useful for determining if individual points are not well fitted by the model. There are two types of residuals which can be considered:

√  ̂ ( ) ̂

̂

√ ̂ ̂

Either of these measurements can be plotted in an Index Plot to identify potential outliers. The general rule of thumb is that a point with absolute value greater than 2 is a poor fit. A normal Q-Q plot can be obtained and checked to see how well the points follow a straight line; however these plots frequently have one or more extreme deviations.

Finally, the goodness-of-fit of the model can be checked using a Hosmer-Lemeshow test which compares the following hypotheses:

H₀: the model is appropriate for the population ---2.2.3.1 H₁: the model is not appropriate for the population

The fitted values are ordered and grouped into g classes of roughly equal size. The observed and expected number in each group is calculated and combined to give a test statistic. This is compared to a Chi-square distribution with g-2 degrees of freedom. H₀ is rejected if the observed statistic is large in comparison and a corresponding p-value can be obtained. The Hosmer-Lemeshow test depends on the number of groups so it is recommended to repeat this test for a range of values (usually between about 6 and 15, and there must be at least 3 groups).

### Deviance Explained

An analogous measure of variation explained to the coefficient of determination is the Deviance Explained. The deviance of a model is defined to be:

where  is the maximised log likelihood for the model. The null deviance is the deviance for a model with just a constant term. The residual deviance is the deviance of the fitted model. Hence the Deviance Explained is:

The range of the Deviance Explained is between 0 and 1 and can be multiplied by 100 to give the percentage of Deviance Explained by a model.

## Reducing Dimensionality

### The Method

Principal Component Analysis (PCA) is a multivariate statistical technique used for reducing the dimensionality of data. It is useful when there are many original variables in a data set with strong correlations between some or all of the variables. The general procedure is to compute the correlation or covariance matrix and perform an eigen-analysis. This provides a set of new variables, or components, made up of uncorrelated linear combinations of the original variables. The resulting eigenvalues give the variance of each component and the eigenvectors give the loadings of each component. The loadings of the components can be considered to establish appropriate interpretation for independent components. By construction the components are independent and the total variance is preserved under the principal component transformation. Dimension reduction is usually achieved without substantial loss of information by working with the first *k* components.

### Correlation or Covariance Matrix?

One of the key decisions to be made in performing PCA is whether to use the correlation or covariance matrix for the eigen-analysis. The general rule is to use the sample correlation matrix if the standard deviations vary or the variables have been recorded on different scales.

### Scree Plot

Another key decision to be made in performing PCA is how many of the new components to retain. A common tool used to aid this decision is a Scree Plot which plots

∑  ( ) against m, where  is an eigenvalue. The point at which the plot ‘levels off’ is usually taken to be the cut-off point for components to be included.

### Application

Once the components have been found and an appropriate number of components to retain have been decided upon, these can be used as new variables in an analysis. For example, the components can be fitted in a linear or logistic regression with the aim of finding a more appropriate model than one including the original variables.

# Chapter 3: Analysis

## 3.1 Ankle Brachial Pressure Index

The first response variable under investigation is Ankle Brachial Pressure Index (ABI) which is the ratio of the systolic blood pressure (sBP) in the ankles to the systolic blood pressure in the arms. Measurements were taken on the right and left side of the body and the variable ABI was then calculated as the worst (lowest) ABI of the two. The following analysis is split into two sections. Firstly, known potential confounding factors identified by experts were modelled using linear regression in order to obtain a basic model for ABI. Then the explanatory variables of interest were added in order to investigate whether there is a relationship between ABI and any lipids, thyroid hormones or sex hormones.

Before fitting a model with confounding factors a number of plots and statistics were produced. A histogram of ABI (Figure 3.1.1) displays a fairly neat bell-curve shape which indicates that the response is roughly normally distributed. The range of values for ABI in this sample sits between 0.270 and 2.340. The average value of ABI for patients in this data

set is 0.986 and the median value is 1. A value of less than 0.5 suggests severe disease and less than 0.3 suggests critical disease (*Al-Qaisi, et al., 2009*). Although a value of above 1.4 is also considered to indicate vascular disease, there are relatively few people in this study who have such a high ABI and expert advice informed us this is unlikely to affect the results. Therefore, on average, the patients in this study have a ‘normal’ ABI. However, the minimum extreme value shows that there are patients within this study who display signs of vascular disease which are severe cause for concern. The definition of a continuous variable is conventionally a variable that can take on any range of

Frequency

0

100

200

300

400

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Min** | **1st**  **Quartile** | **Median** | **Mean** | **3rd**  **Quartile** | **Max** | **Missing**  **Values** |
| 0.270 | 0.900 | 1.000 | 0.986 | 1.090 | 2.340 | 6.000 |

positive or negative

values. ABI is a ratio and therefore does not have

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | |  |  |  |
|  |  |
|  |
|  |

**Figure 3.1.1: Histogram of ABI**



0.5

1 0 1.5

ABI

2.0

this range of possible

**Table 3.1.1: Summary Statistics of ABI (*figures to 3 d.p.*)**

values. However, since ABI can take on any value within its specific range then for the purposes of the following analysis ABI will be considered a continuous response.

Correlation coefficients between the eight continuous variables (Table 3.1.2) do not highlight any strongly correlated confounding factors. There is a moderate positive linear relationship between dBP and sBP (ρ = 0.520). These variables are both measurements of blood pressure, one a measure of maximum pressure (sBP) and the other a measure of minimum pressure (dBP). Therefore a relationship between the two variables would be expected biologically: as sBP increases dBP increases (*Gavish, Ben-Dov and Bursztyn, 2008*). Overall, however, there should not be any issues with multicollinearity between confounding factors.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Age** | **Duration** | **BMI** | **Cholesterol** | **sBP** | **dBP** | **HBA1c** | **HDL** |
| **Age** | 1.000 | 0.049 | -0.187 | -0.038 | 0.093 | -0.102 | -0.126 | 0.089 |
| **Duration** | 0.049 | 1.000 | 0.012 | -0.093 | 0.039 | -0.147 | 0.246 | -0.006 |
| **BMI** | -0.187 | 0.012 | 1.000 | -0.044 | 0.029 | 0.039 | 0.069 | -0.125 |
| **Cholesterol** | -0.038 | -0.093 | -0.044 | 1.000 | 0.120 | 0.153 | 0.086 | 0.273 |
| **sBP** | 0.093 | 0.039 | 0.029 | 0.120 | 1.000 | 0.520 | 0.063 | 0.062 |
| **dBP** | -0.102 | -0.147 | 0.039 | 0.153 | **0.520** | 1.000 | 0.089 | 0.038 |
| **HBA1c** | -0.126 | 0.246 | 0.069 | 0.086 | 0.063 | 0.089 | 1.000 | -0.090 |
| **HDL** | 0.089 | -0.006 | -0.125 | 0.273 | 0.062 | 0.038 | -0.090 | 1.000 |

**Table 3.1.2: Correlations between all continuous confounding factors (*figures to 3 d.p.*)**

Scatterplots between ABI and each continuous confounding factor do not indicate any strong relationships. For example, a scatterplot between ABI and age (Figure 3.1.2) demonstrates no increasing, decreasing, or more complex, patterns. Similar plots can be seen for all other continuous confounders (see Figure A1, in the Appendix) and indeed

correlation coefficients are all very low (Table 3.1.3). Since there is no indication of any non- linear relationships, analysis by simple linear regression appears appropriate and no transformations of the continuous confounders were required. Boxplots between ABI and each confounding factor do not indicate any particularly strong relationships. There are some differences in medians; however, there is considerable overlap in all the plots so the extent of these differences is not clear. For example, a boxplot between ABI and sex (Figure 3.1.3) shows two almost identical shapes. There is a larger range in extreme maximum values in males and the median ABI in females is marginally lower, but overall there is no clear relationship. Similar boxplots can be seen for all other categorical confounders (see Figure A2, in the Appendix). The variable indicating smoking status perhaps displays the strongest relationship (Figure 3.1.4). Current smokers have the lowest median ABI, ex- smokers have a slightly higher ABI and patients who have never smoked have a higher ABI again. Nevertheless, there is again considerable overlap between the three categories.

1.5

2.0

**Figure 3.1.2: Scatterplot between ABI and Age**

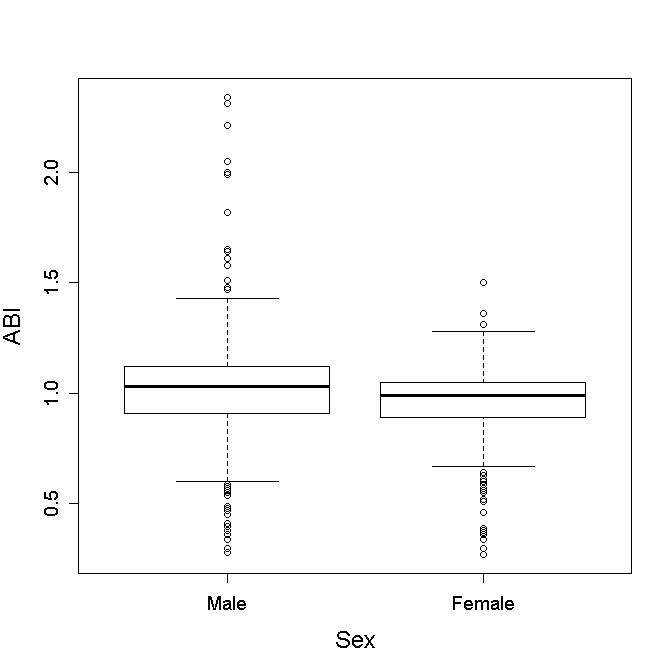


65

70

Age (years)

75



ABI

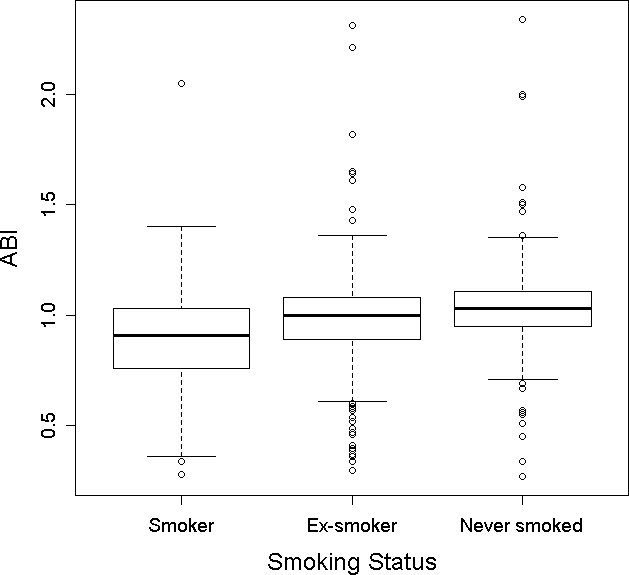
0.5

1.0

|  |  |
| --- | --- |
|  | **ABI** |
| **Age** | -0.007 |
| **Duration** | -0.020 |
| **BMI** | -0.113 |
| **Cholesterol** | -0.058 |
| **sBP** | -0.058 |
| **dBP** | 0.093 |
| **HbA1C** | -0.024 |
| **HDL** | 0.069 |

**Table 3.1.3: Correlations between ABI and continuous confounding factors**

**Figure 3.1.3: Boxplot of ABI by Sex**

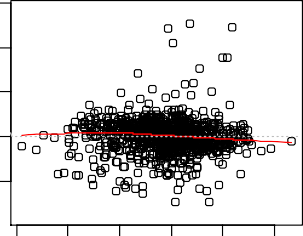


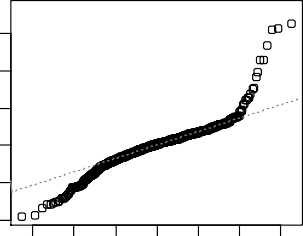
**Figure 3.1.4: Boxplot of ABI by Smoking Status**

Starting with a full additive linear regression model, AIC scores were compared for each possible model with one variable removed. The final reduced model selected includes the terms BMI, cholesterol, sBP, dBP, HDL, sex and smoking status:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Estimate** | **2.5%**  **Lower Interval** | **97.5%**  **Upper Interval** | **Std. Error** | **p-value** |
| **Intercept** | 1.016 | 0.865 | 1.167 | 0.077 | <0.000 |
| **BMI** | -0.003 | -0.005 | -0.001 | 0.001 | 0.011 |
| **Cholesterol** | -0.029 | -0.046 | -0.012 | 0.009 | 0.001 |
| **sBP** | -0.002 | -0.002 | -0.001 | 0.000 | <0.000 |
| **dBP** | 0.003 | 0.002 | 0.005 | 0.001 | <0.000 |
| **HDL** | 0.068 | 0.026 | 0.109 | 0.021 | 0.001 |
| **Sex [Females]** | -0.068 | -0.097 | -0.039 | 0.015 | <0.000 |
| **Smoking [Ex-smoker]** | 0.095 | 0.054 | 0.136 | 0.021 | <0.000 |
| **Smoking [Never smoked]** | 0.164 | 0.122 | 0.206 | 0.021 | <0.000 |

**Table 3.1.4: Summary of Linear Model for ABI including confounding factors (*figures to 3 d.p.*)**

All confounding factors included in this model have some statistically significant effect on the response. BMI has the highest p-value of 0.011 and all other variables are highly significant (p-values <<0.05). The regression coefficient estimates give the size of the effect that each variable has on ABI. The corresponding 95% confidence intervals are all narrow which assures the accuracy of these estimates. A rather surprising set of relationships to find is that as sBP increases, ABI decreases but as dBP increases, ABI increases. The result for sBP is what would be expected from numerous previous studies i.e. that the risk of cardiovascular disease increases as systolic blood pressure increases. The relationship between ABI and dBP may be due to the fact that many diabetic patients with high dBP are prescribed medication to reduce blood pressure. Therefore a seemingly high-risk patient in fact has a lower diastolic blood pressure.



Residuals vs Fitted

940839211

0.7 0.8 0 9 1.0 1.1 1.2

Fitted values

Normal Q-Q

899134201

-3 -2 -1 0 1 2 3

Theore ical Quantiles

Standardized residuals

Residuals

-4 -2 0 2 4 6

-0.5

0 5 1.0 1.5

The adjusted R² for this model is 0.1164 which is very poor. It means that only 11.64% of the variability in the data is explained by this model and that the model has extremely low predictive power. A plot of the residuals versus fitted values (Figure 3.1.5) shows a random scatter of points. Therefore the underlying assumptions of constant variance and linearity are satisfied. A normal Q-Q plot (Figure 2.5) indicates some concern with the assumption of normality as

**Figure 3.1.5: Diagnostic Plots for Linear Model including confounding factors**

there are large deviations at the tails. However, the middle of the data points follows the normal line very well. In the next stage of analysis possible transformations of the response will be explored. Hence the final basic linear regression model, including confounding factors only, is accepted to be one that models ABI by BMI, cholesterol, sBP, dBP, HDL, sex and smoking status.

Before adding the explanatory variable in each category of interest, the relationships between the covariates and the response were investigated. Since all the explanatory variables are continuous, scatterplots were produced. There are no striking relationships highlighted by these graphs, demonstrated by the plot between ABI and Apolipoprotein A1 (Figure 3.1.6). The full set of the plots can be found in Figure A3 in the Appendix. Correlation coefficients between ABI and the explanatory variables are no higher than 0.130 which confirms that there are no moderate or strong linear relationships. Consequently, a linear regression model is appropriate to model the data and no transformations of the explanatories are necessary. There are some

ABI

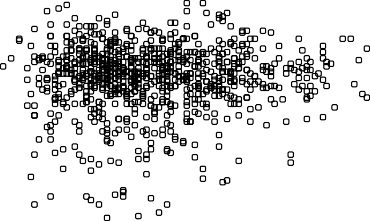
0.5

1.0

1.5

2.0

**Figure 3.1.6: Scatterplot between ABI and Apolipoprotein A1**



0.8

1.0

1.2

1.4

1.6

1.8

2.0

Apolipoprotein A1 (g/L)

strong relationships between all possible variables (the confounding factors and the covariates of interest) to be aware of, in case of issues with multicollinearity (Table 3.1.5). Cholesterol is strongly positively correlated with Apolipoprotein B (ρ=0.850). HDL is strongly positively correlated with Apoliprotein A1 (ρ=0.865). These variables are all measurements of fats so it is unsurprising that there are strong correlations between them. FT3 is strongly correlated with TT3 (ρ=0.740) and FT4 is strongly correlated with TT4 (ρ=0.800), which makes logical sense as both sets of variables measure the same thyroid hormone, just in different ways. Keeping these correlations in mind, all the variables were added in turn and in fact no issues of multicollinearity arose. In relation to the lipids, this is probably because cholesterol and HDL, which are already included in the model, adequately account for the variation in the data.

|  |  |
| --- | --- |
|  | **ABPI** |
| **Apolipoprotein A1** | 0.068 |
| **Apolipoprotein B** | -0.059 |
| **FFA** | -0.032 |
| **FT3** | 0.086 |
| **FT4** | -0.130 |
| **TT3** | 0.045 |
| **TT4** | -0.144 |
| **TSH** | -0.010 |
| **TES** | 0.129 |
| **SHBG** | 0.011 |

**Table 3.1.5: Correlations between ABI and the continuous covariates of interest (*figures to 3 d.p.*)**

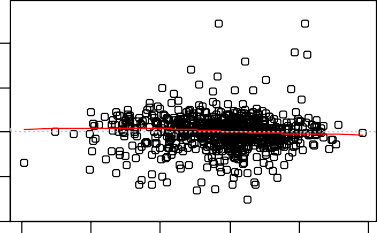
Lastly, to confirm any relationships between ABI and the explanatory variables of interest, individual terms were added to the above basic model in turn. Only variables which had a statistically significant effect on the response (p-value<0.05) were added to the model; otherwise they were discarded. Using this method the final model includes only the additional term FT4, which has a p-value of 0.006. FT4 is the available thyroxine, a thyroid hormone. The following table shows a summary of the final model for ABI:

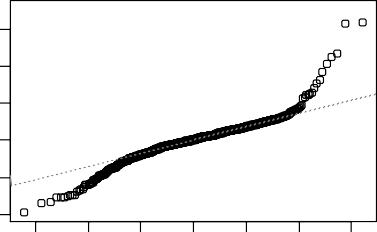
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **2.5%**  **Lower Interval** | **97.5%**  **Upper Interval** | **p-value** |
| **Intercept** | 1.138 | 0.095 | 0.951 | 1.324 | <0.000 |
| **BMI** | -0.003 | 0.001 | -0.006 | -0.001 | 0.009 |
| **Cholesterol** | -0.025 | 0.009 | -0.043 | -0.006 | 0.009 |
| **sBP** | -0.001 | 0.000 | -0.002 | <0.000 | 0.007 |
| **dBP** | 0.003 | 0.001 | 0.001 | 0.004 | 0.005 |
| **HDL** | 0.055 | 0.022 | 0.012 | 0.098 | 0.013 |
| **Sex [Females]** | -0.046 | 0.016 | -0.077 | -0.014 | 0.005 |
| **Smoking [Ex-smoker]** | 0.110 | 0.022 | 0.066 | 0.154 | <0.000 |
| **Smoking [Never smoked]** | 0.171 | 0.023 | 0.127 | 0.216 | <0.000 |
| **FT4** | -0.010 | 0.004 | -0.017 | -0.003 | 0.006 |

**Table 3.1.6: Summary of Final Linear Model for ABI (*figures to 3 d.p.*)**

The p-values for the variables in the final model are highly statistically significant and the 95% confidence intervals do not include zero which both indicate that the covariates all have a significant effect on ABI. The 95% confidence intervals are very narrow which means that there can be a high level of confidence in the accuracy of the regression coefficient estimates. As BMI, cholesterol, sBP and FT4 increase, ABI decreases. As dBP and HDL increase, ABI increases. The magnitudes of these effects are quantified by the estimates in Table. The link between obesity and cardiovascular disease has been extensively researched and highlighted (*Grundy, 2004*) so the decrease in ABI as both BMI and cholesterol increase is to be expected. The relationships between ABI and the two measurements of blood pressure have been previously discussed and demonstrate a combination of what would be expected biologically and a possible treatment effect. HDL carries cholesterol away from cells and back to the liver where it is broken down and passed out of the body as weight. Therefore it is commonly referred to as “good cholesterol” and the relationship displayed here has been previously well documented (*Stern, 1995*). The relationship between FT4 and ABI suggests that the less well researched category of thyroid hormones should be further investigated. As the levels of thyroxine increase ABI decreases i.e. the risk of cardiovascular disease is considered to be higher. The model indicates that females have a lower ABI than males: for a female, ABI is lower by a multiple of -0.046 compared to the baseline, males. This result is unusual as it is generally accepted that males are at a higher risk from cardiovascular diseases than females (*Kardys, et al., 2007*) and therefore we would expect a lower ABI in men. This result is a commonly found phenomenon and a ‘paradox’ to do with the measurement of ABI. Experts advised us that one theory to explain this is related to the differences in height between men and women. The model also indicates a relationship between ABI and smoking. For an ex-smoker, ABI is higher by a multiple of 0.110 compared to the baseline, smokers. For a patient who has never smoked, ABI is higher by a multiple of

0.170 compared to the baseline, smokers. This final result is reasonable as the link between cigarette smoking and cardiovascular disease is common knowledge (*Ambrose and Barua, 2004*).

The underlying assumptions of constant variance, linearity and normality were checked by use of diagnostic plots (Figure 3.1.7). The additional assumption of independence is considered to hold by the study design since measurements are taken on different patients. The residuals versus fitted values plot shows a random scatter of data point with no fanning, suggesting that constant variance of errors is reasonable. There are no trends in the residual plot which confirms that a linear model is appropriate for the data. The assumption of normality is not completely satisfied as there is considerable deviation from the normal line at the tails. A number of transformations of the response were considered in an attempt to improve this assumption. First standard transformations such as log, square root and inverse of ABI were investigated, but none improved the relationship in the normal plot. A Box-Cox transformation was also considered. The 95% confidence interval for λ is approximately (-0.2, 01). Since this interval includes zero, no transformation of the response is recommended by the Box-Cox method (see Figure A4 in the Appendix for a plot). As seen in the histogram of ABI, the response does look to be



Residuals vs Fitted

940 321

133

0.7 0.8 0.9 1.0 1.1 1.2

Fitted values

Normal Q-Q

321940

133

-3 -2 -1 0 1 2 3

Theoretical Quan iles

Standardized residuals

Residuals

-4 -2 0 2 4 6

-1.0

0.0 0 5 1.0

**Figure 3.1.7: Diagnostic Plots for Final Linear Model**

normally distributed. Therefore no

transformation can be found to improve normality. One possible reason for the unsatisfactory assumption of normality is due to the fact that ABI is a ratio. As previously discussed, ABI is not strictly a continuous variable and has a restricted range that values can take. There are a large amount of points clustered in the centre of the distribution of the variable and very small tails which could lead to large deviations in the normal plot. The R² for the final model is unfortunately still very low, 0.114. In fact, this is marginally lower than the previous R² value, which is due to missing values in the data set. Only 11.14% of the variability in the data is explained so the model has extremely poor predictive powers. Since the R² value is so low, it may be that a number of variables are statistically significant because of the large amount of available data. Both an F-test and a likelihood ratio test were carried out in order to check that this is the most appropriate model. Both tests compare the full final model to the reduced confounding factors model comparing the hypotheses stated in 2.1.3.1. The p-values returned were 0.004 for the F-test and 0.003 for the likelihood ratio test. In each case the null hypothesis can be rejected in favour of the alternative hypothesis. Therefore the unrestricted full model is accepted as the best model, compared to the confounders model.

The final check on this model was to investigate whether sBP and dBP were both required in the model. As previously mentioned, the two variables are moderately correlated. There also could be an issue since the response ABI is derived from two measurements of sBP. First sBP was removed from the model and the analysis was repeated. This gave a model where dBP was now just not statistically significant, with a p-value of 0.087. The R² value for

this model was 0.1042, equivalent to 10.42% of the variability explained. This is slightly lower than the previous model and inspection of the corresponding diagnostic plots indicate that they are almost identical (see Figure A5 in the Appendix). An F-test and a likelihood ratio test between this and the fuller model both indicated that the full model is the preferred model (the p-values were 0.015 and 0.014 respectively). Finally, a further restricted model was tested with dBP also removed. The R² value was 0.1005, or 10.05% of the variability explained, which is again slightly lower than the full model. The residual plots were also very similar to the full model (see Figure A5 in the Appendix). An F-test and a likelihood ratio test between the reduced and unrestricted models indicated that the unrestricted model is marginally more appropriate (the p-values were 0.052 and 0.048 respectively).

Therefore, the final model selected is one that models ABI by BMI, cholesterol, sBP, dBP, HDL, sex, smoking status and FT4, the chosen variables significantly affect ABI and the underlying assumptions are mainly satisfied. However, it should be remembered that the final linear model will not be particularly useful for future predictions due to the low R² value.

## Myocardial Infarction

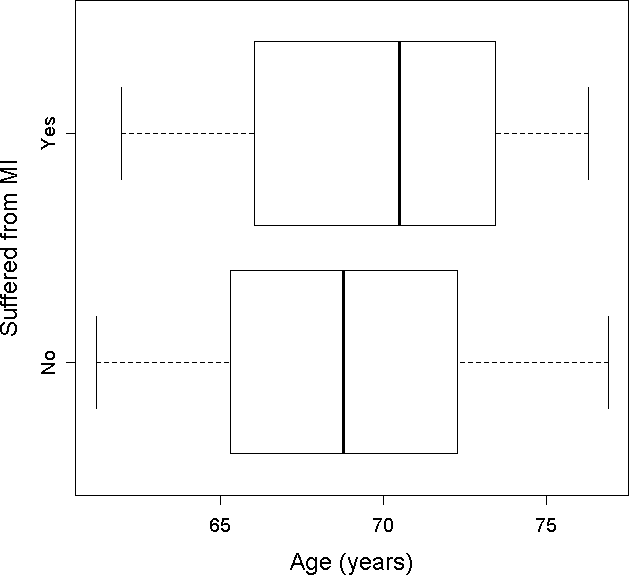
The second response variable under investigation is myocardial infarction (MI) which occurs when the blood supply to a part of the heart is interrupted, causing heart cells to die. It is commonly referred to as a ‘heart attack’, although it is not the only cause of a heart attack. As mentioned in Chapter 1, the mortality rate of diabetic patients after their first myocardial infarction is extremely high. The variable MI is binary, with non-response (or no attack) coded 0 and a response (or an attack) coded 1, and so logistic regression is an appropriate form of analysis to take. As mentioned in the introduction, the information about MI was collected a year before many of the physical measurements were taken, so some time after the event data were collected. In fact, the events generally happened at a time far before the questionnaire was completed. The assumption underpinning the subsequent analysis is that the historical event has not changed any subsequent physical measurements, which is unlikely to be strictly true. However, expert advice informed us that a year is not considered to be a long period of time in terms of change in physical measurements for cardiovascular disease, so the current approach to analysis was little different to that undertaken for other accepted analyses investigating associations between disease prevalence and risk factor levels. The following analysis is split similarly to the analysis for ABI. Firstly, confounding factors have been modelled using logistic regression. Then the explanatory variables of interest have been added in order to investigate any potential relationships between them and the response.

Of the 940 patients who remained in the study after the Year-1 follow-up, 808 (or 86%) had not suffered from MI and 132 (or 14%) had. A series of boxplots highlight any differences in response and non-response for each continuous confounding factor (see Figure A6 in the Appendix). There are a number of shifts in median values which could indicate differences. For example, it appears that older patients are more likely to have suffered from MI,

compared to younger patients (Figure 3.2.1). However, there is extensive overlap in all the plots which means that any true difference is hard to discern at this point in the analysis. In order to obtain an overview of the relationships between MI and the categorical confounding factors proportion tables have been produced (see Tables A2 to A7 in the Appendix). There does appear to be some moderate relationship between MI and sex: of the patients who did have an attack a much higher proportion (77.3%) were male (Table 3.2.1). The distribution of deprivation score across responses and non-responses is very similar and the same can be said for treatment by diet, insulin and tablets. Therefore these tables do not highlight any noticeably strong relationships between MI and SIMD or type of treatment. Finally, there may be a relationship between MI and smoking status (Table 3.2.2). Of the patients who had not suffered from MI, 12.5% were smokers, 46.0% were ex- smokers and 41.5% had never smoked. Of the patients who had suffered from MI, 15.9% were smokers, 51.5% were ex-smokers and 32.6% had never smoked. This suggests that smoking is a higher risk for MI, even for those who have quit.

|  |  |  |
| --- | --- | --- |
|  | **Male** | **Female** |
| **No** | 0.479 | 0.521 |
| **Yes** | 0.773 | 0.227 |

**Table 3.2.1: Proportions Table of response/non-response of MI by Sex (*figures to 3 d.p.*)**



|  |  |  |  |
| --- | --- | --- | --- |
|  | **Current smoker** | **Ex-smoker** | **Never smoked** |
| **No** | 0.125 | 0.460 | 0.415 |
| **Yes** | 0.159 | 0.515 | 0.326 |

**Table 3.2.2: Proportions Table of response/non- response of MI by Smoking Status (*figures to 3 d.p.*)**

**Figure 3.2.1: Boxplot of response/non-response of MI by Age**

Starting with a full additive logistic regression model, AIC scores were compared for each possible model with one variable removed. The final reduced model selected includes the terms age, dBP, HDL, sex, treatment by diet control and treatment by tablets:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **2.5%**  **Lower Interval** | **97.5%**  **Upper Interval** | **p-value** |
| **Intercept** | 0.289 | 2.030 | -3.708 | 4.261 | 0.887 |
| **Age** | 0.053 | 0.025 | 0.005 | 0.102 | 0.031 |
| **dBP** | -0.041 | 0.011 | -0.064 | -0.019 | <0.000 |
| **HDL** | -1.199 | 0.355 | -1.918 | -0.524 | 0.001 |
| **Sex [Females]** | -1.242 | 0.238 | -1.724 | -0.788 | <0.000 |
| **Controls by Diet [Yes]** | -0.912 | 0.416 | -1.719 | -0.080 | 0.028 |
| **Takes Tablets [Yes]** | -1.021 | 0.367 | -1.723 | -0.274 | 0.005 |

**Table 3.2.3: Summary of Logistic Model for MI (*figures to 3 d.p.*)**

Other than that for the intercept, all p-values are less than 0.05 which indicates that all terms have some statistically significant effect on the response, MI, in addition to each other. The confidence intervals are all very narrow which provides a high level of confidence in the coefficient estimates. The log odds, and hence the odds, of response increase as the age of the patient increases. The log odds, and hence the odds, of response decrease as dBP and HDL increase. Females have lower log odds of response compared to the baseline, males. In comparison to patients who do not control their diabetes by diet, those who do have lower log odds of response. Similarly, in comparison to patients who do not control their diabetes with tablets, those who do have lower log odds of response.

In order to check that this first model is appropriate, residuals plots and goodness-of-fit tests were investigated. There are no points higher than two in absolute value in the plots of deviance and Pearson residuals (Figure 3.2.2). This means that no observations have been highlighted as potential outliers which could be influencing the model. A normality plot (Figure 3.2.3) shows considerable deviation from the normal line, particularly at the upper tail of the data, although it should be noted that residual checks for logistic regression are tentative and do not tend to give appropriate results. A series of Hosmer-Lemeshow tests were completed comparing the hypotheses in 2.2.3.1. Tests were performed using different numbers of groups (g=6, 10 and 14) and all returned non-significant p-values much greater than 0.05. Therefore there is insufficient evidence to reject the null hypothesis and it can be concluded that the model is a good fit to the data. Although the usual R² value is not given for a logistic regression, an alternative generalisation is the Deviance Explained. For this model, the Deviance Explained is calculated as 0.1098 which means that 10.98% of the variability in the data is described by the model. Similar to the results for ABI, this is a very low value so the model has poor predictive power. The additional assumption of independence is considered to hold by the study design since measurements are taken on different patients.



0

200

400

600

800

Index

0

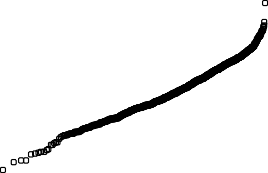
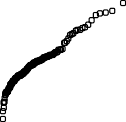
200

400

600

800

Index



**Normal probability plot of standardised deviance residuals**

-3 -2

-1

0

1

2

3

Theoretical Quantiles

Deviance residuals

0.0

0.5

1.0

Sample Quantiles

0.5

1.0

**Figure 3.2.2: Residual Plots for Logistic Model (Top: Deviance Residuals; Bottom: Pearson Residuals)**

**Figure 3.2.3: Normal Q-Q Plot for Logistic Model**

Pearson residuals

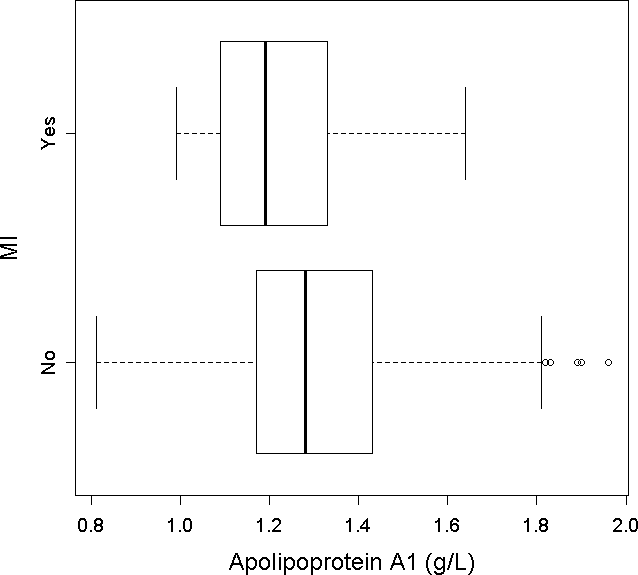
0.0

0.5

1.0

0.0

The next stage of analysis was to explore any potential relationships between the response, MI, and the explanatory variables of interest. Boxplots show some shifts in median values, although there is again considerable overlap meaning that the extent of any difference is not clear. For example, Figure 3.2.4 suggests that Apolipoprotein A1 is lower for response patients compared to non-response patients (see Figure A7 in the Appendix for full set of plots). Finally, the explanatory variables of interest were added individually to the above model with the intention that any statistically significant terms would be retained in the final model. However, no significant relationships were highlighted as all p-values were greater than 0.05. Therefore the final and



**Figure 3.2.4: Boxplot of response/non- response of MI by Apolipoprotein A1**

most appropriate model for MI is the previously discussed model which includes only a selection of known confounding factors. The reason for this result has been considered and is most likely due to the fact that the basic model adequately describes a substantial amount of the variability in the data and so the specified variables of interest do not provide any additional information. Including, for example, Apolipoprotein A1 in the above model may be redundant as HDL explains the variation due to lipids. The only logical reason for including the explanatory variables of interest instead of the confounding factors would be if they are considerably easier or cheaper to measure. The confounding factors age, dBP, HDL, sex and treatment type are routinely and easily measured or determined. Hence the final model is indeed the most suitable.

The following table shows the statistically significant relationships in terms of odds of response (i.e. the exponential of the estimates and confidence intervals in Table 3.2.3):

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Estimate** | **2.5% Lower Interval** | **97.5% Lower Interval** |
| **Intercept** | 1.335 | 0.025 | 70.868 |
| **Age** | 1.055 | 1.005 | 1.107 |
| **dBP** | 0.960 | 0.938 | 0.981 |
| **HDL** | 0.301 | 0147 | 0.592 |
| **Sex [Females]** | 0.289 | 0.178 | 0.455 |
| **Controls by Diet [Yes]** | 0.402 | 0.179 | 0.923 |
| **Takes Tablets [Yes]** | 0.360 | 0.179 | 0.760 |

**Table 3.2.4: Odds of response calculated from the Logistic Model Summary (*figures to 3 d.p.*)**

For every year increase in age, the odds of MI in a diabetic patient increase by 5.5%. Although this is not a drastic increase, it is to be expected given the known link between elderly people and cardiovascular disease (*Lakatta, 2002*). For every unit increase in HDL, the odds of MI decrease by 69.9%, which is realistic given the previous discussion on the protective properties of this ‘good’ cholesterol. For a female, the odds of MI decrease by 71.1%, in comparison to the baseline, males. Again this links back to the previous evidence

that males are at a higher risk of cardiovascular disease than females (*Kardys, et al., 2007*). For a patient who is able to control their diabetes by diet alone, the odds of MI decrease by 59.8%, in comparison to the baseline group of patients who need to use medication (as well as diet) to control their diabetes. For a patient using tablets to treat their diabetes, the odds of MI decrease by 64.0%, in comparison to the baseline group who do not take tablets. Both these results could be due to an intervention, such as medicine or lifestyle changes, which would presumably only be implemented if it is known to reduce the risk of an attack. However, additional information which could be considered but is not available here is the severity of diabetes. Patients who are taking tablets or insulin may be the ones who have the most acute diabetes so an indication of degree of severity of diabetes may provide a different interpretation here, were it available. Finally, for every unit increase in dBP, the odds of MI decrease by 4.1%. Although this is not a strong effect it does seem unusual and could be explained by a similar argument to that discussed in the analysis of ABI: those patients who have had an attack may be on medicine to reduce blood pressure in order to prevent future attacks. Therefore, a treatment effect is being witnessed and although initially it was assumed that there have been no physical changes between the attack and the measurements being taken, this is not strictly true. Hence the model is not ideally suited to prediction, but is an indication that certain interventions that have been implemented are working. Indeed, the relationship between dBP and MI was further investigated through a linear model with dBP as the response and MI as a single categorical explanatory variable. This showed that for patients who have had an attack, dBP decreases by a multiple of 3.034 and corroborates the treatment effect theory. However, the other variables do indicate potential problems which would be expected, for example the increased risk with a higher age.

In order to illustrate the effects of these variables on the MI, a number of probabilities have been calculated for some model patients. Here, a clinically appropriate dBP value is considered to be less than 80 mmHg, a dBP which may be cause for concern is between 80- 89 mmHg and a high dBP is above 90 mmHg (*Kshirsagar, et al., 2006*). A low level of HDL which would be indicative of cardiovascular disease is considered to be less than 1 mmol/L in men and less than 1.2 mmol/L in women (*JBS 2, 2005*).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sex** | **Age** | **dBP** | **HDL** | **Controls**  **Diabetes by diet** | **Takes tablets** | **Probability of MI** |
| Female | 60 | 70 | 1.5 | Yes | Yes | **0.012** |
| Female | 30 | 60 | 2 | Yes | No | **0.006** |
| Male | 60 | 70 | 1.5 | Yes | Yes | **0.042** |
| Male | 80 | 90 | 0.8 | No | No | **0.470** |

**Table 3.2.5: Table of Probabilities for Example Patients**

The probability of an MI in a women aged 30 with low dBP, high HDL, and who treats her diabetes by her diet but does not take tablets, is 0.006. This is extremely low, which is realistic with regards to previous information. The probability of an MI in an older woman aged 60 with a normal dBP value, moderate HDL, and who uses both diet and tablets to treat her diabetes is only marginally higher at 0.012. However, in comparison, a man of the same age, with the same dBP, HDL and treatment regime has a higher probability of an attack at 0.042. Finally, the highest probability of an MI is for an elderly male aged 80 with

high dBP, low HDL and who does not control his diet or take tablets. In this case the probability is 0.470 which would be considered a fairly high risk patient.

In conclusion, the most appropriate model to explain the odds of MI is one that includes the confounding factors age, dBP, HDL, sex and treatment by diet and tablets. Although the predictive power of this model is questionable, the model has been shown to fit the data well and the underlying assumptions have been deemed to hold.

## Principal Component Analysis

The previous models for both ABI and MI are relatively simple and contain variables which are easy to measure and interpret. It was for this key reason that analysis began with the full set of variables. However, since there are a number of original potential covariates and since some of these have previously been shown to be strongly or moderately correlated, it is of interest to investigate whether the dimensionality of the data can be reduced. To do so Principal Component Analysis (PCA) will be used in order to determine if there are linear combinations of the covariates, which can better describe the variability in the data. This approach considers only continuous covariates, which limits the set under analysis to 18 variables (age, duration of diabetes, BMI, cholesterol, sBP, dBP, HDL, HbA1c, Apolipoprotein A1, Apolipoprotein B, FFA, FT3, TT3, FT4, TT4, TSH, testosterone and SHBG). There are alternative methods which can handle discrete or categorical variables, such as Multiple Correspondence Analysis (*Niitsuma and Okada, n.d.*), however these will not be implemented here.

The first check required of the data is to explore the correlations between the 18 covariates. Although there are not groups of variables which are highly correlated, there are a number of pairs of covariates which are correlated. As discussed in Section 3.1, sBP and dBP are moderately positively correlated, with a correlation coefficient of 0.511. Cholesterol and Apolipoprotein B are strongly positively correlated with a correlation coefficient of 0.844. HDL and Apolipoprotein A1 are strongly positively correlated with a correlation coefficient of 0.864. FT3 and TT3 are strongly positively correlated (ρ=0.740) and FT4 and TT4 are strongly positively correlated (ρ=0.800). A series of correlation tests between these pairs of variables compared the null hypothesis that the true correlation is zero to the alternative hypothesis that the true correlation is not zero. In all cases the p-values were highly significant (< 2.2e-16) which provides evidence that the null hypothesis should be rejected. It can be concluded that there is evidence of correlation between each of these pairs. It should be noted that although there are statistically significant correlations between some variables, many of the variables in this set of 18 are extremely weakly correlated with p- values greater than 0.05. Therefore, although it is still hoped that the dimensionality can be reduced, the PCA will not be expected to perform perfectly.

The next check on the specified set of variables is to investigate the standard deviations. In PCA, if the standard deviations are similar then an eigen-analysis on the sample covariance matrix is performed in order to find linear combinations of the variables. However, in this

case the standard deviations are very different, ranging from 0.167 to 19.419. Therefore, here, the sample correlation matrix will be preferred as it will standardise the variables.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Age** | **Duration** | **BMI** | **Cholesterol** | **sBP** | **dBP** | **HDL** | **HbA1C** | **Apo. A1** |
| 4.170 | 6.421 | 5.485 | 0.769 | 17.765 | 8.885 | 0.343 | 1.034 | 0.187 |
| **Apo. B** | **FFA** | **FT3** | **FT4** | **TT3** | **TT4** | **TSH** | **TES** | **SHBG** |
| 0.164 | 0.241 | 0.597 | 2.036 | 0.241 | 18.618 | 1.261 | 6.839 | 19.194 |

**Table 3.3.: Standard Deviations of the 18 continuous variables (*figures to 3 d.p.*)**

A PCA was performed on the 18 variables with the following results:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Component** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** |
| **Standard Deviation** | 1.608 | 1.458 | 1.418 | 1.320 | 1.169 | 1.145 | 1.079 | 1.044 | 0.929 |
| **Proportion of**  **Variation** | 0.144 | 0.118 | 0.112 | 0.097 | 0.076 | 0.073 | 0.065 | 0.060 | 0.048 |
| **Cumulative Proportion** | 0.144 | 0.262 | 0.373 | 0.470 | 0.546 | 0.619 | 0.684 | 0.744 | 0.792 |
| **Component** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** |
| **Standard Deviation** | 0.899 | 0.872 | 0.792 | 0.748 | 0.620 | 0.526 | 0.361 | 0.354 | 0.264 |
| **Proportion of**  **Variation** | 0.045 | 0.042 | 0.035 | 0.031 | 0.021 | 0.015 | 0.007 | 0.007 | 0.004 |
| **Cumulative Proportion** | 0.837 | 0.879 | 0.914 | 0.945 | 0.967 | 0.982 | 0.989 | 0.996 | 1.000 |

**Table 3.3.: Table of PCA Summary (*figures to 3 d.p.*)**

Comp.1 Comp 3 Comp 5 Comp.7 Comp 9

Variances

1 5

2.0

2 5

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Component:** | **1** | **2** | **3** | **4** |
| **Age** |  |  | -0.268 | -0.145 |
| **Duration** |  | -0.202 | -0.161 | 0.109 |
| **BMI** |  | -0.189 | 0.256 | 0.270 |
| **Cholesterol** | -0.425 | 0.270 | 0.179 | 0.138 |
| **sBP** | -0.214 |  | 0.167 |  |
| **dBP** | -0.207 | 0.193 | 0.270 |  |
| **HDL** | -0.427 |  | -0.402 |  |
| **HbA1C** |  |  | 0.160 | 0.181 |
| **Apo. A1** | -0.459 | 0.124 | -0.304 |  |
| **Apo. B** | -0.279 | 0.264 | 0.365 | 0.142 |
| **FFA** | -0.274 | -0.105 |  |  |
| **FT3** |  | -0.115 | 0.257 | -0.602 |
| **FT4** | -0.208 | -0.538 |  | 0.129 |
| **TT3** |  | -0.180 | 0.304 | -0.544 |
| **TT4** | -0.238 | -0.543 | 0.121 |  |
| **TSH** |  | 0.190 |  |  |
| **TES** | -0.212 | 0.114 | -0.119 | -0.293 |
| **SHBG** | -0.147 | -0.166 | -0.320 | -0.168 |

**Figure 3.3.1: Scree Plot of Components**

0.0

0 5

1.0

**Table 3.3.: Loadings of the first four components (*figures to 3 d.p.*)**

One of the main considerations with PCA is deciding how many covariates to retain. Standard methods include choosing a previously specified high percentage of cumulative variation (i.e. 70% or 90%) and including enough components to hit this level; choosing a high percentage of proportional variation (i.e. 10%) and stopping when components explain less variation than this; or deciding when the scree plot ‘levels off’. The first method would lead to retaining at least seven or eight components in order to achieve a satisfactorily high cumulative proportion. However, given that the aim of the PCA is to produce as simplified a model as possible this is not ideal. Furthermore, the variability explained by the previous models in Sections 3.1 and 3.2 was extremely low and therefore it is unlikely that PCA will be able to drastically improve this. Consequently, the decision was taken to follow the second method and retain the first four components which each explain approximately 10% of the variation in the data. Inspection of the scree plot (Figure 3.3.1) confirms that this is a sensible choice as the shape appears to ‘level off’ after a jump from the fourth to the fifth component. This does only provide explanation of 47% of the variability, but given the previous analyses this number was deemed adequate. The fact that many variables are not highly correlated also reassures the acceptance of this level of variation explained. The interpretation of the components can be inferred from their loadings (Table 3.3.?), although often it is difficult to find meaningful interpretations. Component 1 represents a weighted average of most of the variables (cholesterol, sBP, dBP, HDL, Apolipoprotein A1, Apolipoprotein B, FFA, FT4, TT4, testosterone and SHBG). Component 1 is dominated by three measurements of fat, cholesterol, HDL and Apolipoprotein A1 which, which have the largest magnitudes. Component 2 represents a contrast between most of the variables. It is the difference between the average of cholesterol, dBP, Apolipoprotein A1, Apolipoprotein B, TSH and testosterone, and duration of diabetes, BMI, FFA, FT4, TT4, FT3, TT3 and SHBG. Component 2 is dominated by two thyroid hormones, FT4 and TT4, which have the largest magnitudes. Component 3 again represents a contrast between most of the variables. It is the difference between the average of age, duration of diabetes, HDL, Apolipoprotein A1, testosterone and SHBG, and BMI, cholesterol, dBP, sBP, HbA1C, Apolipoprotein B, FT3, TT3 and TT4. There are two measurements of fat that slightly dominate the component (Apolipoprotein B and HDL). Finally, Component 4 represents a contrast between a smaller collection of variables: the difference between the average of age, FT3, TT3, testosterone and SHBG, and duration of diabetes, BMI, cholesterol, HbA1C, Apolipoprotein B and FT4. Component 4 is dominated by two thyroid hormones (FT3 and TT3).

The final stage of the PCA is to re-fit models for the response variables ABI and MI using the chosen components. Firstly, ABI was modelled in a linear regression using the four components, plus any statistically significant categorical variables which could not be included in PCA, with the following results:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **p-value** |
| **Intercept** | 0.853 | 0.023 | <0.000 |
| **Component 1** | 0.009 | 0.005 | 0.078 |
| **Component 2** | 0.021 | 0.005 | 0.000 |
| **Component 3** | -0.012 | 0.006 | 0.030 |
| **Component 4** | -0.022 | 0.006 | <0.000 |
| **Smoking [Ex-smoker]** | 0.126 | 0.026 | <0.000 |
| **Smoking [Never smoked]** | 0.192 | 0.025 | <0.000 |

Unfortunately, not all the components are returned as having a statistically significant effect on ABI. Components 2, 3 and 4 have p-values which are less than 0.05 which means that they do have some significant relationship with ABI, quantified by the regression coefficient estimates. However, Component 1 has a p-value slightly greater than 0.05. The only categorical variable with a statistically significant effect on ABI (comparing full and reduced models using AIC) is smoking status. Both levels have a highly significant p-value: for an ex- smoker, ABI increases by 0.126 compared to a smoker; for a patient who has never smoked, ABI increases by 0.192 compared to a smoker. Furthermore, the R² value for this model is 0.1217 which is only fractionally higher than the R² value for the final model selected in Section 3.1. Only 12.17% of the variability in the data is explained by this model which gives the model poor predictive power. Similar to before, the normal plot shows deviation at the tails of the data (see Appendix). The residual versus fitted values plot shows a random scatter of points which indicates that constant variance and linearity (see Appendix). However, the interpretation of these components is incredibly complex and would not be straightforward to use in a clinical setting, which far outweighs the benefit in the marginal increase in variability explained. Due to the strength of the underlying assumptions and the simplicity of the interpretation, the final model in Section 3.1 is the best model for ABI that has been found.

Secondly, MI was modelled in a logistic regression using the chosen components. No categorical variables were found to have a statistically significant relationship with MI, although it should be noted that sex was only just not significant with a p-value of 0.070. Hence the results are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Estimate** | **Std. Error** | **p-value** |
| **Intercept** | -2.220 | 0.147 | <0.000 |
| **Component 1** | 0.378 | 0.089 | <0.000 |
| **Component 2** | -0.218 | 0.100 | 0.029 |
| **Component 3** | 0.136 | 0.101 | 0.178 |
| **Component 4** | -0.078 | 0.100 | 0.436 |

**Table 3.3.: Summary of Logistic Model for MI including Components from PCA**

Again, not all the components have a statistically significant effect on MI. Components 1 and

2 have p-values less than 0.05, but Components 3 and 4 do not. The effects of the components on the log odds of MI are quantified by the coefficient estimates. The underlying assumption of normality seems equally as reasonable as for the model Section

3.2. However, the residual plots highlight some concerns: there are indications of outliers as a number of data points have an absolute value greater than two in both the deviance and Pearson residual plots (see Appendix). Furthermore, the Deviance Explained for this model is calculated to be 0.0582 which means that 5.82% of the variability in the data is explained by this model. In comparison to the model in Section 3.2, this is far worse, despite the components themselves explaining a fair amount of the variation. Similar to the above argument, the final model selected using the original variables is the most appropriate model which has been found. This model is straightforward to interpret, the underlying assumptions are more reasonable and a higher amount of variation is explained.

# Chapter 4: Conclusion

## Conclusions

The analyses carried out in Sections 3.1 and 3.2 have identified a number of important relationships between the responses and explanatory variables. The results of the linear regression modelling in Section 3.1 showed that there is a statistically significant association between ABI and plasma levels of thyroid hormones, specifically the available thyroxine, FT4. Given this link, further investigation into the relationship between thyroid hormones and cardiovascular disease in diabetic patients would be interesting and should be carried out. There is no statistically significant effect of either sex hormones or the specified lipids of interest on ABI. The confounding factors which are required in a model for ABI are BMI, cholesterol, sBP, dBP, HDL, sex and smoking status. These are known to be risk factors for cardiovascular disease. The final model was deemed to be appropriate in relation to underlying assumptions, particularly the assumptions of linearity, independence and constant variance. The only issue with this model is that it has low predictive power (R²=11.14%). The results of the logistic regression modelling in Section 3.2 showed that there is no statistically significant association between MI and plasma levels of thyroid hormones, sex hormones or the specified lipids of interest. The confounding factors included in the final model are age, dBP, HDL, sex and treatment by diet and tablets. These variables provide sufficient information so that no additional covariates of interest are required. The final model was considered to fit the data well and the underlying assumptions held. Once again the amount of variation explained was low (Deviance Explained=10.98%). Additionally, the initial assumption that the physical measurements had not changed since the attack does not strictly hold as a treatment effect is observed for dBP. However, this does suggest that interventions put in place are working. The remainder of the terms in the model highlight relationships which would be expected and have been previously documented.

These models include variables that are easy to measure and provide results that are easily interpreted. However, out of interest, PCA was applied to the 18 continuous variables in order to discover whether the dimensionality of that data could be simplified. The results of the PCA in Section 3.3 suggested that the first four components should be retained as they each explain a reasonable amount of variation (approximately 10%) and also indicate the levelling off of the Scree Plot. Linear regression and logistic regression were carried out for the responses ABI and MI respectively using these new components, plus the categorical explanatory variables. However, the resulting models are not as satisfactory as the models found using the original variables. The linear model for ABI explains marginally more variability than the model including original variables, although this is not considered to compensate for the complex interpretations resulting from PCA. An interesting result is that smoking status is still highlighted as an important risk factor. The amount of variation explained in the logistic model for MI is lower than the model including original variables and the underlying assumptions do not hold as strongly. Thus, the final and best models that can be found using the techniques described in Methods are the two discussed above which consider the original confounding factors and explanatory variables of interest.

## Discussion and Future Work

The advantages of the data which have been used in the analysis for this project include the large sample size and the availability of data on a wide variety of risk factors, with a satisfactorily low number of missing values. Although there is not a wide range of ages included within the study, it is hopefully representative of the key risk group for type 2 diabetes, elderly patients. The main issue with the data relates to the cross-sectional study design. The time period between the cardiovascular events and the physical measurements being taken may have resulted in changes in risk factors, including treatment effects. Therefore, although most of the explanatory variables highlight expected relationships, the model for MI may also be an indication that certain interventions put in place after an attack are working. The solution to this problem would be complex, expensive and hence unrealistic. It would require a prospective study which follows a large group of diabetic patients for an extended period of time (potentially 20 or 30 years) so that risk factor measurements can be taken over a long period of time prior to any cardiovascular events occurring. There would additionally be no guarantee that many of the chosen study group would suffer from an event, so the study would have to be large to ensure sufficient power. The methods used here and the results obtained are the best that can be expected from data from a cross-sectional study, although additional information which would have been useful to include would have been whether or not patients were taking medication to reduce dBP (and/or cholesterol). This information could have been used to help confirm the treatment effect theory by investigating whether patients who had had an event were receiving medication which reduces dBP, and to create a variable for hypertension (e.g. raised BP or on medication for raised BP) which could have been included in the statistical analysis.

Some further issues which arose during the analyses were the underlying assumption of normality for the linear models and the variability explained by all models. An alternative approach which could be taken with the aim of improving the distributional assumption of the linear models would be to assume a different underlying distribution. However, these alternative distributions are extremely complex to use and hence the approach here was taken to investigate transformations. The variability in the data explained by both linear and logistic regression is low leading to models with poor predictive power. Although the models do provide information on indicators of risk of cardiovascular disease, the models will not be adequate to calculate future predictions. It would be interesting to find out if, from a biological point of view, there are any confounding factors missing which are standard to include in a model and explain a higher amount of variation. Some further suggestions from experts were family history or genetic influences which could be considered. Aside from this there is no real way of improving the variability explained, it should just be noted alongside the models.

In the linear and logistic regression analysis no interactions were investigated in any of the models. This was a conscious decision made before analysis began since interactions typically do not provide meaningful interpretations, particularly in any interactions between two continuous variables. In the papers already published from the Edinburgh Type 2 Diabetes Study some have included analysis of interaction terms. For instance, in a study of the association between depression and complications of diabetes, the interactions

between sex and each variable were tested and significant interactions were included (*Labad, et al., 2009*). Although interactions between a categorical and continuous variables are the most straightforward to interpret, in this case interactions between sex and each variable would not have been sensible. Due to missing values in the measurements of sex hormones in women, after including the full set of variables women essentially drop out of the sample. However, if expert advice suggested other potential interactions of interest then these could be investigated in future work.

On a related note, the method of analysis in this report has been entirely data based. A statistical approach has been taken which considers all possible variables available. Models have then been chosen based on which covariates are statistically significantly related to the response. However, an alternative approach which could be taken would be from a biological point of view. Given prior knowledge some variables may considered to be more important than others. If there are known biologically significant associations then these could be deemed to be essential to be included and modelling could start from this point. Moreover, if the confounding factors are actually mediating factors (factors which might be in a causal pathway between a risk factor and the disease outcome of interest) then adjusting for them may hide what is a biologically important association in relation to, for example, developing new interventions or medications to prevent cardiovascular disease in people with diabetes. However, this approach would involve detailed consideration of which risk factors to include in such a biologically meaningful model and would require extensive medical knowledge.

One point which should be considered in relation to the statistically significant relationship found between FT4 and ABI is the large number of variables available for analysis. Although this result is encouraging and suggests further investigation would be interesting, the statistically significant evidence could have occurred by chance due to the high number of tests carried out.

In the last section of the analysis in this report, PCA was carried in an attempt to reduce the dimensionality of the data. The limitation of PCA in this context is that it only considers continuous variables. As mentioned, there are alternative methods which can handle categorical data, such as Multiple Correspondence Analysis (MCA). MCA represents categorical covariates as an indicator, or dummy, matrix and is applicable to a large set of categorical variables (*Niitsuma and Okada, n.d.*). Further work on this area could include analysis of the full set of covariates using a technique such as MCA in order to obtain improved results in terms of variability explained. It should be remembered, however, that any reduction of dimensionality will result in more complex interpretations so the aim of the analysis should be kept in mind.

Finally, further work which will be very interesting and provide more detailed results is the analysis of the four year follow-up data which was collected in 2010/11. This was not available for the start of this project as it was still being collated, but it would be appropriate to perform a repeated measures analysis. The assumption of independence would no longer hold as there would be a relationship between patients’ Year-1 and Year-4 measurements. However, this data will help to resolve the issue with single time point data encountered in this project.

# References

Al-Qaisi, Nott, King and Kaddoura, 2009. Ankle Brachial Pressure Index (ABPI): An update for practitioners. *Vascular Health and Risk Management*, 5 pp 833-841

Ambrose and Barua, 2004. The Pathophysiology of Cigarette Smoking and Cardiovascular Disease. *Journal of the American College of Cardiology*, 43(10) pp 1731-1737

Diabetes UK, 2004. *Diabetes in the UK 2004: A report from Diabetes UK*. [online] Available at: [http://www.diabetes.org.uk/Professionals/Publications-reports-and-](http://www.diabetes.org.uk/Professionals/Publications-reports-and-resources/Reports-statistics-and-case-studies/Reports/Diabetes_in_the_UK_2004/)

[resources/Reports-statistics-and-case-studies/Reports/Diabetes in the UK 2004/](http://www.diabetes.org.uk/Professionals/Publications-reports-and-resources/Reports-statistics-and-case-studies/Reports/Diabetes_in_the_UK_2004/)

Diabetes UK, 2010. *What is Type 2 diabetes?* [online]

Available at: [http://www.diabetes.org.uk/Guide-to-diabetes/Introduction-to-](http://www.diabetes.org.uk/Guide-to-diabetes/Introduction-to-diabetes/What_is_diabetes/What-is-Type-2-diabetes/) [diabetes/What is diabetes/What-is-Type-2-diabetes/](http://www.diabetes.org.uk/Guide-to-diabetes/Introduction-to-diabetes/What_is_diabetes/What-is-Type-2-diabetes/) [Accessed on 14 March 2012]

Diabetes UK, 2011. *Diabetes in the UK 2011/2012: Key Statistics in Diabetes*. [pdf]

Available at: [http://www.diabetes.org.uk/Documents/Reports/Diabetes-in-the-UK-](http://www.diabetes.org.uk/Documents/Reports/Diabetes-in-the-UK-2011-12.pdf) [2011-12.pdf](http://www.diabetes.org.uk/Documents/Reports/Diabetes-in-the-UK-2011-12.pdf)

Diabetes UK, 2012. *Complications – how to protect yourself*. [online]

Available at: <http://www.diabetes.org.uk/Guide-to-diabetes/Complications/> [Accessed on 14 March 2012]

Gavish, Ben-Dov and Bursztyn, 2008. Linear relationship between systolic and diastolic blood pressure monitored over 24 h: assessment and correlates. *Journal of Hypertension*, 26(2) pp 199-209

Grundy, 2004. Obesity, Metabolic Syndrome and Cardiovascular Disease. *The Journal of Clinical Endocrinology & Metabolism*, 89(6) pp 2595-2600

JBS 2: Joint British Societies’ guidelines on prevention of cardiovascular disease in clinical practice, 2005. *Heart*, 91(5) pp 1-52

Johnson and Duick, 2002. Diabetes and Thyroid Disease: A Likely Combination. *Diabetes Spectrum*, 15(3) pp 140-142

Kardys, Vliegenthart, Oudkerk, Hofman and Witteman, 2007. – The Female Advantage in Cardiovascular Disease: Do Vascular Beds Contribute Equally? *American Journal of Epidemiology*, 166(4) pp 403-412

Kshirsagar, Carpenter, Bang, Wyatt and Colindres, 2006. Blood Pressure Usually Considered Normal is Associated with an Elevated Risk of Cardiovascular Disease. *The American Journal of Medicine*, 119(2) pp 133-141

Labad, Price, Strachan, Fowkes, Ding, Deary, Lee, Frier, Seckl, Walker and Reynolds, 2009. Symptoms of depression but not anxiety are associated with central obesity and cardiovascular disease in people with type 2 diabetes: the Edinburgh Type 2 Diabetes Study. *Diabetologia*, 53(3) pp 467-471

Lakatta, 2002. Age-associated cardiovascular changes in health: impact on cardiovascular disease in older persons. *Heart Failure Reviews*, 7(1) pp 29-49

Lopez and Murray, 1998. The global burden of disease, 1990-2020. *Nature Medicine*, 4(11) pp 1241-1243

Miettinen, Lehto, Salomaa, Mahonen, Niemela, Haffner, Pyorala and Tuomilehto, 1998.

Impact of Diabetes on Mortality After the First Myocardial Infarction. *Diabetes Care*

21(1) pp 69-75

NHS Choices, 2010. *Diabetes.* [online]

Available at: <http://www.nhs.uk/Conditions/Diabetes/Pages/Diabetes.aspx> [Accessed on 14 March 2012]

NHS Scotland, 2010. *Scottish Diabetes Survey 2010.* [pdf] Scottish Diabetes Monitoring Group. Available at:

<http://www.diabetesinscotland.org.uk/Publications/SDS%202010.pdf>

Niitsuma and Okada, n.d. *Covariance and PCA for Categorical Variables* [pdf] Available at: <http://www.dm-lab.ws/www/contents05/covEigGiniRep16.pdf>

Price, Reynolds, Mitchell, Williamson, Fowkes, Deary, Lee, Frier, Hayes and Strachan, 2008.

The Edinburgh Type 2 Diabetes Study: study protocol. *BMC Endocrine Disorders*, 8(18) Available at: <http://www.biomedcentral.com/1472-6823/8/18>

Reynolds, Labad, Strachan, Braun, Fowkes, Lee, Frier, Seckl, Walker and Price, 2010.

Elevated Fasting Plasma Cortisol Is Associated with Ischemic Heart Disease and Its Risk Factors in People with Type 2 Diabetes: The Edinburgh Type 2 Diabetes Study. *Endocrine Care*, 95(4) pp 1602-1208

Stern, 1995. Diabetes and cardiovascular disease. The common soil hypothesis. *Diabetes*, 44(4) pp 369-374

Whiting, Guariguata, Weil and Shaw, 2011. IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030. *Diabetes Research and Clinical Practice*, 94(3) pp 311-321

Wild, Byrne, Smith, Lee and Fowkes, 2006. Low Ankle-Brachial Pressure Index Predicts Increased Risk of Cardiovascular Disease Independent of the Metabolic Syndrome and Conventional Cardiovascular Risk Factors in the Edinburgh Artery Study.

*Diabetes Care*, 29(3) pp 637-642

Williamson, Price, Glancy, Perry, Nee, Hayes, Frier, Look, Johnston, Reynolds and Strachan, 2011. Prevalence of and Risk Factors for Hepatic Steatosis and Nonalcoholic Fatty Liver Disease in People With Type 2 Diabetes: the Edinburgh Type 2 Diabetes Study. *Diabetes Care*, 34(5) pp 1139-1144

Zmuda, Cauley, Kriska, Glynn, Gutai and Kuller, 1997. Longitudinal Relation between Endogenous Testosterone and Cardiovascular Disease Risk Factors in Middle-aged Men. *American Journal of Epidemiology*, 146(8) pp 609-617

# Appendix

|  |  |
| --- | --- |
| **Variable** | **Typical Ranges (*for adults*)** |
| Ankle Brachial Pressure Index | ABI < 0.3 suggests critical vascular disease  0.3 < ABI < 0.5 suggests severe vascular disease  0.5 < ABI < 1.4 considered normal range ABI > 1.4 suggests vascular disease |
| Body Mass Index | BMI < 18.5 may be underweight  18.5 < BMI < 24.9 considered a healthy range 25 < BMI < 35 may be overweight  BMI > 35 classified as obese |
| Cholesterol | Ideally should be ≤ 5 mmol/L |
| Systolic blood pressure | 140 – 210 High blood pressure  110 – 130 Normal blood pressure  50 – 90 Low blood pressure |
| Diastolic blood pressure | 90 – 120 High blood pressure  75 – 85 Normal blood pressure  33 – 60 Low blood pressure |
| Glycated hemoglobin | 4-6% considered a healthy range |
| High-density lipoprotein | Ideally should be ≥ 1 mmol/L |
| Apolipoprotein A1 | 0.9 – 2.0 g/L considered a normal range |
| Apolipoprotein B | 0.6 – 1.33 g/L considered a normal range |
| Free fatty acids | <1.7 mmol/L considered a normal level |
| Thyrotrophin-stimulating hormone | 0.4 – 4.5 mIU/L considered a normal range |

**Table A1: Normal or healthy ranges for a number of key explanatory variables**

ABI

1.0 1.5

ABI

1.0 1.5

ABI

1.0 1.5 2.0

2.0

0.5

2.0

0.5

2.0

0.5

ABI

1.0 1.5

ABI

1.0 1.5

ABI

1.0 1.5 2.0

2.0

0.5

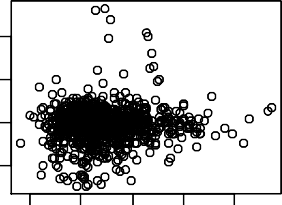
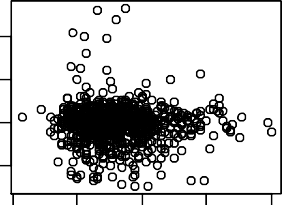
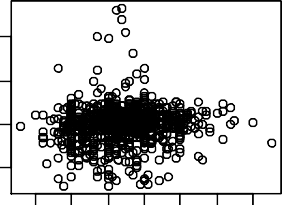
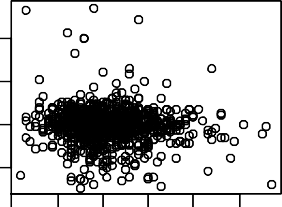
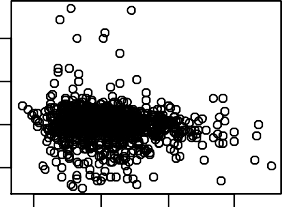
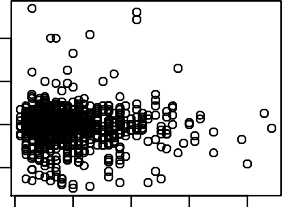
2.0

0.5

2.0

0.5

**Figure A1: Scatterplots between ABI and all continuous confounding factors**



65

70

75

0 10 20 30 40

Age (years) Duration (years)

20 30 40 50

2 3 4 5 6 7

BMI (kg/(m^2)) Cholesterol (mmol/L)

100 140 180 220

50 70 90 110

sBP (mmHg) dBP (mmHg)

4 6

8 10 12

0.5

HbA1C (%)

1.0 1.5 2.0 2.5

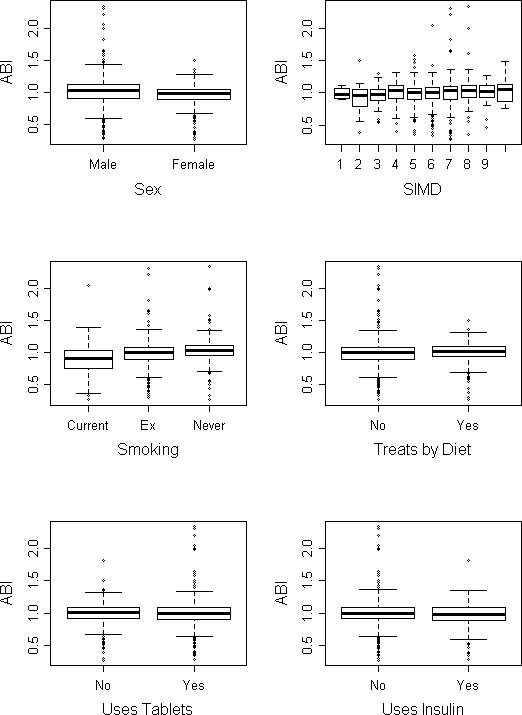
HDL (mmol/L)

ABI

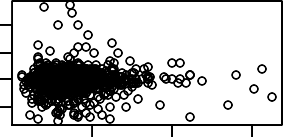
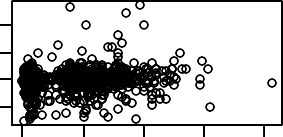
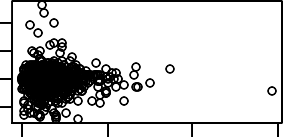
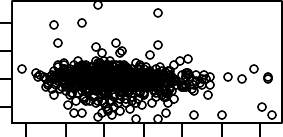
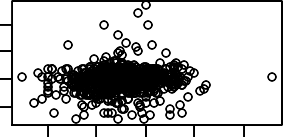
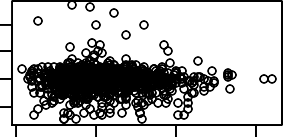
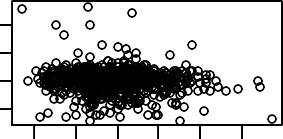
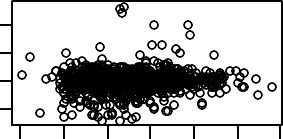
0.5 1.0 1.5

ABI

0.5 1.0 1.5



**Figure A2: Boxplot of ABI by each categorical confounding factor**



0.8

1.2

1.6

2.0

0.4

0.8 1.2

Apolipoprotein A1 (g/L) Apolipoprotein B (g/L)

0.0 0.5 1.0 1.5

3 4 5 6 7

FFA (mmol/L) FT3 (pmol/L)

10 15 20 25

1.0 1.5 2.0 2.5

FT4 (pmol/L) TT3 (nmol/L)

40 80 120 160

0

5

10

15

TT4 (nmol/L) TSH (mIU/L)

0 10 20 30 40

50 100 150

Testosterone (nmol/L) SHBG (nmol/L)

ABI

0.5 1.5

ABI

0.5 1.5

ABI

0.5 1.5

ABI

0.5 1.5

ABI

0.5 1.5

ABI

0.5 1.5

ABI

0.5 1.5

ABI

0.5 1.5

**Figure A3: Scatterplot between ABI and all covariates of interest**

ABI

0.5 1.5

ABI

0.5 1.5

Residuals

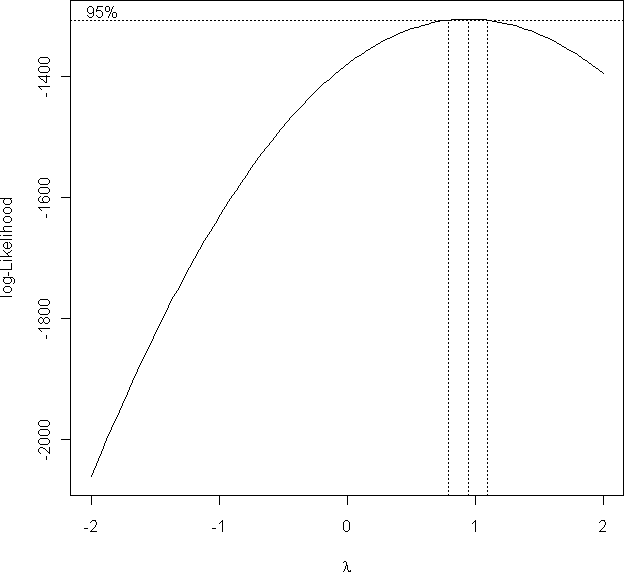
-0.5 0.0 0.5 1.0

-1.0

Standardized residuals

-4 -2 0 2 4 6

-4



**Figure A4: Box-Cox Transformation Graph (*dotted lines indicate 95% CIs for λ*)**

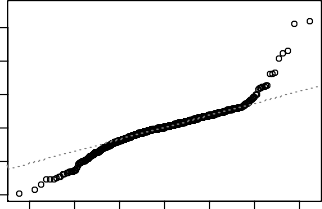
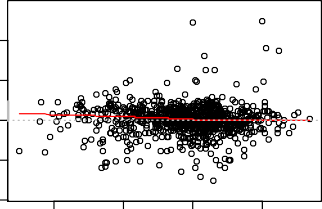
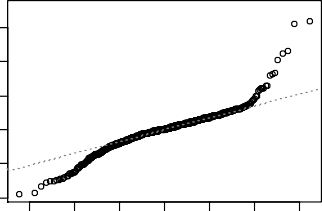
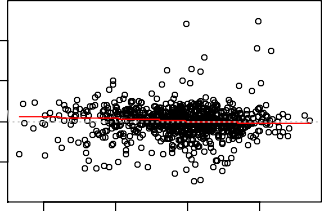
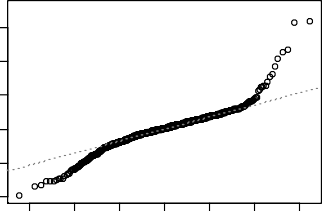
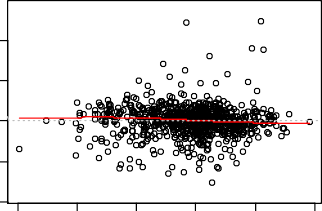
Residuals

-0.5 0.0 0.5 1.0

Standardized residuals

-2 0 2 4 6

**Figure A5: Diagnostic plots for Linear Models of ABI (Top: Full Model; Centre: Reduced Model with sBP removed; Bottom: Reduced Model with sBP and dBP removed)**



Residuals vs Fitted

Normal Q-Q

940

321

32940

133

133

0.7 0.8 0.9 1.0 1.1 1.2

-3 -2 -1 0 1 2 3

Fitted values Theoretical Quantiles

Residuals vs Fitted

Normal Q-Q

940

321

32940

133

133

0.8 0.9 1.0 1.1

-3 -2 -1 0 1 2 3

Fitted values Theoretical Quantiles

Residuals vs Fitted

Normal Q-Q

940

321

32940

133

133

0.8 0.9 1.0 1.1

-3 -2 -1 0 1 2 3

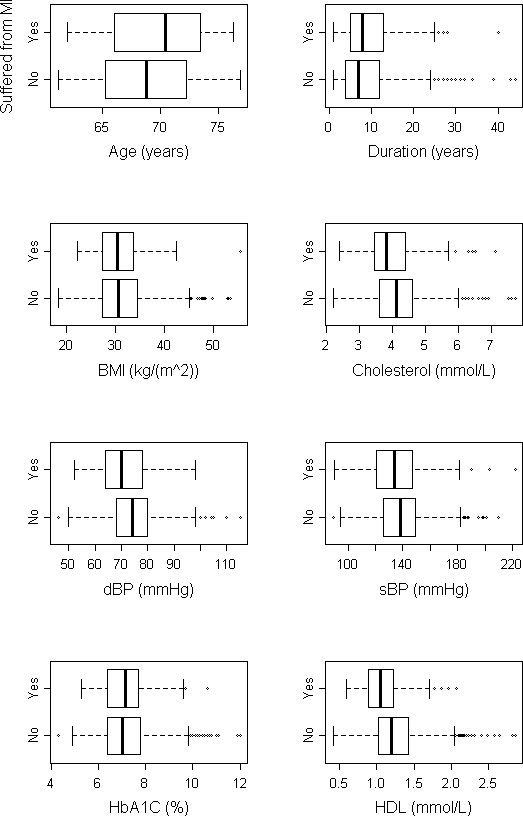
Fitted values Theoretical Quantiles

Residuals

-1.0 -0.5 0.0 0.5 1.0

Standardized residuals

-4 -2 0 2 4 6



**Figure A6: Boxplots of response/non-response of MI by each continuous confounding factor**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Current**  **smoker** | **Ex-smoker** | **Never**  **smoked** |
| **No** | 0.125 | 0.460 | 0.415 |
| **Yes** | 0.159 | 0.515 | 0.326 |

|  |  |  |
| --- | --- | --- |
|  | **Male** | **Female** |
| **No** | 0.479 | 0.521 |
| **Yes** | 0.773 | 0.227 |

**Table A2: Proportions Table of response/non- response by Smoking Status (*figures to 3 d.p.*)**

**Table A3: Proportions Table of**

**response/non-response of MI by Sex (*figures to 3 d.p.*)**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
| **No** | 0.004 | 0.020 | 0.048 | 0.103 | 0.231 | 0.238 | 0.230 | 0.091 | 0.029 | 0.007 |
| **Yes** | 0.008 | 0.017 | 0.099 | 0.099 | 0.198 | 0.289 | 0.198 | 0.066 | 0.017 | 0.008 |

**Table A4: Proportions Table of response/non-response of MI by Deprivation Score (*figures to 3 d.p.*)**

|  |  |  |
| --- | --- | --- |
|  | **Treats by diet** | |
|  | **No** | **Yes** |
| **No** | 0.806 | 0.195 |
| **Yes** | 0.811 | 0.189 |

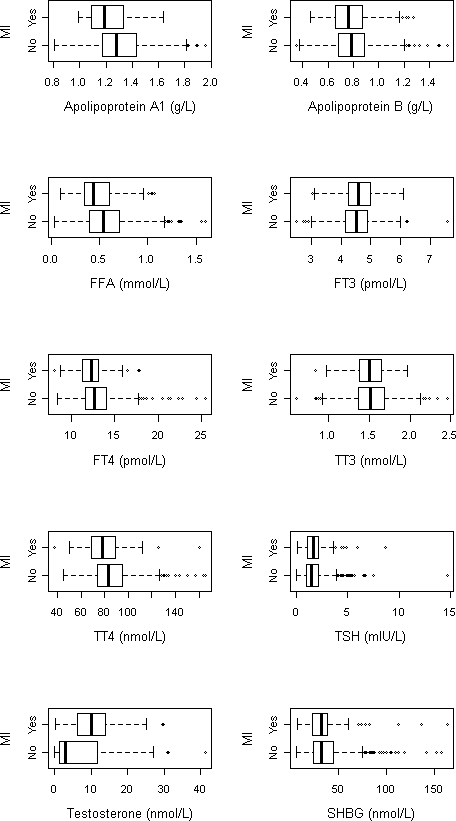
|  |  |  |
| --- | --- | --- |
|  | **Uses tablets** | |
|  | **No** | **Yes** |
| **No** | 0.249 | 0.751 |
| **Yes** | 0.296 | 0.705 |

**Table A5: Proportions Table of response/non-response of MI by Treatment by Diet (*figures to 3 d.p.*)**

**Table A6: Proportions Table of response/non-response of MI by Treatment by Tablets (*figures to 3 d.p.*)**

|  |  |  |
| --- | --- | --- |
|  | **Uses insulin** | |
|  | **No** | **Yes** |
| **No** | 0.849 | 0.151 |
| **Yes** | 0.803 | 0.197 |

**Table A7: Proportions Table of response/non-response of MI by Treatment by Insulin (*figures to 3 d.p.*)**



**Figure A7: Boxplots of response/non-response of MI by all covariates of interest**

**Table 6.8: Correlations between all 18 continuous variables (*figures to 2 d.p.*)**

44

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Dur** | 0.08 |  |  |  |  |  |  |  |  |  | | | | | | | |
| **BMI** | -0.19 | 0.04 |  |  |  |  |  |  |  |
| **Chol** | -0.04 | -0.13 | -0.03 |  |  |  |  |  |  |
| **sBP** | 0.12 | 0.04 | 0.01 | 0.13 |  |  |  |  |  |
| **dBP** | -0.10 | -0.20 | 0.02 | 0.18 | 0.50 |  |  |  |  |
| **HDL** | 0.13 | <0.00 | -0.13 | 0.29 | 0.05 | 0.07 |  |  |  |
| **HbA1C** | -0.12 | 0.22 | 0.09 | 0.03 | 0.04 | 0.04 | -0.10 |  |  |
| **Apo A1** | 0.09 | -0.05 | -0.13 | 0.36 | 0.07 | 0.13 | **0.86** | -0.04 |  |
| **Apo B** | -0.11 | -0.12 | -0.02 | **0.84** | 0.11 | 0.19 | -0.08 | 0.12 | 0.06 |  |  |  |  |  |  |  |  |
| **FFA** | 0.10 | -0.04 | 0.09 | 0.12 | 0.15 | 0.07 | 0.22 | <0.00 | 0.22 | 0.06 |  |  |  |  |  |  |  |
| **FT3** | <0.00 | -0.07 | -0.07 | <0.00 | 0.03 | 0.12 | -0.05 | -0.03 | 0.05 | 0.04 | 0.05 |  |  |  |  |  |  |
| **FT4** | -0.01 | 0.10 | 0.08 | 0.01 | 0.06 | -0.05 | 0.08 | 0.03 | 0.02 | -0.04 | 0.18 | 0.03 |  |  |  |  |  |
| **TT3** | -0.03 | -0.07 | 0.06 | -0.04 | 0.03 | 0.08 | -0.10 | 0.01 | 0.01 | 0.04 | 0.05 | **0.74** | <0.00 |  |  |  |  |
| **TT4** | -0.01 | 0.08 | 0.19 | 0.08 | 0.05 | -0.05 | 0.05 | 0.06 | 0.02 | 0.05 | 0.15 | 0.07 | **0.80** | 0.21 |  |  |  |
| **TSH** | 0.02 | -0.03 | 0.02 | 0.03 | 0.01 | -0.05 | -0.01 | 0.02 | 0.02 | 0.05 | 0.02 | -0.04 | -0.25 | <0.00 | -0.12 |  |  |
| **TES** | 0.04 | 0.06 | -0.29 | -0.12 | -0.11 | <0.00 | -0.14 | -0.04 | -0.18 | -0.07 | -0.17 | 0.10 | -0.13 | <0.00 | -0.25 | -0.09 |  |
| **SHBG** | 0.13 | 0.16 | -0.16 | 0.03 | -0.01 | -0.11 | 0.27 | -0.05 | 0.22 | -0.10 | 0.04 | 0.02 | 0.16 | 0.04 | 0.18 | -0.03 | 0.16 |
|  | **Age** | **Dur** | **BMI** | **Chol** | **sBP** | **dBP** | **HDL** | **HbA1C** | **Apo.**  **A1** | **Apo.**  **B** | **FFA** | **FT3** | **FT4** | **TT3** | **TT4** | **TSH** | **TES** |

0.5

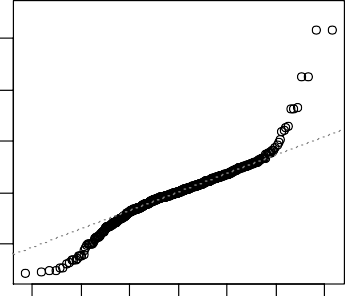
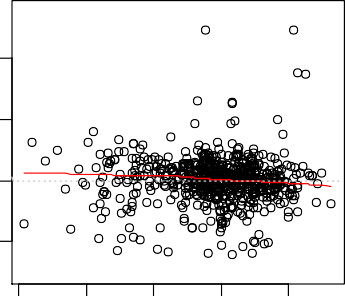
1.0

2

4

6

**Figure 6.8: Diagnostic Plots for Linear Model for ABI including Components from PCA**



Residuals vs Fitted

Normal Q-Q

617 221

22 6 7

98

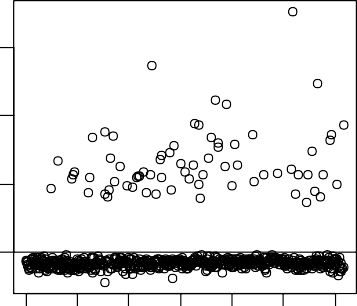
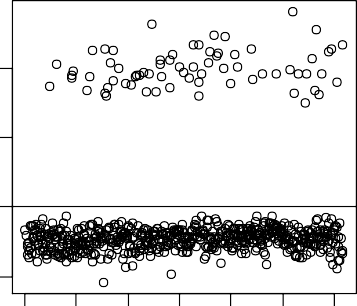
98

0.7 0.8 0.9 1.0 1.1

-3 -2 -1 0 1 2 3

Fitted values

Theoretical Quantiles



0 100

300

500

0 100

300

500

Index

Index

Deviance residuals

Residuals

1

2

-0.5 0.0

Pearson residuals

Standardized residuals

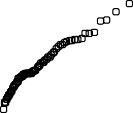
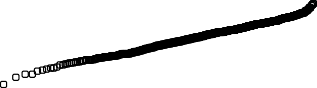
-2

0

4

6

**Figure 6.9: Residual Plots for Logistic Model for MI including Components from PCA**



**Normal probability plot of standardised deviance residuals**

-3

-2

-1

0

1

2

3

Theoretical Quantiles

-1

0

Sample Quantiles

1

2

0

2

**Figure 6.10: Normal Q-Q Plot for Logistic Model for MI including Components from PCA**

-1

0

45