UNIVERSITY
of
GLASGOW

Department of
Statistics

A Statistical Analysis of Low Birthweight in Glasgow

Barbara A Murray

A thesis submitted for a Doctor of Philosophy Degree in Statistics at the University of Glasgow

October 1999

Abstract

The percentage of singleton livebirths resulting in low birthweight deliveries has remained constant in the last 20 years, with between 6 and 10% of singleton pregnancies resulting in such a delivery. Low birthweight infants have been shown to develop medical problems in infancy and childhood, such as visual impairment, lower IQs and neuromotor problems, and as such it is important to identify those pregnancies that may result in low birthweight infants. This thesis considers factors that may be related to low birthweight, and uses these factors in the construction of a model to predict the probability of a woman delivering a low birthweight infant in order to identify high risk mothers.

One factor that may be thought of as being related to low birthweight is deprivation. In this thesis a new deprivation measure is proposed which updates previous work in the area by using the 1991 small area census data to create a continuous deprivation measure, based on postcode area of residence, within the Greater Glasgow Health Board. This new measure of deprivation is included in the model referred to above.

As there are many possible risk factors involved in modelling the probability of delivering a low birthweight infant multiple comparisons are involved in the production of the model and it is important to produce a model that incorporates most of the relevant factors and relatively few of the unimportant factors. The first order Bonferroni bound is one method used to correct for multiple comparisons by giving an upper bound on the actual p-value. This thesis considers the second order Bonferroni bound which gives a lower bound on the p-value and, when used in conjunction with the first order bound, gives a better correction method than the first order bound alone. These two bounds are then extended into logistic regression models.

Acknowledgements

I would like to acknowledge three years of funding from the Engineering and Physical Sciences Research Council during the course of this work.

I would also like to thank the following people for their support over the last few years:

My supervisor, Professor Adrian Bowman, for his encouragement and advice throughout, and the statistics department at Glasgow University, in particular Eileen Wright, Graeme Archer, Val Easton and Mike Ashcroft. Dr. Gill McIlwaine and Dr. John Kingdom for their invaluable medical knowledge, John Womersley and Andrena McMillan for giving me an insight into the production of the Womersley scoring system, and Dr. Susan Cole at the Information and Statistics Division of the Scottish Health Service for making the linked SMR2 data set available. The statistics department at Quintiles have been understanding over the last three years, in particular Sue McKendrick, who was always on hand with comments and advice, both practical and statistical, and Mark Beard, who understood the need for me to see my supervisor regularly. Finally, this would not have been completed without the love and support of Graham Collins, and my parents, Annie and Stan, whose love, understanding and support throughout my university career has been invaluable.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Approximately 6 to 10% of singleton pregnancies result in a low birthweight infant, a proportion that has changed very little in the past 20 years. There are many possible factors affecting birthweight, one of which may be deprivation. This area is of interest as low birthweight may cause an extended stay in hospital for mother and infant, and will affect the resource usage within hospitals. There is also the likelihood of the infant being placed in a special care baby unit (SCBU) and having an increased risk of mortality and morbidity. If the probability of low birthweight can be accurately modelled and factors relating to birthweight identified, then high risk mothers can be identified and monitored carefully throughout the pregnancy. This thesis is concerned with modelling the probability of delivering a low birthweight infant.

1.1.1 Low Birthweight

Low birthweight infants, and in particular those very low birthweight infants born under 1500 g, have been shown to face problems in later life. Examples of these problems are visual impairment, lower IQs, and neuromotor problems, and as such it is important to identify these pregnancies which may result in low or very low birthweight infants.

In order to investigate factors related to low birthweight, all births in the Greater Glasgow Health Board area between 1981 and 1991 are considered. Attention is restricted to the Greater Glasgow Health Board as this area has the largest number of high deprivation sectors in Scotland. It is hoped that this will increase understanding of the role deprivation plays in birthweight. Data on births in this area were collected from all six hospitals in the area on a form known as an 'SMR2' form. This form is completed for each hospital visit during pregnancy and records maternal data, previous pregnancy data, and details of the current pregnancy, including infant data if the hospital visit results in delivery of an infant.

Births shall be categorised as low birthweight (under 2500 g), and normal birthweight (2500 g or more). In some cases low birthweight shall be sub-categorised as very low birthweight (under 1500 g) and low birthweight (1500 g to 2499 g), as very low birthweight may be thought of as being caused by obstetric problems in the mother and therefore as being distributed differently to those births between 1500 g and 2499 g. Factors shown previously to be related to low birthweight will be investigated in order to see if they are also related in the population of interest.

This work differs from that carried out previously in that there is a second data set which has been 'linked'. Each woman is given a unique maternal identification number and as a result all forms for each pregnancy can be identified and linked to other pregnancies to the same mother.

1.1.2 Defining a Glasgow Specific Deprivation Measure

Social class has recently been thought of as an unreliable measurement of deprivation, and various techniques have been explored in order to find an improved method for measuring deprivation. Several measurements have been proposed, the best known being the Carstairs score [10], the Townsend score [45], and the Jarman index [26, 27]. Jarman and Townsend are calculated for areas in England, while the Carstairs score is derived for all postcode sectors in Scotland. The postcode sector can be derived from the postcode by removing the final two letters of the postcode, for example, if the postcode is G74 3HT, then the postcode sector is G74 3. The Carstairs score is calculated from the 1981 small area census data, which records the proportion of households in each postcode sector satisfying various criteria. Examples of these criteria are households with 1 or 2 rooms, with no car, with no children, with pensioners living alone and with male unemployment. For each sector, the Carstairs score is calculated as the sum of the standardised proportions of households with low social class, male unemployment, overcrowding and car ownership. This score is then categorised from 1 (most affluent) to 7 (most deprived). 32 (23.7%) of the post-code sectors in Greater Glasgow Health Board fall into the most deprived Carstairs category.

A Glasgow specific deprivation score will be defined for two reasons. Firstly, in creating a Glasgow specific score it should be possible to separate those 32 post-code sectors in Glasgow which Carstairs categorises 'most deprived'. Secondly, as the population of interest is geographically constrained to Greater Glasgow Health Board (GGHB), the population involved in the deprivation calculations should be similarly constrained in order to give a clearer perspective of the relationship between birthweight and deprivation in the GGHB area.

1.1.3 Variable Selection

In regression models there are often a large number of explanatory variables, some of which may be related to the response variable, some which may be correlated with each other, and some which may be neither. In these cases, it is preferable to fit a model that adequately describes the data without including variables that are unnecessary or misleading.

There are various selection techniques that can be used to produce a model containing a subset of these explanatory variables. The most common techniques are backwards elimination and forward selection. These techniques involve multiple comparisons at each step in the procedure, and a correction method should be employed to produce a bound on the true level of significance for a variable entering or leaving the model, based on the number of variables involved. The first order Bonferroni bound is one possible correction method. This produces an upper bound on the actual level of significance for fitting the variable of interest. If this Bonferroni bound is not used, the final model selected will tend to have too many variables in the model, which may cause unnecessary measurements to be taken if the model is to be used for prediction of further cases.

As there are many possible explanatories involved in the case of low birthweight, variable selection techniques and Bonferroni corrections will be applied in the selection of a model for the probability of delivering a low birthweight infant.

1.2 Aims

The aims of this thesis are:

- To investigate possible factors related to low birthweight using univariate methods. This shall update previous work in the area and be carried out using data from all births in the Greater Glasgow Health Board area between 1981 and 1991.
- To produce a model for the probability of delivering a low birthweight infant based on maternal factors. This shall also use data from all births in the Greater Glasgow Health Board area between 1981 and 1991.
- To create a deprivation measure for the Greater Glasgow Health Board area, and to compare this with current methods. The 1991 small area census data shall be used in the derivation of this measure.
- To investigate the relationship between birthweights of the first two children to each mother. The birthweight of a first child is thought to have an effect on the birthweight of a second child and the linked data set will be used to investigate this.
- To investigate the use of Bonferroni bounds in regression models assuming Normally distributed data. The first order Bonferroni bound is used in regression modelling with variable selection to give an upper bound on the p-value obtained by adding the most significant variable into the model, in order to correct for multiple comparisons. If this bound is not used too many variables may be added to the model. In the same way, the second order Bonferroni bound can be used to give a lower bound on the same p-value and these can be used together to give a better estimate of the p-value.
- To extend the use of Bonferroni bounds into logistic regression models, and use these bounds to model the probability of delivering a low birthweight infant.

1.3 Outline

The contents of each chapter are as follows:

Chapter 2: Overview of Low Birthweight. Low birthweight may be due to pre-term birth or to the infant being small for their gestational age. This chapter gives an overview of work previously carried out in these fields and discusses the factors thought to be related to low birthweight, such as maternal age and height, marital status, the number of previous pregnancies to the mother and their outcomes, and smoking history. In addition, there is a brief overview of the disadvantages low birthweight infants encounter in later life, such as visual impairment, lower IQs and neuromotor problems.

Chapter 3: Analysis of Unlinked Data. Low birthweight within the geographical constrains of Greater Glasgow Health Board is considered. This updates previous work carried out on birthweight data [35]. Various maternal factors, such as maternal age and height, deprivation category of the area lived in, and marital status are considered univariately to investigate their relationship with birthweight. This is carried out by taking a data set where the SMR2 forms have not been linked by maternal identification number, and considering data from those visits where delivery occurred. A regression model of the probability of delivering a low birthweight infant is then fitted.

Chapter 4: Deprivation Measures. This chapter discusses several methods for assigning a deprivation score to a small area. Various scoring systems have been created in order to measure small area deprivation. A new deprivation measure for postcode sectors in the Greater Glasgow Health Board area is proposed using the 1991 small area census data. This new measure is compared with measures that have previously been calculated for the same sectors.

Chapter 5: Analysis of Linked Data. The aim of this chapter is to consider mothers who delivered their first child in 1980 and subsequently had a second child in the twelve year period 1980 to 1991. These pregnancies will be investigated to determine whether the birthweight of the second child is related to the birthweight of the first child. The outcome of previous pregnancies is expected to have an effect on the birthweight of subsequent pregnancies.

Chapter 6: Variable Selection. The aim of this chapter is to consider several methods that can be used in forward subset selection to decide which variables should be included in a regression model. Several types of subset selection and stopping rules that can be used in selection procedures will be discussed. The sizes and powers of these stopping rules and their ability to select the best variable for entry into a model are then investigated and compared using simulation studies.

Chapter 7: Bonferroni Bounds in Variable Selection. This chapter focuses on how to fit the 'best' model to a set of data, taking into account the effect of multiple comparisons on the overall p-value of any tests that are carried out in order to determine which explanatory variables should be included. A method is introduced to calculate both upper and lower bounds on the p-value to test whether a variable should be added when a linear regression model is fitted. This method is then extended to the logistic regression case and is used to fit a model of the probability of delivery a low birthweight baby using 1991 data.

Chapter 8: Discussion and Conclusions. This chapter will discuss the results of this thesis and draw conclusions. Ideas for further work will also be suggested.

Chapter 2

Overview of Low Birthweight

Low birthweight is one if the few obstetrical problems that has not decreased in proportion in the last 20 years, with approximately 6 - 10% of all births resulting in a low birthweight baby. There are several maternal factors that may be though of as having an adverse effect on birthweight, including maternal age, maternal height, and active or passive smoking. This chapter considers previous literature where the relationships between these factors and birthweight have been investigated. In addition, previous literature has discussed the disadvantages which low birthweight infants face in later life and this is also considered here.

2.1 Maternal Risk Factors

2.1.1 Maternal Age

Maternal age is a factor that is often thought of as being a high risk factor for delivering a low birthweight baby. Rosenberg and McEwan [42] summarised the trends and risks of teenage pregnancy in Scotland using data from between 1975 and 1988. Data for the latest of these years, 1988, indicated that low birthweight and pre-term delivery were slightly more common in teenagers (8% and 7%) than in the 20-24 year old age bracket (7% and 6%). However, Rosenberg and McEwan state that no attempt was made to correct for social deprivation in this assessment and that if this was added to the analysis it would be expected that there would be no difference between the age groups.

De Sanjose and Roman [14] examined the effect of various maternal factors on low birthweight, pre-term delivery, and small for gestational age births, using data from 1981 to 1984 in Scotland. Maternal age was split into 4 categories - under 20, 20-29, 30-34 and over 34. Low birthweight was shown to have a 'U'-shaped relationship with age, with the under 20 and over 34 age groups having the highest proportions of low birthweight deliveries, and this was shown to be significantly higher than for those mothers aged 20-29.

Bakketeig et al. [2] attempted to categorise those mothers who repeat small-for-gestational-age (SGA) births, and compare them with mothers who have had either no SGA births, or only one SGA birth. The data used were those women in Norway who had delivered after 16 weeks of gestation between 1967 and 1976. Mothers were identified by a unique maternal identification number, making linkage over successive pregnancies possible. As no information on social conditions was collected by the Medical Registration of Births, record linkage was carried out for births between 1970 and 1973 with the 1970 Norwegian census data. Mothers were excluded from the analyses if, for at least one pregnancy, gestational age or birthweight were unknown, or gestational age was less than 28 weeks or greater than 46 weeks. Births were excluded if congenital abnormalities were recorded in the birth registry, or if the mother had recorded diseases before or during pregnancy. SGA was defined as births with birthweight for gestation below the 10th percentile. Mothers were included in this analysis if they had three successive singleton births in the study period. Maternal age was categorised as under 20, 20-34 and 35 or over. As maternal age increased, the proportion of SGA deliveries decreased from 11.6% to 8.5%. In addition, if age at first delivery was considered, the relative risk of delivering one SGA birth was significantly higher in the under 20 age group than in the reference group of 20-34 year olds. However, the relative risk for more than one SGA birth was significantly higher for both the under 20 and over 35 age groups, and in both cases the risk of more than one SGA birth was 1.2 times higher than in 20-34 year olds.

Rodriquez et al. [41] investigated the effect of different social and demographic factors on low birthweight and used data from the National Institute of Statistics, covering births in Spanish provinces in 1988 and using only those provinces where over 99.5% of births recorded had a record of the infant's birthweight. Low birthweight was split into two categories, pre-term, defined as less than 37 weeks gestation, and term, defined as between 37 and 42 weeks gestation, and these were considered separately. Maternal age was split into 5 categories - under 20, 20-24, 25-29, 30-34 and over 34. For pre-term low birthweight, Rodriquez at al. also showed that after adjustment for variables of interest, low birthweight had a 'U'-shaped relationship with age, with the under 20, 30-34 and over 34 groups having the highest proportions of low birthweight deliveries. Of these the 30-34 and over 34 groups were shown to be significantly higher than mothers aged 25-29. However for those low birthweight deliveries at term there was no evidence of a 'U'-shaped relationship and the proportion of low birthweight deliveries decreased as maternal age increased, with both the under 20 and 20-24 age groups being significantly higher than the 25-29 age group.

Rasheed and Rahman [37] considered sociodemographic, biological, genetic and medical factors as possible predictors of Saudi Arabian birthweight. The data used were those of all singleton livebirths delivered between October 1985 and September 1986 at the King Fahd Hospital of the University in Al-Khobar. Only those cases with a previous sibling birth were considered. Maternal age was split into the same categories as Rodriquez et al. with the under 20 category being described as 12-19. After a multiple regression of various factors on birthweight was carried out, maternal age was shown to have a significant effect on birthweight, which increased as maternal age increased, with the 25-29 and over 35 categories being significantly higher than the 20-24 age group.

Lang et al. [31] estimated the effects of 23 possible risk factors on the prevalence of premature labour and fetal growth retardation, both of which may occur with or without low birthweight. The data used were those of all singleton pregnancies at the Boston Hospital for Women between August 1977 and March 1980. The risk factors were studied among small-for-gestational-age babies born at term (10,889 cases), and premature births with an appropriate size for gestational age outcome (9,490 cases). All livebirths between 22 and 45 weeks of gestation were included if birthweight and sex of the infant were also recorded and if the mother did not suffer from chronic disease before pregnancy. Maternal age was split into 5 categories - under 16, 16-19, 20-24, 25-34 and over 34. Pre-term delivery was defined as being less than 37 completed weeks of gestation, and small-for-gestational-age (SGA) was defined as having a birthweight less than the 10th percentile for gestation using sex-specific criteria. Multivariate regression was used to adjust for possible confounding effects of other risk factors. Using this model, all maternal age groups had a greater risk of pre-term labour than the 25-34 age group, although in no case was this risk statistically significant. The under 16 and 20-24 age groups also had a greater risk of a term SGA delivery than the 24-34 age group, while the 16-19 and over 34 age groups had a smaller risk; again none of these risks were significantly different to the 24-34 age group.

An antenatal screening program was introduced in the African republic of Cape Verde in 1977 and Wessel et al. [46] carried out a prospective study to evaluate antenatal risk screening by relating possible maternal risk factors to pre-term birth and low birthweight. The data used were that of pregnant women presenting for antenatal care between October 1991 and December 1992. A cohort of 358 women was selected, which was reduced to 353 women after restricting the study to singleton births only. Pre-term delivery was defined as less than 37 weeks completed gestation, calculated from last menstrual period. Low birthweight was defined as less than 2500 g. In 30 cases birthweight was determined after 7 days and birthweights for 23 infants were unknown. Maternal age was categorised as 15-19, 20-24, 25-29, 30-34 and 35 or over. Adjustments were made using multiple regression analysis to take into account confounding variables. As there were low numbers of both pre-term birth and low birthweight (44 and 27 respectively), maternal age was analysed as 15-19 and 20 or over. While the adjusted relative risk for pre-term delivery in the 15-19 age group was 1.6, this was not significantly higher than the control group of 20 or over. The adjusted relative risk for a low birthweight delivery for the 15-19 age group was 3.7, significantly higher than that for the 20 and over group.

The above studies seem to suggest that low birthweight is a higher risk in teenage mothers than mothers in their twenties, and also in mothers who are in their thirties, while the risk of small-for-gestational-age infants decreases as maternal age increases. As many of the results are adjusted for known and expected obstetric and socio-economic factors, it appears that maternal age is an important factor in the birthweight of infants and that teenage mothers run a high risk of delivering either and SGA infant or a low birthweight infant.

2.1.2 Previous Pregnancies

Bakketeig et al. [2] investigated the risk of a small for gestational age (SGA) delivery to mother who have delivered 0, 1 or 2 previous SGA babies. For mothers whose first delivery was SGA, the relative risk of delivering a subsequent SGA birth was 3.27, significantly higher than the reference category of those mothers whose first delivery was not SGA. In addition, those mothers with two previous births that were not SGA had a significantly lower risk of delivering a subsequent SGA infant (0.68). The risk of a subsequent SGA baby after a first birth that was SGA and a second that was not SGA was significantly higher than the reference category (2.12), with the risk for those mothers whose first birth was not SGA and whose second was SGA was slightly higher at 2.71. The relative risk for a mother delivering two previous SGA babies was 5.07, significantly higher than all other possible combinations.

When parity and age were combined, some interesting differences were discovered. It appeared that for women aged 35 or over, the proportion delivering an SGA baby was significantly higher than for women in the 20-34 age group for primaparous women, but for those women with 2 previous births, the proportion of women aged 35 or over delivering an SGA baby was significantly lower than the 20-34 age group.

Bratton et al. [5] attempted to estimate the risk of repeat low birthweight deliveries among women whose first infant was born very low birthweight (<1500 g). The data used in this analysis were those women who delivered their first and second singleton births between 1984 and 1991. The analysis was restricted to white women due to the very small number of non-white women whose first infant was very low birthweight (VLBW), and infants with congenital deformities at birth were excluded. 182 women in the population had a first delivery resulting in a VLBW infant. In addition, a control group was selected from the population and consisted of 619 women who had a first birth with a birthweight of over 2500 g. Birthweights for the second infant were categorised as 500-1499 g, 1500-2499 g and 2500 g or more. Using the control group as a reference, women who had a previous VLBW delivery had a relative risk of 8.2 of a low birthweight delivery, and a relative risk of 53.3 of a second VLBW delivery. As there were so few VLBW deliveries in the second pregnancy (15 in total), VLBW and LBW deliveries were combined and the relative risk of a delivery under 2500 g was 11.1.

Goldenberg et al. [20] examined the effect of a previous low birthweight birth on birthweight in the current pregnancy. In this paper, low birthweight is defined as being less than 2750 g. The data used in this study were from multiparous, low income women delivering at the University of Alabama between December 1985 and October 1988. The population was oversampled for risk factors associated with decreased fetal growth. Women were excluded from the study if they delivered preterm (<27 completed weeks gestation), had a multiple pregnancy, or suffered fetal or neonatal death. Infants whose measurements were not completed before discharge were also excluded. The resulting population were categorised as those who had a history of LBW and those who did not. A regression analysis was carried out in order to determine the effect of a history of LBW on the birthweight in the current pregnancy, while correcting for risk factors of gestational age, race, infant sex, hypertension in the mother, body mass index, maternal height, maternal age, maternal weight gain and use of tobacco, alcohol and drugs. After correcting for these factors, history of LBW accounted for a decrease in birthweight of 107 g.

De Sanjose and Roman's [14] paper, mentioned previously, also investigated the effect of previous perinatal death on birthweight and discovered that mothers with at least one previous perinatal death were more than twice as likely to deliver a low birthweight infant than those mothers with no previous perinatal deaths. This was shown to be the strongest predictor of low birthweight in their study.

Rasheed and Rahman [37] considered the birth interval between the birth of interest and the previous birth as a possible predictive factor of low birthweight, and also the parity of the mother. Parity was categorised as 1-2, 3-4, 5-6 and over 6, while birth interval was categorised as less than 1 year, 1-1.99 years, 2-2.99 years and 3 or more years. While neither of these had a statistically significant effect on birthweight, there appeared to be an increase in birthweight with birth interval. There also appeared to be an increase with parity until parity became more than 6, where birthweight then dropped to almost 70 g less than those mothers with a parity of 1-2.

Lang et al. [31] consider parity, outcome of last pregnancy, and previous induced abortions, spontaneous abortions (miscarriages) and stillbirths. Parity was split into primaparous and multiparous, and the outcome of the last pregnancy was categorised as term, pre-term, or less than 22 weeks gestation. Previous induced and spontaneous abortions were categorised as 0, 1, 2 and 3 or more, while previous stillbirths were categorised as 0, 1 and 2 or more. Parity and outcome of last pregnancy had an effect on both pre-term births and term SGA deliveries. Primaparous women had a significantly higher risk of both pre-term delivery (1.8) and SGA delivery at term (also 1.8) than multiparous women. Multiparous women with a last pregnancy outcome of less than 22 weeks gestation had a significantly higher risk of SGA delivery (1.5) then those with a term delivery in their last pregnancy. Mothers with a last pregnancy outcome of pre-term delivery had higher risks than those with a term delivery in their last pregnancy for both pre-term delivery (5.9) and term SGA delivery (2.3).

Previous induced and spontaneous abortions and previous stillbirths had an effect on the risk of pre-term delivery only, and this effect was statistically significant if the mother had suffered at least 2 stillbirths or induced or spontaneous abortions. For induced abortions, the risk of pre-term delivery after 2 or 3 or more abortions was significantly higher than those mothers with no induced abortions (1.9 and 3.6 respectively), while for spontaneous abortions the risk was less than those with induced abortions but still significantly higher than the reference category of no spontaneous abortions (1.8 and 2.7 for 2 and 3 or more spontaneous abortions). For women with at least 2 stillbirths the risk of pre-term delivery was significantly higher than the reference category of none, having a relative risk of 4.7.

Parity, previous perinatal death and previous low birthweight deliveries were possible risk factors considered by Wessel et al. [46]. Parity was categorised as primaparous and multiparous, while previous perinatal death and previous low birthweight deliveries were both categorised as yes or no. None of these had any effect on the risk of pre-term delivery. However, primaparous women had a significantly higher risk of delivering a low birthweight baby (5.2 times the risk of multiparous women). In addition, multiparous women with a previous low birthweight delivery were at a higher risk of delivering a low birthweight baby than those women whose previous pregnancies had not been low birthweight (6.5 times higher).

These studies indicate that an unfavourable outcome in previous pregnancies can have a bearing on the outcome of the current pregnancy. A previous pregnancy that resulted in low birthweight has an increased risk of low birthweight in the current pregnancy, while a previous pre-term delivery, or two or more stillbirths, induced abortions or spontaneous abortions produce a higher risk of pre-term delivery in the current pregnancy. Previous SGA births and pre-term births increase the risk of an SGA birth in the current pregnancy. This indicates that obstetric history is another important factor in the investigation of low birthweight, and that many variables in a woman's obstetric history may affect birthweight.

2.1.3 Maternal Active and Passive Smoking

Eskenazi et al. [19] investigated how maternal exposure to tobacco smoke affects birthweight. Asking women how much they smoke may lead to imprecise or incorrect data and so serum cotinine was used as a marker of tobacco exposure, as this is a by-product of nicotine metabolism. The data used were those of women participating in the Child Health and Development Studies in Oakland, California, between 1964 and 1967. Those women whose smoking status was constant throughout the pregnancy and who delivered a singleton livebirth between 20 and 44 weeks gestation with known birthweight were included. Serum cotinine levels were used to categorise women into three groups, non-smokers, non-smokers exposed to environmental tobacco (passive smokers), and smokers. The smokers were then split into three groups of low, medium and high exposure. 49 women who claimed to be non-smokers but were classified by serum cotinine to be smokers were excluded from the analysis. Relative risks of low birthweight using serum cotinine levels only showed that with non-smokers as the reference, the relative risk of low birthweight increased with cotinine levels, with medium and high exposure smokers having a significantly greater risk of a low birthweight infant (1.6 and 3.3). While the risk of pre-term birth also increased with cotinine, this increase was much less marked and only high exposure smokers had a significantly greater risk of pre-term birth (1.47). A multiple regression model fitting several factors including categorised cotinine levels showed similar results in that birthweight decreased as the level of cotinine increased.

Wisborg et al. [48] investigated the association between smoking during pregnancy and pre-term birth by using data from women presenting for routine antenatal care between August 1989 and September 1991 at Aarhus University Hospital in

Denmark. The women were asked to fill in 2 questionnaires. Primaparous women who had a singleton pregnancy lasting at least 28 weeks, and completed both questionnaires including data on smoking habits and gestational age, were included in the study. Women were categorised as smokers or non-smokers depending on their smoking habits at 16 weeks gestation, and the smokers were also categorised as 1-5 cigarettes per day, 6-10 cigarettes per day, and 11 or more cigarettes per day. Other variables of interest included caffeine intake, maternal height, pre-pregnancy weight, maternal age, marital status, years of education, working status and alcohol consumption. Pre-term delivery was defined as delivery before 37 completed weeks of gestation, where gestation was determined by an ultrasound scan before 21 weeks of gestation (81% of cases), time since last menstrual period (8%), or a combination of late ultrasound scan and menstrual history (11%). Relative risks of pre-term birth using non-smokers as a reference showed that smokers had a significantly greater risk of pre-term birth, and this risk increased as cigarette exposure increased, with the 6-10 and 11 or more groups having a significantly greater risk than non-smokers (1.5 and 1.8). Correcting for the various factors mentioned above showed that none significantly changed the results except caffeine intake. This was categorised as less than 400 mg per day and 400 mg or more per day. There was no difference between smokers and non-smokers for women with a caffeine intake of less than 400 mg per day, but for those women with a caffeine intake of 400 mg or more per day it was shown, using non-smokers as a reference, that smokers had a significantly greater risk of pre-term birth, and this risk increased as cigarette exposure increased, with the 6-10 and 11 or more groups having a significantly greater risk than non-smokers (2.9 and 4.9).

Ellard et al. [18] investigated whether an estimate of nicotine intake based on a urine sample can predict smoking related birthweight deficits with more accuracy than patient reported cigarette use, by using data over a period of 20 months from women presenting for routine hospital visits at the Simpson Memorial Maternity Pavilion in Edinburgh. Data were recorded at 12-13 weeks gestation and at approximately 32 weeks gestation. At each visit the women were asked if they smoked and if so, how many cigarettes they smoked per day. Maternal weight was recorded at both visits and maternal age and parity were recorded at 12-13 weeks only. Gestational age was also confirmed at this visit by ultrasound screening, and maternal social class was determined. Urine samples were taken to determine each woman's nicotine intake. The presence of nicotine metabolites in the urine sample was indicated by pink-red chromophores in a diethylthiobarbituric acid extraction method and this gave an indication that the woman smoked. Concentrations of nicotine metabolites and creatinine were determined in those samples that tested positive for nicotine metabolites, and the ratio of nicotine metabolites to creatinine was calculated. These were then corrected for the mean value taken from a random sample of urine samples that did not indicate the presence of nicotine metabolites. 139 women who claimed to be non-smokers gave positive results when tested by the diethylthiobarbituric acid method, while 20 women who claimed to be smokers gave negative results. These women were excluded from the analyses. Birthweights were adjusted for the effects of maternal weight, maternal age, parity, gestation and infant sex using a multiple regression model. Cigarette consumption was categorised as 0 per day, 1-12 per day and more than 12 per day. The ratio of nicotine metabolites to creatinine was categorised as 0, 0.01-11.0 and greater than 11.0. Placental weight was not affected by maternal smoking status. However, birthweight decreased if the mother was a smoker, with this effect being more apparent if the nicotine metabolites to creatinine ratios were considered. This may be due to inhalation patterns of smokers, in that women who smoke a high number of cigarettes inefficiently may have a lower nicotine intake than mothers who smoke a low number of cigarettes efficiently.

Wilcox et al. [47] investigated the effect of smoking on birthweight after adjustment for the effects of physiological factors, by using data collected from University and City Hospitals in Nottingham and Derby City Hospital between August 1988 and December 1991. Data were recorded at the point of patient contact. Gestational age was calculated from ultrasound measurements made before 25 weeks gestation, and ethnic group, maternal height, booking weight, parity, birthweight and infant sex were also recorded. Data on smoking and alcohol use were obtained at the booking visit. At one hospital, vaginal bleeding data were recorded and this was categorised as 'mild' if there was no pain and hospitalisation was not required or 'moderate' otherwise. Maternal weekly weight gain was calculated from the first and last hospital visits with maternal weight data. Jarman index [26] was assigned by the electoral ward of the patient's address. The individualised birthweight ratio (IBR) was calculated for each infant by dividing the observed birthweight by the predicted birthweight from a multiple regression model containing gestational age, maternal weight, height, parity, ethnic group and infant sex. The adjusted birthweight was calculated by multiplying the IBR by the mean birthweight of a reference infant - male, 40 weeks gestation, European mother of parity 1, height 163 cm and weight 64 kg. Infants who had a gestational age under 259 days or over 300 days were excluded, as were multiple births, stillbirths, infants with congenital abnormalities and in utero transfers. All subjects with complete data were used in the analysis. Smoking was defined by the number of cigarettes smoked per day - none, 1 to 9, 10 to 19 and 20 or more. After correcting for gestational age, maternal weight, height, parity, ethnic group and infant sex in a multiple regression model, smoking was shown to have an effect of a reduction in birthweight as the number of cigarettes consumed increased, the difference between non-smokers and those who smoked 20 or more cigarettes per day being 219 g, assuming all other factors were identical.

Lang et al. [31] also investigated the effect of smoking on both pre-term delivery and SGA births at term. Smoking history was categorised as no smoking during pregnancy, stopped smoking early in pregnancy, started smoking later in pregnancy and smoked throughout pregnancy. Smoking was shown to have an effect of an increased risk of pre-term labour if the mother smoked throughout the pregnancy. The risk of a term SGA delivery increased compared to the reference of no smoking during pregnancy if the mother smoked throughout the pregnancy (relative risk of 2.3) or started smoking later in the pregnancy (relative risk of 2).

Smoking during pregnancy has been shown to increase the risk of low birthweight, pre-term birth and SGA births. In addition, maternal smoking is one of the few factors though to be related to low birthweight that can be changed by the mother. As such, while this may not be the most important factor in low birthweight deliveries, it is a factor that should be stressed to the mother as a risk that can be lessened.

2.1.4 Maternal Height

Lang et al. [31] categorised maternal height as 5' or under, 5' 1" - 5' 4", 5' 5" - 5' 7", and greater than 5' 7". Using the 5' 5" - 5' 7" category as a reference, the categories of 5' or under and 5' 1" to 5' 4" had an increased risk of both pre-term delivery (1.4 and 1.2 respectively) and term SGA delivery (3.0 and 1.6 respectively). In addition those women over 5' 7" had a decreased risk of term SGA delivery (0.7).

Wessel et al. [46] categorised maternal height as 154 cm or under and 155 cm or over. The risk of a pre-term birth was greater in the 154 cm or under group, but this was not statistically significant. Similarly, the risk of a low birthweight infant was less in the 154 cm or under group, but this was also not statistically significant.

2.1.5 Marital Status

De Sanjose and Roman [14] investigated the marital status of women and the effect this had on low birthweight. It was found that mothers who were not married were more likely to deliver a low birthweight baby.

Marital status was also investigated by Rodriquez et al. [41]. The results were similar to those of De Sanjose and Roman in that a higher proportion of unmarried mothers delivered a low birthweight baby in both term and pre-term births.

Lang et al. [31] also investigated marital status, which was categorised as single, married and other. The risk of pre-term delivery was significantly higher for single mothers (2.2) compared to the reference of married mothers, while the risk of a term SGA delivery was increased for both single (2.0) and other (1.8) mothers.

Unmarried mothers appear to have a higher risk of low birthweight babies. However, it is likely that marital status and maternal age are highly correlated, with most of the high risk teenage pregnancies discussed earlier falling into the unmarried category, and as a result this is likely to be restating the result of the investigation of maternal age on birthweight.

2.1.6 Obstetric factors

Wessel et al. [46] investigated the effect of maternal hypertension or convulsions on both birthweight and pre-term delivery. While this had no effect on the risk of a low birthweight delivery, it was the only factor to show a significant effect on the risk of pre-term birth, with women suffering from hypertension or convulsions having a relative risk of 2.6 over those women who did not.

Bakketeig et al. [2] investigated the association between SGA and three types of maternal condition - pre-eclampsia, vaginal bleeding and pathological conditions of the placenta. Pathological conditions of the placenta included abruptio placenta, placenta previa and placental infarctions. For those women with pre-eclampsia, the relative risk of mothers delivering their first child as SGA was 2.1, compared with the reference category of mothers without pre-eclampsia delivering their first child. Similarly, the risks for mothers delivering their second and third children SGA, compared to women of the same parity but without pre-eclampsia, were 1.3 and 1.1. Similar results were obtained for those women with vaginal bleeding during pregnancy, with the relative risks in the first, second and third pregnancies for women with vaginal bleeding delivering an SGA baby compared to those women with the same parity but with no vaginal bleeding being 2.1, 1.6 and 1.6. For mothers who reported pathological conditions of the placenta, the relative risks of SGA deliveries in the first, second and third pregnancies were 2.1, 1.7 and 2.1 compared to women of the same parity but with no placental conditions.

2.1.7 Other Factors

As described previously, Rasheed and Rahman [37] considered many factors that were similar to those in European studies. They also considered whether the fact that the child's parents were first cousins had any effect on birthweight. Of the 278 cases considered, 166 (59.7%) were children born to parents who were first cousins. While this was not statistically significant in the model, children born to parents who were first cousins had a predictive birth weight of 74 g less than those whose parents were unrelated.

2.2 Environmental Risk Factors

2.2.1 Social Class

Deprivation is often measured by social class, which is dependant on the occupation of the mother, or more usually the father. De Sanjose and Roman [14] investigated both maternal and paternal social class. For maternal social class, the risk of delivering a low birthweight baby increased from social class I (professional) to IIIM (the first manual class), and then remained similar between classes IIIM and V (manual). For paternal social class there was a steady rise in the risk of a low birthweight baby from social classes I to V.

Rodriquez et al. [41] also investigated the effect of maternal and paternal occupation. 'Maternal activity' was categorised as outside the home, at home and other, while paternal occupation was categorised as manual, non-manual and other. For pre-term births, the proportion of low birthweight was lowest in mothers who worked outside the home and highest in mothers with 'other' occupations. The proportion of low birthweight deliveries to mothers who stayed at home was significantly higher than those who worked outside the home. The proportion of low birthweight deliveries where the paternal occupation was manual was significantly higher than those where the paternal occupation was non-manual. Similar results were presented for term low birthweight deliveries, although there was no significant difference between maternal activity categories.

Rasheed and Rahman [37] used number of rooms in the home, presence of a home help or housemaid, and mother's education as indicators of socio-economic status. The predictive birthweight of infants increased with the number of rooms, presence of help in the home and educational status of the mother, although none of the variables indicated a significantly different birthweight.

2.2.2 Maternal Education

Bakketeig et al. [2] considered maternal education as the number of years of schooling, categorised as 7, 9, 12 and more than 12. Using mothers with more than 12 years of schooling as the reference category, all other categories had a larger relative risk of delivering an SGA baby, with this risk increasing as the years of education decreased and with 7 and 9 years of education having a significantly higher risk. If the risk of delivering more than one SGA baby is considered then all categories had a significantly higher risk than that of those mothers who had more than 12 years of education, and this risk increased as the number of years of education decreased.

Lang et al. [31] considered health insurance and maternal education, which can be thought of as indicators of socio-economic status. Both public and other health insurance had a significantly higher risk of pre-term labour (1.8 and 2 respectively) and term SGA delivery (2.1 and 1.2 respectively) than the reference of private health care. Using a reference of high school graduate for maternal education, those mothers who were not high school graduates had a higher risk of both pre-term delivery and term SGA delivery (1.3 in both cases), and those mothers who had some college education had a lower risk of both pre-term delivery and term SGA delivery (0.5 and 0.7 respectively).

2.2.3 Deprivation Measures

Wilcox et al. [47] considered the Jarman score [26], as discussed in Chapter 4, as a measure of social deprivation. This was categorised into 6 groups - under -20, -20 to -11, -10 to -1, 0 to 9, 10 to 19 and 20 or more, where a high Jarman score indicated increased area deprivation. As the Jarman score increased, the individualised birthweight ratio decreased significantly from 1.011 to 0.988, and so adjusted birthweight decreased from 3715 g to 3631 g. After correcting for gestational age, maternal weight, height, parity, ethnic group and infant sex in a multiple regression model, there was a significant effect of continuous Jarman score on birthweight, in that as the Jarman score increased, birthweight decreased.

2.3 Conditions Related to Low Birthweight

2.3.1 Infant Mortality and Morbidity

The Scottish Low Birthweight Study Group [21, 22] considered livebirths weighing under 1750 g who were born in Scotland in 1984. 99% (896) of these births were enrolled in a prospective study to document survival and to determine the prevalence of sensory, neuromotor and cognitive impairments, language attainment, cognitive status and behavioural problems and to relate these to morbidity, social circumstances and perinatal experiences. At 4.5 years an assessment of surviving children was carried out. 636 (71%) had survived to 4.5 years and of these 611 were assessed. Of those who had not survived, 217 died before 28 days, with most of these deaths being infants weighing under 1000 g, 36 died in the first year, 6 in the second year and 1 between 2 and 4 years. Of the 204 infants born under 1000 g, only 60 survived to 4 years. Birthweights were categorised into 3 groups, under 1000 g, 1000-1499 g, and 1500-1749 g, and the results of the assessments carried out were tested for a birthweight effect. The proportion of children with neuromotor impairment showed a significant trend with birthweight, with children under 1000 g being almost twice as likely to have some level of impairment compared with children between 1500 and 1749 g. This was mainly due to those children with severe or moderate disability. Of all other impairments assessed, squints had an overall prevalence of 11.6%, and all other impairments had overall prevalences under 5%. Language attainment was assessed by two tests, the Renfrew action picture test (APT) and the bus story test of continuous speech (BSTCS). In the information part of the APT, significantly more children with a birthweight of under 1000 g scored below the 25th centile than in the other birthweight groups, and in the grammar part of the test, significantly less children with a birthweight less than 1000 g scored above the 75th centile than other birthweight groups. In the BSTCS there were very little differences between the birthweight groups, but in the information part of the test significantly more children with a birthweight of under 1000 g scored below the 25th centile than in the other birthweight groups. Cognitive ability was assessed using British ability scales (BAS) which in turn was used to calculate the IQ of the children assessed. There was no difference between the mean IQ in each of the birthweight groups. The mean scores in each section of the BAS were significantly lower than the standards. In the number skills section of the BAS there was a statistically significant effect of birthweight, this being that the number of children performing below the 10th centile decreased as the birthweight category increased.

Hall et al. [24] considered a subgroup of the population discussed by the Scottish low birthweight study group [21, 22] of children born under 1500 g who were still resident in Scotland at the age of eight to nine years. Two control classmates of each child were chosen for a comparison group to identify how these very low birthweight (VLBW) children differ from their classroom peers. Children in both the VLBW population and the control population had growth, blood pressure, respiratory function, cognitive ability, school attainment, visual acuity and hearing measured. The VLBW population was split into two groups dependent on birthweight - under 1000 g at birth, and 1000 - 1499 g at birth. The control patients were also split into two groups dependent on the birthweight of the classmate they were controlling for. Both the children born under 1000 g and those born between 1000 and 1499 g performed significantly worse in movement tests than their control classmates, with many more children than expected falling below the 10th centile of the tests. Similarly the VLBW groups performed less well in the neurological screening tests, with 24% of the under 1000 g group having a normal score compared with the control group of 88%, and 42% of the 1000 - 1499 g group having a normal score compared to 74% of the control group. In cognitive testing, there were significant differences between both VLBW groups and their controls in both verbal and visual IQs, and similarly with word reading and number skills.

Kollee et al. [30] attempted to determine the 5-year outcome of VLBW infants who were referred to tertiary perinatal centres. The data used were of infants born in 1983 in the Netherlands at less than 32 weeks gestation or weighing less than 1500 g, or both. The five-year outcome assessment was carried out on a subset of these infants, where antenatal care had not been received in one of the eight university hospitals serving as tertiary centres, gestation was between 26 and 31 completed weeks, normal cardiotocographic tracings were available and tocolysis had lasted more than 24 hours. This subset of infants was then split into those who were born after maternal transport to a tertiary centre (113) and those who were not, and these infants were then sub-divided into those who were transferred to a tertiary centre (124) and those who were not referred for tertiary care but were only treated in local hospitals (131). Of these 368 infants, 252 were available for investigation at the five-year follow-up, where the infants were assessed for congenital malformation, neuromotor function, mental development, hearing and visual function, language and speech development, musculoskeletal system, respiratory tract, and ear, nose and throat disorders, and from this it was determined whether the child was disabled or handicapped. Kolle et al. then investigated the relationship between mode of referral and outcome (disability, handicap or neither) using logistic regression and correcting for 20 possible risk factors. While the odds ratios for disabilities and handicaps for maternal transport versus infants born in local hospitals were 1.33 and 0.87 respectively, these were not statistically significant. However, it was shown that 31% of the infants were considered to be disabled and 15% were considered handicapped, which was a high percentage of those infants born VLBW.

Kitchen et al. [29] assessed the outcome at 2 years of age of infants who were born with a birthweight between 500 g and 999 g. Two cohorts of infants were considered, those born between 1977 and 1982, and those born between 1985 and 1987. All births occurred at the Royal Women's Hospital in Carlton, Australia. Stillbirths were included in the study if there was a possibility that the foetus may have been alive at the start of labour. The outcome at 2 years of age was categorised as unacceptable if fetal or infant death occurred, or if the infant was severely disabled, and as acceptable if the infant survived to 2 years and was not severely disabled. The survival rates for the two cohorts were quite low (25.3% and 37.4%), and in each cohort the survival rate increased as the birthweight increased from 500 - 599 g to 900 - 999 g. Logistic regression, accounting for obstetric and sociodemographic variables, showed that gestational age, cervical suturing and the presence of antenatal haemorrhaging had a statistically significant effect on the 2 year outcome, as did antenatal steroid therapy, birthweight and sex of the infant. In addition the regression showed that those infants born in the second cohort had a significantly higher chance of an acceptable outcome, indicating the presence of a time effect on the outcome.

2.3.2 Growth Impairment

Powls et al. [36] compared the growth of VLBW children to that of normal birthweight children of the same age group, examining factors that contribute to growth. The data used were of two groups of VLBW children treated at the Merseyside regional neonatal unit. The first group were children with a birthweight of 1200 g or less born between January 1980 and June 1981. The second group were children of 1500 g or less, with a gestational age of less than 31 weeks, born between January 1982 and November 1983. There were 137 children in total, none of who had any major neurodevelopmental handicap. A control population of 160 normal birthweight infants was recruited from classmates of the same sex and similar age to the VLBW children. Standing and sitting height, weight, occipito-frontal circumference (OFC), skinfold thickness and pubertal staging were measured in both populations, and standardised tests of cognitive and educational ability were carried out on all children.

Bone age assessment was carried out on 96 of the VLBW population. There were significant differences between the VLBW and normal birthweight populations in both standing and sitting heights. VLBW boys were on average 3.8 cm shorter standing and 1.3 cm shorter sitting than their normal birthweight controls, while VLBW girls were on average 4.4 cm shorter standing and 2.2 cm shorter sitting than their normal birthweight counterparts. The VLBW children were also lighter than their controls, with the average difference in girls being larger than that in boys (3.8 kg and 1.1 kg respectively). Head circumference was significantly smaller in the VLBW children, being on average 0.7 cm less for boys and 1.1 cm less for girls compared with the control population. A difference between the populations still existed after correction for standing height, indicating that the VLBW children had disproportionately smaller heads. There were no differences between the populations in the stage of pubertal growth reached. OFC was associated with cognitive and educational ability, those with a smaller OFC having poorer results in the educational tests.

Chapter 3

Analysis of Unlinked Data

This chapter updates previous work carried out on birthweight in Scotland by Pickering [35], using more recent data. A univariate approach to those factors that may possibly be related to low birthweight is considered in order to choose variables to use in modelling the probability of delivering a low birthweight infant. Maternal length of stay is considered to investigate whether this differs over hospital or deprivation score, as are the caesarean section rates and the survival rates of infants. All low birthweight infants are considered to investigate whether medical complications can, in some way, explain low birthweight. Deprivation and maternal height are then investigated after correction for gestational age, and then some of the variables investigated are considered as explanatory variables in fitting a possible model for low birthweight.

3.1 Introduction

In this chapter low birthweight within a geographically constrained population - namely Greater Glasgow Health Board (GGHB) - is considered, and previous work carried out on birthweight data is updated using a more recent data set. Specifically, it is intended to see whether the occurrence of low birthweight is in any way related to the Womersley score, also known as neighbourhood type - a score based on the 1981 census data, and allocated to households by postcode sector. This scoring method is discussed in detail in chapter 4. Low birthweight is divided into two categories, very low birthweight, which is defined as a new born baby weighing 1500 g or less, and low birthweight, where the baby weighs between 1501 g and 2500 g. Both categories of low birthweight, and 'normal' birthweight, which is defined here as a birth weight of more than 2500 g, are considered in the following analyses.

The full data set was obtained from the Information and Statistics Division (ISD) of the Scottish Health Service and contained data for all admissions during pregnancy for all mothers either resident in Glasgow, or attending Glasgow hospitals, between January 1981 and December 1991. This data comes from one of the Scottish Morbidity Records (SMR), the SMR2. This form was introduced in 1969 and was designed for use in maternity hospitals to obtain information for every hospital discharge. By 1975 96% of hospital deliveries in Scotland were recorded on it, incorporating the following sections: general information on the mother, including age, occupation and marital status; information on the current pregnancy, including date of admission, date of last menstrual period, and abortion details if relevant; the outcomes of any previous pregnancies, indicating the number of previous abortions, miscarriages, deaths and caesarean sections; maternal discharge data, including the date of discharge; a record of labour, including the number of births, mode of delivery and sex of the infant; postnatal infant details, including whether the infant was sent to a Special Care Baby Unit; and any conditions or complications present. As one record is completed at each discharge, one record per pregnancy contains delivery details but there may be other records indicating periods of antenatal admission [12]. The variables in this data set are shown in appendix A. In order to give a well-defined geographically constrained population, the subset of these data used were those women who were resident in the Greater Glasgow Health Board area, and delivering in or attending Glasgow hospitals. The analyses in this chapter exclude those women who had either a stillbirth or a multiple birth unless otherwise stated.

Four analyses were considered - a descriptive analysis of the type of mother delivering a baby in each of the three birthweight groups; the average length of stay in hospital by neighbourhood type and by hospital; caesarean section rates by neighbourhood type and by hospital, and finally survival rates for livebirths by neighbourhood type and by hospital.

3.2 Descriptive Analysis

Here tabulations of variables that may be thought of as being relevant to the occurrence of low birthweight are provided. Statistical modelling is discussed in later chapters.

As the data set used provided has a large number of records - over 130,000 - it was decided to use data from 1981 and 1991 only. For both years, the birthweight category and 9 other variables pertaining to the mother were considered. These other variables were :

- Maternal age
- Womersley neighbourhood type (1=affluent, 8=deprived)
- Marital status
- Parity
- Hospital
- Previous induced abortions
- Previous spontaneous abortions
- Maternal height
- Gestational age

Maternal age was grouped into the following categories: less than 16, 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49 and 50-54, and maternal height into the following categories: under 150cm, 150-154cm, 155-159cm,160-164cm, 165-169cm, 170-174cm, 175-179cm, 180-184cm and 185-189cm. In order to compare over categories of the above variables, and between the two years, the proportions in each of the three birthweight categories were calculated for the different categories in the variables of interest. All tables not shown here are presented in appendix B.

Maternal age has been shown previously to be an indicator of low birthweight as discussed in chapter 2 [14, 42, 37, 46]. As can be seen from table B.1, most women have a maternal age between 20 and 34, with more 20-24 year old mothers in 1981 than in 1991, while the numbers of mothers in other age categories is reasonably similar over the two years. There does not appear to be much difference for 1981 deliveries in the proportion of very low birthweight, low birthweight, or normal birthweight babies in each age group unless the mother is in her early forties, or is under sixteen. In both of these cases the proportion of very low birthweight babies becomes much higher. In 1991 there was again not much difference in proportions unless the mother was a teenager, in which case the proportions of very low birthweight and low birthweight babies were slightly higher. There does not seem to be a difference in proportions between the two years.

The Womersley neighbourhood type, which is discussed in detail in chapter 4, is an ordered categorical measurement of deprivation, with a Womersley neighbourhood type of 1 indicating the most affluent areas and a Womersley neighbourhood type of 8 indicating the most deprived area. There are similar numbers of mothers in the neighbourhood type areas over 1981 and 1991, with types 5 and 8 having more deliveries and type 6 having less than the other types. It can be seen from table 3.1 that for both years, as the neighbourhood type increases from 1 to 8, i.e. from affluent areas to deprived areas, the proportion of normal birthweight babies steadily decreases. However there is no apparent trend in the proportion of very low birthweight babies, while the proportion of low birthweight babies increases. This suggests that while neighbourhood type does have an effect on birthweight, the effect is not present in those cases where the babies are very low birthweight. One reason for this may be that while neighbourhood type has an effect on low birthweight babies, very low birthweight babies are born as a result of underlying obstetrical problems, which are not related to neighbourhood type.

Marital status has been considered an indicator of the possibility of low birthweight [14, 41]. The results in table B.2 appear to bear this out, as for both years the proportion of very low birthweight and low birthweight babies appears to be higher for those mothers who are either single or separated, and there does not seem to be much difference between the two years. However, the number of single mothers has more than doubled in the 11 year period between 1981 and 1991, and many more women have been categorised as 'other', resulting in a drop in the number of married mothers.

1981				1991			
Womersley	VLBW	LBW	NBW	Womersley	VLBW	LBW	NBW
1 (n=1354)	0.0052	0.0281	0.9668	1 (n=1526)	0.0046	0.0315	0.9640
2 (n=1155)	0.0069	0.0398	0.9532	2 (n=1102)	0.0109	0.0381	0.9510
3 (n=1241)	0.0040	0.0371	0.9589	3 (n=1149)	0.0078	0.0383	0.9539
4 (n=1219)	0.0049	0.0500	0.9450	4 (n=1689)	0.0065	0.0491	0.9443
5 (n=2549)	0.0090	0.0596	0.9313	5 (n=2406)	0.0087	0.0615	0.9298
6 (n=889)	0.0056	0.0641	0.9303	6 (n=713)	0.0056	0.0547	0.9397
7 (n=1827)	0.0088	0.0805	0.9108	7 (n=1339)	0.0082	0.0792	0.9126
8 (n=2307)	0.0069	0.0728	0.9202	8 (n=2323)	0.0121	0.0762	0.9118

Table 3.1: Relationship between birthweight category and Womersley neighbourhood type.

Considering next the parity of the mother, table B.3 shows that if the mother has had less than three previous pregnancies, there seems to be no difference in the proportions in each birthweight category. For those with more than four previous pregnancies, the proportion of low birthweight babies seems to increase with the parity of the mother. Again these proportions do not differ much over the two years.

If the case mix of patients was similar in each hospital, it would seem reasonable to assume there should be no differences between hospitals in the same year, and this is borne out in table B.4. There are 6 hospitals in Greater Glasgow Health Board (GGHB) with data, labelled 1 to 6 here for convenience. Of these one, denoted here as hospital 4, was a private hospital with under 100 births per year which closed down before 1991. In the light of this, it was decided to consider mothers who had delivered in hospital 4 in all other tables, but to remove them from any analyses that were carried out by hospital. In considering all hospitals except hospital 4, the proportions in each birthweight category are similar over hospitals. More deliveries occur in hospitals 1 and 6 than in hospitals 2, 3 and 5.

In looking at the obstetric history of the mother two possibly relevant factors are considered - the number of previous spontaneous abortions (miscarriages), and the number of previous induced abortions. Each has been shown by Lang et al. [31] to affect the length of future pregnancies and hence from this it may be possible that the birthweight in future pregnancies will also be affected. From table B.5 it can be seen that although the proportion of low birthweight babies remains fairly constant as the number of induced abortions increases, the proportion of very low birthweight babies is higher for those mothers with two previous induced abortions. It is difficult to interpret the figures for three and four previous induced abortions, as the number of mothers involved is very small, with under 1% of women having more than one induced abortion. The proportions of both very low birthweight and low birthweight babies increase as the number of spontaneous abortions increases, as shown in table B.6. This may be expected, as a spontaneous abortion may suggest an obstetric abnormality that may be present in further pregnancies, although it may not have the same effect, and this may increase the likelihood of a low birthweight baby.

In considering maternal height, an increase in low birthweight babies as maternal height decreases may be expected, as small mothers would be thought of as more likely to produce small babies. As maternal height increases, table 3.2 shows that the proportion of very low birthweight decreases and that of normal birthweight babies increases. Also, for mothers over 174 cm in 1981, and 179 cm in 1991, no very low birthweight babies were born. Most mothers are between 150 cm and 169 cm in height, with very few women over 175 cm tall.

1981				1991			
Height (cm)	VLBW	LBW	NBW	Height (cm)	VLBW	LBW	NBW
<150 (n=696)	0.0057	0.1049	0.8894	<150 (n=334)	0.0089	0.1228	0.8683
150-154 (n=2162)	0.0065	0.0828	0.9107	150-154 (n=1412)	0.0113	0.0899	0.8987
155-159 (n=3635)	0.0074	0.0572	0.9354	155-159 (n=2936)	0.0078	0.0647	0.9275
160-164 (n=3350)	0.0057	0.0418	0.9525	160-164 (n=3459)	0.0066	0.0442	0.9491
165-169 (n=1653)	0.0042	0.0357	0.9601	165-169 (n=2179)	0.0060	0.0358	0.9582
170-174 (n=429)	0.0070	0.0187	0.9744	170-174 (n=876)	0.0023	0.0285	0.9692
175-179 (n=68)	0.0000	0.0147	0.9853	175-179(n=192)	0.0052	0.0156	0.9792
180-184 (n=7)	0.0000	0.1429	0.8571	180-184 (n=26)	0.0000	0.0000	1.0000
185-189 (n=3)	0.0000	0.0000	1.0000	185-189 (n=1)	0.0000	0.0000	1.0000
unknown (n=538)	0.0223	0.0855	0.8922	unknown (n=832)	0.0264	0.0841	0.8894

Table 3.2: Relationship between birthweight category and maternal height.

Gestational age clearly must have an effect on the birthweight of the baby. Table 3.3 shows that most deliveries take place between 38 and 41 weeks gestation, with less than 1% taking place before 32 weeks. For gestations under 32 weeks, most babies are very low birthweight, and those that are not are mostly low birthweight. Then, as the gestational age increases the proportion of very low birthweight and low birthweight babies decrease.

1981				1991			
Gestation (weeks)	VLBW	LBW	NBW	Gestation (weeks)	VLBW	LBW	NBW
<32 (n=108)	0.5918	0.3571	0.0510	<32 (n=110)	0.7000	0.3000	0.0000
32-36 (n=574)	0.0296	0.5383	0.4321	32-36 (n=628)	0.0414	0.4857	0.4729
37 (n=553)	0.0018	0.1591	0.8391	37 (n=695)	0.0000	0.1482	0.8518
38 (n=1563)	0.0000	0.0678	0.9322	38 (n=1684)	0.0000	0.0635	0.9365
39 (n=2624)	0.0000	0.0282	0.9718	39 (n=2492)	0.0000	0.0261	0.9739
40 (n=4566)	0.0000	0.0160	0.9840	40 (n=4067)	0.0000	0.0140	0.9860
41 (n=2203)	0.0000	0.0104	0.9896	41 (n=2179)	0.0000	0.0073	0.9927
>41 (n=350)	0.0000	0.0200	0.9800	>41 (n=392)	0.0000	0.0026	0.9974

Table 3.3: Relationship between birthweight category and gestational age.

3.3 Rates of Survival

In this section the survival rates of babies born in different birthweight categories, hospitals, and to mothers of different neighbourhood types are investigated. Again all singleton livebirths are considered, and are divided into VLBW, LBW, and NBW as before. Only 1981 and 1991 are considered for LBW and NBW births, and 1981 to 1991 inclusive for VLBW babies. The SMR2 form gives five possibilities for the outcome of a pregnancy, those being:

- Livebirth
- Stillbirth
- Livebirth, died aged <7 days
- Livebirth, died aged 7-28 days

- Livebirth, died aged >28 days

All singleton livebirths, whether or not they subsequently died, are considered. Those babies who were livebirths and were not categorised as dying at any point are considered survivors. Many stillbirths occur at an early gestation and thus the baby is likely to be low birthweight. For this reason stillbirths are not considered here as they may bias the conclusions.

3.3.1 Neighbourhood Type

Womersley	Total Survivals	Total Livebirths	Proportion Surviving
1	57	70	0.8143
2	52	76	0.6842
3	65	85	0.7647
4	95	116	0.8190
5	157	204	0.7696
6	59	73	0.8082
7	117	147	0.7959
8	185	238	0.7773

Table 3.4: Proportion of survivals for VLBW by Womersley neighbourhood type, singleton livebirths between 1981 and 1991.

For each neighbourhood type, the total number of singleton livebirths and the number of those who survived were calculated, and from this the proportion of survivals for each neighbourhood type was calculated. The results for VLBW babies are shown in table 3.4, and are quite similar, with the exception of type 2, which seems to be lower than the others.

The proportions of LBW babies surviving, shown in table 3.5, do not appear to change much with neighbourhood type. The exception to this, in both 1981 and 1991, is type 6, which is slightly lower than the rest. Comparing the two years, the proportion of babies surviving in 1991 is similar to the proportion in 1981. For NBW

1981				1991			
Womersley	Total Surv.	Total Births	Prop. Of Survivals	Womersley	Total Surv.	Total Births	Prop. Of Survivals
1	37	38	0.9737	1	48	48	1.0000
2	45	46	0.9783	2	41	42	0.9762
3	46	46	1.0000	3	43	44	0.9773
4	59	61	0.9672	4	83	83	1.0000
5	150	152	0.9868	5	146	148	0.9865
6	53	57	0.9298	6	37	39	0.9487
7	145	147	0.9863	7	105	106	0.9906
8	162	168	0.9643	8	176	177	0.9944

Table 3.5: Proportion of survivals for LBW by neighbourhood type, singleton live-births in 1981 and 1991.

babies, the figures in table 3.6 are all very similar, and it would be difficult to say that there is a difference anywhere, either between the two years, or between the neighbourhood types. Comparing the three tables together, the proportion of NBW and LBW babies surviving are comparable, but the proportion of VLBW babies surviving is less than the proportion of NBW and LBW survivals, as is expected.

1981				1991			
Womersley	Total Surv.	Total Births	Prop. Of Survivals	Womersley	Total Surv.	Total Births	Prop. Of Survivals
1	1307	1309	0.9985	1	1470	1471	0.9993
2	1099	1101	0.9982	2	1045	1048	0.9971
3	1189	1190	0.9992	3	1094	1096	0.9981
4	1150	1152	0.9983	4	1593	1595	0.9987
5	2372	2374	0.9992	5	2232	2237	0.9978
6	827	827	1.0000	6	670	670	1.0000
7	1661	1664	0.9982	7	1218	1222	0.9967
8	2123	2123	1.0000	8	2114	2118	0.9981

Table 3.6: Proportion of survivals for NBW by neighbourhood type, singleton live-births in 1981 and 1991.

3.3.2 Hospital

For each hospital, the proportion of babies surviving was calculated as with neighbourhood type. For VLBW babies, shown in table 3.7, hospitals 5 and 6 have a slightly higher survival rate than the other hospitals.

Hospital	Total Survivals	Total Births	Proportion Surviving
1	243	320	0.7594
2	106	139	0.7626
3	93	121	0.7686
5	140	170	0.8235
6	204	257	0.7938

Table 3.7: Proportion of survivals for VLBW by hospital, singleton livebirths between 1981 and 1991.

1981				1991			
Hospital	Total Survivals	Total Births	Proportion Surviving	Hospital	Total Survivals	Total Births	Proportion Surviving
1	203	210	0.9667	1	219	224	0.9777
2	121	123	0.9837	2	74	74	1.0000
3	107	109	0.9817	3	126	126	1.0000
5	129	132	0.9773	5	120	121	0.9917
6	135	139	0.9712	6	140	142	0.9859

Table 3.8: Proportion of survivals for LBW by hospital, singleton livebirths in 1981 and 1991.

The proportions of LBW babies surviving in each hospital are quite similar, and there is a slight increase in the proportions between 1981 and 1991 as shown in table 3.8. For NBW babies there is little difference anywhere, either between hospitals or between years, as shown in table 3.9.

1981				1991			
Hospital	Total Survivals	Total Births	Proportion Surviving	Hospital	Total Survivals	Total Births	Proportion Surviving
1	3028	3030	0.9993	1	3184	3191	0.9978
2	1930	1932	0.9990	2	1709	1713	0.9977
3	1895	1898	0.9984	3	1939	1940	0.9995
5	2013	2015	0.9990	5	1841	1844	0.9984
6	2754	2757	0.9989	6	2762	2768	0.9978

Table 3.9: Proportion of survivals for NBW by hospital, singleton livebirths in 1981 and 1991.

3.3.3 Year Effect

If all data from the eleven year period are used with babies separated into VLBW, LBW, and NBW, then table 3.10 shows that there appears to be an upwards trend in the proportion of VLBW babies surviving over the eleven year period.

Year	Total Survivals	Total Births	Proportion Surviving
1981	56	86	0.6512
1982	51	82	0.6220
1983	65	81	0.8025
1984	77	100	0.7700
1985	70	88	0.7955
1986	77	93	0.8280
1987	70	82	0.8537
1988	81	101	0.8020
1989	71	98	0.7245
1990	81	95	0.8526
1991	88	103	0.8544

Table 3.10: Proportion of survivals for VLBW by year, singleton livebirths between 1981 and 1991.

For LBW and NBW babies, as shown in tables 3.11 and 3.12 the numbers are so close to unity that it is impossible to conclude anything about a trend. From this it can be seen that while the proportion of babies weighing over 1500 g who survive does not appear to change, the proportion of VLBW babies surviving increases. This may be due to increased specialist care in special care baby units (SCBUs) of VLBW babies.

Year	Total Survivals	Total Births	Proportion Surviving
1981	697	715	0.9748
1982	657	670	0.9806
1983	635	646	0.9830
1984	649	665	0.9759
1985	716	720	0.9944
1986	713	719	0.9917
1987	712	721	0.9875
1988	687	698	0.9842
1989	652	658	0.9909
1990	584	591	0.9882
1991	679	687	0.9884

Table 3.11: Proportion of survivals for LBW by year, singleton livebirths between 1981 and 1991.

3.4 Maternal Complications

There are various complications that can occur in pregnancy, and some of these are thought to increase the likelihood of an early birth, which in turn increases the risk of low birthweight. In this section only VLBW and LBW births are of interest. All births where the mother suffered from placenta previa, premature placental separation, antepartum haemorrhaging, or hypertension shall be defined here as an 'explained' low birthweight, and the proportion 'explained' by one or more of these

Year	Total Survivals	Total Births	Proportion Surviving
1981	11728	11740	0.9990
1982	11335	11355	0.9982
1983	11203	11215	0.9989
1984	11277	11286	0.9992
1985	11542	11559	0.9985
1986	11538	11546	0.9993
1987	11533	11542	0.9992
1988	11539	11554	0.9987
1989	10983	10994	0.9990
1990	10563	10576	0.9988
1991	11436	11457	0.9982

Table 3.12: Proportion of survivals for NBW by year, singleton livebirths between 1981 and 1991.

maternal conditions is investigated.

3.4.1 Neighbourhood Type

Womersley	Total Explained	Total Cases	Proportion Explained
1	19	70	0.2714
2	23	76	0.3026
3	27	85	0.3176
4	34	116	0.2931
5	69	204	0.3382
6	20	73	0.2740
7	44	147	0.2993
8	64	238	0.2689

Table 3.13: Proportion of explained LBW deliveries by Womersley neighbourhood type, singleton livebirths between 1981 and 1991.

In the case of VLBW babies, shown in table 3.13, there does not appear to be much difference between neighbourhood types. Considering LBW babies, in table 3.14, again there is not much difference between neighbourhood types, except that in 1981, types 1 and 2, and 7 and 8, the most affluent and deprived areas, seem to have a higher proportion of explained low birthweight than the others. Also, for LBW babies, there does not appear to be much difference between the two years.

3.4.2 Hospital

Considering first VLBW babies in table 3.15, the proportion of explained VLBW births is similar for each hospital, with the exception of hospital 3 which is slightly higher. For the LBW babies table 3.16 shows that in 1981 hospitals 3 and 5 had a smaller proportion of explained LBW than the others, and in 1991 hospital 5 had a smaller proportion of explained LBW than the other hospitals. Over the years, the proportion decreases in all hospitals except hospital 3.

1981				1991			
Womersley	Total Expl.	Total Cases	Prop. Expl.	Womersley	Total Expl.	Total Cases	Prop. Expl.
1	11	38	0.2895	1	6	48	0.1250
2	10	46	0.2174	2	9	42	0.2143
3	5	46	0.1087	3	10	44	0.2273
4	12	61	0.1967	4	12	83	0.1446
5	26	152	0.1711	5	21	148	0.1419
6	6	57	0.1053	6	6	39	0.1538
7	36	147	0.2449	7	15	106	0.1415
8	36	168	0.2143	8	22	177	0.1243

Table 3.14: Proportion of explained LBW deliveries by neighbourhood type, singleton livebirths in 1981 and 1991.

Hospital	Total Explained	Total Cases	Proportion Explained
1	100	320	0.3125
2	38	139	0.2734
3	46	121	0.3802
5	46	170	0.2706
6	70	257	0.2724

Table 3.15: Proportion of explained VLBW deliveries by hospital, singleton livebirths between 1981 and 1991.

Although the proportions of explained VLBW and LBW births seem to differ over hospitals, it must be pointed out that there may be several reasons for this. One major reason is the difference in referral strategies of doctors and hospitals themselves. A hospital may have the policy that a life threatening complication may be better dealt with at a neighbouring hospital with better facilities to deal with the problem, in which case the first hospital would have a smaller proportion of explained VLBW and LBW babies than the second. Clearly conclusions cannot be drawn from this data without further investigation of such possible factors.

1981				1991			
Hospital	Total Explained	Total Cases	Prop. Expl.	Hospital	Total Explained	Total Cases	Prop. Expl.
1	55	210	0.2619	1	28	224	0.1250
2	25	123	0.2033	2	11	74	0.1486
3	12	109	0.1101	3	25	126	0.1984
5	12	132	0.0909	5	8	121	0.0661
6	37	139	0.2662	6	29	142	0.2024

Table 3.16: Proportion of explained LBW deliveries by hospital, singleton livebirths in 1981 and 1991.

3.5 Correcting For Gestational Age

As mentioned in the previous section, it is possible for one specific variable to suggest a trend in birthweights, but when other factors are taken into account this trend may no longer exist as it is explained by other factors. Several variables previously discussed in this chapter have appeared to be related to the occurrence of low birthweight. This section discusses two factors that are undoubtedly related to birthweight, gestational age and maternal height, and also investigates Womersley neighbourhood type.

In order to correct the proportions of each birthweight category for gestational age, the data were divided into eight gestational age categories, and for each of these categories, the proportion of VLBW and LBW babies for each neighbourhood type or maternal height category were calculated. The data used in this section are from all Glasgow resident mothers, delivering singleton livebirths in Glasgow hospitals in 1981 to 1991 inclusive.

3.5.1 Neighbourhood Type

The proportions of babies born in each birthweight category and neighbourhood type for each gestational category are shown in table 3.17. This shows that although there seems to be no difference over neighbourhood type for these babies born at less than 32 weeks, or for those babies born at 41 weeks or beyond, for all other gestational ages there appears to be an increase in the proportion of LBW babies as the neighbourhood type increases. This suggests that unless the baby is very pre-term, or late, then those women living in less affluent areas of Glasgow are more likely to give birth to a LBW baby.

3.5.2 Maternal Height

The numbers of babies born in each maternal height category and birthweight category are shown in table 3.18. Excepting those babies born at under 32 weeks gestation, and at 41 weeks and over, there seems to be a clear trend in the rate of LBW. The proportion of babies that are LBW at a given gestational age decreases as the height of the mother increases. This is to be expected, as it would be supposed that smaller mothers would be more likely to give birth to smaller babies.

Under 32 weeks				32-36 weeks			
Womersley	Prop. VLBW	Prop. LBW	Prop. NBW	Womersley	Prop. VLBW	Prop. LBW	Prop. NBW
1 (n=85)	0.612	0.353	0.035	1 (n=582)	0.029	0.469	0.502
2 (n=83)	0.711	0.289	0.000	2 (n=450)	0.033	0.453	0.513
3 (n=100)	0.720	0.260	0.020	3 (n=573)	0.021	0.480	0.499
4 (n=131)	0.687	0.305	0.008	4 (n=774)	0.031	0.465	0.504
5 (n=253)	0.656	0.324	0.020	5 (n=1392)	0.026	0.508	0.466
6 (n=75)	0.667	0.280	0.053	6 (n=461)	0.048	0.469	0.484
7 (n=174)	0.603	0.368	0.029	7 (n=1116)	0.036	0.528	0.436
8 (n=282)	0.652	0.337	0.011	8 (n=1385)	0.038	0.512	0.451
37 weeks				38 weeks			
Womersley	Prop. VLBW	Prop. LBW	Prop. NBW	Womersley	Prop. VLBW	Prop. LBW	Prop. NBW
1 (n=636)	0.000	0.101	0.899	1 (n=2022)	0.000	0.044	0.956
2 (n=461)	0.000	0.126	0.874	2 (n=1516)	0.001	0.043	0.956
3 (n=584)	0.002	0.108	0.890	3 (n=1692)	0.000	0.046	0.954
4 (n=817)	0.001	0.154	0.845	4 (n=2066)	0.000	0.072	0.928
5 (n=1307)	0.001	0.151	0.848	5 (n=3553)	0.000	0.073	0.927
6 (n=488)	0.000	0.129	0.871	6 (n=1264)	0.000	0.080	0.920
7 (n=1047)	0.001	0.172	0.827	7 (n=2628)	0.000	0.080	0.920
8 (n=1284)	0.001	0.166	0.833	8 (n=3460)	0.000	0.070	0.930
39 weeks				40 weeks			
Womersley	Prop. VLBW	Prop. LBW	Prop. NBW	Womersley	Prop. VLBW	Prop. LBW	Prop. NBW
1 (n=3440)	0.000	0.013	0.987	1 (n=5609)	0.000	0.007	0.993
2 (n=2677)	0.000	0.021	0.979	2 (n=4338)	0.000	0.009	0.991
3 (n=2825)	0.000	0.017	0.983	3 (n=4634)	0.000	0.007	0.993
4 (n=3272)	0.000	0.027	0.973	4 (n=5278)	0.000	0.014	0.986
5 (n=5656)	0.000	0.033	0.967	5 (n=9188)	0.000	0.015	0.985
6 (n=2065)	0.000	0.033	0.966	6 (n=2901)	0.000	0.012	0.988
7 (n=3827)	0.000	0.038	0.962	7 (n=5475)	0.000	0.021	0.979
8 (n=5200)	0.000	0.039	0.961	8 (n=8286)	0.000	0.019	0.981
41 weeks				42+ weeks			
Womersley	Prop. VLBW	Prop. LBW	Prop. NBW	Womersley	Prop. VLBW	Prop. LBW	Prop. NBW
1 (n=2733)	0.000	0.004	0.996	1 (n=404)	0.000	0.000	1.000
2 (n=2203)	0.000	0.001	0.999	2 (n=321)	0.000	0.006	0.994
3 (n=2409)	0.000	0.006	0.994	3 (n=325)	0.000	0.000	1.000
4 (n=2832)	0.000	0.005	0.995	4 (n=529)	0.000	0.006	0.994
5 (n=4336)	0.000	0.006	0.994	5 (n=692)	0.000	0.006	0.994
6 (n=1476)	0.000	0.007	0.993	6 (n=252)	0.000	0.008	0.992
7 (n=2369)	0.000	0.011	0.989	7 (n=326)	0.000	0.012	0.988
8 (n=4052)	0.000	0.006	0.994	8 (n=652)	0.000	0.006	0.994

Table 3.17: Proportion of births in each birthweight category by Womersley neighbourhood type and gestational age, singleton livebirths between 1981 and 1991.

Under 32 weeks				32-36 weeks			
Maternal Height (cm)	Prop. VLBW	Prop. LBW	Prop. NBW	Maternal Height (cm)	Prop. VLBW	Prop. LBW	Prop. NBW
<150 (n=74)	0.595	0.405	0.000	<150 (n=393)	0.041	0.583	0.377
150-154 (n=188)	0.622	0.351	0.027	150-154 (n=1148)	0.048	0.524	0.428
155-159 (n=279)	0.685	0.290	0.025	155-159 (n=1759)	0.024	0.512	0.464
160-164 (n=282)	0.638	0.344	0.018	160-164 (n=1702)	0.026	0.484	0.490
165-169 (n=147)	0.619	0.361	0.020	165-169 (n=809)	0.037	0.423	0.540
170-174 (n=40)	0.800	0.200	0.000	170-174 (n=220)	0.032	0.368	0.600
175-179 (n=7)	0.857	0.143	0.000	175-179 (n=45)	0.022	0.222	0.756
180-184 (n=0)	-	-	-	180-184 (n=7)	0.000	0.286	0.714
185-189 (n=0)	-	-	-	185-189 (n=1)	0.000	0.000	1.000

37 weeks				38 weeks			
Maternal Height (cm)	Prop. VLBW	Prop. LBW	Prop. NBW	Maternal Height (cm)	Prop. VLBW	Prop. LBW	Prop. NBW
<150 (n=418)	0.000	0.196	0.804	<150 (n=983)	0.000	0.137	0.863
150-154 (n=1063)	0.001	0.205	0.794	150-154 (n=2856)	0.000	0.089	0.911
155-159 (n=1827)	0.001	0.153	0.846	155-159 (n=4992)	0.000	0.067	0.933
160-164 (n=1720)	0.000	0.122	0.878	160-164 (n=4808)	0.000	0.050	0.950
165-169 (n=834)	0.001	0.096	0.903	165-169 (n=2484)	0.000	0.042	0.958
170-174 (n=274)	0.000	0.106	0.894	170-174 (n=821)	0.000	0.040	0.960
170-179 (n=53)	0.000	0.019	0.981	175-179 (n=146)	0.000	0.027	0.973
180-184 (n=9)	0.000	0.000	1.000	180-184 (n=23)	0.000	0.000	1.000
185-189 (n=0)	-	-	-	185-189 (n=7)	0.000	0.000	1.000
190-194 (n=0)	-	-	-	190-194 (n=1)	0.000	0.000	1.000
195-199 (n=0)	-	-	-	195-199 (n=1)	0.000	0.000	1.000

39 weeks				40 weeks			
Maternal Height (cm)	Prop. VLBW	Prop. LBW	Prop. NBW	Maternal Height (cm)	Prop. VLBW	Prop. LBW	Prop. NBW
<150 (n=1215)	0.000	0.077	0.923	<150 (n=1598)	0.000	0.039	0.961
150-154 (n=4189)	0.000	0.046	0.954	150-154 (n=6229)	0.000	0.023	0.977
155-159 (n=7727)	0.000	0.031	0.969	155-159 (n=11956)	0.000	0.016	0.984
160-164 (n=7998)	0.000	0.023	0.977	160-164 (n=13028)	0.000	0.010	0.990
165-169 (n=4450)	0.000	0.016	0.984	165-169 (n=7274)	0.000	0.006	0.994
170-174 (n=1489)	0.001	0.010	0.989	170-174 (n=2507)	0.000	0.002	0.998
175-179 (n=292)	0.000	0.014	0.986	175-179 (n=485)	0.000	0.004	0.996
180-184 (n=39)	0.000	0.026	0.974	180-184 (n=65)	0.000	0.000	1.000
185-189 (n=5)	0.000	0.000	1.000	185-189 (n=10)	0.000	0.000	1.000
190-194 (n=1)	0.000	0.000	1.000	190-194 (n=0)	-	-	-

41 weeks				42+ weeks			
Maternal Height (cm)	Prop. VLBW	Prop. LBW	Prop. NBW	Maternal Height (cm)	Prop. VLBW	Prop. LBW	Prop. NBW
<150 (n=616)	0.000	0.010	0.990	<150 (n=93)	0.000	0.043	0.957
150-154 (n=2734)	0.000	0.013	0.987	150-154 (n=433)	0.000	0.007	0.993
155-159 (n=5889)	0.000	0.005	0.995	155-159 (n=915)	0.000	0.009	0.991
160-164 (n=6426)	0.000	0.004	0.996	160-164 (n=1038)	0.000	0.003	0.997
165-169 (n=3891)	0.000	0.002	0.998	165-169 (n=593)	0.000	0.002	0.008
170-174 (n=1388)	0.000	0.004	0.996	170-174 (n=214)	0.000	0.000	1.000
175-179 (n=278)	0.000	0.007	0.993	175-179 (n=51)	0.000	0.000	1.000
180-184 (n=33)	0.000	0.000	1.000	180-184 (n=6)	0.000	0.000	1.000
185-189 (n=2)	0.000	0.000	1.000	185-189 (n=1)	0.000	0.000	1.000

Table 3.18: Proportion of births in each birthweight category by maternal height and gestational age, singleton livebirths between 1981 and 1991.

3.6 Modelling

In this section, the probability of a mother giving birth to a baby of low birthweight is modelled using stepwise logistic regression with six possible variables - gestational age, maternal height, maternal condition, maternal age, marital status and Womersley neighbourhood type. As the Womersley neighbourhood type was created using data from the 1981 census, only 1981 births are used here. In addition only first time mothers are considered. Gestational age was used as an explanatory variable as it is obviously related to birthweight. Maternal height and Womersley neighbourhood type have been shown to be related to the occurrence of low birthweight after correction for gestational age. Maternal condition is defined as the presence of placenta previa, premature placental separation, antepartum haemorrhaging, or hypertension during pregnancy. This is an unknown quantity that may have some effect on birthweight and is included here in order to investigate this. Maternal age and marital status are thought to be related to birthweight. Maternal age was categorised as less than 16, 16-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49 and 50-54, and maternal height was categorised as under 150cm, 150-154cm, 155-159cm,160-164cm, 165-169cm, 170-174cm, 175-179cm, 180-184cm and 185-189cm. Maternal condition was categorised as present or absent, gestation was split into four categories, very premature (<32 weeks), premature (32-36 weeks), term (37-41 weeks), and beyond term (42 weeks and beyond), and marital status was categorised as married or unmarried, where unmarried included divorced, separated, widowed and other as well as single. 4755 cases were used in the analysis. Gestational age was the first variable entered into the model with a p-value of $< 10^{-16}$. Neighbourhood type was then entered with a p-value of 5.2×10^{-4}. The next variable to be added was maternal height with a p-value of 0.0065. At the next step the most significant variable was marital status with p=0.058. However this value is not significant at the 5% level, and taking multiple comparisons into account it is not significant at the 10% level either, and as a result this variable is not entered into the model.

The results of the stepwise logistic regression are given in table 3.19. 95 % confidence limits are given for each of the parameter estimates. The baseline categories for the variables were very premature, <150 cm, and neighbourhood type 1. In order to calculate the fitted value of the probability of a mother delivering a low birthweight

Parameter	Estimate	Upper Limit	Lower Limit
Intercept	3.129	5.268	0.990
Premature	-3.073	-1.055	-5.091
Term	-6.919	-4.916	-8.922
Beyond term	-7.368	-5.137	-9.599
150-154 cm	-0.019	0.576	-0.614
155-159 cm	-0.364	0.215	-0.943
160-164 cm	-0.755	-0.152	-1.358
165-169 cm	-0.799	-0.122	-1.476
170-174 cm	-0.975	0.132	-2.082
175-179 cm	-5.689	9.788	-21.167
180-184 cm	-5.616	55.246	-66.478
185-189 cm	-6.016	80.137	-92.169
Neighbourhood 2	0.759	1.578	-0.060
Neighbourhood 3	0.504	1.322	-0.314
Neighbourhood 4	0.919	1.731	0.107
Neighbourhood 5	0.912	1.653	0.171
Neighbourhood 6	1.240	2.062	0.418
Neighbourhood 7	1.378	2.126	0.630
Neighbourhood 8	1.035	1.770	0.300

Table 3.19: Parameter estimates for logistic regression, singleton livebirths in 1981.

baby, the intercept value is added to the parameter estimates for the categories that the mother falls into - for baseline categories there are no parameter estimates. After obtaining this value, t say, for the mother, the probability of a low birthweight baby is calculated from the formula

$$\Pr(\textit{low birthweight baby}) = \exp(t)/(1 + \exp(t)).$$

	Type A	Type B
Neighbourhood 1	0.509	0.011
Neighbourhood 2	0.689	0.022
Neighbourhood 3	0.632	0.017
Neighbourhood 4	0.722	0.026
Neighbourhood 5	0.721	0.026
Neighbourhood 6	0.702	0.035
Neighbourhood 7	0.805	0.040
Neighbourhood 8	0.745	0.029

Table 3.20: Probability estimates for mothers by neighbourhood type, singleton live-births in 1981.

In order to illustrate this, table 3.20 shows the fitted probabilities for two types of mother - mothers who are delivering prematurely (32-36 weeks) and are 150-154 cm tall (Type A), and mothers who are delivering at term and are 160-164 cm tall (Type B), for each of the eight different neighbourhood types. Comparing a mother of type A with a mother of type B, the fitted probabilities of a type A mother having a low birthweight baby are much larger than a type B mother. Comparing neighbourhood types shows a tendency for the probability of a low birthweight baby to increase as the neighbourhood type increases from affluent areas to deprived areas.

3.7 Discussion

The analyses carried out have raised some interesting points. Merely looking at the data from a univariate point of view suggests that several variables may have an effect on the birthweight category of the baby, these being neighbourhood type, marital status, previous induced and spontaneous abortions, maternal height, and gestational age. However considering a multivariate point of view, a model that contains maternal height, gestational age and neighbourhood type seems to best fit the data, with marital status, maternal age, and maternal condition all shown to be insignificant after correcting for these variables.

Work involving a linked data set containing first time pregnancies only, where each pregnancy rather than each visit is represented as a single record, is carried out in chapter 5. Successive pregnancies to the same mother are also linked together in chapter 5 in order to follow women through more than one pregnancy, and to link in any neonatal data recorded on the SMR11 form.

Chapter 4

Deprivation Measures

Various scoring systems have been created in order to measure small area deprivation. This chapter considers four of these methods and discusses them in detail. Of these four measures, two were originally created from data in Scotland, one to explain trends in health inequalities and one to ease health status assessment. The remaining two were created from data in England, one to explain trends in health inequalities and one to improve GP services. A new deprivation measure for postcode sectors in the Greater Glasgow Health Board area is calculated using the 1991 small area census data, and this new measure is compared with those measures created from Scottish data.

4.1 Previous Deprivation Measures

4.1.1 The Townsend Score

The Townsend score [45] was created in an attempt to explain trends in health inequalities within the United Kingdom. Townsend used the North East of England as a case study, and calculated a deprivation measure at local authority ward level for the 678 wards in that area. Townsend does not appear to back up his choice of census variables to be included in the deprivation measure with any statistical techniques. However the variables that he includes - unemployment, car ownership, overcrowding, and home ownership - are variables which one would intuitively expect to have some link with deprivation.

In creating a deprivation score, Townsend firstly carries out a log transformation on two of his chosen variables - unemployment and overcrowding - in order to produce distributions that are closer to Normality. All four variables are then standardised and combined with equal weighting to give a deprivation index for each local authority ward. An advantage of the Townsend measure is that, assuming the four variables are available for the area of interest, it is quite simple to calculate an equivalent Townsend measure for any area of interest.

The main disadvantage of this measurement is that it is not composed of any social variables. Townsend states that he believes social variables, such as colour or single parenthood, should not have a bearing on deprivation. In an ideal world this may be the case, but it appears obvious that some social variables should be considered in deriving a deprivation measure. Townsend also refuses to consider social class for the same reason.

4.1.2 The Carstairs Score

The Carstairs score [10] was also created in an attempt to explain trends in health inequalities. This score is Scotland specific as at the time, work was being carried out in England and Wales to explain these trends, and this work was not extended to Scotland.

A deprivation measure was created for each of the 1010 postcode sectors in Scotland. The choice of census variables is the extension of some previous work, although the work does not appear to be based on any statistical analyses. However the variables used, as those involved in the Townsend score, are variables that would be expected to be linked to deprivation. These variables are overcrowding, male unemployment, low social class, and car ownership. These are standardised and then combined with equal weights to produce a continuous score, and this score is then used to split the postcode sectors into 7 categories, denoted 1 (most affluent) to 7 (most deprived).

As with Townsend, it is relatively easy to calculate an equivalent score to Carstairs for any area of interest. Carstairs also uses social class as a variable, and this is useful as it has been used previously as a deprivation indicator.

The disadvantage of the Carstairs score is that much of the information it gives is lost in the conversion from a continuous score into categories. Also, the method behind the cut-off points for the categories is not apparent, as the categories are equal in neither sector size nor population size.

4.1.3 The Womersley Score

The Womersley score, or neighbourhood type [49], was created in order to make the health status of different communities within Greater Glasgow Health Board (GGHB) easier to assess. If a certain type of event occurs quite rarely, examining together different communities that are quite similar can achieve better results.

Neighbourhood types were calculated by a series of procedures on the 1981 small area census data. Twenty-nine variables, shown in table 4.1, were selected to attempt to identify differences between postcode sectors in GGHB, and the percentage of the population of each postcode sector who were identified with each variable was calculated. The postcode areas of G1 and G2 were merged together because of their small population size, and there was no distinction made between two parts of a postcode sector which fell in different local government districts. Any postcode sector that, after checks for skewness, appeared to contain a noticeable outlier on one or more variables, was removed.

In order to classify the remaining postcode sectors, a principal components analysis was carried out on the correlation matrix of the 29 variables. It was decided to use the first 4 components in the procedure, as together these accounted for 86% of the total variance. A cluster analysis was then carried out on these components, using hierarchical clustering and the complete linkage method, in order to produce 8 clusters of postcode sector, which are numbered 1 for most affluent to 8 for least affluent.

This categorisation system takes into account far more variables than any other established method of calculating deprivation. While this will allow for variables that are not thought to influence deprivation but do, it will also include those variables that do not affect deprivation at all. The number of clusters created in the cluster analysis is arbitrary and could be chosen to be higher or lower.

% resident population less than 4 years old
% resident population aged 4 to 15 years old
% resident population at least 75 years old
% population born in New Commonwealth or Pakistan
% households with 1 or 2 rooms
% households which are owner occupied
% households rented from the Local Authority
% households which are 'other' rented furnished accommodation
% households with spaces that are vacant
% households with no car
% households with at least 2 cars
% households without exclusive use of amenities
% resident population married
% households with no children
% households with at least 3 children under 15 years old
% pensioners living alone
% economically active residents - not employed
% resident population over 16 who are students
% residents in household aged 25 to 44 years old
% residents in household aged 45 to 64 years old
% single parent families
% single non-pensioner households
% working wives in full time employment, aged 16 to 59 years old
% households with over 6 rooms
% occupancy norm of -1 or more
% resident population with low social economic group
% resident population professional and managerial workers
% resident population non-manual workers
% resident population skilled manual workers

Table 4.1: Factors used in calculating Womersley neighbourhood type.

4.1.4 The Jarman Score

The Jarman score [26, 27] was created in order to identify underprivileged areas with a view to improving GP services. Previous work had drawn attention to both geographical variations in problems dealt with by primary care services and variation in services provided between areas, and Jarman's work investigated this further.

Approximately 4000 GPs and organisations involved in health care, all within London, were asked for comments on primary care services. Of these, less than 10% replied - approximately 180 GPs and 190 organisations. Jarman analysed the replies and created 21 possible factors that may have increased a GP's workload or affected their pressure of work. From this a further questionnaire was sent out to 2614 randomly selected GPs in England and Wales, asking them to score each of the 21 factors on a scale from 0 (no problem) to 9 (very problematic), according to the degree to which it affected pressure of work, or increased workload. Of the 2614 questionnaires, 70% were analysed.

Thirteen social factors were used in Jarman's questionnaire. Of these, the proportion of elderly over 65 was not included as this was already weighted for in GP's remuneration, and the proportion of elderly living alone was included. Also excluded were crime rate, as it is not a census variable and is highly correlated with overcrowding which is included; difficulty in being able to visit patients, which again is included in remuneration; non-married couple families, which was difficult to determine from census data; and households lacking basic amenities, which was excluded as this ruled out poor housing estates where basic amenities are a council requirement. The remaining variables together with the average score obtained are shown in table 4.2. The Jarman score was calculated for each London borough by calculating the average score for each variable over all boroughs and using these as the weights in a weighted sum of the proportion of the population of each borough satisfying each social factor. A higher Jarman score indicated more difficulty for primary care services. This scoring system was later extended to each electoral ward in England and Wales, using the weights calculated for London boroughs.

The underprivileged area score can be considered as a measure of deprivation if we assume that those areas that require more primary care services do so because they are more deprived. Under this assumption, there are advantages to using Jarman

Social factor	Weighting
Elderly living alone	6.62
Children under 5	4.64
Lower social class	3.74
Unemployment	3.34
Single parent households	3.01
Overcrowded households	2.88
Highly mobile people	2.68
Ethnic minorities	2.50

Table 4.2: Social factors and weights used in the Jarman scoring system.

as a deprivation measure. It contains census variables, and of these there are some which would be expected to be related to deprivation. In addition it contains social class, which has been used previously as an indicator of deprivation.

The weights used for calculation of the Jarman score are satisfactory for the evaluation of primary care requirements. However, in the case of a deprivation measure, some variables are weighted too highly. For example, the proportion of elderly people living alone is weighted almost twice as much as the other variables, which would suggest that areas with sheltered housing schemes would be classed as deprived. Clearly this is a failing of the scoring system in its application to deprivation measurement. In addition, factors involved in the calculation of the Jarman score are based on the results of a London based questionnaire with a low response rate and this may have weakened the analysis of the responses.

4.1.5 Overview and Discussion

The Jarman index has generated criticism from several authors. Davey Smith [43] states that the index is biased towards London, having originally been created from data on London alone, and relies on out of date census data in its calculations. Talbot [44] also states that the index is London biased, with no Northern region appearing in the list of the twenty most deprived regions. He also criticises the use of electoral wards as geographical areas of interest, as the size of such areas varies greatly. Using psychiatric admissions in the London borough of Islington in 1985, Cotgrove et al. [13] showed that the Jarman index was correlated to neither admission rates nor to mean length of psychiatric stay, and concluded that the index should not be used for planning psychiatric service provision.

Ben-Shlomo et al. [3] compared the Jarman index with both the Townsend and Carstairs scores in order to assess the ability of the three scoring systems to predict GP workload. Workload was defined as the sum of the number of surgery consultations plus 2.5 times the number of consultations elsewhere, excluding preventative procedures. Using 25 practices in England and Wales it was shown that of the three scoring systems, Townsend was best at assessing GP workload, while Jarman was less valid than the others as it did not include car ownership or housing tenure. In addition, the weighting assigned by Jarman to children under five years old underestimated the additional workload created by these patients. It was suggested that using the Townsend score and weighting the capitation fee for children under five would be a better method of allocating payments and would also remove the London bias of the Jarman index.

Reading et al. [39] used the Townsend score to investigate differences in child health in over 21000 children resident in Northumberland between January 1985 and September 1990. Health measures of interest were the proportion of singleton livebirths under 2800 g, the proportion of births to teenage mothers, the proportion of 15 month old children not immunised against whooping cough, the proportion of infants not screened at 6 weeks of age, the proportion of infants not screened at 18 months, and the mean standardised height of children in each enumeration district. It was shown that there were significant differences between the most deprived 10% of areas and the most affluent 10% of areas for all of the health measures of interest except the proportion of infants not screened at 6 weeks, and in each case the proportions were higher in the most deprived areas.

Campbell et al. [9] compared the Jarman index with the Townsend score, unemployment rates in 1981, 1985 and 1990 and two Government departmental scoring methods in the Central Nottinghamshire Health Authority area. The health measures of interest were the standardised mortality ratio for all ages and for ages under 65, age specific hospital admission rates for 0-14 years and over 75 years, average admission rates for 1983-85 and 1989-90, standardised admission rates for 1983-85 and the permanent sickness rate. Correlations of the seven possible deprivation measures with health measures were calculated and ranked from 1 (highest correlation) to 7 (lowest correlation) for each health measure. For each health measure the Jarman index was ranked 7th, with the exception of admission rates for the over 75 year old age group, where it was ranked 6th. For all admission rates except the over 75 age group, the 1990 unemployment rate was ranked first, while Townsend was ranked first for the standardised mortality ratios.

McLoone and Boddy [33] compared mortality experience of Scottish postcode sectors as characterised by Carstairs scores for both 1980-82 and 1990-92. Carstairs scores for 1990-92 were calculated using the 1991 census data and using the same methods as for the 1981 census data. It was shown that areas that were deprived in 1981 were more deprived in 1991, and that the mortality experiences in these areas worsened in comparison with affluent areas and with Scotland as a whole.

Of the four methods described here, the Womersley score uses a very different technique to the others, involving several statistical methods. However, this is also the most computationally intensive method and thus makes the score difficult to reproduce for other areas of interest. Carstairs and Townsend use similar methods in that they combine census variables by way of an unweighted sum to produce a deprivation measure. The only differences between the two methods are the transformations of some of the Townsend variables, and the categorisation of the Carstairs score.

As we have seen, some variables are involved in more than one of the four measurements. Inspection of the data shows that two variables - overcrowding and unemployment - are involved in the calculation of all 4 scores, while car ownership and social class are involved in 3. However, this finding cannot be used to state categorically that these variables are related to deprivation, as one of the measurements uses 29 variables in its calculation and the other 3 involve subjective inclusion. The fact that these variables are included in 3 or 4 of the measurements is more an indication of which variables are thought of as being related to deprivation, rather than showing a relationship between these variables and deprivation.

It appears that while Carstairs and Townsend are both correlated with health measures, the Jarman index has several problems. It has been shown in the above papers to be biased towards London and to be less successful than other deprivation measures in assessing GP workload, an area for which it was specifically created.

4.2 A Continuous Deprivation Score For Greater Glasgow Health Board

Within the Greater Glasgow Health Board (GGHB), there are 136 postcode sectors of interest. While some of these lie in the most prosperous areas of Europe, Glasgow as a whole compares unfavourably with the rest of Scotland. A Scottish score, such as Carstairs, will place many of the sectors in Glasgow into its most deprived category. With GGHB as the area of interest, it seems sensible to calculate a score that will give as much information as possible, yet not be influenced by other areas of Scotland.

A categorised measure brings with it many problems. If a cluster analysis is carried out, there must be a decision made on the number of clusters to be used. If a continuous score is calculated, as in Carstairs [10], and then turned into categories, there is the problem of choosing cut-off points. Also, it is not ideal to calculate a continuous score and then lose information by converting it into categories.

For these reasons the calculation of a continuous, Glasgow specific deprivation measure is considered.

The data used in the calculation of this new deprivation measure come from the 1991 small area census data. There are 29 variables, as detailed previously in table 4.1, including variables that could be expected to have a bearing on deprivation.

A principal components analysis (PCA) was carried out on the data to reduce the dimensionality of the problem. Various transforms of the data were considered, but it was decided to use untransformed data as it would be easier to interpret the results. The analysis was carried out on the correlation matrix of the data. From table 4.3, it can be seen that 78.7% of the variation is explained by the first three principal components.

These principal components are difficult to interpret, as there do not appear to be any obvious contrasts or differences between various census variables. A different approach is suggested here, which is to group the census variables using the principal

Eigenvalue	11.27848	6.52565	5.02632
Percentage	38.89132	22.50225	17.33214
Cumulative	38.89132	61.39357	78.72571

Variable No.	Component 1	Component 2	Component 3
1	0.08093	-0.27403	-0.23781
2	-0.01373	-0.35142	-0.14389
3	-0.00214	0.17730	0.31967
4	-0.02219	0.18146	-0.21000
5	0.15649	0.26511	-0.03488
6	-0.28680	0.04195	-0.04145
7	0.22262	-0.14932	0.13284
8	0.00438	0.29139	-0.23785
9	0.14690	0.02505	-0.23566
10	0.29225	0.03158	0.02269
11	-0.27001	-0.06409	0.00694
12	0.02359	0.22797	-0.24104
13	-0.24749	-0.14114	0.13882
14	0.04744	0.35713	0.14343
15	0.01246	-0.29543	-0.18753
16	0.07193	0.17433	0.34212
17	0.28011	-0.03718	-0.02702
18	0.10809	0.23679	0.14986
19	-0.07499	0.12412	-0.35440
20	-0.02692	-0.02292	0.38505
21	0.27152	-0.02995	-0.03273
22	0.15559	0.26932	-0.18610
23	-0.23165	0.11525	-0.15449
24	-0.26159	-0.05924	0.02563
25	0.24504	0.05127	-0.15296
26	0.26854	-0.10425	0.02445
27	-0.22832	0.14192	-0.05859
28	-0.22496	0.09407	-0.01618
29	0.18906	-0.16989	0.08982

Table 4.3: Analysis of correlation matrix of untransformed data.

components as a basis for deciding which group a variable should be placed in. This is done as follows:

1. Consider the absolute values of the parameter value for every variable for each of the three principal components.

2. If the parameter value for the first component is much larger than the other two then place the variable in group 1, and similarly for the second and third components with groups 2 and 3.

3. For those variables with two or more similar parameter values, leave until all other variables have been considered and then place in whichever group seems to contain variables that are similar in nature.

4. When all variables have been placed into one of the three groups, consider whether there are any patterns of variables and, if there are, consider moving some variables to groups that are more appropriate.

Carrying out this procedure produces the three groups shown in table 4.4. The grouping of variables suggests there is a deprivation group, an age group, and a housing/general group.

In order to compare these results with those from the principal component analysis in terms of amount of variation explained, the parameters must be scaled so that for each group the sum of squares of the parameters is 1, as with principal components. Two possibilities are considered - a weighted score, where the amount of weighting a variable receives is proportional to its parameter, and an unweighted score, where all variables have equal weighting. The weighted score may be expected to fare better than the unweighted score, as this takes into account the size of the corresponding parameters in the principal components, and not just their sign.

The percentage of variation explained by the groups using the unweighted score has dropped from 78% to 59%, as shown in table 4.5. This suggests that while the principal components were too difficult to interpret, this more simplistic method loses some of the explanation that the PCA gave us.

The weighted scores shown in table 4.6 fare similarly to the unweighted scores, suggesting that attempting to correct for the size of the parameters in the princi-

Deprivation

% households with 1 or 2 rooms

% households which are owner occupied

% households with spaces which are vacant

% households with no car

% households with at least 2 cars

% resident population married

% economically active residents - not employed

% single parent families

% working wives in full time employment, aged 16 to 59 years old

% households with over 6 rooms

% occupancy norm of -1 or more

% resident population with low social economic group

% resident population professional and managerial workers

% resident population non-manual workers

% resident population skilled manual workers

Housing/General

% population born in New Commonwealth or Pakistan

% households rented from the Local Authority

% households which are 'other' rented furnished accommodation

% households without exclusive use of amenities

% households with no children

% households with at least 3 children under 15 years old

% single non-pensioner households

Age

% resident population less than 4 years old

% resident population aged 4 to 15 years old

% resident population at least 75 years old

% pensioners living alone

% resident population over 16 who are students

% residents in household aged 25 to 44 years old

% residents in household aged 45 to 64 years old

Table 4.4: Groups created from principal component analysis.

Group	Variation Explained	Cumulative Explained
Deprivation	33.81444%	33.81444%
Housing/General	11.74752%	45.56196%
Age	13.52220%	59.08416%

Table 4.5: Percentage variation explained by unweighted scores.

pal components makes little difference. As the weighted and unweighted scores are similar, the unweighted score shall be considered.

Group	Variation Explained	Cumulative Explained
Deprivation	34.86366%	34.86366%
Housing/General	11.75046%	46.61412%
Age	12.57955%	59.19367%

Table 4.6: Percentage variation explained by weighted scores.

Looking at plots of the three variables, shown in figure 4.1, it can be seen that all are symmetrical, with no obvious outliers in the deprivation (first) variable. Considering only the deprivation score, shown in figure 4.2, it can be seen that although there may be cut-offs to indicate the most affluent and deprived areas, there is nothing to indicate that the centre portion can be split into categories.

This deprivation score is denoted by the 'Murray Score' and shall be considered as a possible deprivation measure for GGHB. However it loses explanatory power as a result of grouping variables after the principal components analysis, and further investigation is needed to discover whether this measure will be of any use in measuring deprivation. In order to do this, it is compared with the Womersley and Carstairs scoring systems.

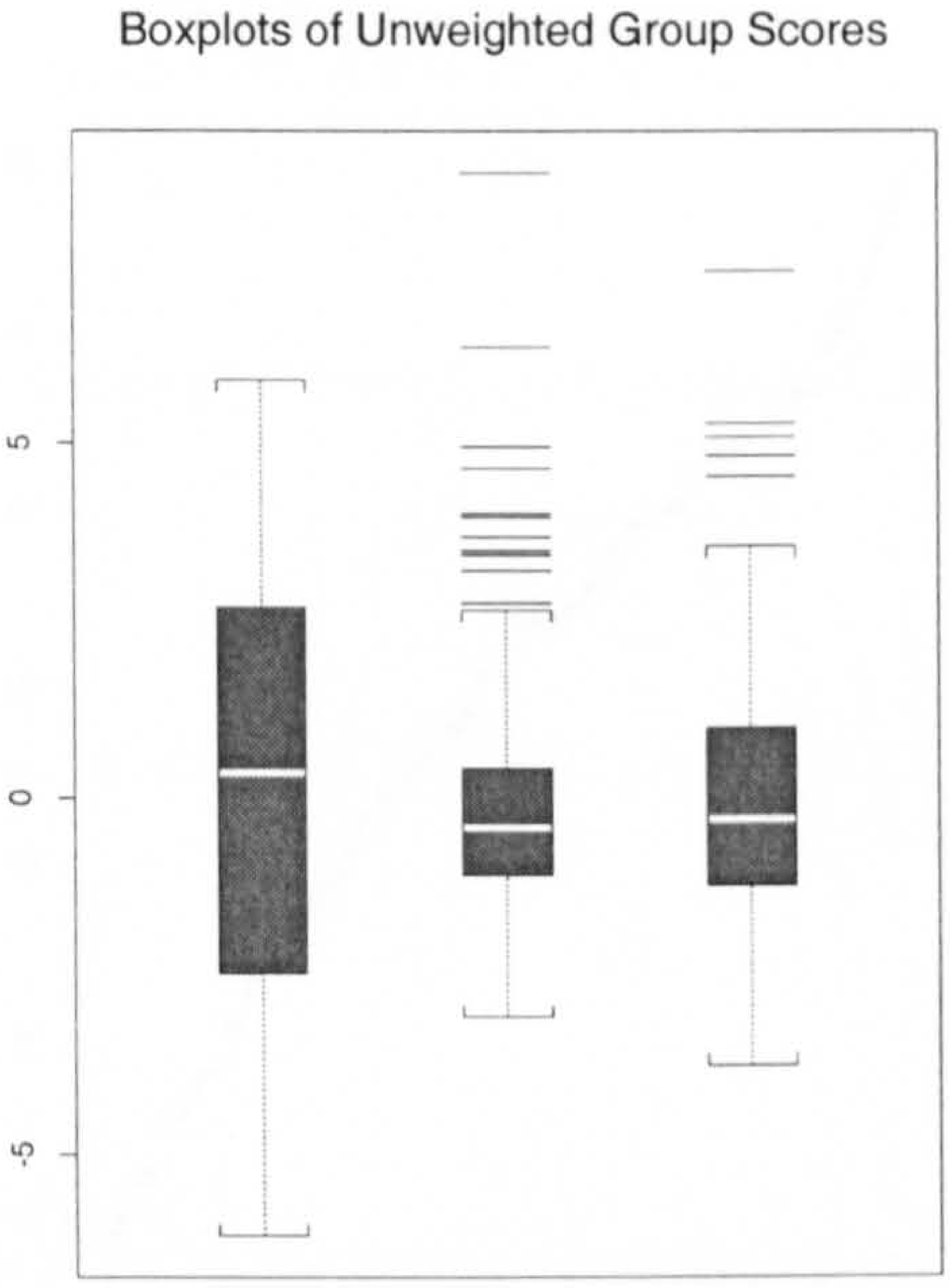

Figure 4.1: Boxplots of all three groups produced.

4.3 Comparison Of Deprivation Measures Calculated for the Greater Glasgow Health Board Area

Here the Murray, Womersley and Carstairs scores are compared for postcode sectors in GGHB. The Carstairs and Womersley categories are compared in table 4.7. There are some sectors that fall within two districts in GGHB, for example Glasgow City and Bearsden, which Carstairs treats separately yet Womersley does not. As a result, these sectors are not compared here. Postcode areas G1 and G2, which Carstairs splits up into postcode sectors but Womersley merges together are also not compared. Altogether 104 postcode sectors are compared.

With the exception of one or two sectors, the table shows scores clustered around the diagonal, suggesting that the two measures are comparable. However, where Carstairs places 54 postcode sectors in the 2 most deprived areas, Womersley has only 45 in the 3 most deprived areas. This suggests that Womersley is able to split up deprived areas in Glasgow better than Carstairs, as Carstairs is Scotland specific

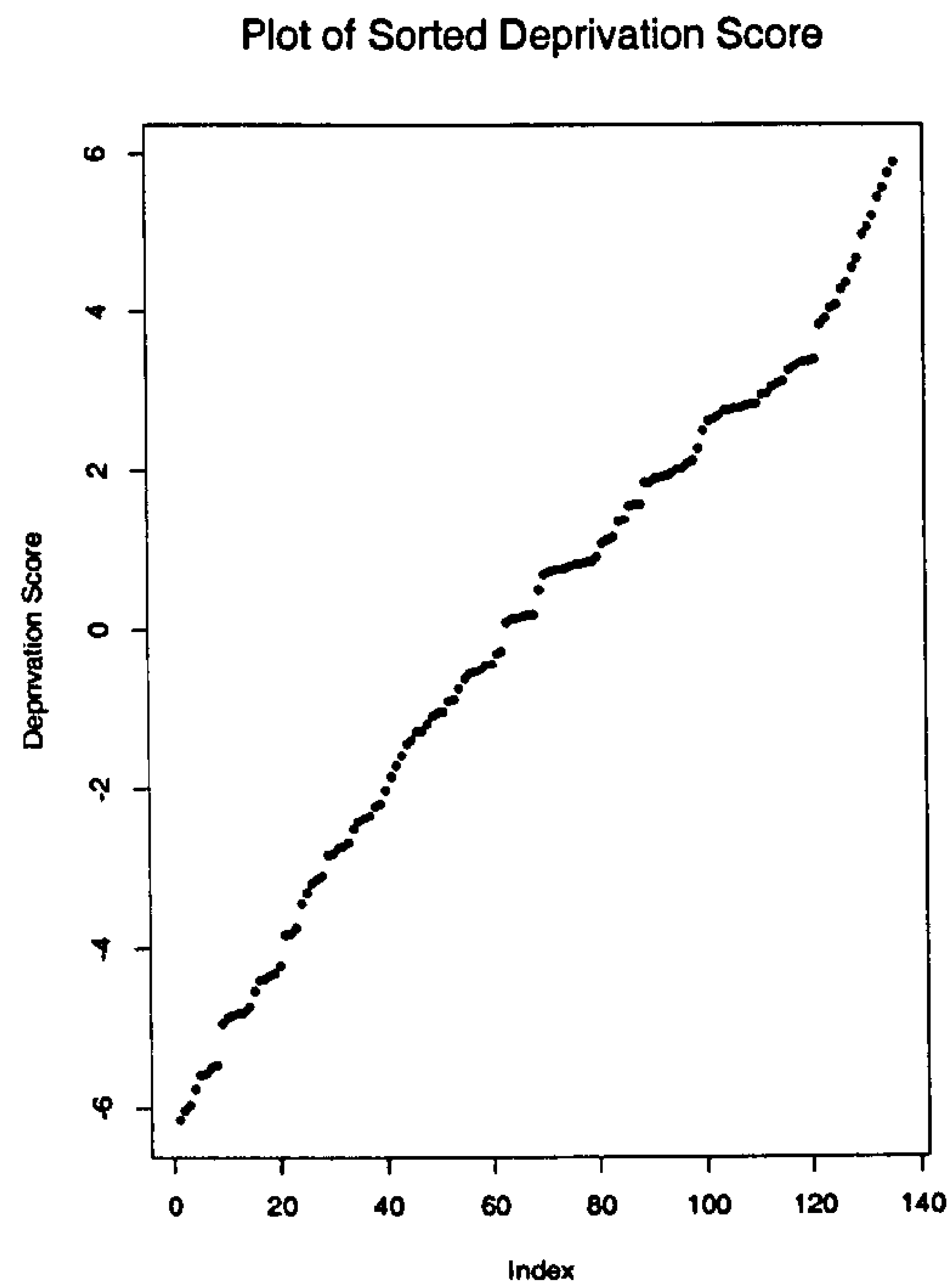

Figure 4.2: Plot of deprivation scores.

Carstairs :	1	2	3	4	5	6	7
Womersley :							
1	9	2	1	0	0	0	0
2	2	3	1	0	1	0	0
3	1	2	5	2	2	0	0
4	0	0	1	3	6	7	0
5	0	0	0	3	1	3	4
6	0	1	0	3	1	3	2
7	0	0	0	0	0	0	8
8	0	0	0	0	0	10	17

Table 4.7: Carstairs score versus Womersley neighbourhood type in GGHB.

and Womersley is Glasgow specific.

Next the Womersley categories and Murray score are compared. The Murray score is created from census data collected ten years after the data used in the calculation of the Womersley score, so some differences are to be expected. A plot of the Murray score for each postcode sector against the Womersley score for the same sector is shown in figure 4.3.

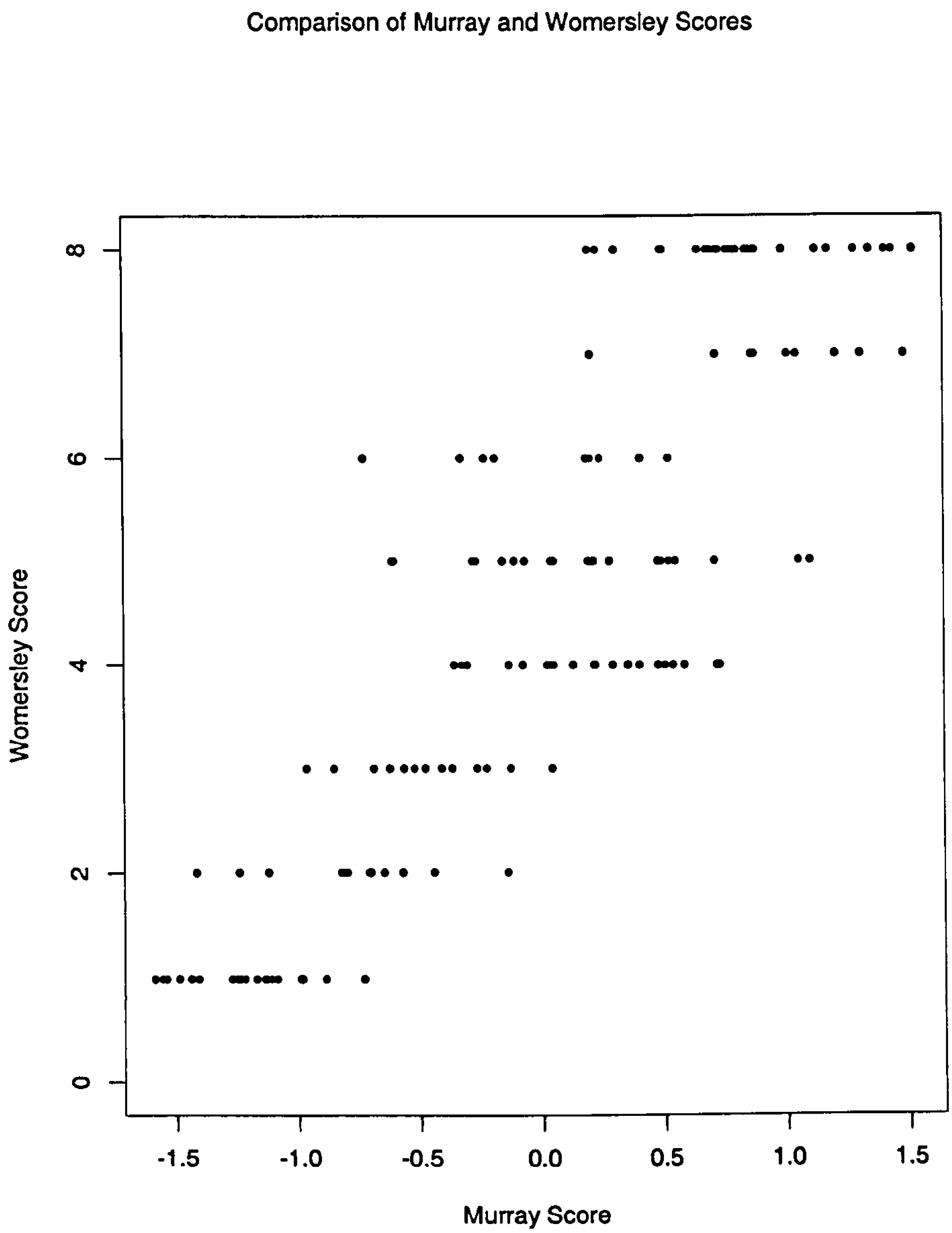

Figure 4.3: Murray score versus Womersley neighbourhood type.

As can be seen from the plot, the Murray score increases along with the Womersley category. The only exception to this is Womersley category 6, where postcode sectors seem to have a lower Murray score than those sectors in Womersley category

5. These sectors lie in the Hillhead and Gorbals areas of Glasgow, which improved over the ten year period 1981-1991 and so this difference is not unexpected.

Finally the Murray score is compared with the Carstairs categories, and also the continuous scores obtained by Carstairs before creating categories. From figure 4.4 it can be seen that as the Murray score increases the Carstairs category does also, and this is more pronounced than the comparison with Womersley, with only a few outliers in the plot. Comparison with the continuous Carstairs score in figure 4.5 shows that the two scores are comparable, with again very few outliers.

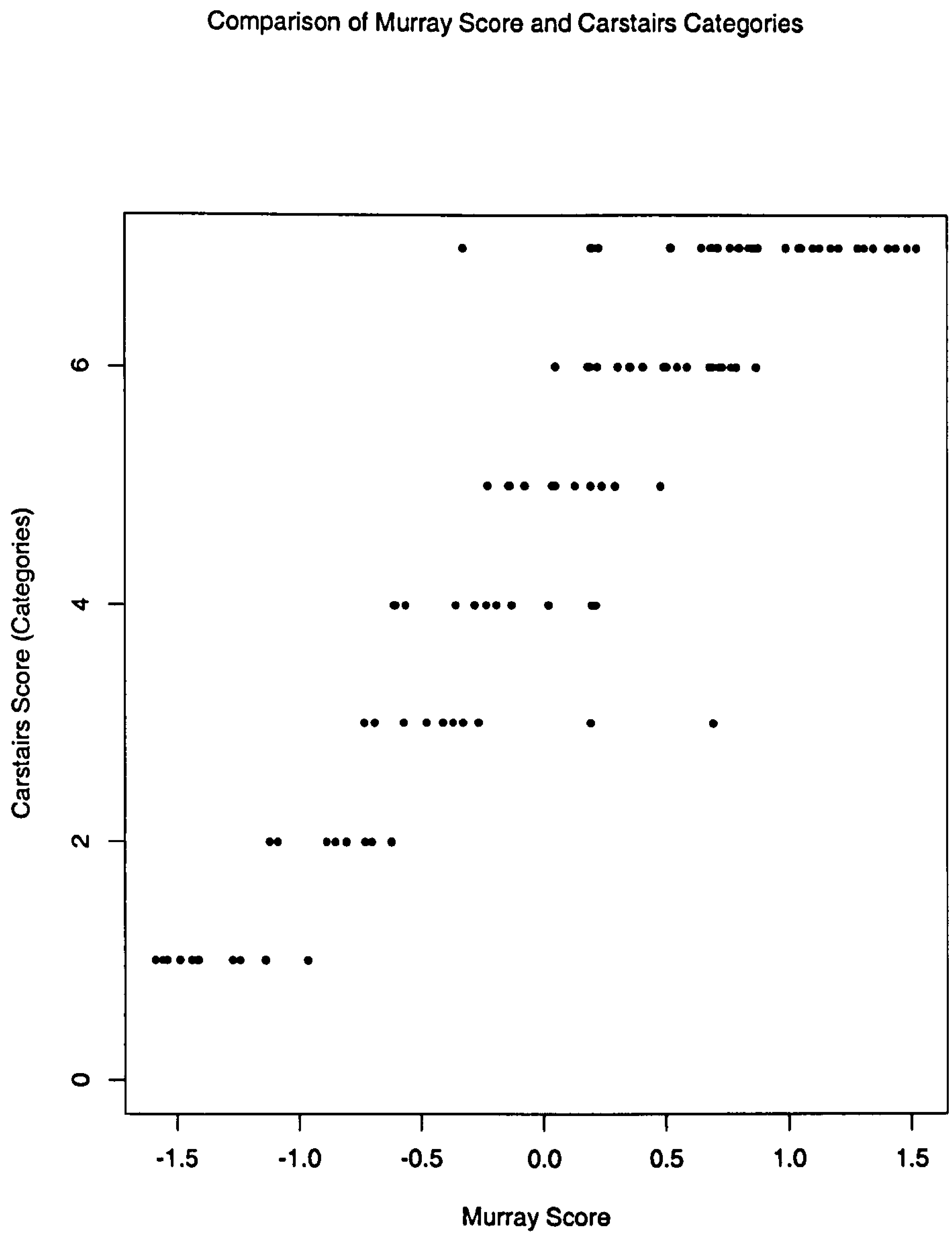

Figure 4.4: Murray score versus Carstairs categories.

4.4 Conclusions

From the discussion on deprivation measures it can be seen that while Carstairs and Townsend can be thought of as indications of deprivation, the Jarman index has generated some criticism, and it has been shown that the Jarman index is not the best method for calculating GP workload, which it was expressly created for.

The derived Murray score, while losing 20% of the variation explained by princi-

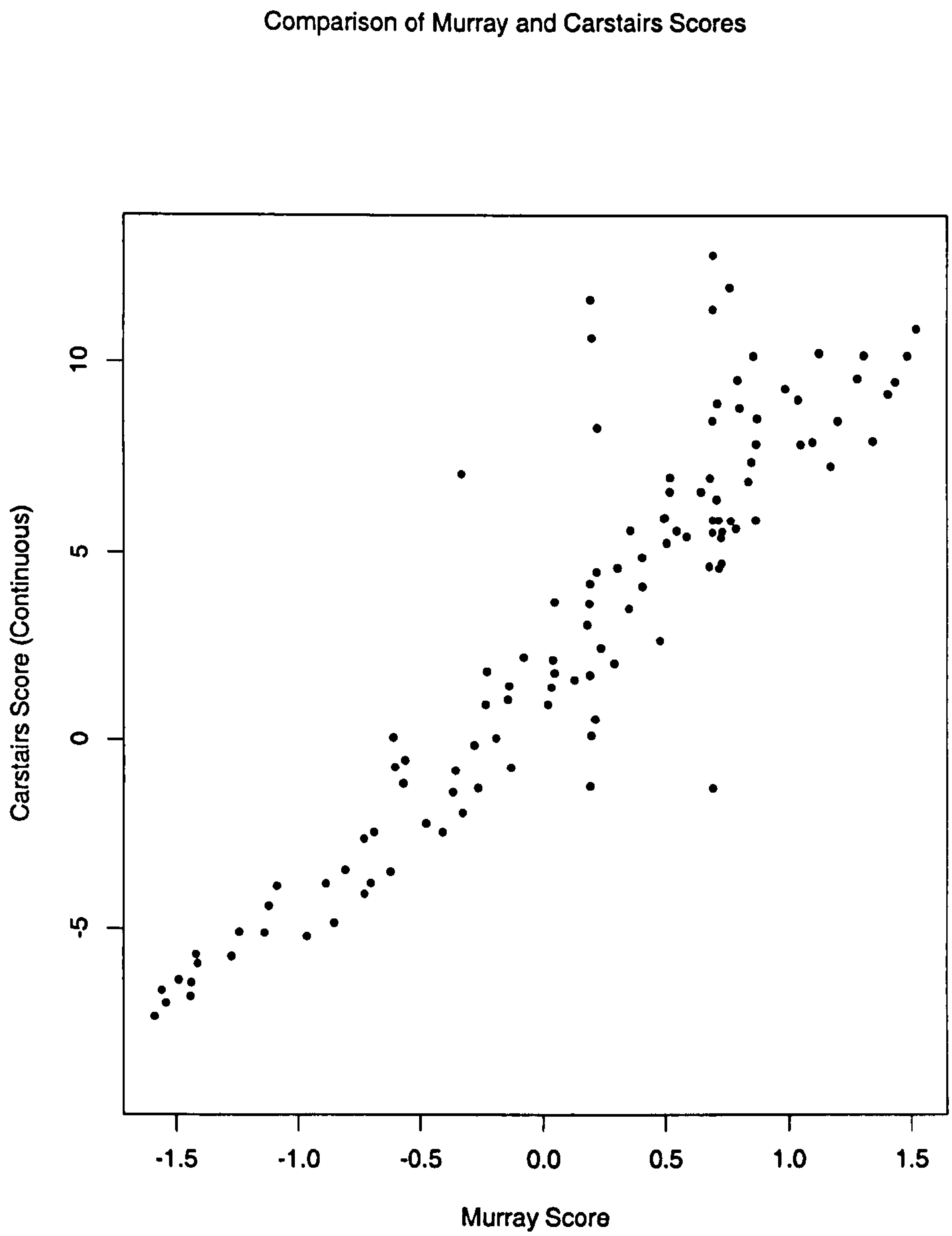

Figure 4.5: Murray score versus continuous Carstairs score.

pal components analysis, is similar to both the Womersley and Carstairs categories, and also to the Carstairs continuous score, which contains more information than the Carstairs categories. The scores are not expected to be identical as the Murray score is calculated from 1991 census data while the others use 1981 data. In addition to being calculated on more recent census data, the Murray score is Glasgow specific. As the main interest of this thesis is GGHB, the remainder of this thesis will use Womersley as a category measure of deprivation, and the Murray score as a continuous measure.

Chapter 5

Analysis of Linked Data

In this chapter a data set is considered that is similar to that used in chapter 3. In this data set each mother is given a unique maternal identification number. Each record in the data set belonging to a specific mother can be found by using this maternal number. Another difference between this and the previous data set is that each record refers to an entire pregnancy rather than one hospital visit. This data set is used to examine hospital resource use in GGHB, in particular to investigate whether there are differences in resource usage between hospitals, between deprivation types, and between the two years 1980 and 1991. Special care baby unit resources are considered, in order to investigate changes between 1983, when this care was first recorded on the neonatal SMR11 forms, and 1991. Finally those mothers who delivered their first child in 1980 and subsequently delivered a second child are considered to investigate whether the birthweight of the second child is in some way affected by the birthweight of the first child.

5.1 Introduction

The data received from the Information and Statistics Division of the National Health Service in Edinburgh has linked maternal SMR2 forms. This is done by using a weighted scoring system between two records in order to determine whether they belong to the same person or not [28]. The patients' surname, first initial, first name if available, sex, date of birth and postcode are used in this scoring system. Surnames are pre-processed to remove spelling inconsistencies and surnames which sound similar are considered to be the same. Records are compared by using a blocking system. Firstly, records are compared on sex, processed surname and initial. The more uncommon a matching characteristic is, the higher weighting it receives. If these three variables are the same for both records, date of birth and postcode are compared. The weights for two records being compared are then combined, and if the total obtained is above a specific threshold value, the two records are considered to belong to the same woman and are given the same maternal identification number. It is estimated that 0.5% of true links are missed using this method, however the multiple comparisons involved in comparing records will increase the false positive rate.

This method is also used to link the SMR2 forms to link to SMR11 forms, which are records of neonatal care, and this can be used to follow care and resource usage through the entire pregnancy and then into the first months of life, rather than focusing on individual hospital visits as in chapter 3.

5.2 Linked Births - Birthweight Follow Up By Womersley

In 1980 there were 4499 livebirth singletons to first time mothers where the Womersley category was known. 5 of these mothers appeared to have had 2 first pregnancies in 1980, so they were removed. A further 32 appeared to have their first pregnancy in 1980, and a subsequent first pregnancy between 1981 and 1991, so they also were removed. On following up the remaining 4457 women, 11 of them appeared to have had 2 second pregnancies, so they were also removed. This left us with 4446 mothers, 2556 of who had second children in the data set. These women are considered in order to see if the birthweight of the first child influences the birthweight of the second.

Firstly data for the 1826 mothers whose Womersley score did not change over the two pregnancies is considered. A contingency table of this data is presented in table 5.1.

A Chisquared test on this table gives a p-value of <0.0001, which implies that there is strong evidence of association between the two. Under a null hypothesis

		Second Baby	
		LBW	NBW
First	LBW	20	82
Baby	NBW	65	1659

Table 5.1: Mothers with first delivery in 1980 and a subsequent second delivery whose Womersley category did not change.

of null hypothesis of no association, 5 women would be expected to have two low birthweight babies. There were 20 such women, indicating more cases of two low birthweight babies than expected.

		Second Baby	
		LBW	NBW
First	LBW	4	9
Baby	NBW	13	301

Table 5.2: Mothers with first delivery in 1980 and a subsequent second delivery whose Womersley category did not change - Womersley category 5.

		Second Baby	
		LBW	NBW
First	LBW	11	25
Baby	NBW	15	288

Table 5.3: Mothers with first delivery in 1980 and a subsequent second delivery whose Womersley category did not change - Womersley category 8.

Splitting up by Womersley score indicates that neighbourhood types 5 and 8 have an association between the two birthweights. The contingency tables are shown in tables 5.2 and 5.3. Chisquared values are 12.964 and 26.2845, giving p-values of 0.0003 and <0.0001 respectively.

All mothers, including those who have moved Womersley category between their

		Second Baby	
		LBW	NBW
First	LBW	37	126
Baby	NBW	89	2304

Table 5.4: All mothers with first delivery in 1980 and a subsequent second delivery.

first and second pregnancies, are now considered, giving 2556 mothers. A contingency table is given in table 5.4. A Chisquared test on this table gives a p-value of <0.0001, which again implies that there is strong evidence of association between the two. In this case the number of women expected to have two low birthweight babies under a null hypothesis of no association is 9, indicating a much larger number of women with two LBW babies than expected.

		Second Baby	
		LBW	NBW
First	LBW	2	11
Baby	NBW	6	289

Table 5.5: All mothers with first delivery in 1980 and a subsequent second delivery - Womersley category 2.

Splitting up by Womersley score of the first pregnancy shows that neighbourhood types 2, 3, 5, 7 and 8 have an association between the two birthweights. The results are shown in tables 5.5-5.9. Chisquared values are 4.2888, 16.3833, 13.2489, 11.604 and 34.3779, giving p-values of 0.0384, 0.0001, 0.0003, 0.0007 and <0.0001 for neighbourhood types 2, 3, 5, 7 and 8 respectively. In all of these cases, the number of cases where both babies are low birthweight is larger than the expected number under a null hypothesis of no association.

		Second Baby	
		LBW	NBW
First	LBW	3	8
Baby	NBW	4	230

Table 5.6: All mothers with first delivery in 1980 and a subsequent second delivery - Womersley category 3.

		Second Baby	
		LBW	NBW
First	LBW	6	20
Baby	NBW	19	415

Table 5.7: All mothers with first delivery in 1980 and a subsequent second delivery - Womersley category 5.

		Second Baby	
		LBW	NBW
First	LBW	6	20
Baby	NBW	18	362

Table 5.8: All mothers with first delivery in 1980 and a subsequent second delivery - Womersley category 7.

		Second Baby	
		LBW	NBW
First	LBW	13	33
Baby	NBW	20	430

Table 5.9: All mothers with first delivery in 1980 and a subsequent second delivery - Womersley category 8.

5.3 Logistic Regressions For Birthweight

A logistic regression model was fitted using the linked dataset with all mothers who had a first delivery in 1980 followed by a subsequent delivery, with the response a binary variable to indicate whether the second baby was low birthweight (1) or not (0). 2556 cases were used in this analysis. Possible explanatories in the models were previous low birthweight (1=yes, 0=no), Womersley neighbourhood type and maternal height. Previous low birthweight was entered first ($p=7.2\times10^{-15}$) and Womersley neighbourhood type was entered second (p=0.024). After these were added, maternal height was not significant (p=0.23). The most deprived category (8) was the baseline category, which is incorporated into the constant.

The model calculated was:

$$\begin{aligned} \textit{logit}(\textit{Pr}(\textit{2nd baby lbw})) &= -2.99038 \\ &+1.970089(\textit{if first baby lbw}) \\ &-1.441946(\textit{if 1980 Womersley} = \textit{1}) \\ &-0.7798942(\textit{if 1980 Womersley} = \textit{2}) \\ &-0.7364242(\textit{if 1980 Womersley} = \textit{3}) \\ &-0.6164404(\textit{if 1980 Womersley} = \textit{4}) \\ &-0.04161894(\textit{if 1980 Womersley} = \textit{5}) \\ &-0.04702384(\textit{if 1980 Womersley} = \textit{6}) \\ &+0.02619255(\textit{if 1980 Womersley} = \textit{7}) \end{aligned}$$

As shown in table 5.10, there is a much higher probability of the second baby being

	First Baby	
Womersley	LBW	NBW
1	0.0785	0.0117
2	0.1418	0.0225
3	0.1472	0.0235
4	0.1629	0.0264
5	0.2569	0.0460
6	0.2559	0.0458
7	0.2701	0.0491
8	0.2650	0.0479

Table 5.10: Probability of delivering a low birthweight baby in the second pregnancy, mothers with first delivery in 1980 and a subsequent second delivery between 1981 and 1991.

low birthweight if the first baby was low birthweight, with this probability being at least six times higher. There is also an increase in the probability of low birthweight as Womersley neighbourhood type becomes more deprived, where the probability of a low birthweight second delivery is three times higher in deprived areas than in the most affluent ones.

5.4 Conclusions

This chapter has considered the association between the birthweight category of the first infant and that of the second infant. A chi-squared test using only women who did not change Womersley neighbourhood type over these pregnancies showed evidence of an association between birthweight in the first and second pregnancies, with a low birthweight first pregnancy more likely to be followed by a second low birthweight infant. Chi-squared tests carried out for each Womersley category indicated that this association was evident in categories 5 and 8, indicating that the association may be more prevalent in more deprived areas.

Similar results were obtained when all mothers, including those whose Womersley category had changed between the two pregnancies, were considered. Chi-squared tests on each Womersley category indicated that there was evidence of association in several of the categories.

In modelling the birthweight of the second infant as a binary response, with deprivation, height and previous low birthweight as possible explanatory variables, it was shown that previous low birthweight and deprivation were significant variables, with the probability of delivering a low birthweight baby being higher in more deprived areas and also if the previous pregnancy resulted in a low birthweight infant. This is consistent with other authors [5, 20] who have discussed the effect of a previous low birtheight infant on subsequent pregnancies.

Chapter 6

Variable Selection

The aim of regression modelling is to investigate how changes in explanatory variables affect the value of a response variable of interest. Regression modelling can be carried out to produce a model from a given set of data. This model can then be used on future data to predict the value of the response variable. This may be done when the value of the response variable is too difficult or too expensive to determine while the explanatories are not, or when the value that a response variable will take needs to be estimated in advance. The aim of this chapter is to consider several methods which can be used in forward subset selection to decide which variables should be included in a regression model. Several types of subset selection and stopping rules which can be used in selection procedures are discussed. The ability of these procedures to select the correct variable for the model is investigated and compared using simulation studies.

6.1 Variable Selection Methods

Often in regression models there are a large number of possible explanatory variables which may or may not be related to the response variable, or may be highly correlated with other variables so that the inclusion of all variables is neither necessary nor helpful. In previous chapters the birthweight dataset has been considered and this contains a large amount of variables which may or may not be related to birthweight. One way of investigating which variables are related to birthweight is to use variable selection techniques to remove the unrelated variables from the model. In this section various methods of selecting the 'best' subset of all possible explanatory variables are discussed, and these methods will later be extended in order to apply them to the data in the birthweight dataset.

6.1.1 All Subsets Selection

All subsets selection [16, 34, 38], as the name suggests, considers all possible combinations of explanatory variables and for each of these combinations, fits a regression model to the data. The best model is then chosen from all possible models by a pre-specified criterion, e.g. Mallow's C_p statistic, the residual sum of squares, or the value of R^2. While this method may be suitable for a small number of possible explanatory variables, the number of possible combinations of regression models rapidly increases with the number of possible explanatories, and 1024 regression models must be computed when there are only 10 possible explanatories.

A refinement of the all subsets technique, discussed by Miller [34], is to compute the models in such an order that many of the models need not be fitted at all. This involves splitting all possible subsets into two 'branches', one which contains all subsets including a specified variable, and one containing all subsets which do not include this variable. These branches are then split into sub-branches using the same technique. Sub-branches can be ignored if the residual sum of squares for all variables in the sub-branch is larger than that of a subset already tested.

6.1.2 Stepwise Regression

There are three types of method which attempt to find the best subset of explanatory variables without computing all possible regression models. At each step of the forward selection method [16, 34, 38] a model is fitted for each of the explanatory variables which are not already in the regression model, consisting of this variable and the variables already entered in the model. The most significant of these variables is then added to the model, assuming it is significant according to a pre-specified stopping rule. Without a stopping rule, this method will carry on until the full model is fitted.

Backwards elimination [16, 34, 38] begins with the full regression model, and at each step the least significant variable is removed from the model, again according to a pre-specified stopping rule. If this method continues without a stopping rule then a null regression model will be produced containing only an intercept.

Stepwise regression, or the Efroymson algorithm [16, 17, 34, 38] is a mixture of both forward selection and backwards elimination. Each step contains two stages. In the first stage, the most significant variable is added to the model using the same method as in forward selection, and in the second stage, each of the variables previously in the model are tested in order to determine which, if any, can be removed from the model. If the first stage does not enter a variable, and the second stage does not delete a variable, then the method stops.

In certain extreme cases all three of these methods may produce different subsets of variables. In these cases it may be impossible to choose between these subsets without considering another method, such as comparing Mallow's C_p statistics for the models.

Broersen [6] discusses a combination of forward selection and backwards elimination, using Mallow's C_p statistic to compare subsets of different sizes. The method used carries out forward selection until all variables are entered into the model, taking note of the C_p value at each stage. The backwards elimination is carried out, again taking note of the C_p value at each stage, until the null model is produced. Of all the models produced, that with the best C_p value for the number of variables in the model is chosen as the 'best' model.

Altman and Andersen [1]investigate the stability of Cox regression modelling by using bootstrap samples to validate the model produced partly by using forward selection on a set of 17 possible explanatory variables. In carrying out this investigation the variables which were most frequently selected in the bootstrap sampling were those which were selected in the original analysis of the data, but prediction intervals were much wider. This indicates that while there is no problem with the variable selection in forward selection, using such a model for prediction on an entirely new set of data may result in estimating confidence intervals as smaller than they actually are.

Similarly, Chen and George [11] investigated the choice of prognostic factors and the regression parameters on a set of 9 prognostic factors in pediatric acute lymphocytic leukemia. A regresison model was seltected using forward selection and bootstrapping methods were used to validate this model. Due to this validation the final variable selected was dropped from the model. 400 bootstrapping samples were then generated to produce regression estimates for the parameters, and these parameters were similar to those in the original model, suggesting again that there was little problem with the forward selection method.

The remainder of this chapter shall consider simulations using the forward selection method.

6.2 Stopping Rules Used in Variable Selection

Four different selection rules are considered - the omnibus F statistic, a maximum R^2 statistic, and two maximum t statistics, one which assumes independence of all possible t statistics, and one which does not.

The theoretical properties of each of the four tests is discussed individually. In each case a response variable, $\underline{y}$, of n observations and k possible explanatory variables, $\underline{x}_1, \ldots \underline{x}_k$ are assumed, as is the fact that each explanatory variable $\underline{x}_i$ has a mean of zero; this can easily be achieved by transforming each $\underline{x}_i$ by subtracting its mean from every element of the variable, and that $\underline{\epsilon} \sim N(\underline{0}, \sigma^2\mathrm{I})$. By using a size of α=0.05, critical values are determined for each of the four tests.

6.2.1 The Omnibus F Test

The omnibus F test is, in effect, a test for any further information in the remaining variables. In using this test at the first stage, it is testing for relevant information in any of the explanatory variables.

In order to carry out the omnibus F test two residual sums of squares must be calculated. The first is under the null model, i.e. $E(\underline{y}) = \gamma\underline{1} + \underline{\epsilon}$, and the second is under the full model, i.e. $E(\underline{y}) = \gamma\underline{1} + \beta_1\underline{x}_1 + \ldots + \beta_k\underline{x}_k + \underline{\epsilon}$

Now, suppose RSS_0 is the residual sum of squares under the null model, and RSS_{full} is the residual sum of squares under the full model, and define

$$F = \frac{(n-p-1)(RSS_0 - RSS_{full})}{p \times Rss_{full}}.$$

It can be shown that, under the null model, $F \sim F(p, n-p-1)$. The size of the F test, α, is

$$\begin{aligned} \alpha &= Pr(F > c| \text{ null model true}) \\ &= Pr(F > c|F \sim F(p, n-p-1)) \\ &= Pr(F(p, n-p-1) > c) \end{aligned}$$

Hence c is chosen to be the upper $100\alpha\%$ point of the $F(p, n-p-1)$ distribution, and if the observed value of F is greater than this, this indicates a significant result and concludes that by the F test, there is a variable amongst the explanatory variables which can be entered into the model.

6.2.2 Maximum R^2

If a number of regression models are fitted with the same number of explanatory variables, in some sense the 'best' of these models can be thought of as that with the maximum R^2 value. Diehr and Hoflin [15] considered a Monte Carlo approach to the approximation of the distribution of the maximum R^2. For 100 simulations, m independent variables and an independent response were sampled and each k variable regression model was fitted. The maximum R^2 was evaluated for each simulation. From these simulations it appeared that the best estimate of the centiles of the maximum R^2 distribution was of the form $R^2(k, m, n, \alpha) = w(1 - v^k)$, where w and v could be determined from $k = 1$ and $k = m$.

Rencher and Pun [40] stated that the distribution function of R^2 is given by the incomplete beta function

$$F(R^2) = \frac{1}{B(a,b)} \int_0^{R^2} (R^2)^{a-1}(1-R^2)^{b-1} dR^2$$

where $B(a, b)$ is the beta function with $a = p/2$ and $b = (n-p-1)/2$, where p is the number of predictors in the model and n is the number of observations. They go on to approximate the extreme value distribution of the maximum R^2, which can be shown to be approximately

$$G(R_s^2) = exp(-N(1 - F(R_s^2))),$$

where $N = \binom{k}{p}$ and k is the total number of possible explanatory variables.

In order to use the maximum R^2 test, k simple linear regressions must be carried out, and in the i-th regression, the model $E(\underline{y}) = \gamma_i + \beta_i x_i, i = 1, \ldots, k$, is fitted. For each of these regressions R_i^2 is calculated. The maximum R^2 test takes the maximum of the R_i^2 values and tests whether this maximum is significantly large to enter the variable which produces it into the model. The size of this test can be determined from tables of the incomplete beta distribution.

$$\begin{aligned}\alpha &= Pr(max_i R_i^2 > c_3| \text{ null model true})\\ &= 1 - Pr(max_i R_i^2 < c_3| \text{ null model true})\\ &= 1 - G(c_3)\\ &= 1 - exp(-N(1 - F(c_3)))\end{aligned}$$

Hence c_3 is the upper $\frac{-100}{N}$ln (1-α) % point of the incomplete beta distribution with a and b as denoted above, So, if the observed maximum R_i^2 statistic is larger than c_3, this is a significant result which concludes that the variable which gives this maximum R_i^2 value should be entered into the model.

6.2.3 The Independence t Test

To carry out the independence t test, k simple linear regressions must be carried out as before. In the i-th regression, fit a model $E(\underline{y}) = \gamma_i + \beta_i x_i, i = 1, \ldots, k$. For each of these regressions the t statistic for testing $\beta_i = 0$ is calculated.

Now, suppose all these t statistics are independent. Again this independence assumption is generally not true. Then the maximum of these t statistics is the value of interest. Under the null model, each t statistic follows a $t(n-2)$ distribution. The size of the test, α, can be calculated as

$$\begin{aligned}\alpha &= Pr(max_i|T_i| > c_2| \text{ null model true})\\ &= 1 - Pr(max_i|T_i| < c_2| \text{ null model true})\\ &= 1 - Pr(all|T_i| < c_2| \text{ null model true})\\ &= 1 - \textstyle\prod_{i=1}^k Pr(|T_i| < c_2|T_i \sim t(n-2))\\ &= 1 - \textstyle\prod_{i=1}^k Pr(-c_2 < t(n-2) < c_2)\\ &= 1 - (2Pr(t(n-2) < c_2) - 1)^k\end{aligned}$$

Hence c_2 is the upper 50(1-(1-α)$^{1/k}$)% point of $t(n-2)$. So, if the observed maximum absolute t statistic is larger than c_2, this implies a significant result and the conclusion from the independence t test is that the variable which gives this maximum absolute t value should be entered into the model.

6.2.4 The Bonferroni t Test

The Bonferroni t test is similar to the independence test although it does not use the false assumption that all of the t statistics are independent. Instead it uses a Bonferroni upper bound [4, 7, 8] on the required probability. This is calculated as follows:

$$\begin{aligned}
\alpha &= Pr(max_i|T_i| > c_3| \text{ null model true}) \\
&= Pr(\textstyle\bigcup_{i=1}^{k} |T_i| > c_3| \text{ null model true}) \\
&\leq \textstyle\sum_{i=1}^{k} Pr(|T_i| > c_3| \text{ null model true}) \\
&= \textstyle\sum_{i=1}^{k} Pr(|T_i| > c_3 | T_i \sim t(n-2)) \\
&= \textstyle\sum_{i=1}^{k} (Pr(-c_3 > t(n-2)) + Pr(t(n-2) > c_3)) \\
&= 2k(1 - Pr(t(n-2) < c_3))
\end{aligned}$$

Hence c_3 is the upper $50\alpha/k\%$ point of $t(n-2)$. So, if the observed maximum absolute t statistic is larger than c_3, we have a significant result and the conclusion from the Bonferroni t test is that the variable which gives this maximum absolute t value should be entered into the model.

6.3 Comparison Of Stopping Rules

In this section, simulations are used to compare the size of these tests, using independent variables, and then the power of these tests, using variables correlated to the response. These simulations also investigate the ability of the selection rules to select the correct variable for entry into the model.

6.3.1 Comparison of Test Size

Here Fortran is used to generate a response variable with n observations, and k possible explanatory variables, each of which is independent of the response and of all other explanatory variables. This is carried out for a specified number of simulations. For each simulation, each of the four selection rules is used to determine whether any of the explanatory variables are significant. As the number of simulations increases, the proportion of times each rule selects a significant variable should tend to the actual size of the test. From tables 6.1 and 6.2, as the simulation size increases, the proportions tend to the 5% level of significance assumed in the selection procedures.

The proportions are similar for each of the four different rules, and these proportions do not seem to vary much as the number of explanatory variables increase.

No. Explanatories = 2 No. Simulations	100	500	1000	5000	10000	50000
F test	0.0600	0.0440	0.0370	0.0506	0.0483	0.0493
Max R^2 test	0.0400	0.0580	0.0430	0.0470	0.0517	0.0492
Independence t test	0.0400	0.0560	0.0410	0.0468	0.0511	0.0485
Bonferroni t test	0.0400	0.0560	0.0400	0.0460	0.5000	0.0481
No. Explanatories = 10 No. Simulations	**100**	**500**	**1000**	**5000**	**10000**	**50000**
F test	0.0500	0.0400	0.0390	0.0486	0.0499	0.0500
Max R^2 test	0.0800	0.0360	0.0410	0.0544	0.0500	0.0495
Independence t test	0.0800	0.0360	0.0410	0.0544	0.0498	0.0493
Bonferroni t test	0.0800	0.0360	0.0400	0.0534	0.0490	0.0484

Table 6.1: Proportion of simulations with significant explanatory variables for 25 observations.

6.3.2 Comparison of Test Powers

In this section the case is considered where there is a correlation of ρ between the response and one of the explanatory variables, but all other correlations are zero. Here each selection rule tests for a significant variable, and determines how often the selection rule selects a variable for entry into the model. The proportion of times that a variable is selected for entry is an estimate of the power of the test. The results of these tests are shown in figures 6.1 to 6.7, where it can be seen that as the correlation between the response variable and the explanatory variable it is correlated with increases from 0 to 1, or decreases from 0 to -1, then the powers of all tests increase. As expected, as the number of explanatory variables increases, the power of the unadjusted t-test becomes much higher than the other tests, due to the fact that this test does not correct for multiple comparisons. Figure 6.4 shows how extreme this test can be, as at a correlation of zero the power of the test would be

No. Explanatories = 5						
No. Simulations	100	500	1000	5000	10000	50000
F test	0.0300	0.0360	0.0510	0.0554	0.0490	0.0485
Max R^2 test	0.0300	0.0280	0.0540	0.0488	0.0501	0.0492
Independence t test	0.0300	0.0280	0.0540	0.0484	0.0499	0.0489
Bonferroni t test	0.0300	0.0280	0.0510	0.0478	0.0496	0.0479
No. Explanatories = 10						
No. Simulations	100	500	1000	5000	10000	50000
F test	0.0200	0.0500	0.0430	0.0530	0.0513	0.0480
Max R^2 test	0.0200	0.0420	0.060	0.0474	0.0543	0.0492
Independence t test	0.0200	0.0420	0.0600	0.0472	0.0539	0.0491
Bonferroni t test	0.0200	0.0420	0.0590	0.0464	0.0522	0.0478
No. Explanatories = 25						
No. Simulations	100	500	1000	5000	10000	50000
F test	0.0400	0.0620	0.0460	0.0502	0.0470	0.0501
Max R^2 test	0.0200	0.0640	0.0550	0.0476	0.0501	0.0489
Independence t test	0.0200	0.0640	0.0550	0.0476	0.0500	0.0488
Bonferroni t test	0.0200	0.0620	0.0530	0.0472	0.0490	0.0479

Table 6.2: Proportion of simulations with significant explanatory variables for 100 observations.

expected to be 0.05, and in this case, with 100 observations and 25 explanatories, the power is approximately 0.74.

6.3.3 Selection of Variables

The previous section does not take into account which explanatory variable is selected to enter the model. There is no distinction made between the variable correlated with the response being selected and any other variable being selected. This section investigates how often the correlated variable is chosen, and how often any other variable is selected. Fortran simulations are carried out to calculate how many times any

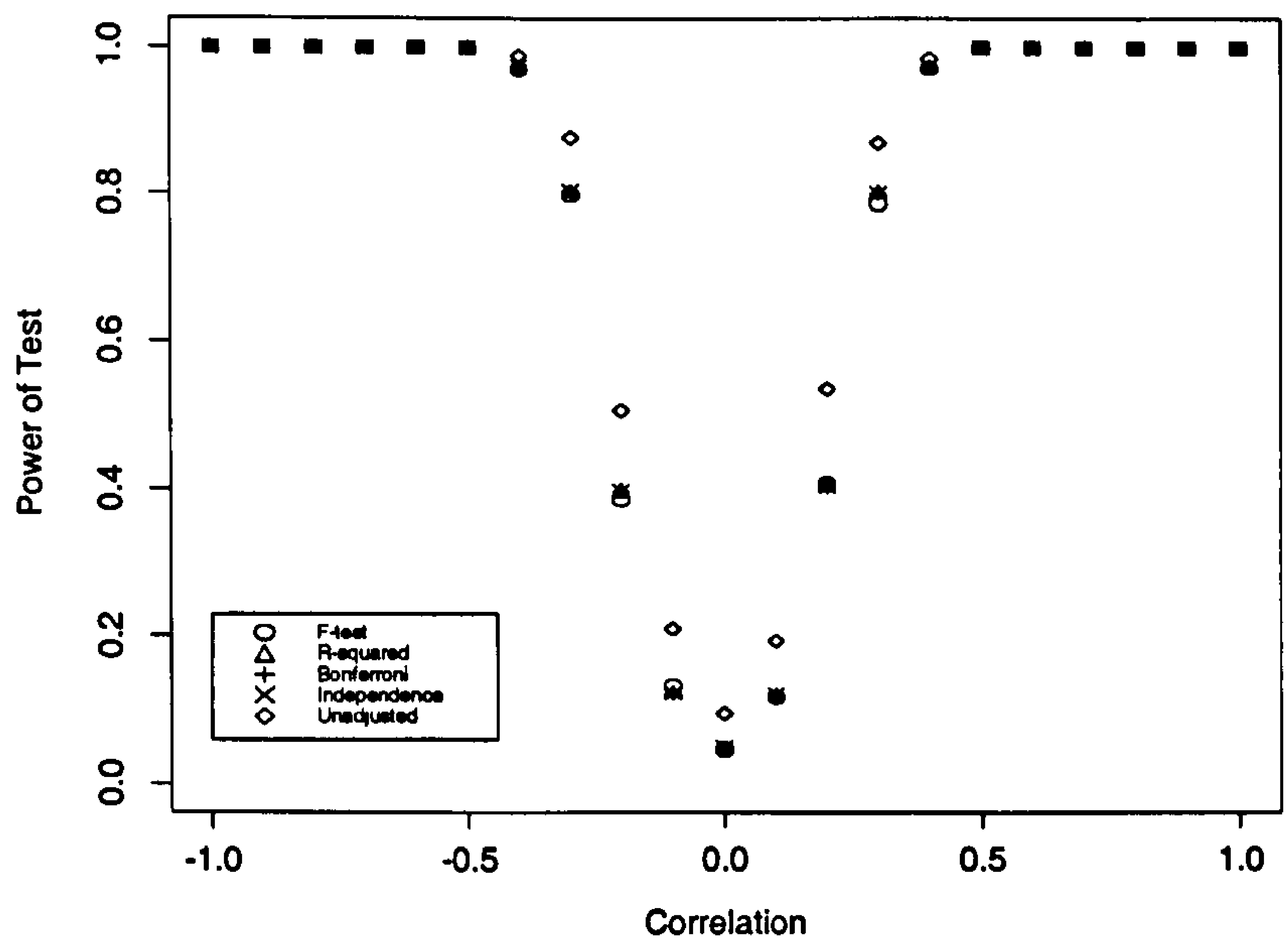

Figure 6.1: Powers of tests with 100 observations and 2 explanatory variables.

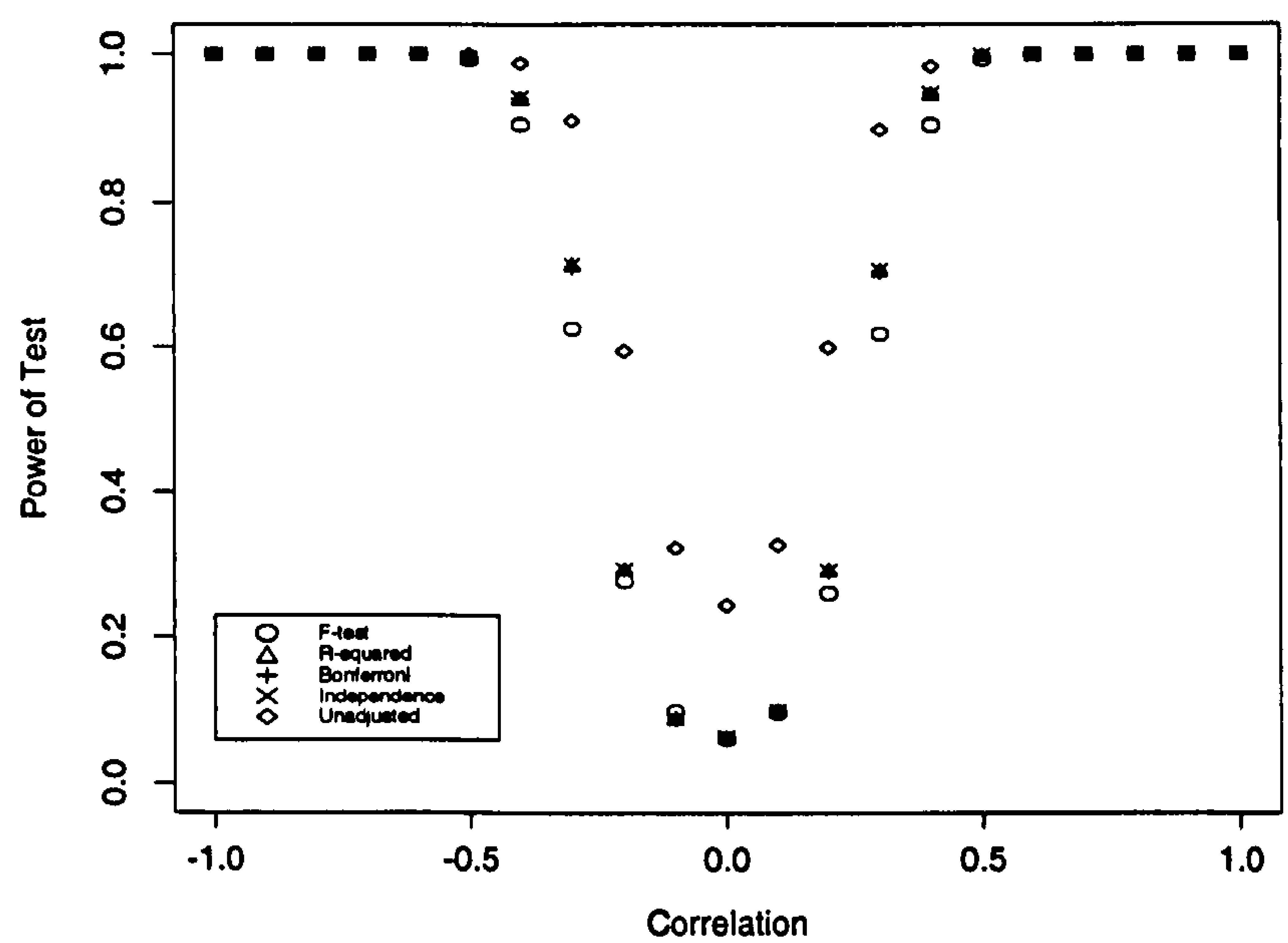

Figure 6.2: Powers of tests with 100 observations and 5 explanatory variables.

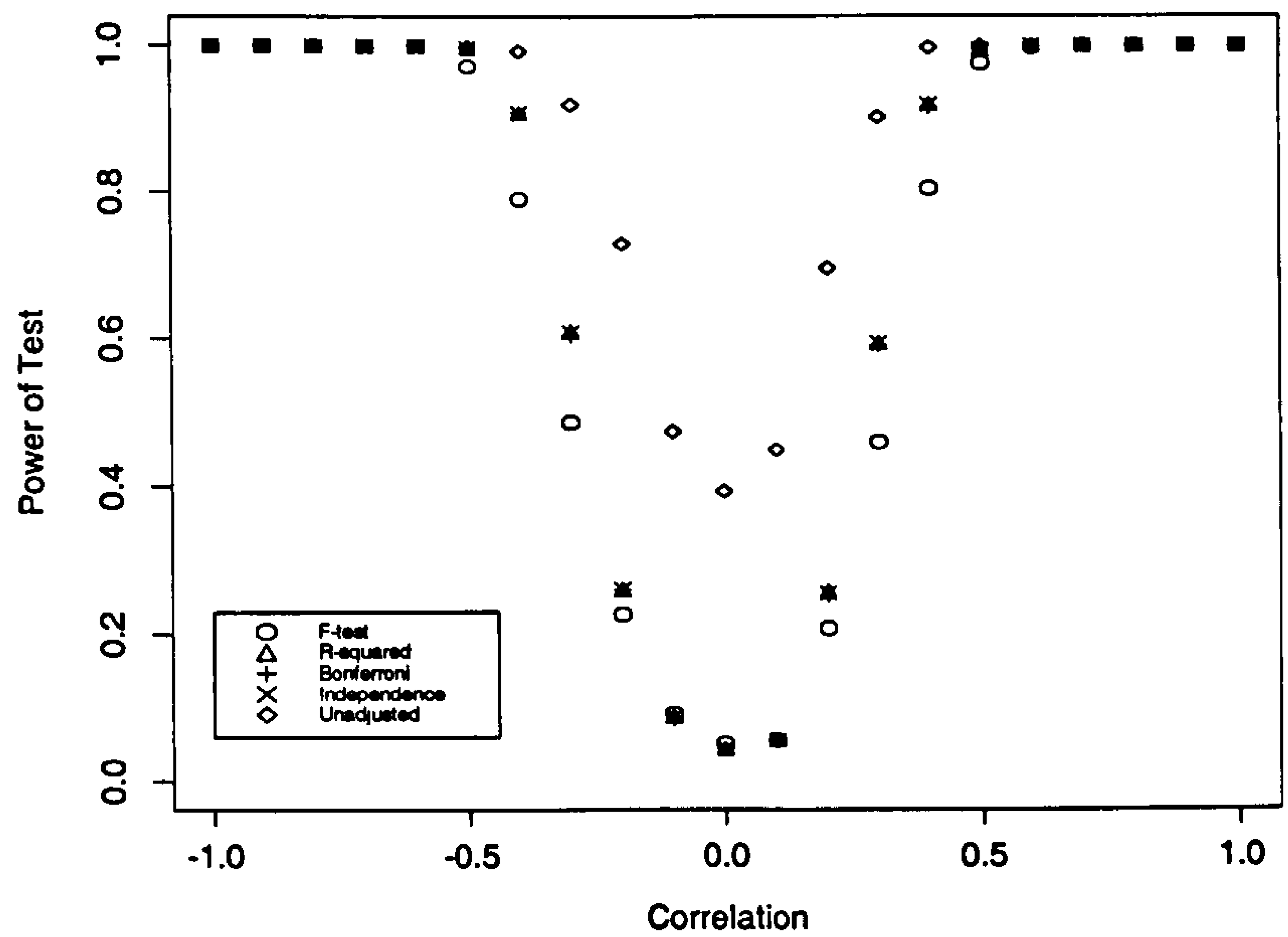

Figure 6.3: Powers of tests with 100 observations and 10 explanatory variables.

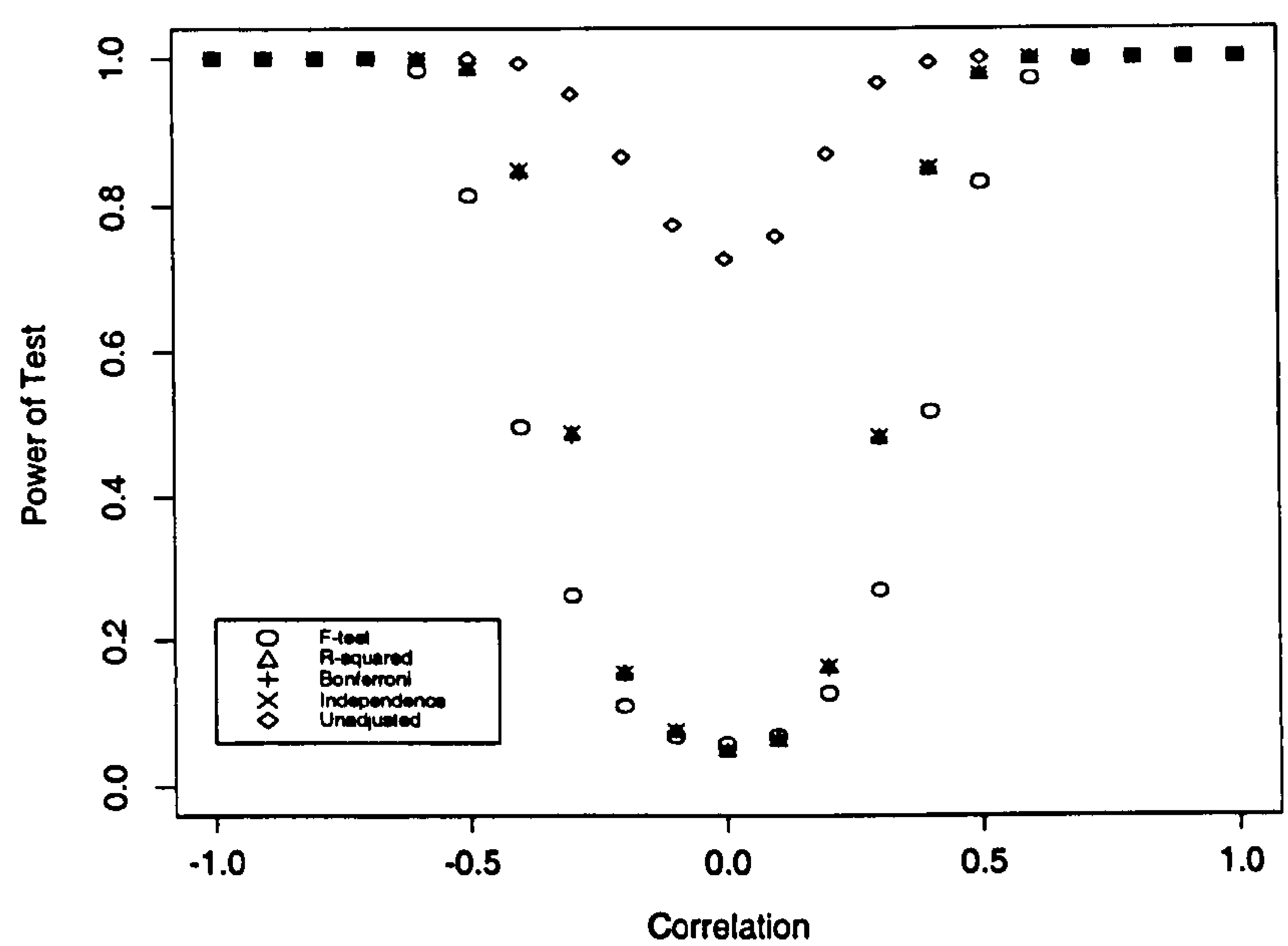

Figure 6.4: Powers of tests with 100 observations and 25 explanatory variables.

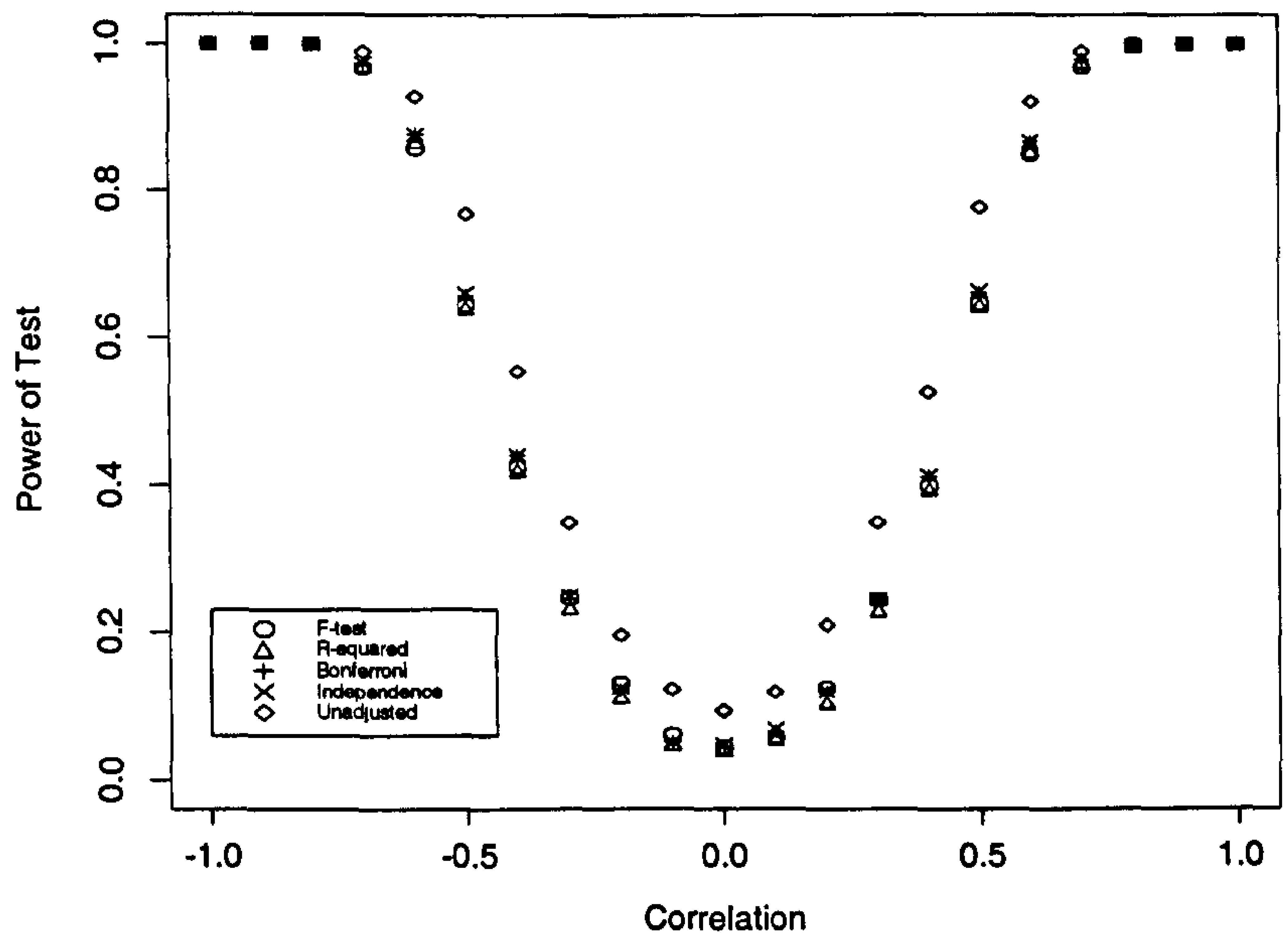

Figure 6.5: Powers of tests with 25 observations and 2 explanatory variables.

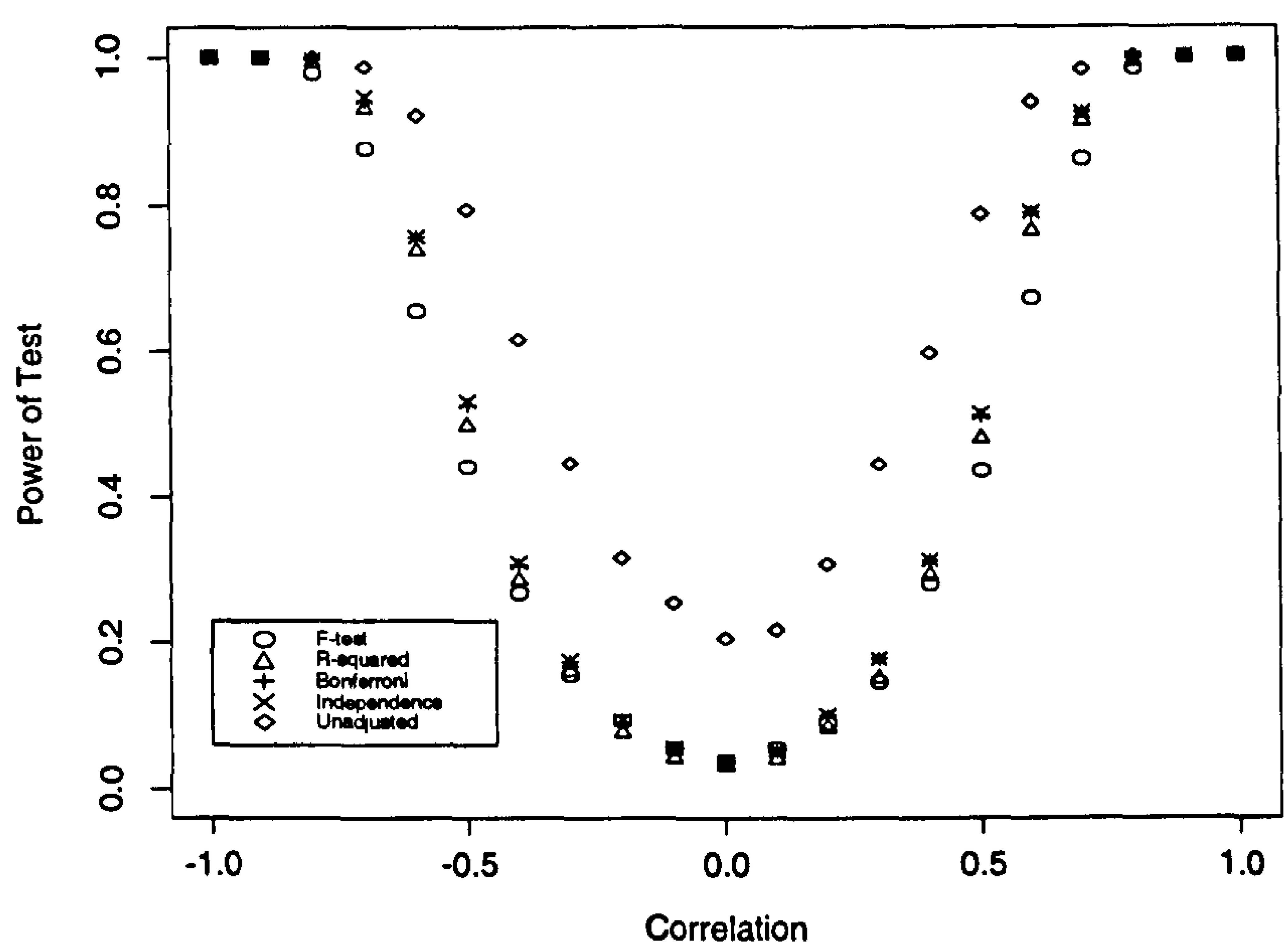

Figure 6.6: Powers of tests with 25 observations and 5 explanatory variables.

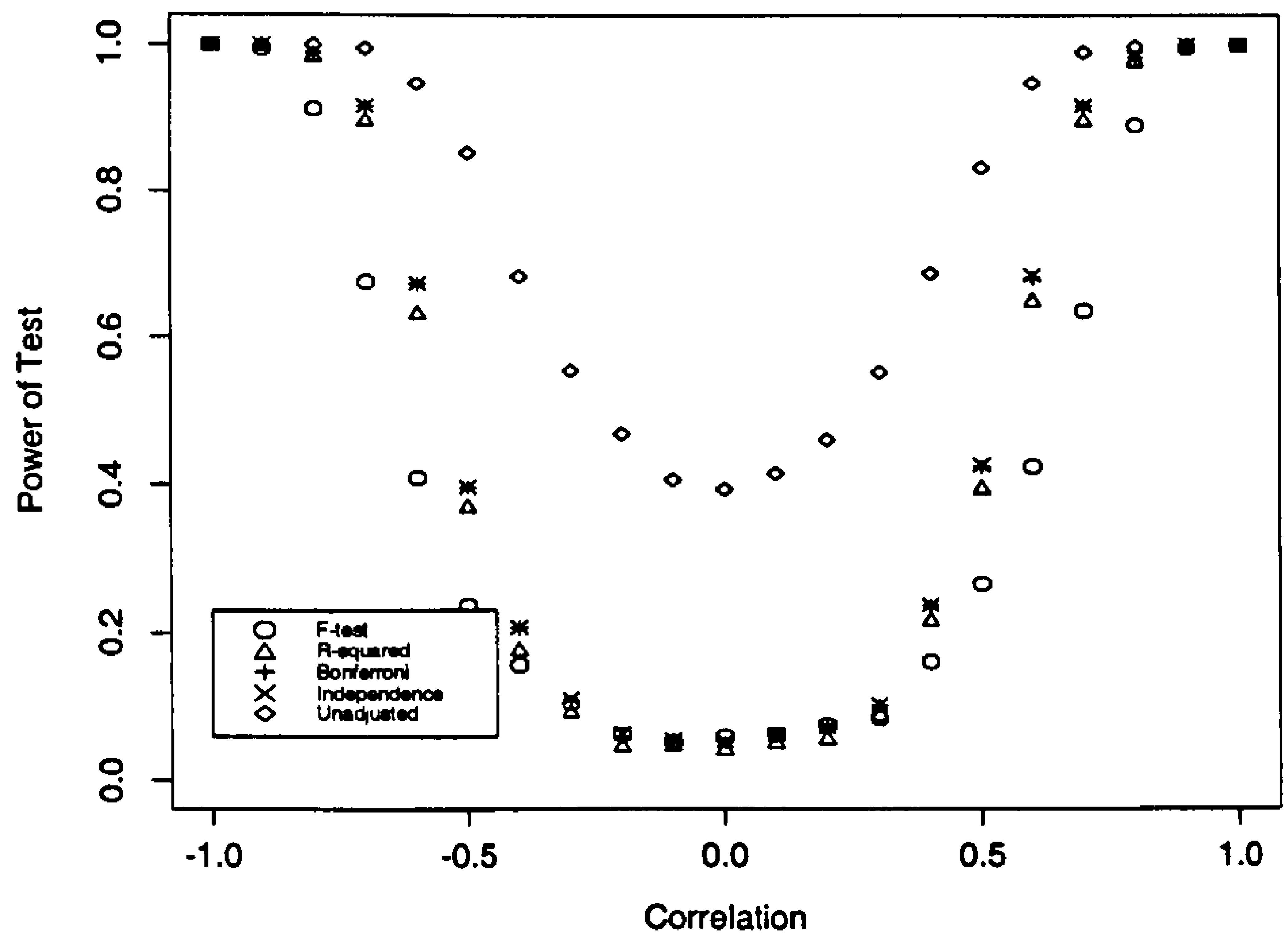

Figure 6.7: Powers of tests with 25 observations and 10 explanatory variables.

variable is selected, and what proportion of those tests where a variable is significant select the correlated variable.

	Maximum R^2 Test		Bonferroni t Test		Independence t Test	
ρ	Number Accepted	Proportion Correct	Number Accepted	Proportion Correct	Number Accepted	Proportion Correct
0.05	53	0.660	51	0.667	51	0.667
0.10	99	0.707	98	0.704	99	0.707
0.15	152	0.829	151	0.828	151	0.828
0.20	228	0.930	227	0.930	227	0.930
0.25	354	0.912	349	0.914	352	0.912
0.30	482	0.969	476	0.968	476	0.968
0.35	638	0.983	635	0.983	636	0.983
0.40	748	0.991	748	0.991	748	0.991
0.45	864	0.991	859	0.991	864	0.991
0.50	937	0.999	937	0.999	937	0.999
0.55	980	1.000	979	1.000	979	1.000
0.60	993	1.000	993	1.000	993	1.000
0.65	1000	1.000	1000	1.000	1000	1.000
0.70	1000	1.000	1000	1.000	1000	1.000
0.75	1000	1.000	1000	1.000	1000	1.000
0.80	1000	1.000	1000	1.000	1000	1.000
0.85	1000	1.000	1000	1.000	1000	1.000
0.90	1000	1.000	1000	1.000	1000	1.000
0.95	1000	1.000	1000	1.000	1000	1.000
1.00	1000	1.000	1000	1.000	1000	1.000

Table 6.3: Proportion of simulations with significant explanatory variables, for 2 explanatories, 50 observations and 1000 simulations.

	Maximum R^2 Test		Bonferroni t Test		Independence t Test	
ρ	Number Accepted	Proportion Correct	Number Accepted	Proportion Correct	Number Accepted	Proportion Correct
0.05	55	0.236	53	0.245	55	0.236
0.10	87	0.345	86	0.337	87	0.345
0.15	101	0.594	99	0.596	101	0.594
0.20	140	0.779	138	0.775	140	0.779
0.25	250	0.864	247	0.862	249	0.863
0.30	370	0.935	370	0.935	370	0.935
0.35	483	0.936	479	0.939	482	0.938
0.40	654	0.974	652	0.974	654	0.974
0.45	778	0.983	774	0.984	777	0.983
0.50	883	0.995	883	0.995	883	0.995
0.55	942	0.996	941	0.996	941	0.996
0.60	990	1.000	990	1.000	990	1.000
0.65	999	0.999	999	0.999	999	0.999
0.70	1000	1.000	1000	1.000	1000	1.000
0.75	1000	1.000	1000	1.000	1000	1.000
0.80	1000	1.000	1000	1.000	1000	1.000
0.85	1000	1.000	1000	1.000	1000	1.000
0.90	1000	1.000	1000	1.000	1000	1.000
0.95	1000	1.000	1000	1.000	1000	1.000
1.00	1000	1.000	1000	1.000	1000	1.000

Table 6.4: Proportion of simulations with significant explanatory variables, for 5 explanatories, 50 observations and 1000 simulations.

From tables 6.3 and 6.4, it is clear that as ρ increases, the number of times each test chooses as significant a variable which is not correlated with the response decreases. This should be intuitively obvious; as the correlation between the response and one of the explanatories increases, then the tests should be more likely to choose this variable as being significant and therefore less like to choose a variable which is not correlated with the response.

6.4 Conclusions

As can be seen in the previous sections, the unadjusted t-test deflates the p-values of interest, resulting in insignificant variables being selected in the above simulation studies. The remaining four tests give similar results in the simulations, with the power of the F-test being slightly smaller than that of the independent t-test and Bonferroni t-test at the first step of the selection procedure, with this difference becoming more obvious as the number of explanatory variables increases. In addition, as the number of observations increases, the power of the R^2 test increases to become closer to the powers of both t-tests.

Chapter 7

Bonferroni Bounds in Variable Selection

In the previous chapter it has been shown that in subset selection, failing to correct for multiple comparisons will produce deflated p-values and result in a model with too many variables entered. In this chapter, fitting the 'best' model to a set of data is considered, taking into account the effect multiple comparisons will have on the overall p-value of any tests carried out. A method which calculates both upper and lower bounds on the p-value calculated to test whether a variable should be added when fitting a linear regression model is described, and this method is extended to the case of logistic regression. Finally a logistic model is fitted to the low birthweight data described previously, using techniques described in this chapter.

7.1 Introduction

Often in statistics, a model is fitted to a set of data in order to relate a response variable to one or more explanatory variables. If there are only a few explanatories, 'best' subsets may be the easiest method to use. This is quite simple to do in the case of, say, three or four possible explanatories. All possible models are fitted using all subsets of explanatory variables, i.e. all possible combinations of possible explanatories, and from this we can decide which is the 'best' subset, using a measure such as Mallow's C_p statistic, or the adjusted R^2 from each model. As the number of explanatories increases, the total number of possible subsets increases rapidly -

with 10 explanatories, there are over a thousand possible subsets - and clearly this procedure will be very computationally intensive. One idea to combat this is stepwise subset selection. There are many types of subset selection, some of the better known of which are described here. In forward selection, the null model is the starting point and at each step the procedure adds the most significant next variable into the model. Backwards selection begins with the full model and at each step removes the least significant variable from the model. A method devised by Efroymson [17] is similar to forward selection in that at each step the most significant variable is added, but it then tests whether any other variable already in the model can be removed. All of these methods stop if a notional significance probability is not attained by any of the explanatory variables. However they all involve the problem of multiple comparisons which, if not corrected for, can give a very large type I error for the problem.

Another problem which often occurs in standard packages when using such selection techniques in regression is that any variable with a small p-value is accepted into the model. This problem suggests that any variable which may possibly be related to the response is added to the model. In order to combat these problems, a Bonferroni bound can be calculated in order to give an upper bound for the overall p-value, and this may then be used to decide whether the most significant variable at any stage of the stepwise procedure should be entered into the model.

There are, however, cases where this upper bound on the p-value may be insignificant where the actual p-value is significant. In order to avoid rejecting a variable which is in fact significant, the calculation of a lower bound on the exact p-value is considered.

7.2 Normal Theory

In using stepwise selection to fit a linear model to a set of data with m explanatory variables, the first stage of the procedure is to fit all simple linear regression models, calculate the p-value for the significance of each of the m variables, and examine whether the variable with the smallest p-value (or largest t-value) is significant enough to be entered into the model. Here m p-values are compared in order to determine statistical significance, which clearly poses a multiple comparisons problem.

This problem can be solved by using Bonferroni bounds.

Butler [8] states that, if we consider the event A_{max}, which is the maximum of the events $A_1, A_2, \ldots, A_m$, then we can show that

$$Pr(A_{max}) = S_1 - S_2 + S_3 - \ldots + (-1)^{m-1} S_m \tag{7.1}$$

where $S_1 = \sum_{i=1}^{m} Pr(A_i)$, $S_2 = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} Pr(A_i, A_j)$, and so on, with $S_m = Pr(A_1, A_2, \ldots, A_m)$. He then states that

$$S_1 - S_2 + \ldots - S_{2j} \leq Pr(A_{max}) \leq S_1 - S_2 + \ldots - S_{2j} + S_{2j+1}$$

for $1 \leq j \leq [(m-1)/2]$, while Bølviken [4] states that also

$$S_1 - S_2 \leq Pr(A_{max}) \leq S_1. \tag{7.2}$$

Suppose now we have identically distributed random variables $T_1, \ldots, T_m$ and our events are $A_i = |T_i| \geq c$, where c is the observed maximum of all the $|T_i|$'s, and we are interested in $Pr(A_{max}) = Pr(|T|_{max} \geq c)$. The form of the Bonferroni bounds in which we are interested can be written as

$$\sum_{i=1}^{m} Pr(|T_i| \geq c) - \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} Pr(|T_i|, |T_j| \geq c) \leq p \leq \sum_{i=1}^{m} Pr(|T_i| \geq c),$$

where $p = Pr(|T|_{max} \geq c)$.

The idea behind the need for a lower bound is quite a simple one. Suppose the actual value of a significance probability is close to the required significance level for a given step in the stepwise procedure. It is possible that the upper bound may be above the required level, suggesting the variable should not be accepted into the model, when the actual significance probability is less than the required level and the variable should be accepted into the model. This may also be helpful when the upper bound is quite close to the required significance value, as calculation of the lower bound may show whether the actual value is significant or not.

We now look at how the Bonferroni bounds are calculated when we use stepwise regression in the normal linear case. We assume that we have a response vector, $\underline{Y}$, and m possible explanatory variables, $\underline{x}_1, \ldots, \underline{x}_m$. At the current step we assume we have $k < m$ variables in the model, and denote the matrix of these variables by X_F, where X_F also contains a constant vector as one of its columns in order to allow for an intercept. We shall denote those variables already in the model by $\underline{x}_{[1]}, \ldots, \underline{x}_{[k]}$, and those $l = m - k$ variables which may be additionally added to the model as $\underline{x}_1, \ldots, \underline{x}_l$. We assume that the model containing the variables $\underline{x}_{[1]}, \ldots, \underline{x}_{[k]}$ can be written as

$$\underline{Y} = X_F \underline{\beta}_F + \underline{\epsilon}, \qquad \underline{\epsilon} \sim N(0, \sigma^2 I)$$

Given that we have a current model which was chosen in the previous step of a stepwise linear regression, we want to test the null hypothesis that nothing else need go into the model. This is equivalent to calculating all the possible models with one extra variable, and then testing that the parameters relating to the extra variables are all zero. In doing this, we control for multiple comparisons at each step.

Consider one of the variables which has not been entered into the model, and denote this $\underline{x}_i, i = 1, \ldots, l$. If we denote $\underline{\beta} = [\underline{\beta}_F | \beta_i]$, and $X = [X_F | \underline{x}_i]$, we fit the model $\underline{Y} = X_F \underline{\beta}_F + \beta_i \underline{x}_i + \underline{\epsilon}$, where $\underline{\epsilon} \sim N(0, \sigma^2 I)$.

We know that $\hat{\underline{\beta}} \sim N(\underline{\beta}, \sigma^2 (X^T X)^{-1})$, and from this we can calculate the distribution of $\hat{\beta}_i$. Now, $\hat{\beta}_i = [\underline{0}^T; 1]\hat{\underline{\beta}}$, where $\underline{0}$ is a $(k+1) \times 1$ vector of zero's, and as $\hat{\underline{\beta}}$ follows a Normal distribution, so then must $\hat{\beta}_i$.

Now,

$$\begin{aligned} E(\hat{\beta}_i) &= E([\underline{0}^T; 1]\hat{\underline{\beta}}) \\ &= [\underline{0}^T; 1] E(\hat{\underline{\beta}}) \\ &= [\underline{0}^T; 1] \underline{\beta} \\ &= \beta_i. \end{aligned}$$

Also,

$$\begin{aligned} var(\hat{\beta}_i) &= var([\underline{0}^T; 1]\hat{\underline{\beta}}) \\ &= [\underline{0}^T; 1] \; var(\hat{\underline{\beta}}) [\underline{0}^T; 1]^T \\ &= [\underline{0}^T; 1] \sigma^2 (X^T X)^{-1} [\underline{0}^T; 1]^T \end{aligned}$$

Now $(X^T X)$ can be written as the partitioned matrix

$$\begin{bmatrix} X_F^T X_F & X_F^T \underline{x}_i \\ \underline{x}_i^T X_F & \underline{x}_i^T \underline{x}_i \end{bmatrix}$$

Then we can write $(X^TX)^{-1}$ as

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

Healy [25] has shown that

$$\begin{aligned} A_{22} &= [\underline{x}_i^T \underline{x}_i - \underline{x}_i^T X_F (X_F^T X_F)^{-1} X_F^T \underline{x}_i]^{-1} \\ A_{21} &= -A_{22} \underline{x}_i{}^T X_F (X_F^T X_F)^{-1} \\ A_{12} &= -(X_F^T X_F)^{-1} X_F^T \underline{x}_i A_{22} \\ A_{11} &= (X_F^T X_F)^{-1} - (X_F^T X_F)^{-1} X_F^T \underline{x}_i A_{21} \end{aligned}$$

Thus we can easily show that $var(\hat{\beta}_i) = \sigma^2 A_{22}$, i. e. $\hat{\beta}_i \sim N(\beta_i, \sigma^2 A_{22})$. However, as σ^2 is unknown, we must estimate it, and so this gives the distribution as being

$$\frac{\hat{\beta}_i - \beta_i}{\sqrt{\hat{\sigma^2} A_{22}}} \sim t(n-(k+2)).$$

Now, under the null hypothesis that all β_i's are zero, i.e. we do not need any more variables after fitting X_F, we have $\frac{\hat{\beta}_i}{\sqrt{\hat{\sigma^2} A_{22}}} \sim t(n-(k+2))$, i.e. $\frac{\hat{\beta}_i}{\sqrt{\widehat{var(\hat{\beta}_i)}}} \sim t(n-(k+2))$, $\widehat{var(\hat{\beta}_i)} = \hat{\sigma^2}[\underline{x}_i^T \underline{x}_i - \underline{x}_i^T X_F (X_F^T X_F)^{-1} X_F^T \underline{x}_i]^{-1}$.

Define T_i as being $\frac{\hat{\beta}_i}{\sqrt{\widehat{var(\hat{\beta}_i)}}}$, and suppose for each possible additional variable $\underline{x}_i$ we calculate the point estimate of T_i, t_i, and let t^* be the maximum absolute value of the $t_i, i = 1; \ldots; l$. Then we can show that the overall significance probability, p, is

$$\begin{aligned} p &= Pr(max|T_i| \geq t^* | \text{Null hypothesis true}) \\ &\leq \sum_{i=1}^{l} Pr(|T_i| \geq t^* | T_i \sim t(n-(k+2))) \\ &= 2l(1 - Pr(t(n-(k+2)) \leq t^*)) \end{aligned}$$

In order to calculate the lower bound, we must calculate the bivariate distributions of each of the T's. As each T follows a t distribution, the most sensible approximation to the joint distribution would be a bivariate t distribution. However, as the variances of the β's from which the T's are derived are not equal, the bivariate t distribution is difficult to calculate and as a result a bivariate Normal approximation will be used instead. As we are dealing with t variables, the univariate means are zero and variances 1. Thus, we need only calculate the correlation between them. To do this, we consider the covariance between the corresponding β's, which we shall denote by β_i and β_j. Denote $\underline{\beta_1}, X_1, \underline{\beta_2}$, and X_2 as the corresponding $\underline{\beta}$'s and X's for the models involving β_i and β_j respectively. Now,

$$\begin{aligned} cov(\hat{\beta}_i, \hat{\beta}_j) &= cov([\underline{0}^T; 1]\underline{\hat{\beta}_1}, [\underline{0}^T; 1]\underline{\hat{\beta}_2}) \\ &= [\underline{0}^T; 1]cov((X_1^T X_1)^{-1} X_1^T \underline{Y}, (X_2^T X_2)^{-1} X_2^T \underline{Y})[\underline{0}^T; 1]^T \\ &= [\underline{0}^T; 1](X_1^T X_1)^{-1} X_1^T cov(\underline{Y}, \underline{Y}) X_2 (X_2^T X_2)^{-1} [\underline{0}^T; 1]^T \\ &= \sigma^2 [\underline{0}^T; 1](X_1^T X_1)^{-1} X_1^T X_2 (X_2^T X_2)^{-1} [\underline{0}^T; 1]^T \end{aligned}$$

which can be shown to be:

$$\begin{aligned} cov(\hat{\beta}_i, \hat{\beta}_j) &= \sigma^2 [\underline{x}_i^T (I - X_F (X_F^T X_F)^{-1} X_F^T) \underline{x}_i]^{-1} [\underline{x}_i^T (I - X_F (X_F^T X_F)^{-1} X_F^T) \underline{x}_j] \\ &\times [\underline{x}_j^T (I - X_F (X_F^T X_F)^{-1} X_F^T) \underline{x}_j]^{-1} \end{aligned}$$

From this we can easily calculate the correlation between T_i and T_j by noting that $T_i = \frac{\hat{\beta}_i}{\sqrt{\widehat{var(\hat{\beta}_i)}}}$, and so

$$\begin{aligned} corr(T_i, T_j) &= \frac{cov(T_i, T_j)}{\sqrt{var(T_i) var(T_j)}} \\ &= \frac{cov(\beta_i, \beta_j)}{\sqrt{\widehat{var}(\hat{\beta}_i) \widehat{var}(\hat{\beta}_j)}} \end{aligned}$$

as $var(T_i) = 1, var(T_j) = 1$. We shall define ρ as $corr(T_i, T_j)$.

Now, we know that

$$Pr(max|T_i| \geq t^*) \geq \sum_{i=1}^{l} Pr(|T_i| \geq t^*) - \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} Pr(|T_i|, |T_j| \geq t^*),$$

where $\sum_{i=1}^{l} Pr(|T_i| \geq t^*)$ has already been calculated in the previous section. Denote the bivariate distribution function of T_i and T_j by f_{ij}. Now,

$$\begin{aligned} Pr(|T_i|, |T_j| \geq t^*) &= \int_{t^*}^{\infty} \int_{t^*}^{\infty} f_{ij}(T_i, T_j) dT_i dT_j + \int_{t^*}^{\infty} \int_{-\infty}^{-t^*} f_{ij}(T_i, T_j) dT_i dT_j \\ &+ \int_{-\infty}^{-t^*} \int_{t^*}^{\infty} f_{ij}(T_i, T_j) dT_i dT_j + \int_{-\infty}^{-t^*} \int_{-\infty}^{-t^*} f_{ij}(T_i, T_j) dT_i dT_j \end{aligned}$$

As both T_i and T_j follow $t(n - (k + 2))$ distributions, this can be shown to be $3 - 4Pr(t(n - (k + 2)) \leq t^*) + \int_{-t^*}^{t^*} \int_{-t^*}^{t^*} f_{ij}(T_i, T_j) dT_i dT_j$

So,

$$
\begin{aligned}
S_2 &= \sum_{i=1}^{l-1}\sum_{j=i+1}^{l} Pr(|T_i|, |T_j| \geq t^*) \\
&= \sum_{i=1}^{l-1}\sum_{j=i+1}^{l} (3 - 4Pr(t(n-(k+2)) \leq t^*)) + \sum_{i=1}^{l-1}\sum_{j=i+1}^{l} \int_{-t^*}^{t^*}\int_{-t^*}^{t^*} f_{ij}(T_i, T_j) dT_i dT_j \\
&= (3 - 4Pr(t(n-(k+2)) \leq t^*)) \sum_{i=1}^{l-1}\sum_{j=i+1}^{l} 1 + \sum_{i=1}^{l-1}\sum_{j=i+1}^{l} \int_{-t^*}^{t^*}\int_{-t^*}^{t^*} f_{ij}(T_i, T_j) dT_i dT_j \\
&= (3 - 4Pr(t(n-(k+2)) \leq t^*)) l(l-1)/2 + \sum_{i=1}^{l-1}\sum_{j=i+1}^{l} \int_{-t^*}^{t^*}\int_{-t^*}^{t^*} f_{ij}(T_i, T_j) dT_i dT_j
\end{aligned}
$$

Thus we can say that

$$
\begin{aligned}
Pr(max|T_i| \geq t^*) &\geq l(7-3l)/2 + 2l(l-2)Pr(t(n-(k+2)) \leq t^*) \\
&- \sum_{i=1}^{l-1}\sum_{j=i+1}^{l} \int_{-t^*}^{t^*}\int_{-t^*}^{t^*} f_{ij}(T_i, T_j) dT_i dT_j
\end{aligned}
$$

where we approximate f_{ij} by a standard bivariate Normal distribution function, with correlation ρ.

Suppose that we have a special case, where there are only two possible explanatory variables which can be added to the model. Then from equation(7.2) we know that

$$S_1 \geq p \geq S_1 - S_2,$$

where p is the significance probability for adding the most significant of the two into the model.

Now, from equation(7.1) we know that $p = S_1 - S_2 + \ldots + (-1)^{m-1} S_m$. However, it is clear from the definition that if we only have two possible variables then $m = 2$, and hence $p = S_1 - S_2$. But this is the value that we have calculated for the lower bound. Hence we see here that in the case of only two explanatory variables, the calculated lower bound is, in fact, the exact p-value.

7.3 Simulations

In order to determine how these bounds work in practice, simulations are generated in FORTRAN. Firstly a number of explanatory variables are generated with a response variable which is independent of all the explanatories. All the simple linear regression models are calculated, and the maximum absolute t-statistic is determined. This is used to calculate upper and lower bounds on the p-value. This is repeated for a set number of simulations.

7.3.1 Simulations with 2 Variables

In generating simulations in FORTRAN, 2 explanatory variables, $\underline{x}_1$ and $\underline{x}_2$, and an independent response $\underline{y}$, are generated from a Normal distribution. For each simulation, the two linear regression lines $\underline{y} = \alpha + \beta\underline{x}_1$ and $\underline{y} = \alpha + \beta\underline{x}_2$ are fitted and the variable with the maximum t-statistic is used to calculate upper and lower bounds on the p-value. This is repeated for a set number (1000) of simulations and the results are shown. As there are only two variables, the lower bound is the exact p-value, and so bounds which either lie completely below the 5% level, or straddle it, give significant results.

Number of Observations	Simulations with Lower<0.05	Simulations with Upper<0.05	Simulations Straddling 0.05 Line
50	0.059	0.046	0.013
75	0.054	0.048	0.006
100	0.057	0.049	0.008
150	0.047	0.045	0.002

Table 7.1: Limits on p-values from simulations with 2 variables

In each set of simulations, there are some, but not many, cases where the upper bound alone would have rejected variables. However, as the lower bound is the exact p-value in this case it should still be calculated. The number of significant simulations, when testing at the 5% level, is approximately 5% of the number of simulations, as expected.

7.3.2 Simulations with 5 Variables

In generating these simulations the same procedure is followed as in the two variables case, except for each simulation five simple linear regression lines are fitted. Again the most significant t-statistic is used to calculate upper and lower bounds on the p-value. Due to the nature of the calculation, some of the upper and lower bounds range from less than zero to more than one. These are indicated in table 7.2.

Number of Observations	Simulations with Lower<0.05	Simulations with Upper<0.05	Simulations Straddling 0.05 Line	Simulations with Upper>1 and Lower<0
50	0.171	0.045	0.126	0.029
75	0.144	0.044	0.100	0.036
100	0.116	0.045	0.071	0.033
150	0.118	0.063	0.055	0.032

Table 7.2: Limits on p-values from simulations with 5 variables

There are many cases which straddle the significance value for each set of simulations. Here it is no longer the case that those cases which straddle the 0.05 line are significant. As there are quite a large number of these cases, we shall take the simple step of locating the mid-point of the range and examine whether this lies above or below 0.05, as it seems reasonable to assume that those cases with more than half of the range less than 0.05 are more likely to be significant.

Number of Observations	Simulations with Mid-point>0.05	Simulations with Mid-point<0.05	Simulations Straddling with Mid-point<0.05
50	0.923	0.077	0.032
75	0.937	0.063	0.019
100	0.941	0.059	0.014
150	0.928	0.072	0.009

Table 7.3: Mid-points of p-value range from simulations with 5 variables

In doing this we see that the percentage of simulations with midpoint below the significance level is slightly more than 5%, indicating that the true p-value may be closer to the lower bound than the upper bound.

7.3.3 Simulations with 10 Variables

Again the same procedure is followed as in the two variables case, except for each simulation ten simple linear regression lines are fitted. The most significant t-statistic is used to calculate upper and lower bounds on the p-value.

Number of Observations	Simulations with Lower<0.05	Simulations with Upper<0.05	Simulations Straddling 0.05 Line	Simulations with Upper>1 and Lower<0
100	0.407	0.049	0.358	0.131
150	0.260	0.046	0.214	0.105
200	0.230	0.041	0.189	0.104
400	0.183	0.055	0.128	0.085

Table 7.4: Limits on p-values from simulations with 10 variables

In this case, the lower bound is again no longer the exact p-value and cases which straddle the 0.05 line can no longer be said definitely to be significant or non-significant. The mid-point of the range is again considered to determine whether it lies above or below 0.05.

Number of Observations	Simulations with Mid-point>0.05	Simulations with Mid-point<0.05	Simulations Straddling with Mid-point<0.05
100	0.876	0.124	0.075
150	0.920	0.080	0.034
200	0.934	0.066	0.025
400	0.927	0.073	0.018

Table 7.5: Mid-points of p-value range from simulations with 10 variables

Here the number of cases with mid-point below the significance level is again greater than would be expected to be significant, indicating that the true p-value may be closer to the lower bound than the upper bound.

7.4 Hald Data Set

This data set comes from Hald [23] and is reproduced by Draper and Smith [16]. These data consider the heat evolved, in calories, per gram of cement, and the amounts of four chemicals measured as a percentage of the weight of the clinkers from which the cement was made. This is a small data set, in that there are only 13 observations and 4 possible explanatory variables. The variables are described below.

- x_1 - Amount of tricalcium aluminate
- x_2 - Amount of tricalcium silicate
- x_3 - Amount of tetracalcium alumino ferrite
- x_4 - Amount of dicalcium silicate
- Y - Heat evolved in calories per gram of cement after 180 days curing

As mentioned previously, the amount of each chemical is measured as a percentage of the weight of the clinkers from which the cement was made.

Here the two Bonferroni bounds are used to attempt to fit a suitable model. Using stepwise linear regression, the most significant variable at the first step is x_4, with a t-statistic of 4.77478. From calculation of the upper and lower bound on the p-value it can be concluded that this variable is significant. Carrying out the procedure to its conclusion gives the following results.

Step	Most Significant	t^*	Upper	Lower	Accept?
1	x_4	4.77	0.0023	<0.001	Yes
2	x_1	10.40	$3.6e^{-6}$	<0.001	Yes
3	x_2	2.24	0.1034	0.032	Yes
4	x_3	0.14	0.896	-	No

The variable at the third step, x_2, is entered into the model because although the bounds straddle 0.05, at that step we have only 2 variables left and so the lower bound is the exact p-value. If only the upper Bonferroni bound were considered, this variable would not have been entered into the model as the upper bound indicates this variable is not significant, yet the lower bound, in this case the exact p-value, is significant.

This data set has previously been analysed by Draper and Smith [16] who showed, using a variety of techniques based on an upper bound on the p-value, that the best model consisted of both x_4 and x_1. However, the above method shows that while this is the best method if only the upper bound is considered, the lower bound, which is the exact p-value in this case, indicates that the x_2 variable should also be added.

7.5 Logistic Theory

It has been shown that it is possible to calculate a second order Bonferroni bound which gives a lower bound on the significance value of the most significant test, when using linear regression. In this section these results are extended to the case of logistic regression, where the estimates are no longer Normal, but are asymptotically Normal.

For the case of logistic regression, we shall suppose that we have unknown probabilities π_i, where $\pi_i = Pr(\text{"success"})$, under conditions described by the subscript i, i=1 ...p. If we assume that, for each i, we know the number of "successes", r_i, and the number of trials, n_i, then the model we are attempting to fit has three parts:

i) $R_i \sim Bi(n_i, \pi_i)$,

ii) $\eta_i = g(\pi_i)$, where $g()$ is the logistic transform.

iii) $\eta_i = \underline{x}_i^T \underline{\beta}$.

It can be shown that, asymptotically,

$$\hat{\underline{\beta}} \sim N((X^TWX)^{-1}X^TW\underline{\eta}, (X^TWX)^{-1}), \tag{7.3}$$

where W is a diagonal matrix $\{w_{ij}\}$ with $w_{ii} = n_i\pi_i(1-\pi_i)$, $X = [\underline{1}; \underline{x}_1; \ldots; \underline{x}_p]$, and $\underline{1}$ is a vector with every element being 1.

As with the linear regression case, we are interested in finding the most significant variable to add to our model, and testing whether or not it should be added. To do this, we consider the case where the model contains a constant and variables $\underline{x}_{[1]}; \ldots; \underline{x}_{[k]}$, and we wish to test whether we can add one of the remaining $m - k$

variables into the model. Let $l = m - k$, and for convenience, label the variables $\underline{x}_1; \dots; \underline{x}_l$.

Suppose we consider adding variable i to the model, $i = 1; \dots; l$. If we define the matrix $X_F = [\underline{1}; \underline{x}_{[1]}; \dots; \underline{x}_{[k]}]$, then we are interested in the case where our design matrix $X = [X_F; \underline{x}_i]$. Then from equation (7.3), $\underline{\beta}_i$ has the asymptotic distribution

$$\hat{\underline{\beta}}_i \sim N((X^T W X)^{-1} X^T W \underline{\eta}, (X^T W X)^{-1}),$$

where $\underline{\beta}_i = [\underline{\beta}; \beta_i]$.

Now, we are interested in the distribution of $\hat{\beta}_i$, as we wish to test the null hypothesis that we do not need any additional variables in the model, i. e. for $i = 1, \dots, l$, β_i is zero, $i = 1; \dots; l$.

Clearly, as $\hat{\underline{\beta}}_i$ is asymptotically Normal, then $\hat{\beta}_i$ must be also. Now, $\hat{\beta}_i = [\underline{0}^T; 1]\hat{\underline{\beta}}_i$, where $\underline{0}$ is an $(m+1) \times 1$ vector of zeros. Let μ_i be the expectation of $\hat{\beta}_i$, and σ_i^2 be the variance. Then

$$\begin{aligned} \mu_i &= [\underline{0}^T; 1] E(\hat{\underline{\beta}}_i) \\ &= [\underline{0}^T; 1](X^T W X)^{-1} X^T W \underline{\eta} \end{aligned}$$

and

$$\begin{aligned} \sigma_i^2 &= [\underline{0}^T; 1] V(\hat{\underline{\beta}}_i)[\underline{0}^T; 1]^T \\ &= [\underline{0}^T; 1](X^T W X)^{-1}[\underline{0}^T; 1]^T \end{aligned}$$

In order to calculate these, we need to look more closely at the matrix $X^T W X$. Now, $X = [X_F; \underline{x}_i]$, so we can show that $X^T W X$ can be written as the partitioned matrix

$$\begin{bmatrix} X_F^T W X_F & X_F^T W \underline{x}_i \\ \underline{x}_i^T W X_F & \underline{x}_i^T W \underline{x}_i \end{bmatrix}$$

Then we can write $(X^T W X)^{-1}$ as

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

where

$$B_{22} = [\underline{x}_i^T W \underline{x}_i - \underline{x}_i^T W X_F (X_F^T W X_F)^{-1} X_F^T W \underline{x}_i]^{-1}$$

$$\begin{aligned}
B_{12} &= -(X_F^T W X_F)^{-1} X_F^T W \underline{x}_i B_{22} \\
B_{21} &= -B_{22} \underline{x}_i{}^T W X_F (X_F^T W X_F)^{-1} \\
B_{11} &= (X_F^T W X_F)^{-1} - (X_F^T W X_F)^{-1} X_F^T W \underline{x}_i B_{21}
\end{aligned}$$

as shown by Healy [25]. Then

$$\begin{aligned}
\mu_i &= B_{21} X_F^T W \underline{\eta} + B_{22} \underline{x}_i{}^T W \underline{\eta} \\
&= B_{22} \underline{x}_i [I - W X_F (X_F^T W X_F)^{-1} X_F^T] W \underline{\eta}
\end{aligned}$$

and $\sigma_i^2 = B_{22}$. Thus the asymptotic distribution of $\hat{\beta}_i$ is

$$\hat{\beta}_i \sim N(\mu_i, \sigma_i^2),$$

where

$\mu_i = (\underline{x}_i^T W \underline{x}_i - \underline{x}_i^T W X_F (X_F^T W X_F)^{-1} X_F^T W \underline{x}_i)^{-1} \underline{x}_i^T (I - W X_F (X_F^T W X_F)^{-1} X_F^T) W \underline{\eta}$,
$\sigma_i^2 = (\underline{x}_i^T W \underline{x}_i - \underline{x}_i^T W X_F (X_F^T W X_F)^{-1}) X_F^T W \underline{x}_i]^{-1}$.

Now, under the null hypothesis that all β_i's are zero, i.e. we do not need any more variables after fitting X_F, we have $\underline{\eta} = X_F \underline{\gamma}$, where $\underline{\gamma}$ is the vector of parameters when fitting X_F. Then we notice that

$$\begin{aligned}
(I - W X_F (X_F^T W X_F)^{-1} X_F^T) W \underline{\eta} &= W \underline{\eta} - W X_F (X_F^T W X_F)^{-1} X_F^T W \underline{\eta} \\
&= W X_F \underline{\gamma} - W X_F (X_F^T W X_F)^{-1} X_F^T W X_F \underline{\gamma} \\
&= W X_F \underline{\gamma} - W X_F \underline{\gamma} \\
&= 0.
\end{aligned}$$

Hence μ_i is also zero and asymptotically, $\hat{\beta}_i \sim N(0, \sigma_i^2)$ and

$$\frac{\hat{\beta}_i}{\sigma_i} \sim N(0,1).$$

Define T_i as being $\frac{\hat{\beta}_i}{\sigma_i}$, and suppose for each possible additional variable $\underline{x}_i$ we calculate the point estimate of T_i, t_i, and let t^* be the maximum absolute value of the $t_i, i = 1; \ldots; k$. Then we can show that the significance probability, p, is

$$\begin{aligned}
p &= Pr(max|T_i| \geq t^* | \text{null hypothesis true}) \\
&\leq \sum_{i=1}^{l} Pr(|T_i| \geq t^* | T_i \sim N(0,1)) \\
&= 2l(1 - \Phi(t^*)) \qquad (7.4)
\end{aligned}$$

Note that this is very similar to the linear regression case; the only difference is that here we are not estimating σ_i^2 and thus have a Normal distribution rather than a t-distribution.

We now show how the lower bound can also be extended. From the previous expressions for linear regression we have:

$$Pr(max|T_i| \geq t^*) \geq \sum_{i=1}^{l} Pr(|T_i| \geq t^*) - \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} Pr(|T_i|, |T_j| \geq t^*) \tag{7.5}$$

Thus we must consider the bivariate distribution of (T_i, T_j). We know that

$$\begin{aligned} Pr(|T_i|, |T_j| \geq t^*) &= Pr(T_i, T_j \geq t^*) + Pr(T_i \geq t^*, T_j \leq -t^*) \\ &+ Pr(T_i, T_j \leq -t^*) + Pr(T_i, \leq -t^*, T_j \geq t^*) \end{aligned} \tag{7.6}$$

Suppose that the joint density function of T_i and T_j is f_{ij}. Then equation (7.6) becomes

$$\begin{aligned} Pr(|T_i|, |T_j| \geq t^*) &= \int_{t^*}^{\infty} \int_{t^*}^{\infty} f_{ij}(T_i, T_j) dT_i dT_j + \int_{t^*}^{\infty} \int_{-\infty}^{-t^*} f_{ij}(T_i, T_j) dT_i dT_j \\ &+ \int_{-\infty}^{-t^*} \int_{t^*}^{\infty} f_{ij}(T_i, T_j) dT_i dT_j + \int_{-\infty}^{-t^*} \int_{-\infty}^{-t^*} f_{ij}(T_i, T_j) dT_i dT_j \end{aligned}$$

This can easily be simplified to:

$$\begin{aligned} Pr(|T_i|, |T_j| \geq t^*) &= \int_{t^*}^{\infty} \int_{-\infty}^{\infty} f_{ij}(T_i, T_j) dT_i dT_j - \int_{t^*}^{\infty} \int_{-t^*}^{t^*} f_{ij}(T_i, T_j) dT_i dT_j \\ &\quad + \int_{-\infty}^{-t^*} \int_{-\infty}^{\infty} f_{ij}(T_i, T_j) dT_i dT_j - \int_{-\infty}^{-t^*} \int_{-t^*}^{t^*} f_{ij}(T_i, T_j) dT_i dT_j \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{ij}(T_i, T_j) dT_i dT_j - \int_{-\infty}^{\infty} \int_{-t^*}^{t^*} f_{ij}(T_i, T_j) dT_i dT_j \\ &\quad - \int_{-t^*}^{t^*} \int_{-\infty}^{\infty} f_{ij}(T_i, T_j) dT_i dT_j = \int_{-t^*}^{t^*} \int_{-t^*}^{t^*} f_{ij}(T_i, T_j) dT_i dT_j \\ &= 1 - \int_{-t^*}^{t^*} f_i(T_i) dT_i - \int_{-t^*}^{t^*} f_j(T_j) dT_j \\ &\quad + \int_{-t^*}^{t^*} \int_{-t^*}^{t^*} f_{ij}(T_i, T_j) dT_i dT_j \end{aligned} \tag{7.7}$$

where f_i and f_j are the density functions of T_i and T_j. This can be shown to simplify to

$$Pr(|T_i|, |T_j| \geq t^*) = 1 - 4\Phi(t^*) + 2 + \int_{-t^*}^{t^*} \int_{-t^*}^{t^*} f_{ij}(T_i, T_j) dT_i dT_j$$

and substitution into equation(7.5) gives us

$$\begin{aligned} Pr(max|T_i| \geq t^*) &\geq 2l(1 - \Phi(t^*)) \\ &\quad - \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} (3 - 4\Phi(t^*) + \int_{-t^*}^{t^*} \int_{-t^*}^{t^*} f_{ij}(T_i, T_j) dT_i dT_j) \\ &= \frac{l}{2}(7 - 3l) + 2l(l-2)\Phi(t^*) \\ &\quad - \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} \int_{-t^*}^{t^*} \int_{-t^*}^{t^*} f_{ij}(T_i, T_j) dT_i dT_j \end{aligned} \tag{7.8}$$

Now as T_i and T_j are both univariate Normal it seems reasonable to approximate the joint distribution of T_i and T_j by a bivariate Normal distribution. As we know the means and variances, all that remains is to calculate the correlation between them. We know that T_i is a linear function of β_i, so we shall calculate the correlation between β_i and β_j.

Using the same notation as before, it can be shown (for example, McCullagh & Nelder [32]) that

$$\underline{\hat{\beta}}_i = (X_i^T W_i X_i)^{-1} X_i^T W_i \underline{z}_i,$$

where $X_i = [X_F; \underline{x}_i]$, $W_i = diag(\hat{\pi}_{ik}(1 - \hat{\pi}_{ik})n_k)$, $\underline{\hat{\beta}}_i = [\underline{\hat{\beta}}; \beta_i]$, and
$z_{ik} = logit(\hat{\pi}_{ik}) + (y_k - n_k\hat{\pi}_{ik})/(n_k\hat{\pi}_{ik}(1 - \hat{\pi}_{ik})$.
So, $\text{cov}(\underline{\hat{\beta}}_i, \underline{\hat{\beta}}_j) = (X_i^T W_i X_i)^{-1} X_i^T W_i cov(\underline{z}_i, \underline{z}_j) W_j X_j^T (X_j^T W_j X_j)^{-1}$.
Now, we can show that

$$cov(z_{ik}, z_{jl}) = \begin{cases} 0 & k \neq l \\ \frac{\pi_l(1-\pi_l)}{n_l \pi_{il} \pi_{jl} (1-\pi_{il})(1-\pi_{jl})} & k = l \end{cases}$$

Thus $\text{cov}(\underline{z}_i, \underline{z}_j)$ is a diagonal matrix with elements as above. Call this matrix C_{ij}. Then $\text{cov}(\underline{\hat{\beta}}_i, \underline{\hat{\beta}}_j) = (X_i^T W_i X_i)^{-1} X_i^T W_i C_{ij} W_j X_j (X_j^T W_j X_j)^{-1}$. Thus

$$\begin{aligned} cov(\hat{\beta}_i, \hat{\beta}_j) &= [0; \ldots; 0; 1] cov(\underline{\hat{\beta}}_i, \underline{\hat{\beta}}_j) [0; \ldots; 0; 1]^T \\ &= [0; \ldots; 0; 1] (X_i^T W_i X_i)^{-1} X_i^T W_i C_{ij} W_j X_j^T (X_j^T W_j X_j)^{-1} [0; \ldots; 0; 1]^T \end{aligned}$$

Define ρ as $\frac{cov(\hat{\beta}_i,\hat{\beta}_j)}{\sqrt{var(\hat{\beta}_i)var(\hat{\beta}_j)}}$. Then ρ is the correlation between $\hat{\beta}_i$ and $\hat{\beta}_j$. Now,

$$\begin{aligned} cov(T_i,T_j) &= cov(\frac{\hat{\beta}_i}{\sqrt{var(\hat{\beta}_i)}},\frac{\hat{\beta}_j}{\sqrt{var(\hat{\beta}_j)}}) \\ &= \frac{1}{\sqrt{var(\hat{\beta}_i)var(\hat{\beta}_j)}}cov(\hat{\beta}_i,\hat{\beta}_j) \\ &= \rho \end{aligned}$$

And as $var(T_i) = var(T_j) = 1$, the correlation between T_i and T_j is also ρ.

Then the lower bound on the significance value will be

$$Pr(max|T_i| \geq t^*) \geq \frac{l}{2}(7-3l)+2l(l-2)\Phi(t^*)-\sum_{i=1}^{l-1}\sum_{j=i+1}^{l}\int_{-t^*}^{t^*}\int_{-t^*}^{t^*} f_{ij}(T_i,T_j)dT_idT_j \quad (7.9)$$

where f_{ij} is the joint density function of T_i and T_j; approximated by a bivariate Normal with mean $\underline{0}$, and variance-covariance matrix $\begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$.

The two bounds determined in equations (7.4) and (7.9) shall be used in the following section to determine wheher the calculation of a lower bound in logistic regression

7.6 Simulations

In generating simulations in FORTRAN, the same procedure is followed as for the normal case. 2 explanatory variables, $\underline{x}_1$ and $\underline{x}_2$, are generated. A response variable, with success probability δ at each observation is also generated. For each simulation, the two logistic regressions $\frac{\delta}{(1-\delta)} = exp(\alpha + \beta x_1)$ and $\frac{\delta}{(1-\delta)} = exp(\alpha + \beta x_2)$ are fitted and the variable with the maximum t-statistic is used to calculate upper and lower bounds on the p-value. This is repeated for 1000 simulations, which is large enough to avoid spurious results, but small enough for a reasonable run time for the simulations. The results are shown below. As there are only two variables, the special case holds and so the lower bound is the exact p-value, and so bounds which either lie completely below the 5% level, or straddle it, give significant results.

In this case, very little is learned from the lower bound. In each set of simulations, only one case straddles the significance level. This suggests that for a logistic model

Number of Observations	Simulations with Lower<0.05	Simulations with Upper<0.05	Simulations with Upper<0.1	Simulations Straddling 0.05 Line
50	0.036	0.035	0.088	0.001
75	0.042	0.041	0.093	0.001
100	0.048	0.047	0.102	0.001

Table 7.6: Limits on p-values from simulations with 2 variables - logistic model

with only two possible variables, it is sufficient to calculate the upper bound on the p-value.

7.7 Low Birthweight Data Set

7.7.1 Low Birthweight in 1991

The low birthweight data set described in chapter 3 has been used to investigate the probability of delivering a low birthweight infant. As the response variable of interest is birthweight, specifically, whether the baby is low birthweight (under 2500 g) or not, and there are eight possible explanatory variables, the data shall be analysed using the upper and lower Bonferroni bounds for logistic regression.

The data used in this regression model are from 1991 and consist of singleton livebirths to first time mothers in the Greater Glasgow Health Board area. There are 4272 cases used in the analysis. The variables are described below.

- x_1 - Gestational age (measured in weeks)
- x_2 - Maternal height (measured in cm)
- x_3 - Maternal age (measured in years)
- x_4 - Marital status (married/not married)
- x_5 - Maternal condition (yes/no)
- x_6 - Neighbourhood deprivation score

- x_7 - Neighbourhood age score
- x_8 - Neighbourhood house score
- Y - 1 if baby is low birthweight, 0 if not

The three neighbourhood scores are calculated from the 1991 small area census data as described in chapter 4, and are calculated at postcode sector level. Maternal condition is defined as the presence of placenta previa, premature placental separation, antepartum haemorrhaging, or hypertension during pregnancy.

The two Bonferroni bounds are used to fit a suitable model. Using stepwise logistic regression, at the first step the most significant variable is x_1, gestational age. Calculation of the upper and lower bounds indicate that this variable is significant. Carrying out the procedure to its conclusion gives the following results.

Step	Most Significant	Upper	Lower	Accept?
1	x_1	<0.001	<0.001	Yes
2	x_2	<0.001	<0.001	Yes
3	x_6	<0.001	<0.001	Yes
4	x_4	0.1346	0.1228	No

Although the lower bound does not have an effect on the results, if there were no correction for multiple comparisons then x_4, marital status, would have been added into the model also. The next variable which would be added to the model, x_7, neighbourhood age score, has an uncorrected p-value of 0.29 and so would not be added to the model if multiple comparison corrections were ignored.

The model calculated is

$$\begin{aligned} logit(Pr(\text{low birthweight})) &= 45.02953 + 0.407 \text{ deprivation} \\ &\quad -0.075 \text{ height} - 0.935 \text{ gestation} \end{aligned}$$

This suggests that as gestation and height increase, the probability of a low birthweight baby decreases, but that as deprivation increases, the probability of a low birthweight baby increases.

The appropriateness of using linear terms in this model is considered by fitting three logistic regressions, one for each explanatory variable in the model. This model is then compared with the proportion of low birthweight infants in each of several categories of the explanatory variable, by comparing these proportions, with appropriate standard errors, against the fitted logistic curve. This is shown graphically in figures 7.1 to 7.3. Each of these figures appears to indicate that the linear term is appropriate.

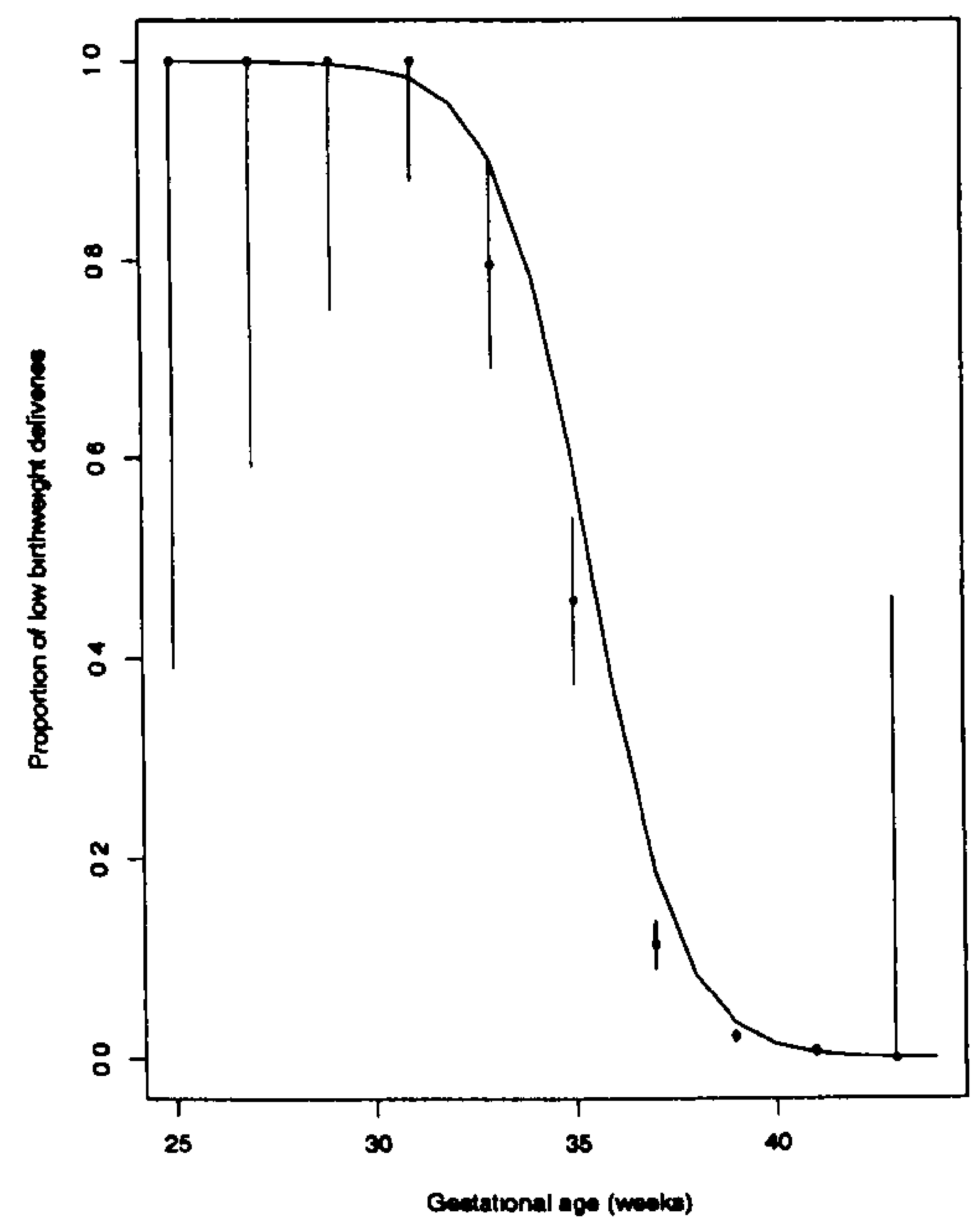

Figure 7.1: Comparison of fitted logistic curve against proportion of low birthweight infants in each gestational age category.

The height range used in determining this model was 137 cm to 180 cm. It would be inadvisable to use this model to determine the probability of delivering a low birthweight infant for mothers outside these extreme heights as the model was constructed using this height range.

For a mother with a Murray score of 0 and a gestation of 38 weeks, a difference in height between 155 cm and 160 cm indicates a difference in probability of delivering a low birthweight baby from 0.119 for 155 cm to 0.076 for 160 cm. For a mother with

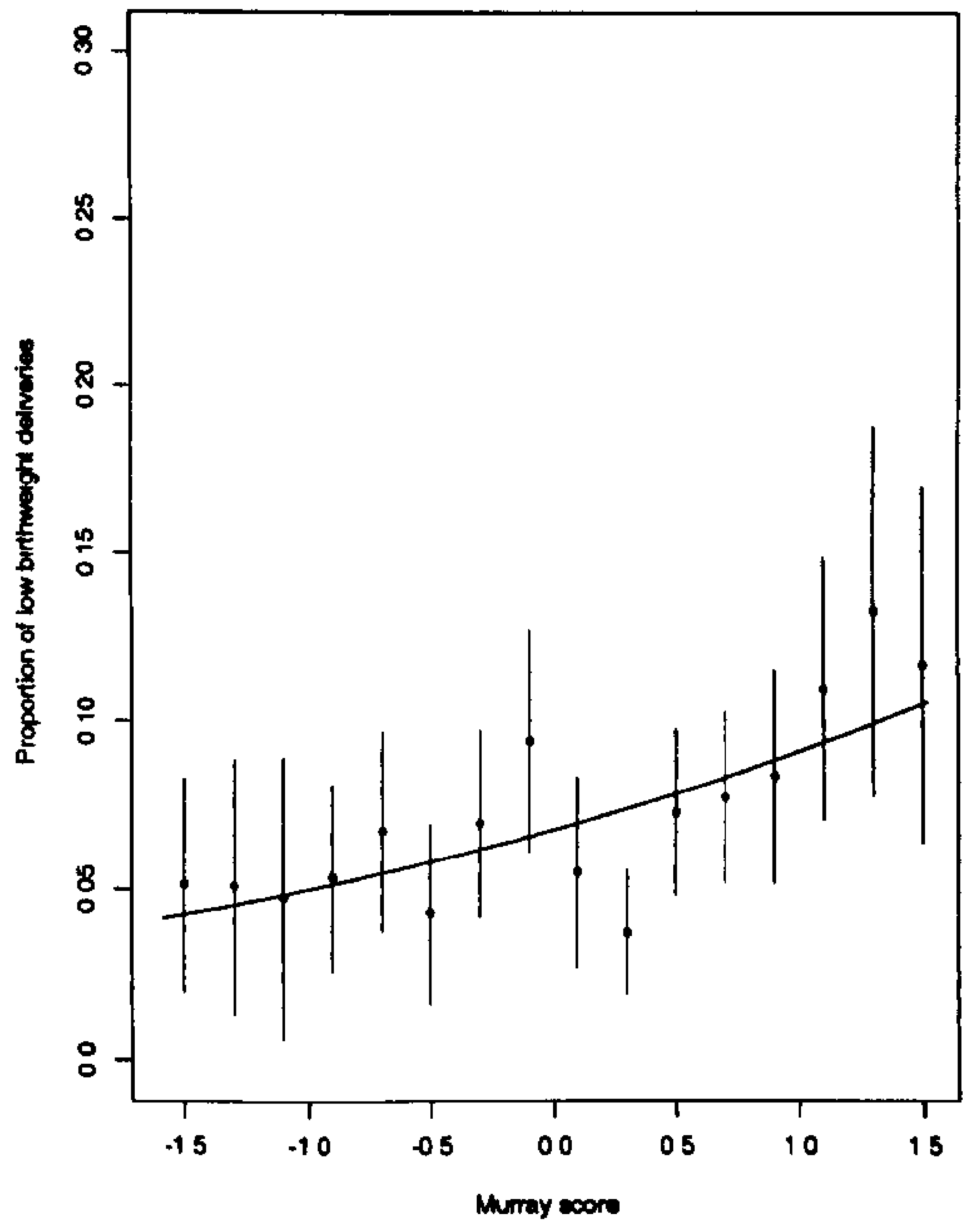

Figure 7.2: Comparison of fitted logistic curve against proportion of low birthweight infants in each Murray score category.

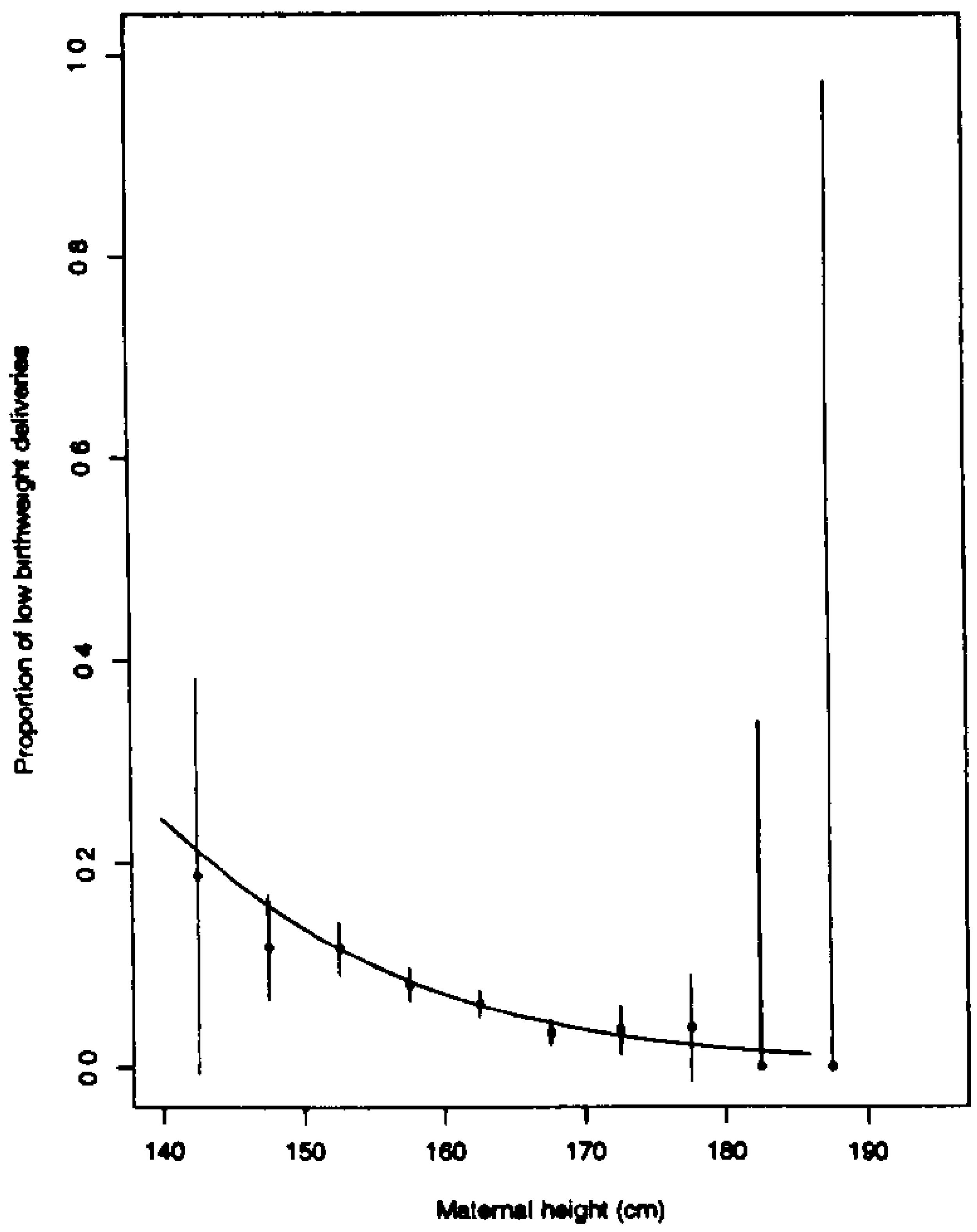

Figure 7.3: Comparison of fitted logistic curve against proportion of low birthweight infants in each height category.

a Murray score of 0 and a gestation of 40 weeks, the same height difference indicates a difference in probability of delivering a low birthweight baby from 0.018 for 155 cm to 0.012 for 160 cm. This suggests that for mothers with the same deprivation and gestation even a difference in height of 5 cm has a difference in the probability, with this difference being greater at shorter gestations.

7.7.2 Effect of Previous Low Birthweight Infant

Here the linked data set described in chapter 5 is used to investigate the probability of delivering a low birthweight infant using previous low birthweight as an explanatory variable. Again, as the response variable of interest is birthweight and there are a large number of possible explanatory variables, the data shall be analysed using the upper and lower Bonferroni bounds for logistic regression.

The data used in this regression model consist of singleton livebirths to second time mothers in the Greater Glasgow Health Board area where the first birth occurred in 1980. There are 2556 cases used in the analysis. The variables are described below. All refer to the second pregnancy unless otherwise indicated.

- x_1 - Gestational age (measured in weeks)
- x_2 - Maternal height (measured in cm)
- x_3 - Maternal age (measured in years)
- x_4 - Neighbourhood deprivation score
- x_5 - Neighbourhood age score
- x_6 - Neighbourhood house score
- x_7 - Birthweight from first pregnancy
- Y - 1 if second baby is low birthweight, 0 if not

The three neighbourhood scores are calculated from the 1991 small area census data as described in chapter 4, and are calculated at postcode sector level. Maternal condition is defined as the presence of placenta previa, premature placental separation, antepartum haemorrhaging, or hypertension during pregnancy.

The two Bonferroni bounds are used to fit a suitable model. Using stepwise logistic regression, at the first step the most significant variable is x_1, gestational age. Calculation of the upper and lower bounds indicate that this variable is highly significant. Carrying out the procedure to its conclusion gives the following results.

Step	Most Significant	Upper	Lower	Accept?
1	x_1	<0.0001	<0.0001	Yes
2	x_7	<0.0001	<0.0001	Yes
3	x_4	<0.0006	<0.0004	Yes
4	x_5	0.404	0.334	No

The model calculated is

$$\begin{aligned} logit(Pr(\text{low birthweight})) = {} & 30.481 + 0.613 \text{ deprivation} \\ & -0.001 \text{ previous birthweight} - 0.798 \text{ gestation} \end{aligned}$$

This suggests that as gestation and previous birthweight increase, the probability of a low birthweight baby decreases, but that as deprivation increases, the probability of a low birthweight baby increases.

The appropriateness of using linear terms in this model is considered by fitting three logistic regressions, one for each explanatory variable in the model. This model is then compared with the proportion of low birthweight infants in each of several categories of the explanatory variable, by comparing these proportions, with appropriate standard errors, against the fitted logistic curve. This is shown graphically in figures 7.4 to 7.6. Each of these figures appears to indicate that the linear term is appropriate. One possible exception is that of previous low birthweight, where the model does not appear to fit well at under 1300 g. However, this may be due to the small number of cases, which here is 12.

For a mother with a Murray score of 0 and a gestation of 38 weeks, a difference in previous birthweight between 2500 g and 2600 g indicates a difference in probability of delivering a low birthweight baby from 0.088 for 2500 g to 0.080 for 2600 g. For a mother with a Murray score of 0 and a gestation of 40 weeks, the same previous birthweight difference indicates a difference in probability of deliver-

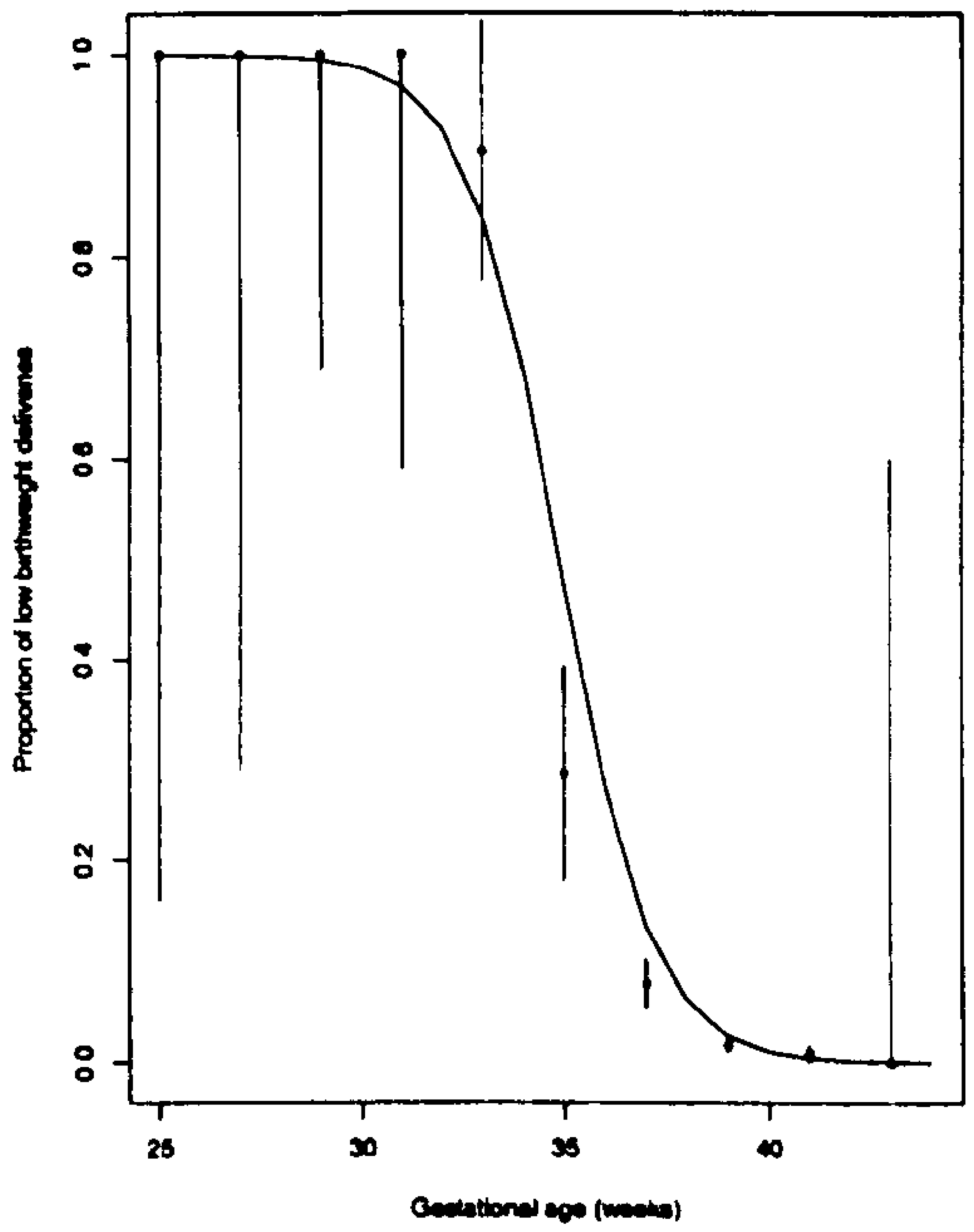

Figure 7.4: Comparison of fitted logistic curve against proportion of low birthweight infants in each gestational age category.

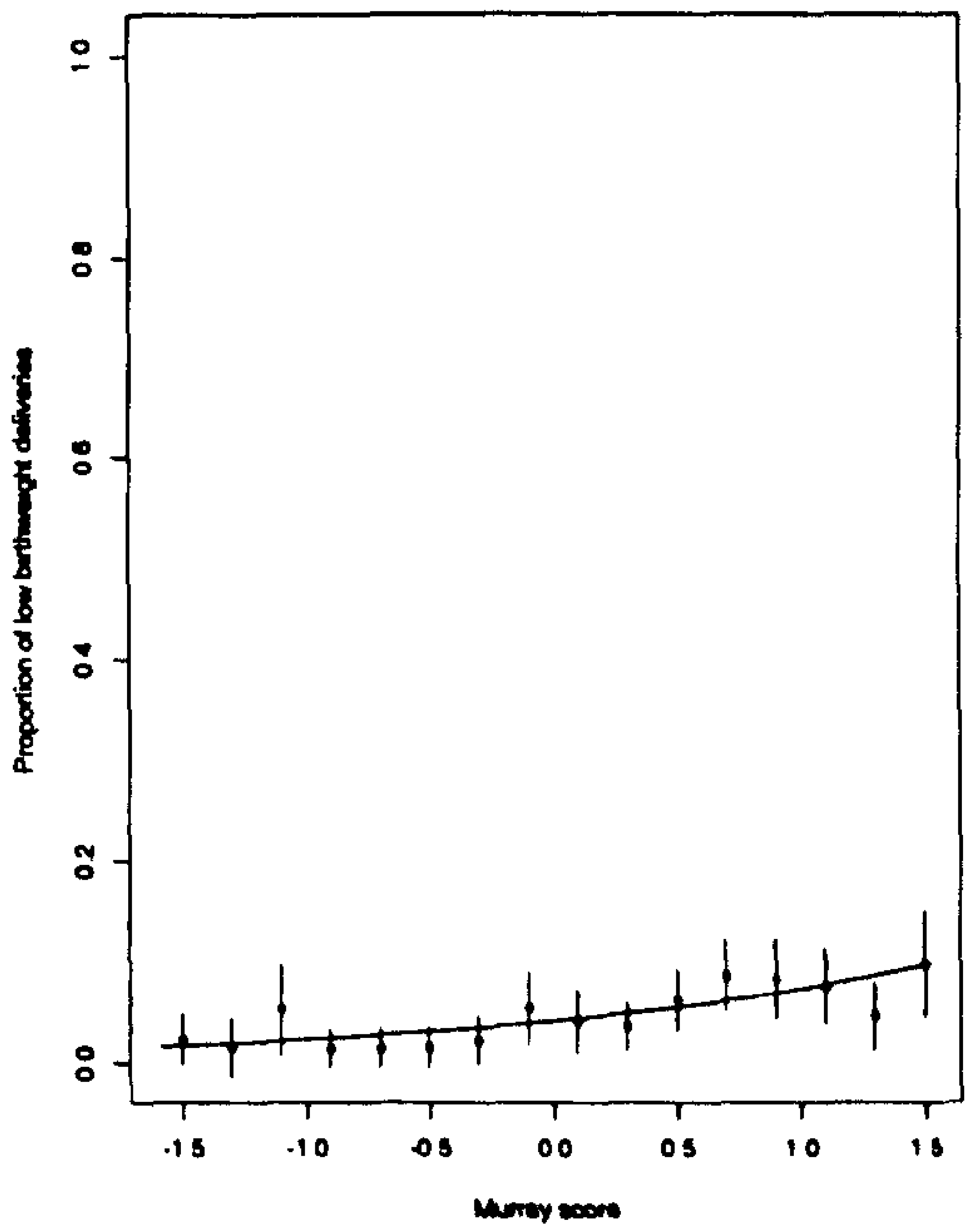

Figure 7.5: Comparison of fitted logistic curve against proportion of low birthweight infants in each Murray score category.

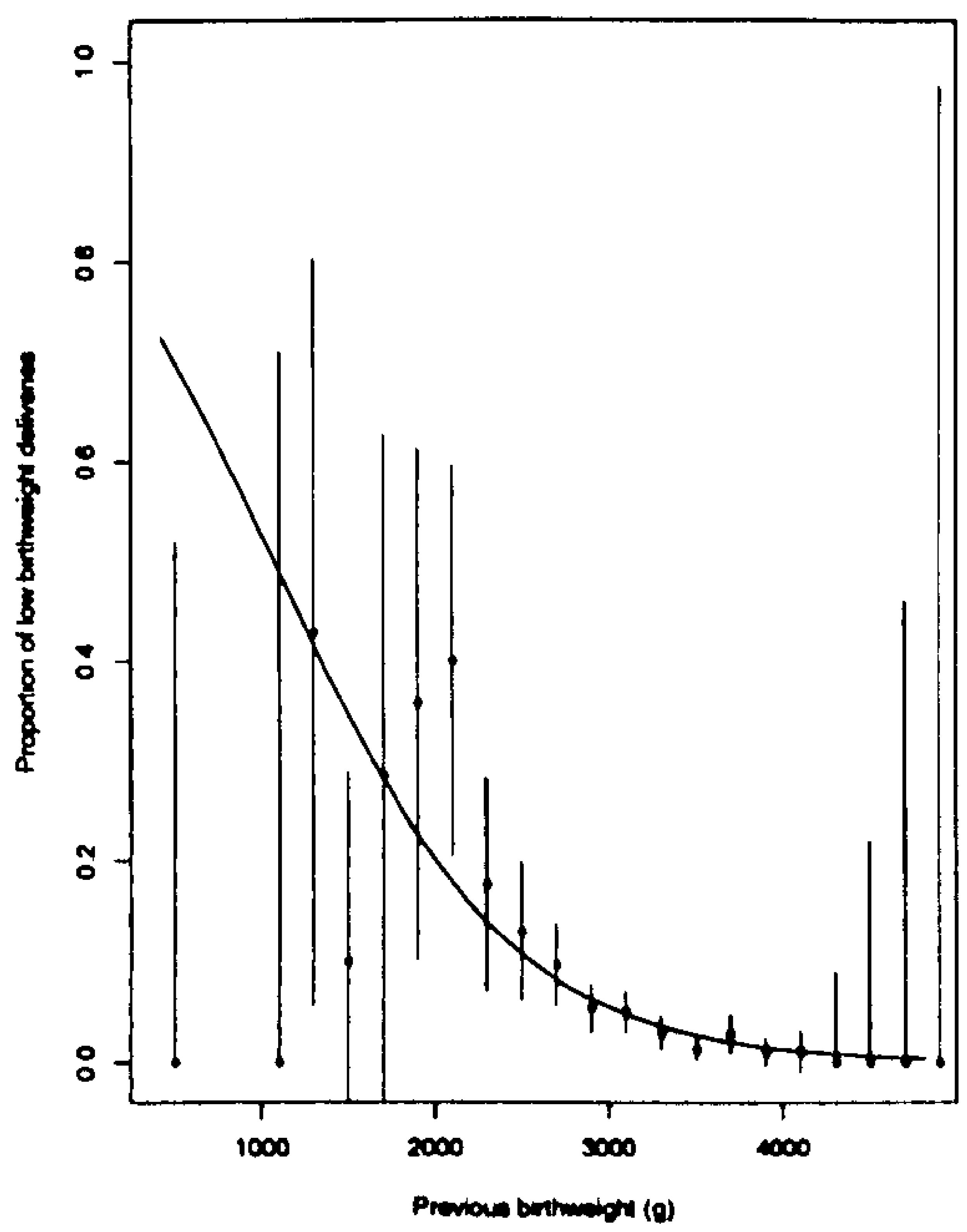

Figure 7.6: Comparison of fitted logistic curve against proportion of low birthweight infants in each previous birthweight category.

ing a low birthweight baby from 0.019 for 2500 g to 0.017 for 2600 g. Similarly to the previous example, there is a difference between the probabilities of delivering a low birthweight infant for mothers with the same gestation and deprivation with a 100 g difference in previous birthweight, with this difference being larger for shorter gestations.

7.8 Conclusions

This chapter has shown that the calculation of both upper and lower bounds on the p-value in stepwise subset regression gives a better estimation of the p-value than calculation of an upper bound only. This is usefully better if Normal regression models as, if the upper Bonferroni bound had been used on its own, the third step in the Hald example would have rejected adding any more variables into the model, yet calculation of the lower bound has shown that one of the variables was in fact significant.

For larger numbers of possible variables, where the actual p-value is unknown, calculation of both bounds may help in those cases where the upper bound is insignificant, yet still close to the required significance level, for example where the upper bound is 0.06, and testing takes place at the 5% level.

The extension of Normal theory to the case of logistic regression is reasonably straightforward, due to the asymptotically Normal nature of the parameter estimates. This implies that extensions to other generalised linear models may be routine.

There are some possible avenues for further work which may be derived from this chapter. The first, mentioned above, is that the theory for logistic models may be extended to other generalised linear models. A second possibility is that it may be possible to extend some of the theory into survival analysis. Simulations investigating the power of tests may also be carried out, similar to those for size mentioned in this chapter. Simulations could also be carried out based on fitted models from real data, to investigate the proportion of simulations where the correct model is chosen. Finally, there is no firm conclusion about the treatment of those sets of bounds which straddle the significance level. In these cases the variable may or may not be accepted into the model. In this chapter, the view taken has been that for larger numbers of variables acceptance or rejection should be decided on whether more than 50% of the bounds' range lies below or above the significance level. A more theoretical idea would be to note that from Butler [8],

$$S_1 - S_2 \leq Pr(A_{max}) \leq S_1 - S_2 + S_3$$

and as the S_i, i=1,...,n are monotonically decreasing, $S_1 - S_2 + S_3 \leq S_1$. Then $S_1 - S_2 + S_3$ is a tighter upper bound on the p-value. Calculation of this bound would remove some, but not all, of the cases which straddle the significance level.

Chapter 8

Conclusions

In this thesis, we have been interested in modelling the probability of delivering a low birthweight baby. Low birthweight is defined as a birthweight of less than 2500 g. Very low birthweight, which was also discussed in this thesis, is defined as a birthweight of less than 1500 g. The probability of delivering a low birthweight baby was modelled using logistic regression while correcting for multiple comparisions, and also by using upper and lower Bonferroni bounds to determine which variables should be entered into the model. In addition, a regression model was considered using previous low birthweight as an explanatory variable along with two variables that were significant in these models.

Possible factors affecting low birthweight in the Greater Glasgow Health Board area were considered using univariate methods in order to determine those factors that had an effect on birthweight. Of the factors investigated, gestational age, maternal age, maternal height, marital status, Womersley neighbourhood type, previous spontaneous abortions (miscarriages) and previous induced abortions had an effect on birthweight when each was considered separately. These factors were then used as possible explanatory variables in a regression modelling the probability of low birthweight. Maternal height, gestational age and Womersley neighbourhood type were found to be statistically significant.

A Glasgow specific deprivation measure by postcode sector was derived from the 1991 small area census data by carrying out a principal components analysis. From this analysis the first three components were manipulated to produce three scores, the first of which contained variables linked to deprivation, and this was denoted the

Murray score. This score differs from others currently used in Scotland in that it is a continuous score that is not converted into categories.

The birthweights of the first two infants to each mother whose first infant was born in 1980 were considered. It was shown that there was evidence of an association between the two birthweights, with a low birthweight first pregnancy being more likely to be followed by a second low birthweight pregnancy.

Finally, the use of first and second order Bonferroni bounds was investigated in regression models assuming Normally distributed data. The theory behind these two bounds was discussed and a model fitted to the Hald data set. This theory was then extended to the case of logistic regression, and first and second order bounds were produced for this case. Two models for the probability of delivering a low birthweight infant were fitted using these techniques, and it was found that gestational age and Murray deprivation score were significant in both models using Bonferroni bounds, along with height in the model from the unlinked dataset and previous birthweight on the model from the linked dataset.

The aim of this thesis was to investigate possible factors related to low birthweight using univariate methods and to produce a model for the probability of delivering a low birthweight infant based on maternal factors. The four models produced used different sets of explanatory variables. In each model, a measure of deprivation was found to be significant, either the Womersley measure or the Murray score. Also, when gestational age was an explanatory variable, it was the first variable to be added to the model. These models are based on data from different years and as such it is inappropriate to make direct comparisons between them. However, the Bonferroni method gives a more reassuring model as both upper and lower bounds are produced and this can indicate definitely if a variable should not be added to the model as it will not significantly improve the model.

Each distinct part of this thesis is now considered in greater detail. The following sections discuss new work carried out, results obtained and suggestions for future work.

8.1 Low Birthweight

Much of the previous work carried out in the field of low birthweight and related obstetric problems has found that unmarried mothers, teenage mothers and those over 35 carry an increased risk of delivering a low birthweight infant. In addition, an unfavourable outcome in previous pregnancies, be it a stillbirth, an induced abortion or a spontaneous abortion, increases the risk of pre-term delivery in successive pregnancies, while previous low birthweight deliveries carry an increased risk of low birthweight in the current pregnancy.

The most recent unlinked Glasgow data has corroborated the findings of other authors. Low birthweight was split into two categories, very low birthweight (under 1500 g), and low birthweight (1500 g - 2499 g). Using these definitions, it was possible to investigate factors related to low birthweight in greater depth as it has been thought that low birthweight and very low birthweight may be related to different factors.

Consideration of the Womersley deprivation score shows that as the deprivation score increases the proportion of low birthweight deliveries increases. However, the proportion of very low birthweight infants remains constant over deprivation type, for both 1981 and 1991, implying that low birthweight may be related to socio-economic factors, while very low birthweight may be due to obstetric factors.

As the number of previous pregnancies ending in spontaneous abortions increases, the proportions of low and very low birthweight infants subsequently delivered increases. This is to be expected, as a spontaneous abortion may indicate an obstetric abnormality that may manifest itself differently in subsequent pregnancies. Similar results for very low birthweight infants occur as the number of previous induced abortions increases, although the proportion of low birthweight infants remains constant. This may indicate that an induced abortion may cause severe problems with subsequent pregnancies, or that factors linked with the induced abortion may make the mother predisposed to problems in any further pregnancies.

Fitting a model to the probability of delivering a low birthweight infant showed that marital status, maternal age and maternal condition were not significant after gestational age, neighbourhood type and maternal height had been entered into the model. In this thesis a woman is defined as having a maternal condition if the data collected indicates that she suffered from placenta previa, premature placental separation, antepartum haemorrhaging, or hypertension during the pregnancy. These are conditions that may cause a pre-term birth likely to result in low birthweight. Both marital status and maternal age may be correlated with deprivation type, and maternal condition and gestational age may also be correlated, indicating that when one of these variables is already fitted there is very little additional information to be obtained from the correlated variables. The model indicated that as gestational age and maternal height increased, and neighbourhood type became less deprived, the probability of delivering a low birthweight infant decreased. As maternal age and marital status have been shown to have an impact on low birthweight, it may be the case that the available data here are insufficient to demonstrate this. While a larger data set would have been possible from the data given, this was thought to be too computationally intensive to be viable.

There were some possible factors, for example smoking, which had been shown by other authors to be related to low birthweight but could not be considered in this work due to the restrictions of the SMR2 form used for data collection. Further work may involve the setting up of a clinical study where data on alcohol, drug and tobacco consumption of women would be collected in addition to data on the SMR2 form.

8.2 Deprivation Measures

Of the four deprivation measures discussed in the first part of chapter 4, the Townsend score and the Carstairs score are correlated with health measures and have been extended to geographical areas other than those used in their creation with no serious problems. The Jarman index has been shown to be biased towards London and to compare unfavourably with the Townsend and Carstairs scores in the prediction of GP workload, an area for which it was specifically created.

A new deprivation measure, denoted the Murray score, was derived for the Greater Glasgow Health Board area, using 29 socio-demographic variables collected in the 1991 census and summarised by postcode sector of residence. This is based on the Womersley neighbourhood type method, differing in that the Murray score does not carry out a cluster analysis after a principal components analysis of 29 census variables per postcode sector. It re-groups the census variables, based in part on the principal components analysis, into three distinct groups of variables, one of which is denoted the 'Deprivation' group. The Murray score is then derived for each postcode sector from the values of the variables in this deprivation group.

This new score was derived in order to produce a deprivation measure derived from the most recent data available. The Carstairs score is derived for the whole of Scotland, however as this thesis concentrated on the Greater Glasgow Health Board area it made more sense to focus on that area alone in the production of a new measurement. In addition, as the majority of the postcode sectors falling in the most deprived category of the Carstairs score are in Glasgow, it was felt that creating a score for Glasgow only would indicate how these postcode sectors differed.

This method of creating a deprivation measure differs from those others derived from Scottish data in that a continuous score is produced, rather than a number of categories, as with both Carstairs and Womersley. In a categorised scoring system, two postcodes at the maximum and minimum values for a category will be given the same score, however with a continuous score it will be easier to see how these postcodes differ.

While the Murray score is restricted to the Greater Glasgow Health Board area, the methods used in its calculation have a wider use. These methods can easily be used to create similar scores for other Health Board areas, or the whole of Scotland. At the moment, the Murray score is calculated for each postcode sector. Future work in this area may involve determining the Murray score by full postcode, or for each household, using similar methods.

8.3 Effect of Low Birthweight on Successive Pregnancies

A linked data set was provided by the Information and Statistics Division of the Scottish Health Service. This consisted of a record for each pregnancy in the Greater Glasgow Health Board area between 1980 and 1991, combining all hospital visits and antenatal visits for that pregnancy. The data were linked in that mothers could be followed throughout pregnancy, and also through successive pregnancies, by using a unique maternal identification number. This enables us to consider all hospital visits during pregnancy. In order to investigate the effect of a low birthweight delivery on subsequent pregnancies, women were identified who delivered a singleton livebirth in their first pregnancy, and later were identified to have a second pregnancy.

Infants were categorised as low birthweight or normal birthweight, and a chi-squared test on all women who did not change Womersley neighbourhood type in their first two pregnancies showed evidence of an association between birthweight in the first and second pregnancies, in particular, a low birthweight first pregnancy is more likely to be followed by a second low birthweight infant. A similar association was shown in those women who changed Womersley type between their first two pregnancies.

Modelling the birthweight of the second infant as a binary response showed that as deprivation increased, the probability of a low birthweight infant in the second pregnancy increased, and this probability increased at least five-fold if the first pregnancy was also low birthweight. This is not unexpected, as if birthweight is related to socio-economic and obstetric factors, these factors may be expected to be similar over pregnancies for a woman, and as a result the factors which result in the first child being born low birthweight will also be present in subsequent pregnancies.

As many factors that may be indicators of low birthweight are similar over pregnancies, it would be of interest to investigate those factors that do change. One such variable is smoking status. Future work may involve investigating whether women who give up smoking between their first and second pregnancies have a lower risk of delivering a subsequent low birthweight baby than those women who do not give up smoking.

8.4 Theory of Bonferroni Bounds

Calculation of both upper and lower bounds on the p-value in stepwise subset regression using a Normal regression model was found to be more useful than calculation of only an upper bound. In fitting a model to the Hald data using upper and lower bounds it was shown that a variable that would have been rejected if only the upper bound had been calculated was in fact significant. In this case the lower bound was shown to be the exact p-value as there were only two remaining explanatory variables.

This thesis has shown that the theory of upper and lower Bonferroni bounds can be extended from Normal models into the area of logistic regression modelling. A model using the upper and lower Bonferroni bounds as applied to the logistic regression case was fitted to model birthweight - specifically, whether an infant is low birthweight or not. The data are from 1991 and consist of singleton livebirths to first time mothers in the Greater Glasgow health Board area. The possible explanatories used were gestational age, maternal height, maternal age, maternal condition as described in chapter 3, marital status, and the three scores derived in chapter 4, a deprivation score, age score and house score. If the upper and lower bounds had not been considered, marital status would also have been entered into the model although age score, the next most significant variable, would not have been added.

It was found that after fitting gestational age, deprivation score and maternal height into the model none of the remaining variables were significant. While the lower bound does not affect us rejecting marital status from the model, the difference between the upper and lower bounds on the p-value for this variable is quite small, indicating that for variables close to significance the lower bound may make a difference. The model calculated suggested that as gestation and height increase, the probability of a low birthweight baby decreases, but that as deprivation increases the probability of a low birthweight baby increases.

A second model was fitted to linked data of mothers who delivered their first infant in 1980 and subsequently had a second delivery. The possible explanatories were gestational age, maternal age, height, previous birthweight, and the deprivation, age and house scores derived in chapter 4. After firring gestational age, previous birthweight and deprivation score into the model no remaining variables were significant. The model calculated suggested that as gestation and previous birthweight increase, the probability of a low birthweight baby decreases, but that as deprivation increases the probability of a low birthweight baby increases.

The theory of the upper and lower Bonferroni bounds has been extended from Normal models to logistic regression models. In addition these bounds have been applied to both Normal and logistic examples, showing that the variables selected would differ if the bounds had not been calculated. It would be interesting to investigate whether these bounds can be extended into other areas of statistics, for example survival analysis, and if this extension was possible, whether it would affect the model produced. Some other future work would be to investigate further, using simulation studies, the size of studies where the lower bound is useful and should be calculated, and the size of studies where the calculation of this bound is unlikely to add any further information.

Appendix A

Data Recorded on SMR2

The following variables are recorded for each visit during pregnancy.

1. General Information

Hospital code

Hospital case record number

Surname

Forename

Middle initial

Maiden name

Maternal age

Date of birth

Marital status

Address

Postcode

Occupation

Husband's occupation

Marriage date

Obstetrician

Family doctor

Type of antenatal care

2. Previous Pregnancies

Number of previous pregnancies

Number of previous spontaneous abortions (miscarriages)

Number of previous therapeutic abortions

Number of previous caesarean sections

Number of previous perinatal deaths

Number of children now living

3. Current Pregnancy

Date of admission

Where admitted from

Number of previous admissions this pregnancy

Type of admission

Date of booking

Original booking for delivery

Blood group

Maternal height

Type of abortion

Management of abortion

Sterilisation after abortion

Principal complication of abortion

Date of last menstrual period

Estimated gestation at abortion or delivery

Certainty of gestation

4. Maternal Discharge Data

Date of discharge

Condition on discharge

Discharged to

Category of patient

Unit on discharge

5. Record of Labour

Method of induction

Presentation at delivery

Mode of delivery, for baby 1 and baby 2

Duration of labour

Sterilisation after delivery

Date of delivery

Number of births this pregnancy

Outcome, for baby 1 and baby 2

Birthweight, for baby 1 and baby 2

Apgar score at 5 minutes, for baby 1 and baby 2

Sex, for baby 1 and baby 2

6. Postnatal Record of Infants

Special care baby unit, for baby 1 and baby 2

Baby discharged to, for baby 1 and baby 2

Case record number, for baby 1 and baby 2

Underlying cause of death or stillbirth, for baby 1 and baby 2

7. Main Condition

The main condition or complication suffered

8. Other Conditions

Any other conditions or complications

9. Operation

Any operations carried out

For multiple births, data are only recorded for the first two babies of the pregnancy. In addition, there is space on the form to record smoking history of the mother. However, this is not routinely filled in and cannot be used in any of the analyses carried out.

Appendix B

Descriptive Analysis

1981				1991			
Mat. Age	VLBW	LBW	NBW	Mat. Age	VLBW	LBW	NBW
<16 (n=15)	0.0667	0.0000	0.9333	<16(n=26)	0.0385	0.0769	0.8846
16-19 (n=1484)	0.0094	0.0842	0.9063	16-19 (n=1068)	0.0234	0.0815	0.8951
20-24 (n=4254)	0.0073	0.0609	0.9318	20-24 (n=2956)	0.0074	0.0673	0.9252
25-29 (n=3927)	0.0051	0.0456	0.9493	25-29 (n=4279)	0.0065	0.0483	0.9451
30-34 (n=2122)	0.0071	0.0419	0.9510	30-34 (n=2826)	0.0060	0.0464	0.9476
35-39 (n=621)	0.0032	0.0805	0.9163	35-39 (n=935)	0.0107	0.0578	0.9316
40-44 (n=112)	0.0268	0.1161	0.8571	40-44 (n=150)	0.0000	0.0467	0.9533
45-49 (n=5)	0.0000	0.0000	1.0000	45-49 (n=7)	0.0000	0.0000	1.0000
50-54 (n=1)	0.0000	0.0000	1.0000	50-54 (n=0)	-	-	-

Table B.1: Relationship between birthweight category and maternal age.

1981				1991			
Status	VLBW	LBW	NBW	Status	VLBW	LBW	NBW
Single (n=1381)	0.0094	0.0985	0.8921	Single (n=3438)	0.0151	0.0765	0.9084
Married (n=10782)	0.0065	0.0508	0.9427	Married (n=7822)	0.0059	0.0446	0.9495
Widowed (n=33)	0.0000	0.0606	0.9394	Widowed (n=10)	0.0000	0.2000	0.8000
Divorced (n=95)	0.0105	0.0842	0.9053	Divorced (n=123)	0.0000	0.0650	0.9350
Separated (n=129)	0.0078	0.1008	0.8915	Separated (n=108)	0.0000	0.0833	0.9167
Other (n=101)	0.0099	0.0495	0.9406	Other (n=714)	0.0070	0.0756	0.9174
Unknown (n=20)	0.0000	0.1500	0.8500	Unknown (n=32)	0.0000	0.0625	0.9375

Table B.2: Relationship between birthweight category and marital status.

1981				1991			
Prev. Pregnancies	VLBW	LBW	NBW	Prev. Pregnancies	VLBW	LBW	NBW
0 (n=4755)	0.0074	0.0616	0.9310	0 (n=4541)	0.0106	0.0639	0.9256
1 (n=3751)	0.0061	0.0483	0.9456	1 (n=3698)	0.0059	0.0468	0.9473
2 (n=2128)	0.0060	0.0449	0.9491	2 (n=2178)	0.0078	0.0436	0.9486
3 (n=1029)	0.0087	0.0554	0.9359	3 (n=1038)	0.0058	0.0626	0.9316
4 (n=456)	0.0132	0.0921	0.8947	4 (n=442)	0.0181	0.0656	0.9163
5 (n=197)	0.0000	0.1015	0.8985	5 (n=197)	0.0051	0.0812	0.9137
6 (n=77)	0.0000	0.1299	0.8701	6 (n=88)	0.0000	0.1477	0.8523
7 (n=46)	0.0000	0.1739	0.8260	7 (n=36)	0.0278	0.1111	0.8611
8 (n=43)	0.0000	0.1395	0.8605	8 (n=16)	0.0000	0.1250	0.8750
9 (n=5)	0.0000	0.0000	1.0000	9 (n=13)	0.0000	0.0000	1.0000

Table B.3: Relationship between birthweight category and parity.

1981				1991			
Hospital	VLBW	LBW	NBW	Hospital	VLBW	LBW	NBW
1 (n=3266)	0.0080	0.0643	0.9277	1 (n=3456)	0.0119	0.0648	0.9233
2 (n=2071)	0.0077	0.0594	0.9329	2 (n=1798)	0.0061	0.0412	0.9527
3 (n=2014)	0.0035	0.0541	0.9424	3 (n=2080)	0.0067	0.0606	0.9327
4 (n=110)	0.0000	0.0182	0.9818	4 (n=0)	-	-	-
5 (n=2162)	0.0069	0.0611	0.9320	5 (n=1976)	0.0056	0.0612	0.9332
6 (n=2918)	0.0075	0.0476	0.9448	6 (n=2936)	0.0089	0.0484	0.9428

Table B.4: Relationship between birthweight category and hospital.

1981				1991			
Abortions	VLBW	LBW	NBW	Abortions	VLBW	LBW	NBW
0 (n=11957)	0.0066	0.0570	0.9364	0 (n=11215)	0.0077	0.0556	0.9368
1 (n=543)	0.0092	0.0589	0.9319	1 (n=941)	0.0149	0.0616	0.9235
2 (n=39)	0.0513	0.0513	0.8974	2 (n=86)	0.0349	0.0581	0.9070
3 (n=2)	0.0000	0.0000	1.0000	3 (n=4)	0.0000	0.2500	0.7500
4 (n=0)	-	-	-	4 (n=1)	0.0000	0.0000	1.0000

Table B.5: Relationship between birthweight category and previous induced abortions.

	1981				1991		
Miscarriages	VLBW	LBW	NBW	Miscarriages	VLBW	LBW	NBW
0 (n=10571)	0.0063	0.0538	0.9398	0 (n=10053)	0.0084	0.0534	0.9382
1 (n=1529)	0.0092	0.0667	0.9241	1 (n=1680)	0.0071	0.0655	0.9274
2 (n=320)	0.0125	0.0844	0.9031	2 (n=390)	0.0103	0.0795	0.9103
3 (n=79)	0.0000	0.1519	0.8481	3 (n=84)	0.0357	0.0833	0.8810
4 (n=23)	0.0435	0.0870	0.8696	4 (n=28)	0.0000	0.0714	0.9286
5 (n=11)	0.0000	0.1818	0.8182	5 (n=8)	0.0000	0.0000	1.0000
6 (n=3)	0.0000	0.0000	1.0000	6 (n=4)	0.0000	0.0000	1.0000
7 (n=3)	0.0000	0.0000	1.0000	7 (n=0)	-	-	-
8 (n=2)	0.0000	0.5000	0.5000	8 (n=0)	-	-	-

Table B.6: Relationship between birthweight category and previous spontaneous abortions.

Bibliography

[1] D. G. Altman and P. K. Andersen. Bootstrap Investigation of the Stability of a Cox Regression Model. *Statistics in Medicine*, 8:771–783, 1989.

[2] L. S. Bakketeig, T. Bjerkedal, and H. J. Hoffman. Small-for-Gestational Age Births in Successive Pregnancy Outcomes: Results from a Longitudinal Births in Norway. *Early Human Development*, 14:187–200, 1986.

[3] Y. Ben-Shlomo, I. White, and P. M. McKeigue. Prediction of General Practice Workload from Census Based Social Deprivation Scores. *Journal of Epidemiology and Community Health*, 46(5):532–536, 1992.

[4] E. Bølviken. Some Probability Bounds Relevant for Bonferroni Significance Levels Under Gaussian Models. *Scandinavian Journal of Statistics*, 15:281–297, 1988.

[5] S.L. Bratton, D.L. Shoultz, and M.A. Williams. Recurrence Risk of Low Birthweight Deliveries Among Women with a Prior Very Low Birthweight Delivery. *American Journal of Perinatology*, 13(3):147–150, 1996.

[6] P. M. T. Broersen. Subset Regression with Stepwise Directed Search. *Applied Statistics*, 35(2):168–177, 1986.

[7] R. W. Butler. Bounds on the Significance Attained by the Best-Fitting Regressor Variable. *Applied Statistics*, 31(3):290–292, 1982.

[8] R. W. Butler. The Significance Attained by the Best-Fitting Regressor Variable. *Journal of the American Statistical Association*, 79:341–348, 1984.

[9] D. A. Campbell, J. M. C. Radford, and P. Burton. Unemployment Rates : An Alternative to the Jarman Index? *British Medical Journal*, 303(6805):750–755, 1991.

[10] V. Carstairs and R. Morris. *Deprivation and Health in Scotland.* Aberdeen University Press, Aberdeen, UK, 1991.

[11] C-H. Chen and S. L. George. The Bootstrap and Identification of Prognostic Factors via Cox's Proportional Hazards Regression Model. *Statistics in Medicine*, 4:39–46, 1985.

[12] S. Cole. Scottish Maternity and Neonatal Records. In I. Chalmers and G. McIlwaine, editors, *Perinatal Audit and Surveillance - Proceedings of the Eigth Study Group of the Royal College of Obstetricians and Gynaecologists*, pages 39–51. Royal College of Obstetricians and Gynaecologists, London, 1980.

[13] A. J. Cotgrove, G. Bell, and C. L. E. Katona. Psychiatric Admissions and Social Deprivation : Is the Jarman Underprivileged Area Score Relevant? *Journal of Epidemiology and Community Health*, 46(3):245–247, 1992.

[14] S. de Sanjose and E. Roman. Low Birthweight, Preterm, and Small For Gestational Age Babies in Scotland, 1981-1984. *Journal of Epidemiology and Community Health*, 45:207–210, 1991.

[15] G. Diehr and D. R. Hoflin. Approximating the Distribution of the Sample R^2 in Best Subset Regression. *Technometrics*, 16(2):317–320, 1974.

[16] N. R. Draper and H. Smith. *Applied Regression Analysis.* Wiley and Sons, New York, 1966.

[17] M. A. Efroymson. Multiple Regression Analysis. In A. Ralston and H. S. Wolf, editors, *Mathematical Methods for Digital Computers*, pages 191–203. Wiley, New York, 1962.

[18] G.A. Ellard, F.D. Johnstone, R.J. Prescott, W. Ji-Xian, and M. Jian-Hua. Smoking During Pregnancy : The Dose Dependence of Birthweight Deficits. *British Journal of Obstetrics and Gynaecology*, 103:806–813, 1996.

[19] B. Eskenazi, A. W. Prehn, and R. E. Christianson. Passive and Active Maternal Smoking as Measured by Serum Cotinine: The Effect on Birthweight. *American Journal of Public Health*, 85(3):395–398, 1995.

[20] R. L. Goldenberg, H. J. Hoffman, S. P. Cliver, G. R. Cutter, K. G. Nelson, and R. L. Copper. The Influence of Previous Low Birth Weight on Birth Weight, Gestational Age, and Anthropomorphic Measurements in the Current Pregnancy. *Obstetrics and Gynecology*, 79(2):276–280, 1992.

[21] The Scottish Low Birthweight Study Group. The Scottish Low Birthweight Study: I. Survival, Growth, Neuromotor and Sensory Impairment. *Archives of Disease in Childhood*, 67:675–681, 1992.

[22] The Scottish Low Birthweight Study Group. The Scottish Low Birthweight Study: II. Language Attainment, Cognitive Status, and Behavioural Problems. *Archives of Disease in Childhood*, 67:682–686, 1992.

[23] A. Hald. *Statistical Theory with Engineering Applications*. Wiley, New York, 1952.

[24] A. Hall, A. McLeod, C. Counsell, L. Thompson, and L. Mutch. School Attainment, Cognitive Ability and Motor Function in a Total Scottish Very-Low-Birthweight Population at Eight Years: A Controlled Study. *Developmental Medicine and Child Neurology*, 37:1037–1050, 1995.

[25] M. J. R. Healy. *Matrices for Statistics*. Oxford University Press, Oxford, 1986.

[26] B. Jarman. Identification of Underprivileged Areas. *British Medical Journal*, 286(6379):1705–1709, 1983.

[27] B. Jarman. Underprivileged Areas : Validation and Distribution of Scores. *British Medical Journal*, 289(6458):1587–1592, 1984.

[28] S. Kendrick and J. Clarke. The Scottish Record Linkage System. *Scottish Health Bulletin*, 51(2):72–79, 1993.

[29] W. H. Kitchen, M. J. Permezel, L. W. Doyle, G. W. Ford, A. L. Rickards, and E. A. Kelly. Changing Obstetric Practice and 2-Year Outcome of the Fetus of Birth Weight Under 1000 g. *Obstetrics and Gynecology*, 79(2):268–275, 1992.

[30] L. A. A. Kollee, R. Brand, A. Am Schreuder, M. H. Ens-Dokkum, S. Veen, and S. P. Verloove-Vanhorick. Five-Year Outcome of Preterm and Very Low Birth Weight Infants: A Comparison Between Maternal and Neonatal Transport. *Obstetrics and Gynecology*, 80(4):635–638, 1992.

[31] J. Lang, E. Lieberman, and A. Cohen. A Comparison of Risk Factors for Preterm Labor and Term Small-for-Gestational-Age Birth. *Epidemiology*, 7(4):369–376, 1996.

[32] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, New York, 1983.

[33] P. McLoone and F. A. Boddy. Deprivation and Mortality in Scotland, 1981 and 1991. *British Medical Journal*, 309(6967):1465–1470, 1994.

[34] A. J. Miller. *Subset Selection in Regression*. Chapman and Hall, London, 1990.

[35] R. Pickering. *Analysis of Categorical Data on Pregnancy Outcome*. PhD thesis, Department of Statistics, Glasgow University, 1987.

[36] A. Powls, N. Botting, R. Cooke, D. Pilling, and N. Marlow. Growth Impairment in Very Low Birthweight Children at 12 Years : Correlation With Perinatal and Outcome Variables. *Archives of Disease in Childhood*, 75(3):F152–F157, 1996.

[37] P. Rasheed and J. Rahman. Predictors of Saudi Birthweights: A Multiple Regression Analysis. *Saudi Medical Journal*, 16(1):23–29, 1995.

[38] J. O. Rawlings. *Applied Regression Analysis: A Research Tool*. Wadsworth and Brooks/Cole, Pacific Grove, California, 1988.

[39] R. Reading, S. Jarvis, and S. Openshaw. Measurement of Social Inequalities in Health and Use of Health Services Among Children in Northumberland. *Archives of Disease in Childhood*, 68(5):626–631, 1993.

[40] A. C. Rencher and F. C. Pun. Inflation of R^2 in Best Subset Regression. *Technometrics*, 22(1):49–53, 1980.

[41] C. Rodriquez, E. Regidor, and J. L. Gutierrez-Fisac. Low Birth Weight in Spain Associated with Sociodemographic Factors. *Journal of Epidemiology and Community Health*, 49:38–42, 1995.

[42] K. Rosenberg and H. P. McEwan. Teenage Pregnancy in Scotland : Trends and Risks. *Scottish Medical Journal*, 36:172–174, 1991.

[43] G. Davey Smith. Second Thoughts on the Jarman Index. *British Medical Journal*, 302(6773):359–360, 1991.

[44] R. J. Talbot. Underprivileged Areas and Health Care Planning : Implications of Use of Jarman Indicators of Urban Deprivation. *British Medical Journal*, 302(6773):383–386, 1991.

[45] P. Townsend, P. Phillimore, and A. Beattie. *Health and Deprivation*. Croom Helm Ltd, Kent, UK, 1988.

[46] H. Wessel, S. Cnattingius, S. Bergstrom, A. Dupret, and P. Reitmainer. Maternal Risk Factors for Preterm Birth and Low Birthweight in Cape Verde. *Acta Obstetricia et Gynecologica Scandinavica*, 75(4):360–366, 1996.

[47] M. A. Wilcox, S. J. Smith, I. R. Johnson, P. V. Maynard, and C. E. D. Chilvers. The Effect of Social Deprivation on Birthweight, Excluding Physiological and Pathological Effects. *British Journal of Obstetrics and Gynaecology*, 102:918–924, 1995.

[48] K. Wisborg, T.B. Henriksen, M. Hedegaard, and N. J. Secher. Smoking During Pregnancy and Preterm Birth. *British Journal of Obstetrics and Gynaecology*, 103:800–805, 1996.

[49] J. J. Womersley. Neighbourhood Types. Personal correspondence, 1996.