

**THE RELATIONSHIP BETWEEN MODIFIABLE HEALTH RISK
FACTORS AND HEALTH CARE COSTS FOR INDIVIDUALS WHO
HAVE COMPLETED A HEALTH RISK ASSESSMENT
QUESTIONNAIRE FOR A SOUTH AFRICAN HEALTH INSURANCE
SCHEME**

Dinesh Hari Lall Garbharran

A research report submitted to the Faculty of Science, University of the
Witwatersrand, in partial fulfilment of the requirements for the degree of Master
of Science

Johannesburg, 2011

DECLARATION

I declare that this research report is my own, unaided work. It is being submitted for the Degree of Master of Science in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other University.

26th day of August 2011

ABSTRACT

Health care funders are seeking managed health care interventions to contain medical inflation. The purpose of this study is to assess the relationship between three health risk factors (smoking status, physical activity and body mass index (BMI)) and inpatient costs among health risk assessment (HRA) respondents at a South African health insurer. The results could inform the design of wellness programmes by ensuring that appropriate health risk factors are being targeted to reduce inpatient costs. This study utilises a two-part regression model to explore the relationships between the health risk factors and inpatient costs. The combined results of the two-part regression model indicate that increasing levels of physical activity and decreasing levels of BMI are associated with lower likelihoods of hospitalisation and lower magnitudes of inpatient costs for those that had a non-zero claim. The results of this study indicate no association between smoking cessation and lower inpatient costs.

CONTENTS	Page
DECLARATION.....	ii
ABSTRACT.....	iii
LIST OF FIGURES.....	vi
LIST OF TABLES.....	vii
CHAPTER 1: INTRODUCTION.....	1
1.1 Managed Health Care and Wellness Programmes.....	1
1.2 The Relationship between Wellness Programmes and Modifiable Health Risk Factors.....	2
1.3 The Relationship between Modifiable Health Risk Factors and Health Care Costs.....	3
1.4 The Vitality Wellness Programme.....	4
1.5 The Health Risk Appraisal (HRA).....	4
1.6 Limitations of the Data.....	5
1.7 Problem Statement.....	6
1.8 Causal Relationships.....	7
1.9 Objective.....	7
CHAPTER 2: LITERATURE REVIEW.....	8
2.1 Literature Review: Modelling of Health Care Cost Data.....	8
2.2 Literature Review: The Relationship between Modifiable Health Risk Factors and Health Care Costs.....	11
2.3 Literature Review: The Relationship between Smoking Status and Health Care Costs.....	13
2.4 Literature Review: The Relationship between Obesity and Health Care Costs.....	14
2.5 Literature Review: The Relationship between Physical Activity and Health Care Costs..	15
CHAPTER 3: STATISTICAL THEORY.....	16
3.1 Regression Methods.....	16
3.2 Generalised Linear Models.....	17
3.3 A Two-Part Regression Model.....	19
CHAPTER 4: DATA SOURCE AND STUDY DESIGN.....	22
4.1 Data Source and Sample.....	22
4.2 Design.....	23
4.3 The Health Risk Appraisal (HRA) Questionnaire.....	23
4.4 Inpatient health care costs.....	24
CHAPTER 5: DATA DESCRIPTION.....	26
5.1 Variables.....	26
5.1.1 Age.....	26
5.1.2 Gender.....	28
5.1.3 Health Insurance Plan Choice.....	30
5.1.4 Chronic Status.....	32
5.1.5 Physical Activity.....	34
5.1.6 Smoking Status.....	36
5.1.7 Body Mass Index.....	38
CHAPTER 6: STATISTICAL METHODS.....	41
6.1 Descriptive statistics and simple inference.....	41
6.2 A Two-Part Regression Model.....	41
6.2.1 Part One: Logistic regression.....	42
6.2.2 Part Two: Linear regression.....	48

CHAPTER 7: RESULTS.....	53
7.1 Results: Logistic regression.....	54
7.1.1 Parameter estimates.....	54
7.1.2 Odds ratios in respect of the modifiable health risk factors.....	56
7.2 Results: Generalised linear model.....	64
CHAPTER 8: CONCLUSION.....	66
CHAPTER 9: POSSIBLE EXTENSIONS TO THIS STUDY.....	70
REFERENCES.....	72
APPENDIX.....	79

LIST OF FIGURES	Page
Figure 5.1: Plot of the adjusted mean annual inpatient cost per life per age category....	28
Figure 5.2: Plot of the adjusted mean annual inpatient cost per life per gender category	30
Figure 5.3: Plot of the adjusted mean annual inpatient cost per life per health plan category.....	32
Figure 5.4: Plot of the adjusted mean annual inpatient cost per life per chronic condition category.....	33
Figure 5.5: Plot of the adjusted mean annual inpatient cost per life per physical activity category.....	36
Figure 5.6: Plot of the adjusted mean annual inpatient cost per life per smoking status...	38
Figure 5.7: Plot of the adjusted mean annual inpatient cost per life per BMI category....	40
Figure 6.1: The Receiver Operating Characteristic (ROC) Curve for the estimated logistic regression model.....	45
Figure 6.2: Influence diagnostic plots.....	47
Figure 6.3: Plot of the studentised residuals against the predicted values.....	50

LIST OF TABLES

Page

Table 4.1:	Summary statistics for inpatient health care utilisation and inpatient health care costs (ZAR).....	25
Table 5.1:	Number of respondents and adjusted mean inpatient health care cost per age band.....	27
Table 5.2:	Number of respondents and adjusted mean inpatient health care cost for Males and Females.....	29
Table 5.3:	Number of respondents and adjusted mean inpatient health care cost per plan type category.....	31
Table 5.4:	Number of respondents and adjusted mean inpatient health care cost per chronic status category.....	33
Table 5.5:	Number of respondents and adjusted mean inpatient health care cost per physical activity category.....	35
Table 5.6:	Number of respondents and adjusted mean inpatient health care cost per smoking status.....	37
Table 5.7:	Number of respondents and adjusted mean inpatient health care cost per BMI category.....	39
Table 7.1:	Parameter estimates from the estimated logistic regression model (Part One of the two-part regression model).....	55
Table 7.2:	Odds ratios for significant 2-way interactions with smoking status variable..	58
Table 7.3:	Odds ratios for significant 2-way interactions with the physical activity variable.....	61
Table 7.4:	Odds ratios for significant 2-way interactions with the BMI variable.....	63
Table 7.5:	Parameter estimates from the estimated ordinary least squares regression model (Part Two of the two-part regression model).....	65

CHAPTER 1: INTRODUCTION

1.1 Managed Health Care and Wellness Programmes

In an environment of increasing health care costs, health care funders around the world are expanding and refining their managed health care interventions in an effort to contain medical inflation. The term *managed health care* (MHC) is used to describe various structures and strategies aimed at improving the performance of the health care system. These range from alternative health delivery systems, to financial incentives, to detailed protocols and guidelines governing clinical behaviour (Hornbrook, Goodman 1991).

An MHC intervention that is gaining popularity in South Africa is wellness programmes (Still 2007). A wellness programme is defined as a structured intervention focused on achieving wellness in the physical, psychological or spiritual realm (Watt *et al.* 1998). There was an exponential growth in the number of health promotion or wellness programmes in the United States between 1975 and 1990 with research showing a causal relationship between a reduction in health care costs and the implementation of a comprehensive wellness programme (Gebhardt, Crump 1990). These wellness programmes promoted healthier lifestyles by encouraging behavioural change in respect of certain health risk factors.

1.2 The Relationship between Wellness Programmes and Modifiable Health Risk Factors

The South African system of health care delivery and funding is reactive in nature and encourages people to assume the role of passive recipients of health care as directed by health care providers. This does little to modify health-related behaviours. A few large health care insurers in South Africa have shifted the paradigm to a more holistic management of health care costs through MHC interventions. This has seen the introduction of wellness programmes to the South African private health care insurance sector. In addition to health education, wellness programmes promote healthier lifestyles, healthier dietary habits and regular physical activity (Chapman 1998).

One such example of an incentive-based wellness programme that was introduced by a South African health insurance company is the Vitality programme. Research has shown that engagement in the Vitality programme was associated with lower health care costs (Patel *et al.* 2010). The study showed that inpatient service costs were 18.6% higher for lowly engaged members when compared to highly engaged members.

The positive correlation between healthy lifestyles and good health has been modelled by economists over an extensive period of time; beginning with Grossman's pioneering work (Cawley 2004; Fuchs 1974; Grossman 1972; Kenkel 1995).

There have been various studies that have concluded that the modification of behaviour in respect of health risk factors leads to a reduced demand for health care services, and hence, results in lower health care costs and improved health outcomes. For example, healthy dietary habits have been clearly linked with a decreased risk of type 2 diabetes and related health problems (Knowler *et al.* 2002). Inadequate physical activity has shown a strong correlation with cardiovascular disease (Kohl 2001), depression (Galper *et al.* 2006) and high medical costs (Pratt *et al.* 2000). In a 2003 study, Serxner *et al.* concluded that participants ($n = 13,048$) in a wellness programme, cost on average USD212 less per annum, than eligible non-participants ($n = 13,363$).

1.3 The Relationship between Modifiable Health Risk Factors and Health Care Costs

Researchers have found that higher medical costs are attributable, to a significant extent, to modifiable health risk factors (Brink 1987; Goetzel *et al.* 1998; Thorpe 2005).

There have been numerous modifiable health risk factors associated with increased health care costs and health outcomes (Wilkerson *et al.* 2008). These include obesity, sedentary lifestyles, stress, poor dietary habits, smoking and excessive alcohol consumption. The modifiable health care factors that are of interest in this study are smoking, sedentary lifestyles and poor dietary habits.

1.4 The Vitality Wellness Programme

A major national health care insurer in South Africa, Discovery Health, offers an incentive-based wellness programme, *Vitality*, as one of its MHC strategies. *Vitality* aims to empower its participants to improve their health by giving them the knowledge, tools and motivation to set and meet their health goals. *Vitality* offers incentives for increased engagement in the form of discounted rates with lifestyle, retail and leisure partners. There are currently 1.3 million¹ beneficiaries registered for *Vitality* (as at 30/11/2010). Through its *Vitality* programme, Discovery Health incentivises its members to complete voluntary and self-reported Health Risk Appraisals on an annual basis.

1.5 The Health Risk Appraisal (HRA)

The Health Risk Appraisal (HRA) is a questionnaire that was originally designed by the Centres for Disease Control and Prevention/Carter Centre. It includes questions on health-related behaviours, psychological risks, biometric measures, and personal and family medical history (Wang *et al.* 2006). In general, a health risk appraisal questionnaire may be designed to create awareness or for purposes of risk assessment. The *Vitality* health risk appraisal was originally designed by the Discovery Health Medical Scheme, in conjunction with academic consultants.

¹ **Source:** Discovery Vitality Management Information Report, November 2010

1.6 Limitations of the Data

Self-reported and voluntary data is a limitation of this analysis. The measurements of BMI, smoking status and physical activity in this study are self-reported. The literature indicates the inaccuracies of self-reported height and weight, in particular the tendency of respondents to overstate their height and understate their weight (Durden *et al.* 2008; Palta *et al.* 1982; Rowland 1990).

Social desirability bias is a term used in scientific research to describe the tendency of respondents to reply in a manner that will be viewed favourably by others. This will generally take the form of over-reporting good behaviour or under-reporting bad behaviour. Social desirability is a potential source of bias in the self-reporting of physical activity. Adams *et al.* concluded that social desirability was associated with over-reporting of physical activity (Adams *et al.* 2005).

This potential bias suggests that caution should be exercised in generalising results. The reporting bias may result in the estimates of the effects of modifiable health risk factors on health care costs being conservative. This potential reporting bias would result in the estimated models predicting lower health care costs at higher levels of BMI than is the case in reality.

In a 1994 study, Patrick *et al.* concluded that self-reported smoking status was accurate in most studies. They did, however, recommend that smoking status should be confirmed with biochemical assessments in order to improve accuracy. The smoking statuses utilised in this analysis are self-reported and have not been confirmed by any biochemical assessments. Hence, the accuracy may be questionable and it is prudent to report it as a limitation of this study.

This study does not distinguish between health care costs that were incurred in respect of elective and non-elective hospitalisations. Instead, it considers all inpatient health care costs that were incurred by HRA respondents in 2009. It is recommended that a future study be considered in which the basis of the study is limited to emergency hospitalisations. This may more accurately explain the association between inpatient health care costs and modifiable health care costs as it will, for example, exclude costs incurred in respect of elective inpatient dental procedures and childbirth.

1.7 Problem Statement

An interesting research question is which health risk factors should wellness programmes, like *Vitality*, focus their efforts on in order to potentially reduce inpatient health care costs? Health risk modification interventions are typically costly and resource-intensive (Pronk *et al.* 1999) so this kind of insight would be invaluable in developing an efficient and cost-effective health risk factor modification strategy.

1.8 Causal Relationships

Although causal relationships between health status, health care costs and health outcomes are fairly intuitive, the “health” construct is complex and is a function of numerous physiological, psychological, behavioural, social and economic factors. This complexity results in any extrapolation of the results of other studies being less credible and, therefore, a study based on the Vitality dataset is warranted.

1.9 Objective

The purpose of this study is to assess the relationship between modifiable health risk factors, such as smoking, obesity and exercise, on the likelihood and magnitude of inpatient health care costs.

A review of the literature indicates that up to now, there has been one similar study in South Africa. In 2009, Lambert *et al.* considered the impact of a single modifiable health risk factor on health care costs. That study focused on fitness-related activities measured by gym visits. This study extends the literature by considering the effect of multiple modifiable health risk factors (namely body mass index (as an indicator of nutritional habits), smoking status and physical activity) on annual inpatient health care costs.

CHAPTER 2: LITERATURE REVIEW

This Chapter provides a review of the literature in respect of studies that have assessed the relationships between smoking status, obesity and/or exercise-related activities, and health care costs and the statistical methods that have been used to assess these relationships. It also presents some of the empirical challenges that these studies faced and which statistical methods were appropriate to overcome these.

2.1 Modelling of Health Care Cost Data

There are a number of typical characteristics inherent to health care cost data that need to be addressed in order to develop a valid model.

Health care cost data is challenging to work with as the cumulative distribution typically has a ‘spike’ at zero (Blough *et al.* 1999; Pronk *et al.* 1999). This ‘spike’ corresponds to individuals who have not incurred health care costs. The health care cost data for individuals who have incurred health care costs also poses a challenge (Blough *et al.* 1999). The distribution is highly skewed to the right with non-constant variance. There tends to be a larger variance amongst health care costs for individuals when costs are high than when they are low.

Distributions with these ‘spikes’ at one or more points are called mixed distributions. The literature describes numerous methods to model health care cost data. One method is to use ordinary least squares and to ignore the mixed

character of the underlying distribution of health care costs (Blough, Ramsey 2000).

A second approach that caters for health care costs being extremely right-skewed and heteroscedastic (variance increases with mean), is to exploit the fact that the likelihood divides the model into two ‘parts’ (Wang *et al.* 2006).

The two-part regression model involves two stages of estimation. The first ‘part’ of such a model is binary, either an individual has incurred health care costs or he/she has not. Typically, in the first stage, a probability model, such as probit or logit, will be used to estimate the relationship between health risk factors and the likelihood of incurring health care costs.

The second ‘part’ of the model is concerned only with those individuals who have incurred health care costs. In the second stage, a generalised linear model (ordinary regression) is used to estimate the model parameters in order to explore the relationship between modifiable health risk factors and the magnitude of health care costs.

Alternatively, given the non-constant variance that is often encountered in the second stage, it may be appropriate to apply some sort of variance stabilising transformation (typically a lognormal transformation) and then perform ordinary regression on the transformed data (Wilkerson *et al.* 2008). However, if the sample size is large enough and the research objective is to explore the net effect

of covariates on health care costs, the original scale can be used in the regression model and the coefficient is a more accurate estimate of the true value than that of the transformation (Buntin & Zaslavsky 2004; Diehr *et al.* 1999).

In order to obtain an overall estimate of health care costs for an individual, one need only multiply the two estimates calculated by each part of the two-part model.

The literature shows that a two-part regression model has been used to explore the association between modifiable health risk factors and health care costs (Duan *et al.* 1983; Lin 2008). An advantage of using a two-part model is that it provides insight into the health care utilisation process, and it deals with two common problems related to health care claims data – a large number of non-claimants and a right-skewed distribution among claimants (Jones 2000; Manning 1998; Mullahy 1998).

The literature shows that generalised linear models are appropriate for the analysis of health care cost data because they provide parametric analysis methods where non-normal distributions may be specified and the link with the covariates may be altered (Barber, Thompson 2004). Unlike the use of data transformation in ordinary least-squares regression, generalised linear models used in the analysis of health care cost data make direct inferences about the population mean cost possible, resulting in more reliable health economic decision-making.

2.2 The Relationship between Modifiable Health Risk Factors and Health Care Costs

Goetzel *et al.* estimated the relationship between ten modifiable health risk factors and health care expenditures and found that seven of the ten were significantly related to an increased likelihood of incurring health care expenditures and to an increased magnitude of those expenditures (Goetzel *et al.* 1998). In this study, a two-part regression model was used. The first part included a logistic model that estimated the likelihood of incurring any health care cost after controlling for demographic and risk factors. The second part of the two-part model included an ordinary least squares regression equation that estimated the magnitude of health care expenditures for individuals that had incurred a cost during the study period. The expenditure response variable used in the second part was log-transformed to satisfy regression assumptions and all outlier cases of the expenditure variable were included in the analysis. The study showed that the likelihood of incurring a health care expenditure was significantly higher for individuals at high risk for depression, having stress, high blood glucose levels, obesity or serious underweight, high cholesterol and lack of exercise. The likelihood of incurring a health care expenditure was found to be unrelated to smoking status, alcohol consumption and poor dietary habits. Furthermore, the study showed that the magnitude of health care expenditure, conditional on an expenditure having been incurred, was significantly higher for individuals at high risk for depression, having stress, having high blood glucose levels, obesity or serious underweight, being a current or former smoker, lack of exercise and high blood pressure.

Lin conducted a study to examine the relationship between modifiable health risk factors and health care costs in Taiwan (Lin 2008). The health risk factors that were considered were alcohol consumption, smoking status, obesity, diet, exercise, healthy lifestyle and betel nut chewing. The covariates in this study were gender, marital status, age, education, income and paternal race. A two-part regression model was used in this study. The first part used a logistic regression model and the second part used an ordinary least squares regression estimation with a log transform of inpatient costs. This was necessary to overcome heteroscedasticity issues. The study concluded that former smokers and people, who followed an exercise regime, are less likely to incur inpatient expenses and also, incur lower inpatient expenses.

Pronk *et al.* explored the relationships between three modifiable health risk factors (physical inactivity, smoking status and obesity) and short-term health care charges (Pronk *et al.* 1999). Again, a two-part regression was used with the first part estimating the probability of utilisation of health care services and the second part estimating the expected level of health care utilisation, conditional on positive utilisation. The first part estimated a logistic model and the second part estimated a semilogarithmic model. In this study, collinearity was assessed using variance inflation factors. In the logistic regression, overall model fit was assessed using the Hosmer-Lemeshow test and the performance of the individual coefficients was assessed using a test. The median annual health care charges for the total study group was USD600. The study concluded that a unit increase in BMI resulted in a 1.9% (USD11.26) increase in median health care charges ($p < 0.01$);

an additional day of physical activity yields a 4.7% (USD27.99) reduction in median health care costs ($p < 0.01$); and current smokers incur 18% (USD107.81) higher health care costs than non-smokers ($p < 0.01$) while former smokers incur 26% (USD154.86) higher costs than 'never-smokers' ($p < 0.01$).

These three studies all use two-part regression models to explore the relationships between modifiable health risk factors and health care charges. Their findings show various correlations between the health risk factors and health care charges.

2.3 The Relationship between Smoking Status and Health Care Costs

In 2003, Fishman *et al.* estimated the long-term health care costs of former smokers compared with continuing and never smokers using a retrospective cohort study of enrollees of a health insurance plan. The methods used in this study needed to overcome the empirical challenge caused by the skewed distribution of health care costs. To address this, health care costs over time were estimated using a Generalised Linear Model multivariate regression. The regression residuals were modelled based on the gamma distribution as this had previously been demonstrated to be an appropriate distribution for health care costs. This study concluded that smoking cessation does not decrease long-term health care costs.

A study of National Health insurance beneficiaries in Japan (Izumi *et al.* 2001) showed that male smokers incurred 11% more health care costs (after adjusting for age, physical functioning status, alcohol consumption, BMI and average time

spent walking) than ‘never smokers’ but for female smokers and ‘never smokers’ the costs were almost equal.

2.4 The Relationship between Obesity and Health Care Costs

In a 2006 study, Wang *et al.* showed that for the BMI range 25 to 45 kg/m², medical costs and pharmaceutical costs increased by USD119.70 (4%) and by USD82.60 (7%) per unit increase in BMI, respectively, adjusted for age and gender. Due to the large sample size that was available for this study, the original scale of the cost data was used as opposed to using a log transformation. Adjusted means of medical costs were obtained using the PROC GLM procedure of the SAS software program, with BMI as a categorical variable controlling for age and gender.

In a study conducted by Durden *et al.*, overall adjusted direct and indirect health care costs were shown to be higher for workers with higher BMI measures relative to those of normal weight. Incremental direct health care costs associated with being overweight, obese and severely obese were estimated to be USD147.11, USD712.34 and USD1,977.43, respectively (Durden *et al.* 2008). This study assessed the differences in health care costs across the BMI groups using a two-part regression analysis consisting of a logistic regression model that estimated the likelihood of incurring a health care cost and a log-gamma generalised linear model that estimated the amount of the health care cost. Both models adjusted for age, age-squared, sex, geographic region, salary, union status, industry type and health plan type.

2.5 The Relationship between Physical Activity and Health Care Costs

Lambert *et al.* explored the effect of participation in fitness-related activities on medical claims and hospital admissions (Lambert *et al.* 2009). The population for this study was members of a major South African health insurer. The study used a multivariate analysis of covariance, with the Tukey-Kramer *t* test to determine if there were significant differences in costs among the fitness activity categories. The study adjusted for age, sex, health plan option and chronic illness benefits. The study concluded that participation in fitness-related activities resulted in lower health care costs. Highly active individuals incurred ZAR1,535 less in annual health care costs than inactive individuals.

As this review of the literature indicates, there have been many studies that have been conducted to assess the relationship between health care costs and a variety of modifiable health risk factors. The literature shows that the results obtained from such studies are not always consistent or intuitive. This, once again, confirms the complexity that is inherent to the study of the relationships between medical costs and modifiable health risk factors. This research is an attempt to determine the nature of the relationships that exist between modifiable health risk factors and annual inpatient health care costs for a population of HRA respondents at a South African health insurance funder.

CHAPTER 3: STATISTICAL THEORY

3.1 Regression Methods

The literature shows that regression methods are an important component of any study associated with the analysis of health care cost data. The choice of regression method for health care cost data is not simple (Barber, Thompson 2004). The assumptions that are necessary for ordinary least-squares linear regression to be reliable for health cost data are unlikely to be met. Health care costs are usually non-normal and heteroscedastic (i.e. not of constant variance) and the relationships may not be linear (Briggs, Gray 1998). In these instances, the literature suggests that the use of a two-part regression model is appropriate with a logistic regression model in the first part and a log-gamma generalised linear model in the second part.

In instances where the health care cost variable is normal and homoscedastic, the literature again indicates that the use of a two-part regression model is appropriate with a logistic regression model in the first part and an ordinary least squares regression in the second part.

In this study, the log of the annual inpatient health care cost variable is found to normally distributed and homoscedastic. Therefore, a two-part regression model is used. The first part uses a logistic regression model to estimate the likelihood of incurring a health care cost. The second part uses an ordinary least squares

general linear model with a log transformation to estimate the magnitude of health care costs incurred.

The next section presents the statistical theory of generalised linear models as compared to the theory of general linear models. It also presents the statistical theory of two-part regression models.

3.2 Generalised Linear Models

Generalised linear models are a generalisation of the general linear model that is typically described in standard statistics textbooks (Hill, Lewicki 2007).

In its simplest form, a linear model specifies the (linear) relationship between a dependent (or response) variable Y , and a set of predictor (or explanatory) variables, the X s such that

$$Y = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k + e \quad (3.1)$$

where b_0 is the regression coefficient for the intercept, the b_i values are the regression coefficients (for the variables 1 to k) computed from the data and e denotes the error variability that remains unexplained by the predictor variables (McCullagh, Nelder 1983; McCullagh 1984; Hill, Lewicki 2007).

However, there are many instances where the relationship cannot be adequately described by a simple linear equation. This is attributable to two major reasons. The first reason is that the dependent variable Y may have a non-continuous distribution and therefore, the predicted values should also follow the same non-

continuous distribution. This is not always mathematically possible. The second reason that a general linear model may be inadequate to describe a relationship is that the effect of the predictor variables on the dependent variable may be non-linear. One likely example of this is the relationship between age and health status. It is conceivable that increments in age at older ages may be associated with greater deterioration in health status, as compared to increments in age in childhood or early adulthood. The generalised linear model overcomes these shortcomings of the general linear model. Generalised linear models can be used to describe relationships where the dependent variable follows a discrete distribution and in instances where the relationship (link) between the dependent variable and the predictor variables is non-linear. In summary, the use of generalised linear models permits the distribution of the dependent or response variable to be (explicitly) non-normal and non-continuous; secondly, the dependent variable values are estimated from a linear combination of predictor variables, which may be related to the dependent variable via a link function.

In the general linear model, the dependent variable Y is linearly associated with the predictor variables X by (1) above. In the generalised linear model, the relationship between the dependent variable and the predictor variables is assumed to be

$$Y = g(b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k) + e \quad (3.2)$$

where e is the error, and $g(\dots)$ is a function. Formally, the inverse function of $g(\dots)$, say $f(\dots)$, is called the link function; so that

$$f(\mu_y) = b_0 + b_1X_1 + b_2X_2 + \dots + b_kX_k \quad (3.3)$$

where μ_y stands for the expected value of y .

3.3 A Two-Part Regression Model

This section presents the theory with respect to each part of the two-part regression model and explains the two-part model predicts the annual inpatient health care cost for an individual by multiplying the product of the likelihood of an individual incurring a cost by the estimate of the magnitude of the health cost (conditional on an individual incurring a cost).

Let Y = annual inpatient health care costs for an individual.

Let X be a vector of covariates.

Let θ be a vector of parameters.

Given $X = x$, let $I =$

Then

$$f_Y(y; \theta | x) = \quad (3.4)$$

Given n observations $(I_1, x_1, y_1), (I_2, x_2, y_2), \dots, (I_n, x_n, y_n)$, where the data has been ordered such that $y_i = 0$ (and $I = 0$) for the first n_I observations and the remaining $(n - n_I)$ observations have $y_i > 0$ ($I = 1$), the likelihood for a parameter vector θ is

$$= (\text{Part 1 likelihood}) \times (\text{Part 2 likelihood}). \quad (3.5)$$

Hence, the likelihood is represented by two parts and the model parameters may be estimated by separately maximising each likelihood.

The interpretation of the two parts of the probability model is fairly straightforward. Part 1 represents the probability that an individual will incur any inpatient health care cost. Part 2 represents the probability distribution of inpatient health care costs given that an individual has incurred a cost.

Part 1 of the model may be used to predict the probability that an individual will incur inpatient health care costs in the subsequent year (with estimated prediction error). Given that an individual has incurred inpatient health care costs, Part 2 may be used to predict the monetary amount of inpatient health care costs that might be incurred (with estimated prediction error). The combined prediction of subsequent year inpatient health care costs for an individual is obtained by simply multiplying the predictions from the two parts of the model. This is shown below.

Given covariate X , let $\hat{p}$ represent the estimated probability of an individual incurring an inpatient health care cost, let $\hat{\mu}$ be the estimated mean inpatient health care cost, given that a cost was incurred. Then the overall estimate of inpatient health care cost is given by

$$\hat{Y} = \hat{p} \times \hat{\mu} \quad (3.6)$$

The standard errors of Parts 1 and 2 of the model may be combined to estimate an overall prediction error. The estimated standard error of $\hat{Y}$ is calculated as the square root of

$$\begin{aligned} \text{Var}(\hat{Y}) &= \text{Var}(\hat{p} \times \hat{\mu}) \\ &= \hat{p}^2 \text{Var}(\hat{\mu}) + \hat{\mu}^2 \text{Var}(\hat{p}). \end{aligned} \quad (3.7)$$

CHAPTER 4: DATA SOURCE AND STUDY DESIGN

4.1 Data Source and Sample

Discovery Health is the largest administrator and funder of medical schemes in South Africa. It covers over 2 million lives and manages 13 medical schemes. This accounted for approximately 40% of the market share of the private medical scheme industry as at June 2010².

Discovery Health pioneered consumer-driven health care in South Africa through many innovations, including its wellness programme, Vitality. Through its Vitality programme, Discovery Health incentivises its members to complete voluntary and self-reported Health Risk Appraisals (HRA's) on an annual basis. Between January 2006 and December 2009, 201,341 members of the Vitality programme completed HRA's. There were 58,244 HRA's completed in 2009. This analysis only included data for HRA respondents in 2009 that had a full 12 month exposure in that year. Only respondents with a BMI greater than 18 are included in the analysis. The J-shape relationship between BMI and health care costs suggests that the inclusion of individuals with a BMI < 18 would confound the potentially linear relationship between BMI and health care costs (Wang *et al.* 2006). Given that one objective of this study is to explore correlations between BMI and inpatient health care costs, it is considered necessary to exclude underweight respondents from the study as this confounds the linear relationship between BMI and inpatient health care costs (Wang *et al.* 2006).

² **Source:** Discovery Health Management Information Report, June 2010

4.2 Design

This study is a cross-sectional study of Discovery Health members who completed a HRA between 1 January 2009 and 31 December 2009. Only respondents who had a full 12 months exposure in 2009 were included in the analysis. Analyses were restricted to individuals with complete data on all analysis variables.

The independence of errors assumption needs to be maintained for normal and logistic regression. It means that the cases of data should not be related; for example, you cannot measure the same people at different points in time. Violating this assumption produces overdispersion. In order to avoid overdispersion, only the most recent HRA was included in the analysis in instances where an individual had completed more than one HRA in 2009.

4.3 The Health Risk Appraisal (HRA) Questionnaire

Individuals belonging to Discovery Health were incentivised (in the form of points for engagement in health promoting activities) to voluntarily complete the HRA online or as part of a corporate wellness day. The HRA questionnaire used for the Vitality programme is comprised of 24 questions related to health and lifestyle factors, readiness for lifestyle change and health-related quality of life. The health and lifestyle questions included smoking status, habitual alcohol consumption, fruit and vegetable intake and weekly physical activity habits. Physical activity questions targeted frequency, relative intensity and minimum

and maximum duration per session, resulting in an estimated minimum or maximum minutes of moderate-to-vigorous activity per week. In addition, respondents provided self-reported serum cholesterol, blood pressure, height and weight.

4.4 Inpatient health care costs

In this study, the main dependent variables are whether a HRA respondent incurred any inpatient health care service costs and the amount that was funded by Discovery Health in respect of these services. The annual inpatient health care cost for an HRA respondent represents all costs funded by Discovery Health in respect of the respondent for all hospital admissions during the year (2009) in which the HRA was completed. It includes costs payable to the hospital facility and to medical professionals (e.g. anaesthetists, surgeons, general practitioners, etc.). Table 2 shows that only 17.97% of HRA respondents incurred inpatient health care costs with the mean cost of inpatient costs was ZAR3,912 (with a standard deviation of ZAR14,245) and a median of ZAR0. Given that inpatient costs were incurred, the mean cost was ZAR21,765 (with a standard deviation of ZAR27,212).

Table 4.1

Summary statistics for inpatient health care utilisation and inpatient health care costs

	Mean	Std Dev.	Minimum	Maximum
Probability of at least one admission	0.1797	0.3840	0	1
Inpatient health care costs per covered life (ZAR)	3,912	14,245	0	499,967
Inpatient health care costs given that health care costs were incurred (ZAR)	21,765	27,212	504	499,967

CHAPTER 5: DATA DESCRIPTION

5.1 Variables

The variables that were used in this study are described below. Importantly, these variables are not exhaustive; that is they do not represent all covariates and predictor variables that might enhance this study, but they are those risk factors that were made available for analysis by Discovery Health. The adjusted mean inpatient health care costs that are presented have been adjusted for all other covariates. The covariates of interest for this study were those variables in the dataset that are likely to independently influence inpatient health care costs irrespective of physical activity levels, smoking status or body mass index. For this analysis, the covariates are Age (in 10 year bands); Gender; No, single or multiple chronic conditions; and Plan type (Classic/Executive; Essential/Coastal; KeyCare). Note that the mean costs that are presented for the different categories of each covariate represent the mean annual inpatient cost per life and not per claimant.

5.1.1 Age

Age is calculated as the age last birthday as at 31 December 2009. All respondents who completed a HRA in 2009 and that were between the ages of 20 and 59 (inclusive) were included in the study. For the purposes of this study, age was grouped into 10-year age bands. The summary statistics are calculated using the data in age bands and not the raw age data. The mean age is 36.70 years (standard deviation 8.47) and the median age is 35 years. This indicates a

skewness to the younger ages. The data for respondents aged greater than 60 were sparse and the exploratory analysis showed that many of these observations were exerting undue influence on the model. It is therefore considered prudent to restrict this analysis to respondents aged less than 60.

The results show that the adjusted mean annual inpatient cost is the highest for the 50-59 age band and the lowest for the 40-49 age band. The higher costs in the 20-29 and 30-39 age bands may be associated with childbirth-related costs.

Table 5.1

Number of respondents and adjusted mean inpatient health care cost per age band

Age Group	Number of Respondents	% of Respondents	Cumulative % of Total Respondents	Adjusted Mean Inpatient Cost	95% Confidence Interval	
20 - 29	10,817	21.62%	21.62%	6,179 [*]	5,760	6,598
30 - 39	23,061	46.09%	67.71%	6,465 [†]	6,087	6,844
40 - 49	11,123	22.23%	89.94%	5,769 [‡]	5,360	6,179
50 - 59	5,032	10.06%	100.00%	6,534	6,058	7,010

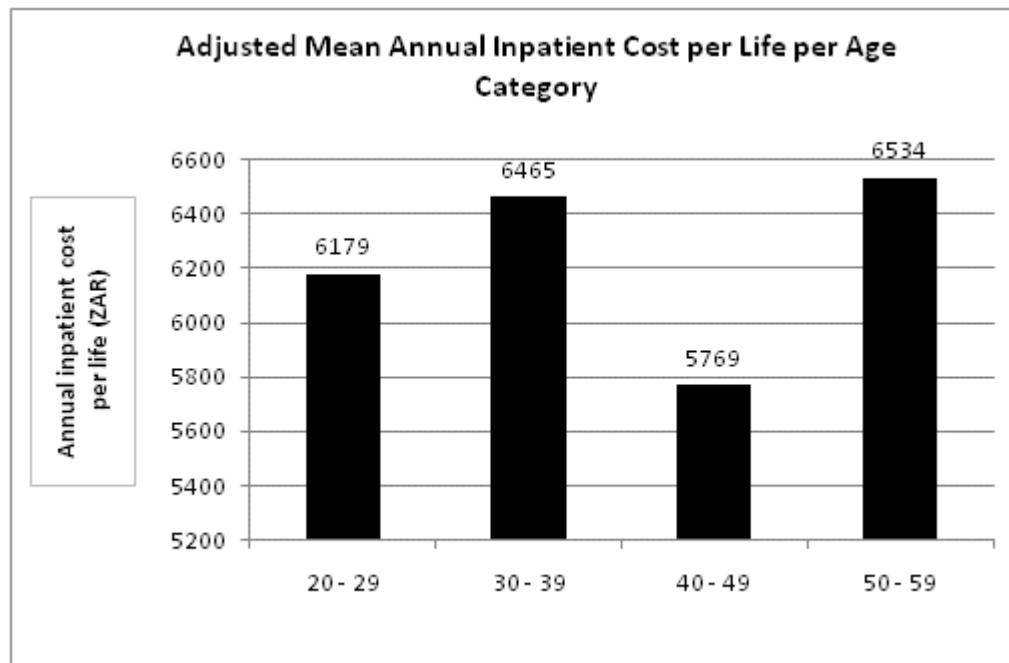
^{*} Significantly different to mean inpatient costs for the other age bands at the 10% level

[†] Significantly different to mean inpatient costs for the 40-49 age band at the 1% level and for the remaining age bands at the 10% level

[‡] Significantly different to mean inpatient costs for the 30-39 and 50-59 age bands at the 1% level and for the 20-29 age band at the 10% level

Figure 5.1

Plot of the adjusted mean annual inpatient cost per life per age category



5.1.2 Gender

Gender is reported as Male or Female. There are slightly more females in the dataset than males. The adjusted means per gender show that females in the dataset incur higher inpatient health care costs than males. The literature suggests that one contributory factor to this result may be health care costs in respect of childbirth.

Table 5.2

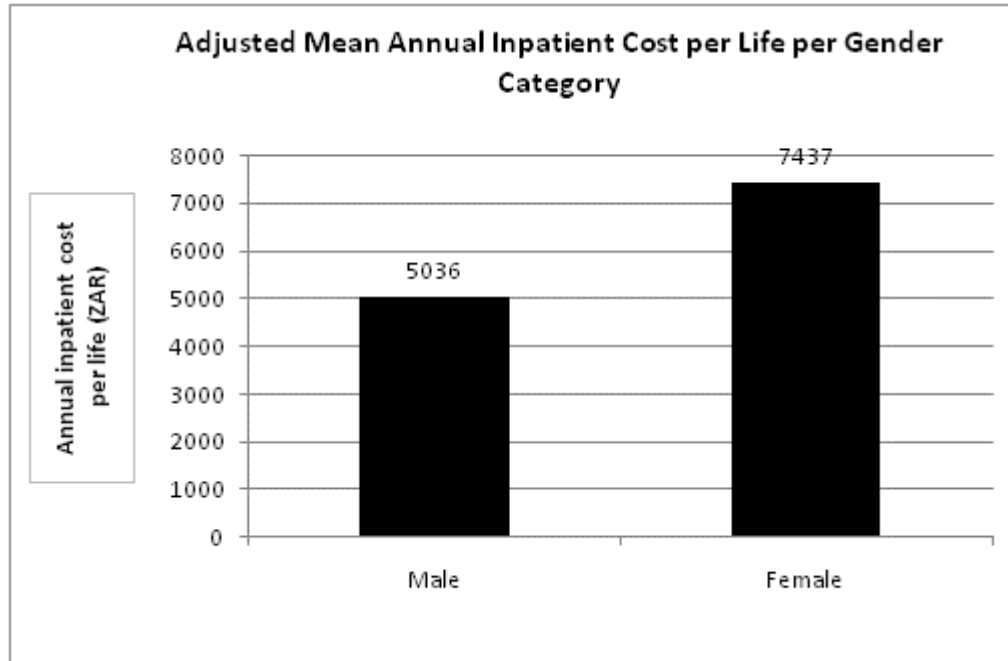
**Number of respondents and adjusted mean inpatient health care cost for
Males and Females**

Gender	Number of Respondents	% of Respondents	Cumulative % of Total Respondents	Adjusted Mean Inpatient Cost	95% Confidence Interval	
Male	24,346	48.66%	48.66%	5,036	4,677	5,396
Female	25,687	51.34%	100.00%	7,437*	7,072	7,803

* Significantly different to mean inpatient costs for Males at the 1% level

Figure 5.2

Plot of the adjusted mean annual inpatient cost per life per gender category



5.1.3 Health Insurance Plan Choice

The Discovery Health plan types offer different levels of inpatient benefits. Therefore, it is expected that HRA respondents belonging to plan types that offer richer benefits will have higher inpatient costs. For this reason, Plan Type was included as a covariate in this study. A review of the Discovery Health benefit schedules indicated that there were three distinct levels of inpatient cover available to members. The Plan Type variable was therefore designed as a categorical variable with three categories. HRA respondents belonging to the Classic and Executive plans were assigned to Category 1, those belonging to Coastal and Essential plans were assigned to Category 2 and respondents

belonging to KeyCare plans were assigned to Category 3. A comparison of the inpatient, outpatient, chronic and oncology benefits offered by Discovery Health per plan type option has been included in the Appendix of this report.

The majority of HRA respondents belong to the Classic and Executive plans. The adjusted mean annual inpatient cost per life was not significantly different for the KeyCare and Essential/Coastal plans. The adjusted mean inpatient cost per plan type category suggests that there is selection against the Classic and Executive plans.

Table 5.3

Number of respondents and adjusted mean inpatient health care cost per plan type category

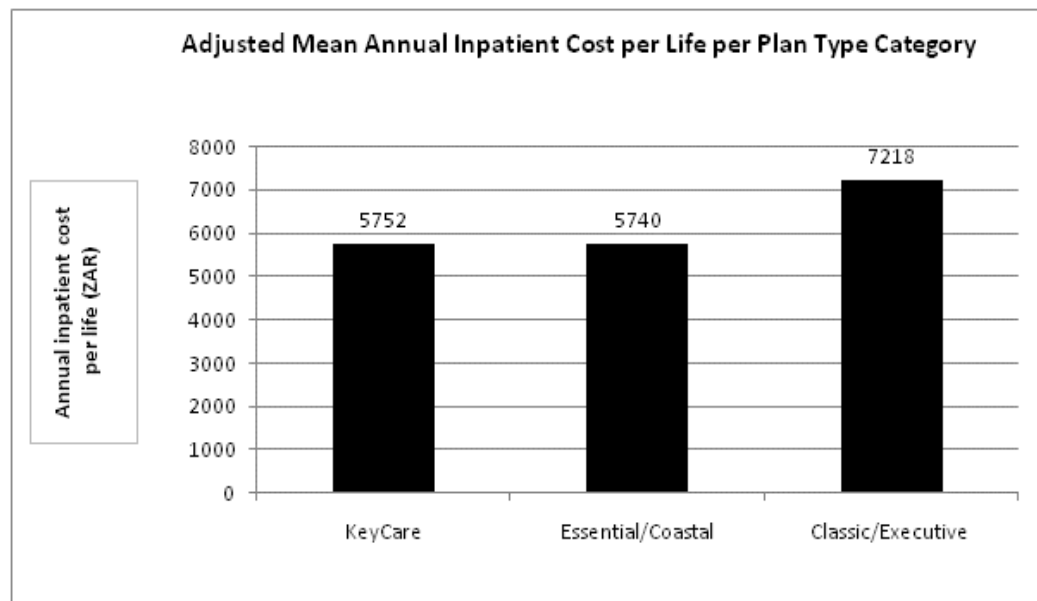
Plan Category	Number of Respondents	% of Respondents	Cumulative % of Total Respondents	Adjusted Mean Inpatient Cost	95% Confidence Interval	
KeyCare	1,961	3.92%	3.92%	5,752 [*]	5,077	6,427
Essential/Coastal	16,130	32.24%	36.16%	5,740 [†]	5,403	6,077
Classic/Executive	31,942	63.84%	100.00%	7,218	6,928	7,509

^{*} Significantly different to mean inpatient costs for Classic/Executive plans at the 1% level but not for Essential/Coastal plans

[†] Significantly different to mean inpatient costs for Classic/Executive plans at the 1% level but not for KeyCare plans

Figure 5.3

Plot of the adjusted mean annual inpatient cost per life per health plan category



5.1.4 Chronic Status

HRA respondents have been divided among three categories of chronic condition categories for the purposes of this analysis: No chronic conditions; Single chronic condition; and Multiple chronic conditions.

There is a significant increase in adjusted mean inpatient costs per chronic status category. This suggests that chronic status is a significant determinant of inpatient health care costs.

Table 5.4

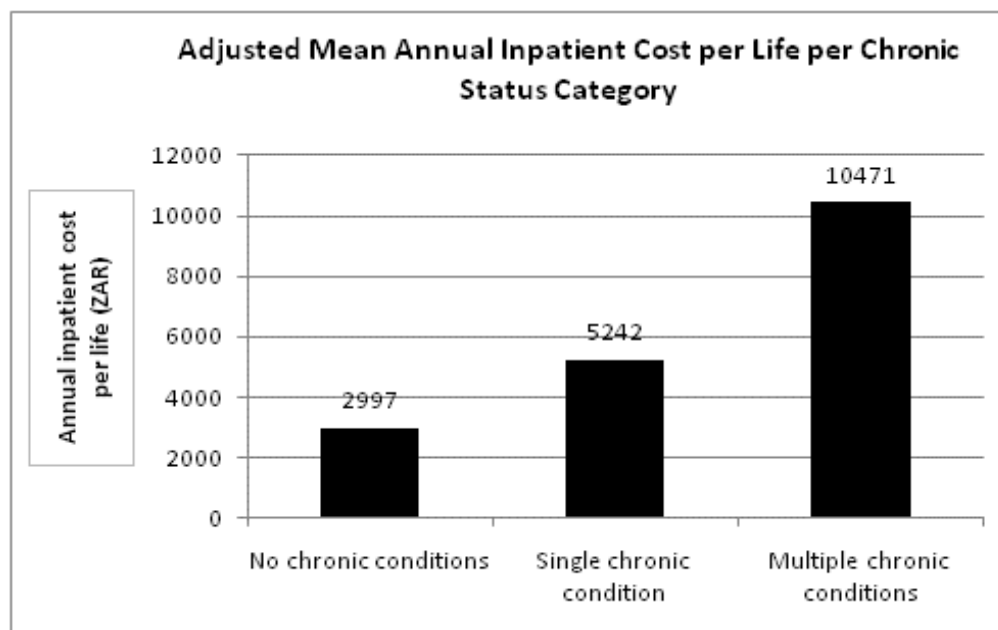
Number of respondents and adjusted mean inpatient health care cost per chronic status category

Chronic status	Number of Respondents	% of Respondents	Cumulative % of Total Respondents	Adjusted Mean Inpatient Cost	95% Confidence Interval	
No chronic conditions	43,035	86.01%	86.01%	2,997	2,720	3,275
Single chronic condition	5,080	10.15%	96.17%	5,242	4,792	5,692
Multiple chronic conditions	1,918	3.83%	100.00%	10,471	9,797	11,146

Note: All pairwise comparisons of differences in mean inpatient costs for the various combinations of chronic status are significant at the 1% level

Figure 5.4

Plot of the adjusted mean annual inpatient cost per life per chronic condition category



5.1.5 Physical Activity

Physical activity is defined with respect to the number of hours per week that HRA respondents reported as being engaged in a fitness-related physical activity. The following categories are defined in the Discovery Health data set: Inactive (0 – 74 minutes per week), Low Active (75 – 149 minutes per week), Active (150 – 225 minutes per week), and High Active (> 225 minutes per week).

The adjusted mean inpatient health care cost decreases as the number of reported hours engaged in physical activity increases with the exception of the comparison between Active and High Active respondents. However, the difference in mean annual inpatient costs between these categories is not significant. The Active physical activity category is associated with the lowest adjusted mean annual inpatient cost (ZAR5,795). The Inactive physical activity category is associated with the highest adjusted mean annual inpatient cost (ZAR6,947).

Table 5.5

Number of respondents and adjusted mean inpatient health care cost per physical activity category

Physical activity category	Number of Respondents	% of Respondents	Cumulative % of Total Respondents	Adjusted Mean Inpatient Cost	95% Confidence Interval	
Inactive	12,674	25.33%	25.33%	6,947*	6,557	7,337
Low Active	10,928	21.84%	47.17%	6,289†	5,879	6,699
Active	10,564	21.11%	68.29%	5,795‡	5,378	6,213
High Active	15,867	31.71%	100.00%	5,915	5,524	6,308

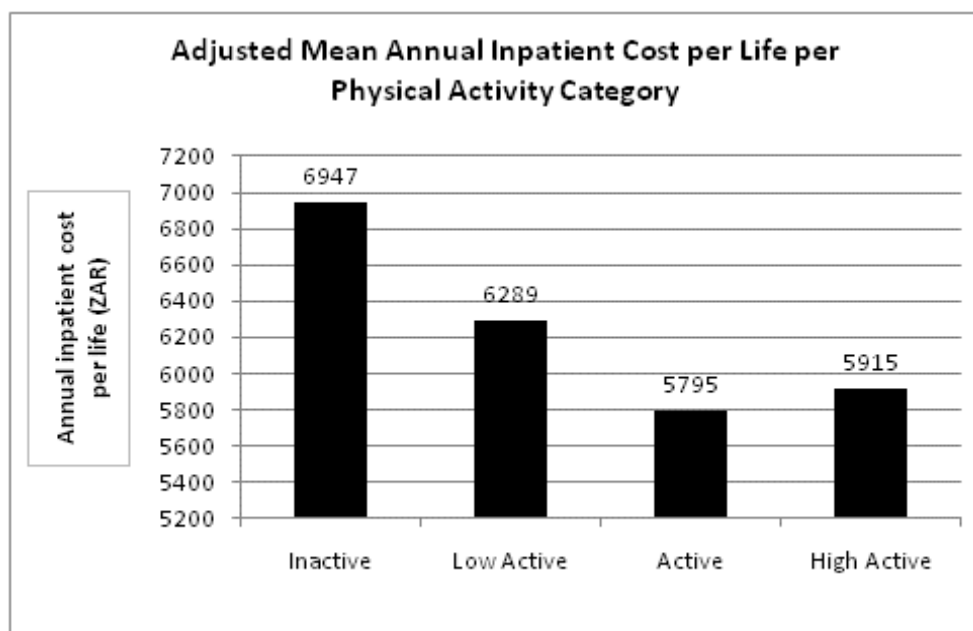
* Significantly different to mean inpatient costs for all other physical activity categories at the 1% level

† Significantly different to mean inpatient costs for Active respondents at the 5% level and for High Active respondents at the 10% level

‡ The difference in mean inpatient costs for Active and High Active respondents is not statistically significant

Figure 5.5

Plot of the adjusted mean annual inpatient cost per life per physical activity category



5.1.6 Smoking status

The HRA questionnaire includes a declaration with respect to smoking status with the following response options: Never smoked, Ex-smoker and Current smoker. The majority of HRA respondents are non-smokers (67.98%) with only 11.53% of respondents reporting that they are current smokers. 20.49% of HRA respondents are ex-smokers. The low percentage of current smokers may be a function of Discovery Health not awarding any points to HRA respondents who smoke and this therefore, reduces the incentive to complete an HRA. Secondly, smoking status may be incorrectly reported. Discovery Health did not perform biochemical assessments (randomly or otherwise) to verify reported smoking status.

Table 8 shows that the adjusted mean inpatient healthcare cost is highest amongst ex-smokers than amongst non-smokers and current smokers. The results indicate that current smokers incur the lowest annual inpatient health care costs of the three defined smoking statuses.

Table 5.6

Number of respondents and adjusted mean inpatient health care cost per smoking status

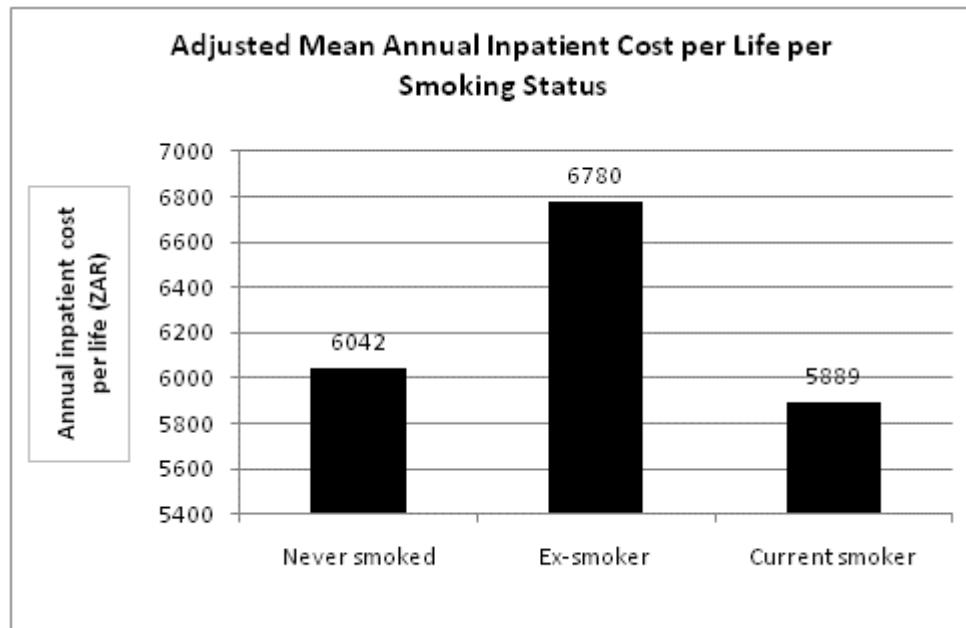
Smoking status	Number of Respondents	% of Respondents	Cumulative % of Total Respondents	Adjusted Mean Inpatient Cost	95% Confidence Interval	
Never smoked	34,330	68.61%	68.61%	6,042*	5,704	6,379
Ex-smoker	10,525	21.04%	89.65%	6,780†	6,380	7,179
Current smoker	5,178	10.35%	100.00%	5,889	5,407	6,371

* Significantly different to mean inpatient costs for ex-smokers at the 1% level but not for current smokers

† Significantly different to mean inpatient costs for ex-smokers and current smokers at the 1% level

Figure 5.6

Plot of the adjusted mean annual inpatient cost per life per smoking status



5.1.7 Body Mass Index

The HRA questionnaire asked respondents to declare their height (in metres) and their weight (in kilograms). These responses were used to calculate a Body Mass Index (BMI) for each respondent using the formula $[\text{Weight}/(\text{Height} \times \text{Height})]$.

The World Health Organisation defines the following weight ranges for individuals: Underweight ($< 18.0 \text{ kg/m}^2$), Normal weight ($18.0 - 24.9 \text{ kg/m}^2$), Overweight ($25.0 - 29.9 \text{ kg/m}^2$) and Obese ($\geq 30.0 \text{ kg/m}^2$) (World Health Organisation 1997). The J-shape relationship between BMI and health care costs suggests that the inclusion of individuals with a BMI < 18 would confound the

potentially linear relationship between BMI and health care costs (Wang *et al.* 2006). Given that one objective of this study is to explore correlations between BMI and inpatient health care costs, it is considered necessary to exclude underweight respondents from the study as this confounds the linear relationship between BMI and inpatient health care costs (Wang *et al.* 2006).

The results show that the adjusted mean inpatient health care cost is lowest for normal weight individuals and highest for obese individuals. Adjusted mean inpatient health care costs increase as BMI increases through the normal, overweight and obese ranges.

Table 5.7

Number of respondents and adjusted mean inpatient health care cost per BMI category

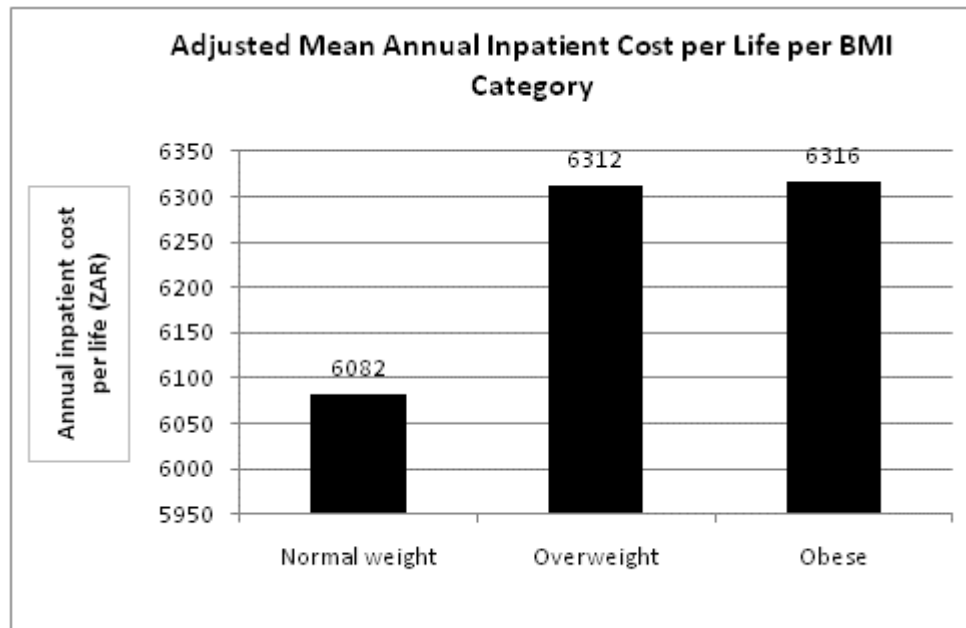
BMI category	Number of Respondents	% of Respondents	Cumulative % of Total Respondents	Adjusted Mean Inpatient Cost	95% Confidence Interval	
Normal weight	24,955	49.88%	49.88%	6,082*	5,712	6,452
Overweight	17,908	35.79%	85.67%	6,312†	5,939	6,686
Obese	7,170	14.33%	100.00%	6,316	5,878	6,753

* Significantly different to mean inpatient costs for overweight and obese respondents at the 10% level

† Significantly different to mean inpatient costs for normal weight respondents at the 10% level but not for obese individuals

Figure 5.7

Plot of the adjusted mean annual inpatient cost per life per BMI category



CHAPTER 6: STATISTICAL METHODS

This section presents a detailed description of the statistical procedures that were used in order to conduct the statistical analysis that is required to achieve the research objective.

6.1 Descriptive statistics and simple inference

Initial examination of the data involved histograms and normal probability plots, in order to assess the distributional properties of the variables in the dataset.

6.2 A two-part regression model

Regression is an iterative process, a process in which the outputs are used to diagnose, validate, criticise, and possibly modify the inputs (Chatterjee, Hadi 2006). The process is repeated until a satisfactory model is estimated. A satisfactory model is one that satisfies the standard regression assumptions and that fits the data reasonably well.

A conditional probability approach (two-part model) was used to analyse the inpatient health care costs. This approach recognises that health care costs are the product of the probability of incurring a health care cost multiplied by the expected level of cost conditional on positive utilisation.

The two-part model involved two stages of estimation. In this study, the first part used a logistic regression model to estimate the relationship between modifiable health risk factors (BMI, smoking status and level of physical activity), after

adjusting for socio-demographic and health status variables (age, sex, number of chronic conditions and plan type), and the likelihood of incurring an inpatient health care cost.

6.2.1 Part One: Logistic regression

Logistic regression is similar to multiple regression but with an outcome variable that is a categorical variable and predictor variables that are continuous or categorical.

The steps that were followed for the logistic regression analysis were (Field, Miles 2010):

- i. A logistic regression model was estimated using the LOGISTIC procedure of the SAS software program with a backward selection method (SAS Institute 1990).
- ii. The overall fit of the model was assessed using the Hosmer-Lemeshow test.
- iii. The performance of individual regression coefficients was assessed using a test.
- iv. Odds ratios were computed for all modifiable health risk variables using the ODDSRATIO statement in the LOGISTIC procedure of the SAS software program. Customised odds ratios were computed using the CONTRAST statement in the LOGISTIC procedure of the SAS software program.

Model Selection

An iterative approach was utilised in estimating the logistic regression model. First, a crude model that included only the main effects was fitted in the hope that a satisfactory fit could be obtained without violating any of the standard assumptions. It was thought that this would be sufficient to explain the association between the modifiable health risk factors of interest and the likelihood of incurring an inpatient health care cost given that this study is an explanatory rather than a predictive one. The estimated model was unsatisfactory.

Next, the model selection process considered a backward selection method. The initial model included all main effects and all two-way interactions. It was further specified that the variables that represent the three modifiable health risk factors that are of interest for this study, must be included in the fitted model. A satisfactory model was estimated. The model converged and the optimal model included all main effects and eight 2-way interactions. The optimal model was estimated after fourteen steps.

Goodness-of-fit

Overall goodness-of-fit of the estimated model was assessed using the Hosmer-Lemeshow test. The Hosmer-Lemeshow statistic has approximately a chi-square distribution under the null hypothesis that the fitted model is correct (Allison 2006). It was concluded that the model is a significant fit of the data

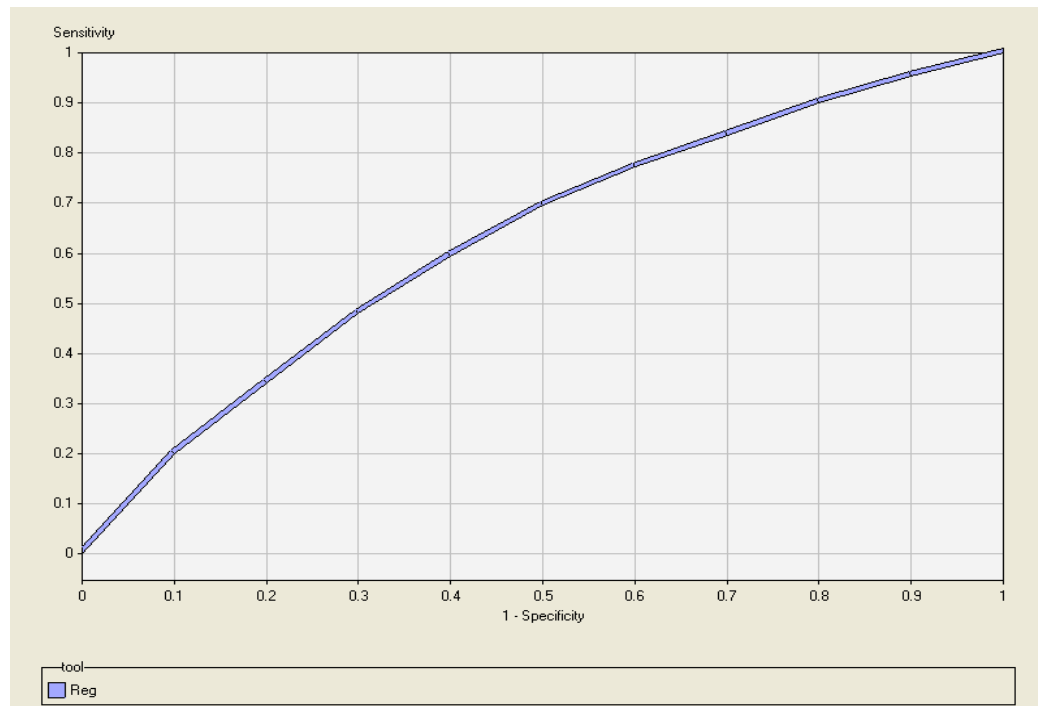
based on the Hosmer-Lemeshow statistic ($p = 0.7152$) and the model likelihood ratio and Wald statistics ($p < 0.0001$ for both statistics).

The R-squared statistic represents the percentage of variation in the outcome that is explained by the model. The Cox and Snell R-squared of the fitted model is 0.0398 and the Nagelkerke (max rescaled) R-squared is 0.0653.

A useful summary of the logistic regression model is provided by the c-statistic (Concordance Index) which is based on the receiver operating characteristic (ROC) curve (Figure 8). If the model is very poor, the c-statistic is 0.5 (Stokes *et al.* 2006). If the c-statistic is 1.0, it means that the model was capable of classifying all responses correctly. The c-statistic for the estimated logistic regression model was computed to be 0.648 indicating that the model was marginally better than chance at predicting the likelihood of incurring an inpatient cost over the period of a year.

Figure 6.1

The Receiver Operating Characteristic (ROC) Curve for the estimated logistic regression model



Diagnostics

Regression diagnostics are used to indicate observations that might have undue influence on the model fit or that might be outliers. They tell you how influential each observation is to the fit of the model. An observation is said to be influential to the fit of a model if its removal substantially changes the estimates of the model coefficients.

Pregibon suggests the use of several diagnostic statistics to identify influential observations and to quantify the impact on the maximum likelihood fit (Pregibon

1981; Sarkar *et al.* 2010): They include the Pearson residual, the deviance residual and the leverage (the hat value). For large enough samples, the following cut-off points are proposed to identify influential observations:

- i. Leverage greater than 2 or 3 times the average leverage.

Average leverage = $\frac{k+1}{n}$ where k is the number of covariates in the model and n is the sample size.

- ii. The absolute value of the Pearson residual greater than 2.
- iii. The absolute value of the Deviance residual greater than 2.

The index plots and data that were produced by specifying the INFLUENCE and IPLOTS options in the LOGISTIC procedure of the SAS software program were examined in an effort to identify if there were any observations that were exerting undue influence on the fitted model.

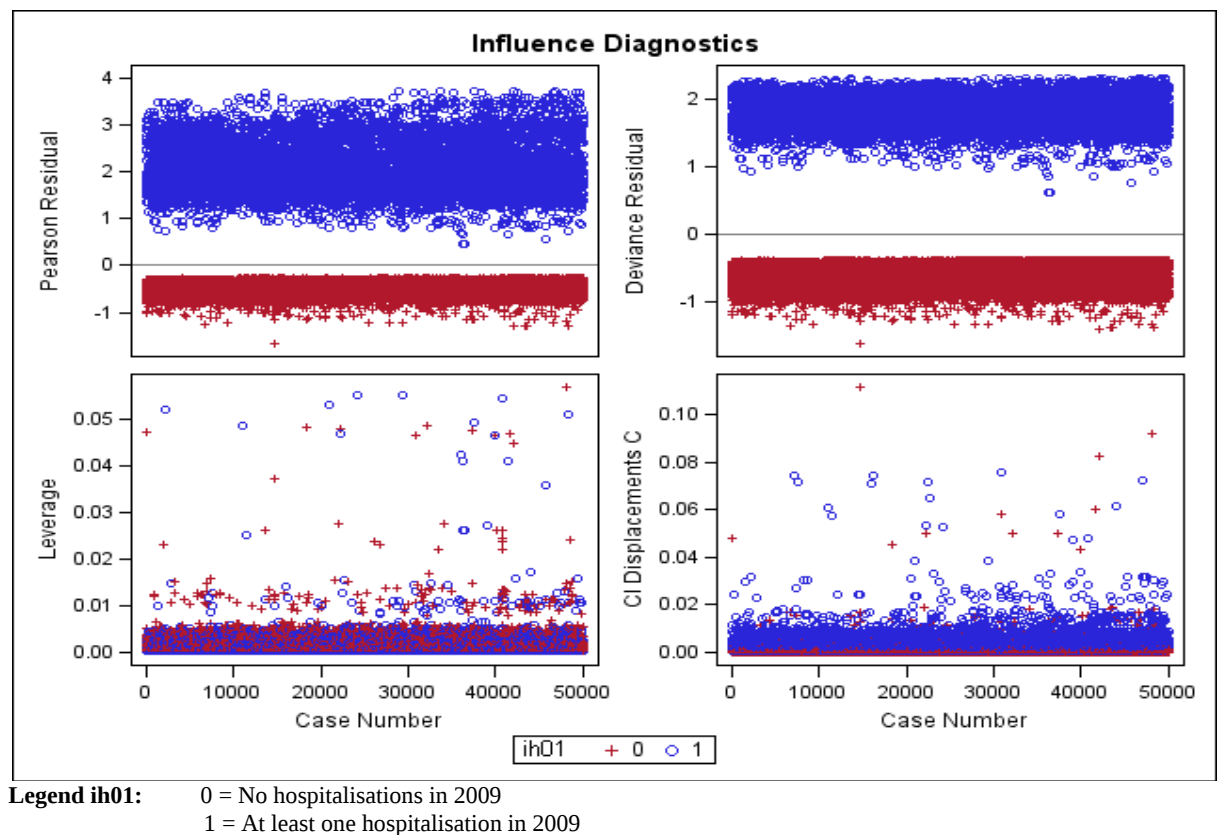
Scatter plots of residuals against fitted values can indicate the appropriateness of the error function that has been assumed. The top-right panel in Figure 9 is the scatter plot of deviance residuals against fitted values for our estimated logistic regression model. Moving from left to right of the plot, it is clear that the general mean and variability of the deviance residuals is reasonably constant thereby suggesting that the assumed variance function is appropriate.

In addition to being needed to calculate the standardised residuals, the leverage statistic is also a useful diagnostic. It can identify particular observations which may have an undue influence on the model (Field, Miles 2010). In the case of our

model, the bottom-left panel in Figure 9 shows there are twenty observations with leverage greater than 0.04 that may be exerting undue influence on the fitted model. Observations with high leverage were flagged and examined to see if they were also influential. The level of influence that was exerted by an observation was assessed using Cook's distance. It was determined that these twenty observations were not influential of the model and were therefore retained in the data set.

Figure 6.2

Influence diagnostic plots



Odds ratios

The odds ratio is indicative of the strength of association between a predictor variable and the outcome of interest. In the case of this study, it is used to indicate the strength of association between each of the modifiable health risk factors of interest and the likelihood of incurring an inpatient health care cost in the calendar year. If the odds ratio is 1, there is no association.

6.2.2 Part Two: Linear regression

For the second part of the two-part regression model, I use ordinary least squares regression to explore the relationship between the modifiable health risk factors and inpatient health care costs controlling for demographic and other risk factors, given that an inpatient cost had been incurred by the HRA respondent in 2009. The annual inpatient cost outcome variable used in the second part of the two-part regression model was log-transformed to satisfy regression assumptions. Outlier cases were included as part of the annual inpatient cost outcome variable.

The steps involved in the linear regression analysis are:

- i) Univariate analysis of the dependent variable representing inpatient costs to explore its distributional characteristics.
- ii) Transformation of the dependent variable to achieve normality.
- iii) Estimation of an ordinary least squares regression model that included all the main effects using the PROC REG procedure of the SAS software program.

- iv) Assess the goodness-of-fit of the model using general linear model diagnostics and by analysing the residuals to ensure that the normality and linearity assumptions have not been violated.

Diagnostics

In regression data analysis, the process of examining a preliminary model fit and using information about any lack of fit to improve the model specification is known as “diagnostic analysis”. Residual analysis is an important type of diagnostic analysis. Residuals are defined as the observed value minus the fitted value, and the analysis involves both the numeric and graphical inspection of the estimated models.

The following plots of the studentised residuals were used to check that the normality and linearity assumptions had not been violated:

- a) *Normal probability plot of the studentised residuals*

Under normality assumption, this plot should resemble a (nearly) straight line.

- b) *Scatter plots of the studentised residuals against each of the predictor variables*

Under the standard assumptions, the studentised residuals are uncorrelated with each of the predictor variables. If the assumptions hold, this plot should be a random scatter of points.

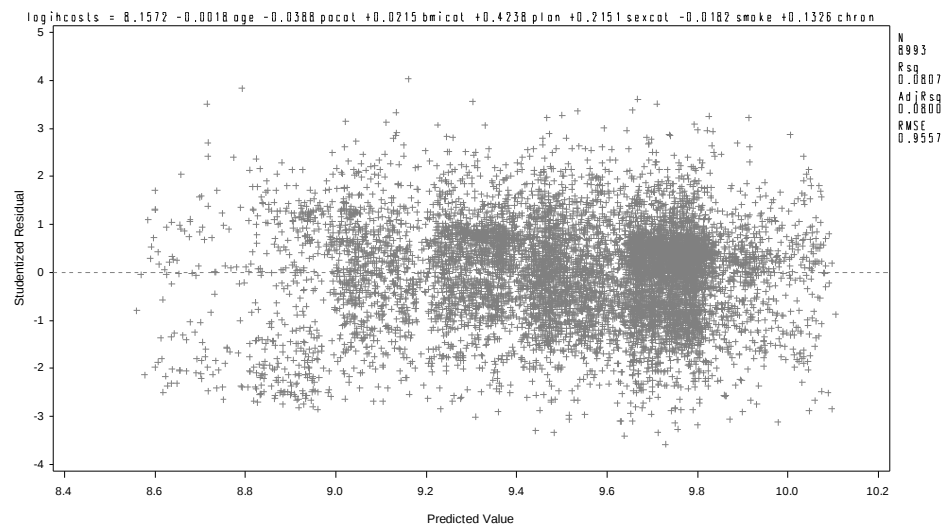
c) *Scatter plots of the studentised residuals versus the predicted values*

Under the standard assumptions, the studentised residuals are uncorrelated with the predicted values; therefore, this plot should also be a random scatter of points.

When the plot of the studentised residuals versus the predicted values does not exhibit any marked trend, it is indicative of an estimated model that is an adequate fit of the data and that has not violated any model assumptions. Figure 10 confirms that this is indeed the case in the instance of the estimated linear model.

Figure 6.3

Plot of the studentised residuals against the predicted values



The next step in the diagnostics process is concerned with the detection of outliers and influential observations. Outliers in the response variable were explored in order to determine an appropriate corrective action. Outliers were those observations with studentised residuals larger than 2 or 4 times the standard deviation (Chatterjee, Hadi 2006). Outliers in the predictors were flagged as those having high leverage. Observations with high leverage were flagged and examined to see if they were also influential. The level of influence that was exerted by an observation was assessed using Cook's distance. The Welsch and Kuh measure of influence (DFITS) was also available but it is sufficient to consider only one measure of influence (Chatterjee, Hadi 2006).

Collinearity

Multicollinearity occurs when there are strong linear dependencies among the explanatory variables. If two or more explanatory variables are highly correlated with one another, the estimates of their distinct effects on the dependent variable may not be reliable. The consequences of multicollinearity only apply to variables that are collinear. Therefore, multicollinearity was assessed by estimating the model using PROC REG and specifying the collinearity options (Allison 2006). Collinearity amongst the dependent variables is assessed using variance inflation factors produced by the PROC REG procedure in SAS (Belsley *et al.* 1980). This ensures that there are no correlations between variables that will burden the model. All of the variance inflation factors were less than 10.0 thereby indicating that collinearity is not an issue and no changes to the explanatory variables that are included in the estimated model, are necessary.

The size of the study sample in the multiple regression is sufficiently large to expect robust parameter estimates. Therefore, alternative statistical methods such as bootstrapping techniques were not applied.

CHAPTER 7: RESULTS

This section presents detailed descriptions of the results that were obtained from the statistical analyses that were conducted. The results are presented per step of the analysis plan.

There were 58,244 HRA respondents in 2009. Of these, 50,033 (85.9%) had complete data on all study variables, and provide the basis for this research report. Of the 50,033 HRA respondents, 8,993 (18.0%) respondents incurred an inpatient health care cost in 2009.

The distribution of 2009 inpatient health care costs incurred by HRA respondents was highly skewed with a small proportion of respondents accounting for a large proportion of the claims. The median annual inpatient cost is ZAR0 (interquartile range, ZAR0 – ZAR0) compared with a mean of ZAR3,912 (standard deviation ZAR14,245). Again, this is consistent with a significant proportion of HRA respondents (82.0%) not incurring an inpatient health care cost in 2009. For HRA respondents who had incurred an inpatient health care cost in 2009, the median cost is ZAR15,965 (interquartile range, ZAR7,293 – ZAR26,656) and the mean cost is ZAR21,765 (standard deviation ZAR27,212).

An analysis of the distributional properties of the positive inpatient health care costs indicated the log-transformed costs were normally distributed. Normality was assessed using the Kolmogorov-Smirnov statistic ($p < 0.0001$). This suggests

that the two-part log-transformed ordinary least squares (OLS) model could be a good estimator (Matsaganis *et al.* 2009).

7.1 Results: Logistic regression

This section presents the results of the logistic regression model that is used in the first part of the two-part regression model to estimate the likelihood of incurring an annual inpatient health care cost taking into account BMI, smoking status and physical activity after controlling for age, gender, number of diagnosed chronic conditions and health insurance plan type.

7.1.1 Parameter estimates

Table 10 presents the estimates of the coefficients for the predictor variables in the fitted model that represent the modifiable health risk factors that are of interest for this study. The important statistic is the Wald chi-square. It reflects whether the coefficient for the given predictor variable is significantly different from zero i.e. whether the predictor variable is making a significant contribution to the likelihood of incurring an inpatient health care cost (Field, Miles 2010). The results include all 2-way interactions that were found to be significant during the backward selection process. The parameter estimates are used to compute the odds ratios in the next section. Odds ratios are computed for ease of interpretation of the estimated logistic regression model.

Table 7.1

Parameter estimates from the estimated logistic regression model (Part One of the two-part regression model)

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	-1.0519	0.0549	366.9966	<.0001
Overweight	-0.0495	0.0225	4.8310	0.0280
Obese	-0.00732	0.0235	0.0968	0.7557
Plan: KeyCare	0.5344	0.1030	26.9400	<.0001
Plan: Essential/Coastal	-0.4223	0.0594	50.5650	<.0001
Female	0.3167	0.0256	152.6096	<.0001
Never smoked	-0.0699	0.0201	12.0473	0.0005
Ex-smoker	0.0671	0.0242	7.6801	0.0056
Inactive	0.1032	0.0345	8.9564	0.0028
Low active	0.0321	0.0371	0.7491	0.3867
Active	-0.1065	0.0395	7.2559	0.0071
Age: 30 - 39	0.0219	0.0422	0.2691	0.6039
Age: 40 - 49	0.1141	0.0399	8.1912	0.0042
Age: 50 - 59	-0.0969	0.0527	3.3752	0.0662
Never smoked*Overweight	-0.0142	0.0253	0.3140	0.5752
Never smoked*Obese	0.0401	0.0268	2.2371	0.1347
Ex-smoker*Overweight	-0.00286	0.0310	0.0085	0.9265
Ex-smoker*Obese	0.1182	0.0319	13.7293	0.0002
No chronic conditions	-0.6097	0.0562	117.6107	<.0001
Single chronic condition	0.0202	0.0640	0.1000	0.7519
KeyCare*Age: 30 - 39	0.1576	0.0765	4.2468	0.0393
KeyCare*Age: 40 - 49	0.1077	0.0750	2.0622	0.1510
KeyCare*Age: 50 - 59	-0.0938	0.1010	0.8636	0.3527
Essential/Coastal*Age: 30 - 39	-0.1629	0.0483	11.3641	0.0007

Parameter	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Essential/Coastal*Age: 40 - 49	-0.0817	0.0446	3.3597	0.0668
Essential/Coastal*Age: 50 - 59	0.1140	0.0582	3.8308	0.0503
Female*Age: 30-39	0.1914	0.0250	58.5811	<.0001
Female*Age: 40 - 49	0.2349	0.0193	148.8124	<.0001
Female*Age: 50 - 59	-0.1470	0.0233	39.8543	<.0001
No chronic conditions*Inactive	0.0227	0.0363	0.3897	0.5325
No chronic conditions*Low active	-0.0523	0.0396	1.7473	0.1862
No chronic conditions*Active	0.0522	0.0420	1.5512	0.2130
Single chronic condition*Inactive	-0.1073	0.0472	5.1712	0.0230
Single chronic condition*Low active	0.0762	0.0503	2.2933	0.1299
Single chronic condition*Active	-0.0645	0.0544	1.4061	0.2357
No chronic conditions*KeyCare	-0.4208	0.1071	15.4249	<.0001
No chronic conditions*Essential/Coastal	0.2574	0.0626	16.9215	<.0001
Single chronic condition*KeyCare	-0.00793	0.1214	0.0043	0.9479
Single chronic condition*Essential/Coastal	0.0595	0.0721	0.6818	0.4090
Female*Inactive	0.0291	0.0218	1.7893	0.1810
Female*Low active	0.0436	0.0231	3.5575	0.0593
Female*Active	-0.0164	0.0233	0.4961	0.4812
Female*KeyCare	0.0994	0.0420	5.5896	0.0181
Female*Essential/Coastal	-0.0692	0.0263	6.9571	0.0083
Female*Never smoked	0.0284	0.0188	2.2791	0.1311
Female*Ex-smoker	0.0581	0.0228	6.5189	0.0107

7.1.2 Odds ratios in respect of the modifiable health risk factors

This section presents the odds ratios that are calculated from the estimated logistic regression model. Note that they are presented in this manner because of the significant 2-way interactions that were found amongst the predictor variables.

The odds ratios represent the odds of incurring an annual inpatient health care cost at the reported levels of the predictor variables with all other variables constant. The odds ratios are presented per modifiable health risk factor with a brief explanation of the interpretation thereof. It must be highlighted that claims related to childbirth may be significantly influencing the results in respect of females and it is recommended that this be considered in future studies that may be based on this dataset. One needs to treat the odds ratio estimates with circumspect in instances where the 95% confidence limit spans the value 1 (Field, Miles 2010) as it indicates that the model cannot reasonably predict whether or not an individual will incur an annual inpatient health care cost at the 5% significance level.

a) Smoking status

The estimated model includes significant 2-way interaction variables that represent interactions between Smoking status, and Sex and BMI Category, respectively. These interactions emerged as significant during the model estimation process. Table 11 presents the odds ratios and their 95% confidence intervals (in brackets) for the specified covariate categories.

Table 7.2

Odds ratios for significant 2-way interactions with smoking status variable

		BMI Category		
		Normal	Overweight	Obese
Never Smoked vs. Ex-Smokers	Females	0.837 (0.767-0.913)	0.783 (0.698-0.878)	0.926 (0.787-1.089)
	Males	0.888 (0.785-1.005)	0.831 (0.745-0.926)	0.982 (0.836-1.154)
Never Smoked vs. Current Smokers	Females	1.011 (0.894-1.143)	1.272 (1.083-1.494)	0.883 (0.719-1.083)
	Males	0.803 (0.687-0.940)	1.011 (0.894-1.143)	0.701 (0.571-0.861)
Ex-Smokers vs. Current Smokers	Females	1.208 (1.052-1.387)	1.625 (1.357-1.945)	0.953 (0.753-1.207)
	Males	0.904 (0.755-1.083)	1.216 (1.022-1.449)	0.714 (0.566-0.901)

Never Smoked vs. Ex-Smokers

The results show that HRA respondents (across both genders and across all BMI categories) who never smoked are less likely to incur an inpatient health care cost in a calendar year than those who are ex-smokers.

Never Smoked vs. Current Smokers

The results show no discernible pattern in respect of the likelihood of incurring an inpatient health care cost when comparing HRA respondents who never smoked to those who are currently smoke across the three BMI categories.

Normal weight and overweight females, and overweight males who never smoked are more likely (1.011, 1.272 and 1.011 times, respectively) to be hospitalised in a calendar year than respondents in those BMI categories who currently smoke.

Obese females, and normal weight and obese males who never smoked are less likely (0.883, 0.803 and 0.701 times, respectively) to incur an inpatient health care cost in a calendar year than HRA respondents in those BMI categories who currently smoke.

Ex-Smokers vs. Current Smokers

The results do not indicate a meaningful trend when comparing the likelihood of hospitalisation in a calendar year for ex-smokers and current smokers across the different BMI categories.

Normal weight and obese males, and obese females who are ex-smokers are less likely (0.904, 0.714 and 0.953 times, respectively) to be hospitalised in a calendar year than their counterparts who are current smokers.

Overweight males, and normal weight and overweight females who are ex-smokers are more likely (1.216, 1.208 and 1.625 times, respectively) to be hospitalised in a calendar year when compared to respondents in those BMI categories who currently smoke.

All smoking statuses

The results indicate a counter intuitive outcome in that the likelihood of hospitalisation is lower for obese males and females when comparing respondents across all combinations of smoking statuses.

b) Physical activity

Two-way interactions between Physical Activity and Sex, and Physical Activity and Number of Chronic Conditions, emerged as significant during the logistic regression process. Hence, the following presentation of the odds ratios of the likelihood of hospitalisation as they relate to the reported levels of physical activity. Table 12 presents the odds ratios and their 95% confidence intervals (in brackets) for the specified covariate categories.

Table 7.3

Odds ratios for significant 2-way interactions with the physical activity variable

		Chronic conditions		
		None	Single	Multiple
Inactive vs. Low Active	Females	1.141 (1.049-1.240)	0.881 (0.727-1.067)	1.180 (0.893-1.559)
	Males	1.174 (1.034-1.333)	0.907 (0.738-1.114)	1.214 (0.917-1.608)
Inactive vs. Active	Females	1.253 (1.147-1.369)	1.237 (1.006-1.520)	1.388 (1.036-1.859)
	Males	1.144 (1.010-1.296)	1.129 (0.908-1.405)	1.267 (0.944-1.701)
Inactive vs. High Active	Females	1.300 (1.195-1.416)	1.015 (0.844-1.220)	1.455 (1.114-1.901)
	Males	1.096 (0.983-1.222)	0.856 (0.706-1.036)	1.227 (0.940-1.601)
Low Active vs. Active	Females	1.099 (1.001-1.206)	1.404 (1.135-1.737)	1.177 (0.860-1.610)
	Males	0.974 (0.856-1.110)	1.245 (0.996-1.557)	1.043 (0.763-1.427)
Low Active vs. High Active	Females	1.140 (1.042-1.247)	1.152 (0.951-1.396)	1.233 (0.922-1.649)
	Males	0.934 (0.833-1.047)	0.944 (0.775-1.149)	1.010 (0.758-1.346)
Active vs. High Active	Females	1.038 (0.945-1.139)	0.821 (0.668-1.008)	1.048 (0.774-1.420)
	Males	0.958 (0.857-1.072)	0.758 (0.615-0.934)	0.968 (0.717-1.307)

Inactive HRA respondents

The results of comparisons between all combinations of physical activity categories and inactive individuals with no chronic conditions show that these inactive respondents are more likely to be hospitalised in a calendar year than

respondents who had a higher level of physical activity. The same is true for inactive individuals suffering from multiple chronic conditions.

Low Active HRA respondents

The results indicate that Low Active respondents with multiple chronic conditions are more likely to incur an inpatient health care cost than their Active and High Active counterparts.

For respondents with no or a single chronic condition, the model estimates an increased likelihood of hospitalisation in 5 of the 8 odds ratios presented in Table 12 that compare the likelihood of hospitalisation for Low Active individuals with Active and High Active individuals.

Active HRA respondents

Four of the six odds ratios that are presented in Table 12 indicate that Active HRA respondents (in the different Chronic Status categories and across both genders) are less likely to be hospitalised than their High Active counterparts.

c) Body mass index

The results of the backward selection logistic regression analysis indicated that there was a significant interaction between BMI Category and Smoking Status. Table 13 presents the odds ratios and their 95% confidence intervals (in brackets) for the specified covariate categories.

Table 7.4**Odds ratios for significant 2-way interactions with the BMI variable**

	Smoking Status		
	Never Smoked	Ex-Smoker	Current Smoker
Normal Weight vs. Overweight	0.908 (0.851-0.969)	0.849 (0.759-0.951)	1.142 (0.967-1.350)
Normal Weight vs. Obese	0.910 (0.834-0.992)	1.006 (0.864-1.171)	0.794 (0.649-0.972)
Overweight vs. Obese	1.002 (0.916-1.096)	1.185 (1.018-1.378)	0.695 (0.561-0.861)

Normal weight HRA respondents

Four of the six computed odds ratios in respect of normal weight respondents across the different combinations of smoking statuses indicate that they are less likely to be hospitalised than their overweight and obese counterparts.

Overweight HRA respondents

Four of the six computed odds ratios in respect of overweight respondents across the different combinations of smoking statuses indicate that they are more likely to incur an inpatient health care cost in a calendar year than their normal weight and obese counterparts.

Obese HRA respondents

Three of the six computed odds ratios in respect of obese respondents across the different combinations of smoking statuses indicate that they are more likely to incur an inpatient health care cost in a calendar year than their normal weight and overweight counterparts.

7.2 Results: Generalised linear model

This section presents the results of the ordinary least squares general linear regression model that is used in the second part of the two-part regression model to estimate the magnitude of annual inpatient health care costs incurred taking into account BMI, smoking status and physical activity after controlling for age, gender, number of diagnosed chronic conditions and health insurance plan type.

In the linear regression estimation, among HRA respondents who have incurred an inpatient health care cost in 2009, there is a negative and significant correlation between levels of physical activity and the log of annual inpatient health care costs. The model predicts that the log of annual inpatient costs decreases as the time spent on physical activity increases.

The estimated regression coefficient for smoking status variable (smoke) indicates that the log of annual inpatient costs is estimated to be lower for current smokers than for ex-smokers and respondents who have never smoked. Also, the log of

annual inpatient costs is estimated to be lower for ex-smokers than for respondents who have never smoked.

The estimated linear regression model indicates that there is a positive correlation between BMI and the log of annual inpatient costs. The model predicts that the log of annual inpatient costs increases as one moves up the ordinal BMI scale, from normal weight to overweight to obese.

Table 7.5

Parameter estimates from the estimated ordinary least squares regression model (Part Two of the two-part regression model)

Variable	Parameter Estimate	Standard Error	Pr > t 	95% Confidence Limits	
Intercept	8.15723	0.09447	<.0001	7.97205	8.34241
Age	-0.00179	0.00128	0.1627	-0.00431	0.00072457
Physical activity	-0.03883	0.00876	<.0001	-0.05600	-0.02166
BMI	0.02155	0.01440	0.1346	-0.00668	0.04977
Plan	0.42381	0.01745	<.0001	0.38960	0.45801
Sex	0.21513	0.02256	<.0001	0.17090	0.25936
Smoking status	-0.01816	0.01510	0.2290	-0.04776	0.01143
No. of chronic conditions	0.13265	0.01857	<.0001	0.09624	0.16905

CHAPTER 8: CONCLUSION

The objective of this study was to assess the relationship between selected modifiable health risk factors and annual inpatient health care costs by analysing the responses of individuals who had completed a health risk assessment (HRA) questionnaire for a South African health insurer in 2009. Importantly, the study is an explanatory study as opposed to a predictive study. Three modifiable health risk factors were selected, namely smoking status, body mass index (as an indicator of nutritional habits) and levels of physical activity. This is the first known attempt at exploring the relationship between multiple modifiable health risk factors and annual inpatient costs in South Africa. The research findings from this study, therefore, add to the existing body of literature in this field and could complement the development of health funding policy in South Africa.

Goetzel *et al.* explored the relationship between modifiable health risk factors and health care costs and found that 7 of 10 health risk factors were significantly related to higher health care costs (Goetzel *et al.* 1998). In this study, 1 of 3 modifiable health risk factors was significantly related to higher inpatient health care costs. The results of this study showed that increasing levels of physical activity are significantly correlated with lower annual inpatient health care costs ($p < .0001$). This is consistent with the literature. Lambert *et al.* concluded that participation in fitness-related activities within an incentive-based health insurance wellness programme was associated with significantly lower health care costs (Lambert *et al.* 2009).

The logistic regression analysis showed that HRA respondents (across both genders and across all BMI categories) who never smoked are less likely to incur an inpatient health care cost in a calendar year than those who are ex-smokers.

Inactive respondents are more likely to be hospitalised in a calendar year irrespective of whether they suffered from multiple chronic conditions or no chronic conditions. Furthermore, the results indicate that the likelihood of hospitalisation does not decrease if an individual increases their physical activity to more than 225 minutes per week. Managed care interventions aimed at increasing physical activity to between 75 minutes and 225 minutes per week will reduce the likelihood of hospitalisation. The results suggest that investment in interventions to increase physical activity to beyond 225 minutes per week may yield poor returns or no return at all.

This study indicates that normal weight individuals are less likely to be hospitalised in a calendar year than overweight and obese individuals, irrespective of their smoking status.

When the focus of the analysis is limited to respondents who had incurred an inpatient health care cost in 2009 ($n = 8,993$), it indicates that of the three modifiable health risk factors that are of interest in this study, two of them are correlated with the magnitude of annual inpatient health care costs even after controlling for age, sex, diagnosed chronic conditions and level of inpatient health

care indemnity insurance. The results indicate that increasing levels of physical activity ($p < .0001$) and decreasing levels of body mass index ($p = 0.1346$) are associated with lower annual inpatient health care costs.

The combined results of each of the models in the two-part regression model indicate that increasing levels of physical activity and decreasing levels of BMI are associated with lower likelihoods of being hospitalised in a calendar year and with lower magnitudes of inpatient health care costs given that an individual has been hospitalised.

Consequently, managed care interventions, for example wellness programmes, that promote and incentivise increased physical activity and healthier dietary regimes could prove significant in curbing health care costs. However, this study does not quantify the reduction in health care charges that may be achieved by modifying behaviour to affect these health risk factors.

The results of this study do not provide evidence to support that smoking cessation is associated with lower annual inpatient health care costs. The results show, for example, that ex-smokers are likely to incur higher annual inpatient health care costs than current smokers. Pronk *et al.* reached a similar conclusion in respect of the association between smoking status and health care costs (Pronk *et al.* 1999). They suggest that the higher health care costs incurred by smokers as compared to former smokers may be attributed to health issues that prompted smoking cessation, for example, following a myocardial infarction. The higher

costs for non-smokers may also be attributed to the childbirth costs which may be higher among non-smoking females.

Interventions to influence modifiable health risk factors are costly and resource intensive (Pronk *et al.* 1999). The results of this study indicate that resources would be better utilised if such interventions were focused on weight loss and physical activity rather than smoking cessation.

A limitation of this study is that the BMI measure, smoking status and the level of physical activity were all obtained via the voluntary, self-reported HRA. The literature indicates the inaccuracies of self-reported height and weight, in particular the tendency of respondents to overstate their height and understate their weight (Durden *et al.* 2008; Palta *et al.* 1982; Rowland 1990). The over-reporting of physical activity due to social desirability is also indicated in the literature (Adams *et al.* 2005). Such potential bias in the reporting of the data suggests that the results be interpreted with caution.

In summary, this study suggests that managed care interventions aimed at positively influencing modifiable health risk factors will provide an effective mechanism to health care funders in curbing the increasing cost of health care.

CHAPTER 9: POSSIBLE EXTENSIONS TO THIS STUDY

Several extensions to this analysis are possible. Stepwise regression techniques, as applied in this study, are influenced by random variation in the data and hence seldom yield replicable results if the model is retested (Field, Miles 2010). Further studies could consider cross-validation of the estimated models, that is estimating the model on one part of the data to assess predictive accuracy on the remaining part, as this will be useful in providing assurance that the conclusions of the study may be generalised beyond the current dataset.

Future studies should consider the association between modifiable health risk factors and non-elective health care costs. This would then exclude, for example, costs that are associated with childbirth and elective dental procedures.

In the estimation of the logistic regression model, only a subset of all 2-way interactions of the covariates proved to be significant. Further studies that investigate the 2-way interactions that were not significant could yield useful insight into the nature of the relationships between these covariates. In so doing, the results may suggest alternate intervention strategies to influence modifiable health risk factors.

A longitudinal study of the data that tracks individuals, who have completed HRA questionnaires in multiple years, would be useful in quantifying the time lag for

intervention strategies, such as wellness programmes, to influence behaviour in respect of modifiable health risk factors.

REFERENCES

- Adams SA, Matthews CE, Ebbeling CB, Moore CG, Cunningham JE, Fulton J, Hebert JR. 2005. The Effect of Social Desirability and Social Approval on Self-Reports of Physical Activity. *American Journal of Epidemiology*, **161(4)**: 389 – 398.
- Allison PD. 2006. *Logistic Regression using SAS: Theory and Application*. Cary, North Carolina: SAS Institute.
- Barber J, Thompson S. 2004. Multiple regression of cost data: use of generalised linear models. *Journal of Health Services Research & Policy*, **9(4)**: 197 – 204.
- Belsley DA, Kuh E, Welsch RE. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons, Inc.
- Blough DK, Madden CK, Hornbrook MC. 1999. Modeling risk using generalized linear models. *Journal of Health Economics*, **18(2)**: 153 – 171.
- Blough KB, Ramsey SD. 2000. Using generalized linear models to assess medical care costs. *Health Services & Outcomes Research Methodology*, **1(2)**: 185 – 202.
- Briggs A, Gray A. 1998. The distribution of health care costs and their statistical analysis for economic evaluation. *Journal of Health Services Research & Policy*, **3(4)**: 233 - 245.

- Brink SD. 1987. *Health risks and behaviour: The impact on medical costs*. Seattle, WA.: Milliman & Robertson.
- Buntin MB, Zaslavsky AM. 2004. Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures. *Journal of Health Economics*, **23**: 525 – 542.
- Cawley J. 2004. An economic framework for understanding physical activity and eating behavior. *American Journal of Preventive Medicine*, **27(3)**: 1 – 9.
- Chapman LS. 1998. The role of incentives in health promotion. *American Journal of Health Promotion*, **2**: 12 – 23.
- Chatterjee S, Hadi AS. 2006. *Regression Analysis by Example*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Diehr P, Yanez D, Ash A, Hornbrook M, Lin DY. 1999. Methods for analyzing health care utilization and costs. *Annual Review of Public Health*, **20**: 125 – 144.
- Duan N, Manning W, Morris CN, Newhouse JP. 1983. A comparison of alternative models for the demand for medical care. *Journal of Business & Economic Statistics*, **1(2)**: 115 – 126.
- Durden ED, Huse D, Ben-Joseph R, Chu B. 2008. Economic Costs of Obesity to Self-Insured Employers. *Journal of Occupational and Environmental Medicine*, **50(9)**: 991 – 997.
- Field A, Miles J. 2010. *Discovering Statistics using SAS*. London: SAGE Publications Ltd.

- Fishman PA, Khan ZM, Thompson EE, Curry SJ. 2003. Health care costs among smokers, former smokers, and never smokers in an HMO. *Health Services Research, 38(2)*: 733 – 749.
- Fuchs V. 1974. *Who shall live? Health, economics & social choice*. New York: Basic Books.
- Galper DI, Trivedi MH, Barlow CE, Dunn AL, Kampert JB. 2006. Inverse association between physical inactivity and mental health in men and women. *Medicine & Science in Sports & Exercise, 38*: 173 – 178.
- Gebhardt DL, Crump C. 1990. Employee fitness and wellness programs in the workplace. *The American Psychologist, 45(2)*: 262 – 272.
- Goetzel RZ, Anderson DR, Whitmer RW, *et al.* 1998. The relationship between modifiable health risks and health care expenditures. An analysis of the multi-employer HERO health risk and cost database. The health enhancement research organization (HERO) research committee. *Journal of Occupational & Environmental Medicine, 40(10)*: 843 – 854.
- Grossman M. 1972. On the concept of health capital and the demand for health. *Journal of Political Economy, 80*: 223 – 255.
- Hill T, Lewicki P. 2007. *STATISTICS Methods and Applications*. Tulsa, OK.: StatSoft.
- Hornbrook MC, Goodman MJ. 1991. Managed care: Penalties, autonomy, risk and integration. In H. Hibberd, P.A. Nutting and M.L. Grady (eds). *Primary Care Research: Theory and Methods*, pp. 107-26. Washington DC: US Department of Health and Human Services, Public Health Service, Agency for Health Care Policy and Research.

- Izumi Y, Tsuji I, Ohkubo T, Kuwahara A, Nishino Y, Hisamichi S. 2001. Impact of smoking habit on medical care use and its costs: a prospective observation of National Health Insurance beneficiaries in Japan. *International Journal of Epidemiology*, **30**: 616 – 621.
- Jones AM. 2000. Health econometrics. In JP Newhouse, & AJ Culyer (eds). *Handbook in health economics*, Vol. 1A, pp 265 – 344. The Netherlands: Elsevier Science B.V.
- Kenkel D. 1995. Should you eat breakfast? Estimates from health production functions. *Health Economics*, **4(1)**: 15 – 29.
- Knowler WC, Barrett-Conner E, Fowler SE, *et al.*; for the Diabetes Prevention Program Research Group. 2002. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *New England Journal of Medicine*, **346**: 393 – 403.
- Kohl HW. 2001. Physical activity and cardiovascular disease: evidence for a dose response. *Medical Science Sports Exercise*, **33**: S472 – S483.
- Lambert EV, da Silva R, Patel D, Fatti L, Kolbe-Alexander, T *et al.* 2009 Fitness-related activities and medical claims related to hospital admissions – South Africa, 2006. *Preventing Chronic Disease*, **6(4)**: 1 – 9.
- Lin T. 2008. Modifiable health risk factors and medical expenditures – The case of Taiwan. *Social Science & Medicine*, **67**: 1727 1736.
- Manning WG. 1998. The logged dependent variable, heteroscedasticity, and the transformation problem. *Journal of Health Economics*, **17**: 283 – 295.

- Matsaganis M, Mitrakos T, Tsakloglou P. 2009. Modelling health expenditure at the household level in Greece. *European Journal of Health Economics*, **10**: 329 – 336.
- McCullagh P, Nelder JA. 1983. *Generalized Linear Models*. London: Chapman and Hall.
- McCullagh P. 1984. Generalized linear models. *European Journal of Operational Research*, **16**: 285 – 292.
- Mullahy J. 1998. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics*, **17**: 247 – 281.
- Palta M, Prineas RJ, Berman R, Hannan P. 1982. Comparison of self-reported and measured height and weight. *American Journal of Epidemiology*, **115**: 223 – 230.
- Patel DN, Lambert EV, da Silva R, Greyling M. 2010. The association between medical costs and participation in the Vitality health promotion program among 948,974 members of a South African health insurance company. *American Journal of Health Promotion*, **24(3)**: 199 – 204.
- Patrick DL, Cheadle A, Thompson DC, Diehr P, Koepsell T, Kinne S. 1994. The validity of self-reported smoking: a review and meta-analysis. *American Journal of Public Health*, **84(7)**: 1086 – 1093.
- Pratt M, Macera CA, Wang G. 2000. Higher direct medical costs associated with physical inactivity. *Physician Sports Medicine*, **28**: 63 – 70.
- Pregibon D. 1981. Logistic regression diagnostics. *The Annals of Statistics*, **9(4)**: 705 – 724.

- Pronk NP, Goodman MJ, O'Connor PJ, Martinson BC. 1999. Relationship between modifiable health risks and short-term health care charges. *Journal of the American Medical Association*, **282(23)**: 2235 – 2239.
- Rowland ML. 1990. Self-reported weight and height. *American Journal of Clinical Nutrition*, **52**: 1125 – 1133.
- SAS Institute. 1990. SAS Procedures Guide, Version 6, 3rd ed. Cary, NC: SAS Institute.
- Sarkar SK, Midi H and Rana S. 2010. Detection of outliers and influential observations in binary logistic regression: An empirical study. *Journal of Applied Science*, **11**: 26 – 35.
- Serxner SA, Gold DB, Grossmeier JJ, Anderson DR. 2003. The relationship between health promotion program participation and medical costs: A dose response. *Journal of Occupational and Environmental Medicine*, **45(11)**: 1196 – 1200.
- Still L. 1974. *Health Care in South Africa 2007*. Johannesburg: Profile Media.
- Stokes ME, Davis CS, Koch GG. 2006. *Categorical Data Analysis using the SAS System*. Cary, North Carolina: SAS Institute.
- Thorpe KE. 2005. The rise in health care spending and what to do about it. *Health Affairs*, **24(6)**: 1436 – 1445.
- Wang F, McDonald T, Bender J, Reffitt B, Miller A, Edington DW. 2006. Association of Healthcare Costs with Per Unit Body Mass Index Increase. *Journal of Occupational and Environmental Medicine*, **48(7)**: 668 – 674.
- Watt D, Verma S, Flynn L. 1998. Wellness programs: a review of the evidence. *Canadian Medical Association Journal*, **158(2)**: 224 – 230.

- Wilkerson GB, Boer NF, Smith CB, Heath GW. 2008. Health-related factors associated with the health care costs of office workers. *Journal of Occupational and Environmental Medicine*, **50(5)**: 593 – 601.
- World Health Organisation. 1997. *Obesity: preventing and managing the global epidemic*. Geneva: WHO. [WHO Technical Report Series: No. 894.]

APPENDIX:

HEALTH PLAN OPTIONS OFFERED BY DISCOVERY HEALTH

Series	Plan	In-hospital cover	Out-of-hospital cover	Chronic and oncology cover
Executive	Executive Plan	★ ★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★
Comprehensive	Classic Comprehensive	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★
	Essential Comprehensive	★ ★ ★ ★ ★	★ ★ ★ ★	★ ★ ★ ★ ★
	Classic Delta Comprehensive network option	★ ★ ★ ★ ★	★ ★ ★ ★ ★	★ ★ ★ ★ ★
	Essential Delta Comprehensive network option	★ ★ ★ ★ ★	★ ★ ★ ★	★ ★ ★ ★ ★
	Priority	Classic Priority	★ ★ ★ ★ ★	★ ★ ★ ★
Priority	Essential Priority	★ ★ ★ ★ ★	★ ★ ★	★ ★ ★
Saver	Classic Saver	★ ★ ★ ★ ★	★ ★ ★	★ ★ ★
	Essential Saver	★ ★ ★ ★ ★	★ ★	★ ★ ★
	Coastal Saver	★ ★ ★ ★ ★	★ ★ ★	★ ★ ★
	Classic Delta Saver network option	★ ★ ★ ★ ★	★ ★ ★	★ ★ ★
	Essential Delta Saver network option	★ ★ ★ ★ ★	★ ★	★ ★ ★
Core	Classic Core	★ ★ ★ ★ ★		★ ★ ★
	Essential Core	★ ★ ★ ★ ★		★ ★ ★
	Coastal Core	★ ★ ★ ★ ★		★ ★ ★
	Classic Delta Core network option	★ ★ ★ ★ ★		★ ★ ★
	Essential Delta Core network option	★ ★ ★ ★ ★		★ ★ ★
KeyCare	KeyCare Plus	★ ★ ★ ★ ★	★ ★	★ ★ ★
	KeyCare Core	★ ★ ★ ★ ★		★ ★ ★
In-hospital cover	★	The blue star indicates that you have cover in any hospital. The number of stars refers to the level of cover you have in hospital.		
	★	A circled star indicates that to get five-star cover in hospital, you need to use a provider in one of Discovery's networks or payment arrangements – either a network hospital or a doctor participating in our direct payments arrangements, or both.		
Out-of-hospital cover	All plans except KeyCare To show the level of cover each plan offers for out-of-hospital and day-to-day benefits, we've added together the star-ratings for each of the following:			
	★	A Medical Savings Account equal to 15% of your total medical scheme contribution.		
	★★	A Medical Savings Account equal to 25% of your total medical scheme contribution.		
	★	An additional star if the plan offers the Insured Network Benefit to extend your day-to-day cover.		
	★	An additional star if the plan offers a limited Above Threshold Benefit.		
	★★	Two additional stars if the plan offers an unlimited Above Threshold Benefit.		
	KeyCare Plus			
	★ ★	Your plan offers cover in a network for certain day-to-day healthcare costs.		
Chronic illness benefit and oncology	★ ★ ★	You have cover for the conditions listed in the Prescribed Minimum Benefits, but you must use our designated service provider to get your approved chronic medicines, or you will have a co-payment. You have full cover for medicines on our list – if you choose a different medicine, you have cover up to a monthly rand amount, except on KeyCare. You must use our designated service provider for cancer treatment.		
	★ ★ ★	You have cover for the conditions listed in the Prescribed Minimum Benefits – you can get your approved chronic medicine from any provider. You have full cover for medicines on our list – if you choose a different medicine, you have cover up to a monthly rand amount. If one of the three stars is circled, you must get your approved medicine from our designated service provider or you will have a co-payment.		
	★ ★ ★ ★ ★	You have cover for the conditions listed in the Prescribed Minimum Benefits, as well as for conditions on an additional disease list. You have full cover for medicines on our list – if you choose a different medicine, you have cover up to a higher monthly rand amount than on other plans. You also have higher cover for cancer treatment.		