THE UNIVERSITY OF CAPE TOWN

SCHOOL OF ECONOMICS

MCOM(ECONOMICS)

MINI-THESIS

“Gender And Life-Cycle differentials In The Correlates Of Adult Ill Health In South Africa”

BY MASHEKWA MABOSHE

SUPERVISOR: A/PROF INGRID D. WOOLARD

FEB, 2011

ABSTRACT

This study investigates the gender and life-cycle differentials in the socio-economic covariates of adult self-evaluated ill health in South Africa using data from the 2008 National Income Dynamics Study wave 1 survey. The study employs the Grossman (1972) model on health capital and the demand for health as the underlying theoretical framework and estimates the results using an ordered probit adjusted for survey design. Results show that among adults aged 14-49 years, the impact of schooling on reducing the probability of reporting ill health is greater for women than for men, but that among those aged 50 years and above, schooling has a greater gradient for men than for women. These results persist even after controlling for other socioeconomic variables. Results further show that employment reduces the probability of reporting ill health, with employed men having the greatest benefits compared to women. The gender differences in this effect are significant, suggesting that males may have better access to health augmenting employment related benefits such as better medical aid compared to their female counterparts. Location of residency matters. Adults living in informal townships are more likely to report ill health relative to those living in urban planned settlements or in rural areas. This finding may suggest that poor water and sanitation and a lack of health services in informal areas may be the intermediary variables through which residential differentials affect health status.

Finally, as in other countries, we find that South African women in general report higher cases of ill health compared to their male counterparts, despite controlling for all other socio-economic variables.

1.0 INTRODUCTION

Demographic and epidemiological projections show that South Africa's population structure is aging; and that the disease profile is evolving rapidly, in common with other low and middle income countries. Over the period 2000 to 2010, the proportion of the South African population 15 years and older increased from 66% to 71%, with overall mortality increasing by 118%. While mortality due to infectious diseases and injuries dropped by 9% and 3% in 2010 respectively, mortality due to non-communicable chronic diseases – the leading cause of mortality in 2000 - increased by 18%. Mortality due to HIV/AIDS was estimated to have quadrupled, thereby becoming the leading cause of mortality in 2010 (US Bureau of the Census, 2010; Steyn et al, 2006). This demographic and epidemiological transition suggests that the incidence of ill health among adults is likely to increase as the population ages and as the disease burden falls more on the adults. Adult ill health could pose significant direct productivity losses and indirect costs in an economy, as well as place large demands on the already over-stretched health systems in developing countries (Strauss et al, 1993). Yet, despite the policy relevance of adult ill health, little research has been done to understand the socio-economic covariates of health in developing countries and South Africa in particular.

This paper contributes to filling this gap by investigating the gender and life-cycle determinants of adult self-reported health status using data from the 2008 National Income Dynamics Study (NIDS). The paper follows the Grossman (1972) theoretical model on health and adopts the framework in Strauss (1993) to investigate the gender and life-cycle correlates of adult health status in South Africa. We use self-reported health status as the measure of health. Epidemiological studies have shown that self-perceived health status is a good indicator of objective tests and that the measure reflects a person's overall perception of health including the biological, psychological and social dimensions that are inaccessible to physicians (Miilunpalo et al, 1997).

Self-reported health status was chosen as a proxy for health status because this measure is easier to collect in practice, relative to conducting physical medical tests. This is especially relevant in the context of a developing country such as South Africa where formal medical tests would imply unaffordable costs. Questions which simply ask how a respondent perceives or evaluates their health provide an easy and economical measure of health status.

This paper is organised as follows: The next section reviews the theoretical and empirical literature on the determinants of self-evaluated health status in developing countries. Section 3 specifies the model and estimation procedures while section 4 describes the data and variables. Section 5 presents and discusses the results, and the conclusion is presented in section 6.

2.0 LITERATURE REVIEW

2.1 Review of theory

This study adopts Grossman's (1972) model on health capital and the demand for health.[1] The theoretical model is particularly appealing to our study because it provides the relevant economic theory on the determinants of health status.

The central proposition of the Grossman model (1972) is that health can be viewed as a durable capital stock that produces an output of healthy time. The model asserts that individuals' demand for health services is derived from the more fundamental demand for "good health". The ensuing health is then used for consumption and investment purposes. As a consumption commodity, health directly enters an individual's preference function thereby affecting utility. As an investment commodity, health determines the total amount of time available for market activities, thereby affecting the monetary value of returns on investments in health (Grossman, 1972). In other words, Grossman argues that individuals have an incentive to invest in their stock of health so that they produce more time for leisure and productivity purposes.

In the model, an individual's health stock is produced by a household production function that is determined by one's *initial stock of health*, *behavioural choices* (i.e. work activity and exercise, or choice of health services) and *"environmental variables"* (factors such as education that influence the efficiency of the production process) (Grossman, 1972). More compactly, the general Grossman health production function is given as;

$$H_t = H(H_{t-1}, X_{ht}, \quad _{ef}, \mathcal{E}_t)$$

[1] The Grossman model is an extension of Schultz's (1961) and Becker's (1965) human capital theories. These theories attempt to explain the importance of human capital investment in increasing labour productivity and augmenting the role of nonhuman conventional capital in economic growth and development.

where H_t is an individual's current health stock at time t; H_{t-1} is the inherited health stock; X_{ht} - is behavioural choices and health services; $_{ef}$ is a vector of environmental factors; and $\mathcal{E}_t$ represents unobserved individual endowments (Grossman, 1972).

A key assumption of the model is that individuals aim to maximise their utility, investing in their health production until the marginal cost of health production equals the marginal benefit of improved health status. Given that the outcome of the investment in health capital is healthy time and longevity, a lack of investment leads to loss of quality time and reduced length of life. The model employs this conception in discussing the theoretical determination of health. Grossman (1972) discusses the effects of education on the health production function. An increase in education *(*an *"environmental variable")* is assumed to increase productivity in the production of health, leading to a higher health stock[2] and therefore increased utility (satisfaction) and income earnings (due to labour productivity) across the lifetime. From this view, more education would therefore improve an individual's health stock - and therefore health status. The effect of age on health stock is discussed in the Grossman model through the depreciation effects of age on the health production function.[3] As individuals age, their physical and mental capacity naturally deteriorates thereby reducing their net health stock over time. From the model, age is therefore negatively correlated with health stock and therefore health status.

Income (wage) effects were analysed by Grossman (1972). The model posits that individuals with higher incomes have a greater incentive to invest in health stock and produce healthy time. The proposition is two-fold -Firstly, an individual's wage rate measures his market efficiency and earnings potential. High wage earners have a greater incentive to invest in health to reduce lost time due to illness, thereby maximizing their earnings in the market sector. A second perspective is that a high wage would induce the substitution of market goods for own non-market leisure activities. Therefore, whether by directly benefiting from a higher market return, or from non-market leisure utility, Grossman's model demonstrates that higher incomes induce investment in health stock, thereby improving health status (Grossman, 1972).

[2] In the model, stock of health capital embodies the concept of health – a higher stock implies better health status.

[3] Grossman assumes an exogenous depreciation rate that increases over time. Net investment in health capital is negatively affected by depreciation $H_{i+1} - H_i = I_t - \delta H_{i.}$ were I_t, δ, H_{i+1} and H_i are investment, depreciation and health stock in periods *t and t+1* respectively.

By the same analogy, Grossman's (1972) model can be extended to analyse the effect of any socioeconomic variables on health status. This study applies and extends the Grossman (1972) theoretical framework in explaining the direct or indirect effects of socioeconomic variables on health status.

2.2 Empirical Literature

Empirical studies on the socioeconomic determinants of self-evaluated health status in Africa, and South Africa in particular are very scanty. However, related studies done in other developing economies were reviewed and are discussed below.

Strauss et al (1993) studied the socioeconomic determinants of adult ill health in Jamaica, specifically focussing on the gender and life-cycle differentials in the patterns and determinants of adult-health. The study used Grossman's (1972) model as the underlying theoretical framework and employed an ordered probit estimation technique. The study used data from the 1989 Jamaican survey of living conditions. Self-reported health status and limitations to physical functioning were used as measures of health with selected socioeconomic variables as explanatory variables.

Among the key results were the findings that, for both men and women, own education reduces the probability of reporting a health problem. However, education effects seemed to dissipate as age increased. Residential effects were also found to be significant in reporting health status. Adults living in urban areas, particularly women, reported better health status compared to those living in rural areas. This finding seemed to indicate an explanatory role for the heterogeneous influence of health services and infrastructure for health outcomes of women and men. The most robust finding of the study was that strong gender differentials exist over the lifetime - with women reporting significantly more health problems with physical functioning and general health across all ages. This is despite the greater life expectancy of women compared to men. For both general health and problems with physical functioning at varying levels, Strauss et al (1993) found that women begin to report significantly greater problems by the time they are 25 to 29 years, both relative to men and younger women. These reporting differentials tend to grow thereafter, until about 65 years, after which they decline (Strauss et al 1993).

In 1998, Handa conducted a study that used the same methodology and dataset as Strauss et al (1993) (see above), but specifically focussing on the gender and life-cycle differences in the impact of schooling on chronic disease in Jamaica. The objective of the study was to test whether the determinants and patterns of physical health as reported by Strauss et al (1993) were the same for chronic disease. In particular, the study sought to find out whether schooling played the same role in determining chronic disease as it did in determining physical health status (Handa, 1998). The study used the same methodology and independent variables as Strauss et al (1993) but used an indicator variable for chronic disease as a measure of health. Six chronic diseases; asthma, arthritis, diabetes, fits (epilepsy), heart disease and hypertension were used in creating the dummy variable. As in Strauss et al (1993), Handa (1998) found that schooling significantly reduces the probability of reporting chronic disease, with the impact slightly higher for females compared to males, although the null hypothesis of gender differences could not be rejected. The effect of schooling persisted even after controlling for other socio-economic factors including income. When stratified by broader age groups, results showed that the impact of schooling was higher for younger adults (14-49 years) compared to the elders (50 years and above). Income and partner effects were not found to be significant, although living with a partner seemed to reduce the probability of reporting chronic disease. As in Strauss et al (1993), the study found that the incidence of chronic disease rises with age, but for women the increase begins much earlier around age 20 compared to men who begin reporting chronic illness at about 40 years (Handa, 1998).

Another study reviewed was one by Power et al (1997) whose main objective was to determine whether social differences in health persist or widen during early adulthood. The study used a longitudinal follow-up of the 1958 British Birth Cohort at age 23 (in 1981) and 33 (in 1991). Analysis of health inequalities was done using logistic regressions of the probability of ill health on social status at birth. Health was measured by six variables; self-rated health, long-standing illness limiting daily activity, psychological distress, respiratory symptoms, asthma (or wheezing) and body mass index. Social status at birth was measured by father's occupation. The social classes were ranked on a 0-1 scale - with 0 representing professional and managerial groups and 1 representing unskilled manual labour (other non-manual and skilled manual groups took intermediary values on the scale). Logistic regressions were then analysed at age 23 and at age 33 for each of the health measures and the results were stratified by gender. Results at age 23 showed that prevalence of poor health increased with decreasing social positions, and these results were evident for all health measures except for long-standing limiting illness and asthma. The social gradients seemed to persist to age 33 for those health measures showing gradients at age 23 but for measures such as self-reported health, the increase was insignificant. Health inequalities in overweight and obesity, and malaise tended to reduce although the reduction was statistically insignificant. The results from the logistic analyses led to the conclusion that social gradients in health were evident in the British Birth Cohort by age 23 and persisted to age 33, but the inequalities did not appear to widen consistently (Power et al, 1997).

In Malawi, Doctor (2001) conducted a study that focussed on the determinants of self-reported health in rural Malawi. The study used multivariate ordinary least squares (OLS) as the estimation method[4], with the categorical dependant variable transformed into logarithms. The paper estimated the effects of nine covariates of health; gender, age, education, household possessions, value of livestock (proxy for household income), number of surviving children, house type (proxy for wealth), residency and health symptoms on reported health status. The study found that age and being a woman were negatively associated with reported health status, a finding consistent with results from Strauss et al (1993). Contrary to expectation, education, number of surviving children and household income were found to be insignificant in rural Malawi. A key finding of the paper was that the effects of reported symptoms were robust before and after controlling for socio-demographic factors. In fact, results showed that all types of symptoms were negatively associated with health status. Regional differentials were significant in rural Malawi, with the central region reporting relatively better health status compared to the southern region (Doctor, 2001).[5]

Finally, another study reviewed was by Ramirez et al (2004) on the determinants of health status in Colombia- following a major health policy reform from a public to a market based insurance system[6]. The paper follows the Grossman theoretical framework and uses the ordered probit as the estimation technique. Various individual and household level, geographical and institutional (insurance scheme) variables were estimated. Results on age, gender, education and household income were consistent with theory and

[4] Reasons for using linear regression on a model involving a discrete outcome are not mentioned in this paper. However, according to Green (2000), the use of linear regression for non-linear discrete choice models may be inappropriate and could lead to inconsistent estimations (Greene 2000).

[5] Doctor (2001) notes that the central region is characterized by large-scale tobacco farming with higher incomes relative to the southern region - that is characterized by fishing activities whose productivity and incomes has been declining due to over fishing.

[6] Following the 1994 reforms, two types of regimes were implemented; i) *A contribution regime* – A scheme for all beneficially employed or self employed citizens , and ii) *A subsidized regime* – A scheme for the financially needy citizens which is fully or partly subsidized depending on need.

empirical findings from Doctor (2001) and Strauss et al (1993). Employment had a positive impact on health status with the employed having a lower probability of reporting regular or bad health compared to the unemployed, and students. Results on socioeconomic status indicated that respondents whose houses had more rooms, or whose homes used electricity or natural gas as opposed to other energy sources, had higher probabilities of reporting good health. The effects of residential effects were similar to results from studies by Doctor (2001) and Strauss et al (1993) with respondents in Bogotá and Antioquia(– the capital and richest state), having higher probabilities of reporting better health compared to respondents from the rest of the country. The most important finding of the study was that the type of health insurance scheme membership matters. People with affiliation to the contributory scheme were associated with higher probabilities of reporting good health compared to people under the subsidised scheme. Ramirez et al (2004) suggests that this may be due to the different medical packages under the two systems. Contributory systems typically provide much wider and better quality health services compared to subsidised regimes (Ramirez et al 2004).

From the review of the above theoretical and empirical literature, this study will follow the Grossman (1972) model as the explanatory framework and will use the ordered probit estimation to understand the socioeconomic determinants of adult self-evaluated health in South Africa. Further, the study will adopt the general framework in Strauss et al (1993).

3.0 SPECIFICATION OF THE MODEL AND ESTIMATION METHOD

3.1 Model

The model in this paper is adopted from Grossman's (1972) household production model of health production; and the stylized dynamic version in this paper is adopted from Strauss et al (1993).

Let an individual's health stock be more elaborately presented by the following health production function, which transforms inputs such as individual behaviours, into health:

$$H_t = H(H_{t-1}, X_{ht}, \mu_{ft}, \mu_{ct}, \varepsilon_t)$$

Where H_t is health at time t, X_h is a vector of health related inputs, μ_f is a vector of individual and family characteristics such as age, education etc; μ_c is a vector of community characteristics such as quality of health infrastructure, ε is a vector of unobserved individual endowments. An individual in period t is assumed to maximize a weakly-time separable utility function, U_t, defined over a vector of health stocks, H_t, consumption goods, X_{ct} and leisure $l_{t.}$ given the household characteristics ν ;

$$U_t = U(H_t, X_{ct}, L_t; \nu)$$

Subject to a current period budget constraint which relates current wealth to the present value of wealth from the previous periods, plus savings and net borrowing;

$$W_t = W_{t-1} + (1+r)(Y_t - P_t X_t) + B$$

Where Y is household income, P_t is a vector of prices, X the vector of goods purchased; r is the time-invariant interest rate while B is the net borrowing. Applying the standard assumption that individuals maximize their utility subject to their resource and environmental constraints, the reduced form health equation can then be derived as;

$$H_t = H(H_o, W_o, u_f, u_c)$$

Where H_o, W_o , μ_f and μ_c represent the initial health and wealth endowments, the individual, family and community characteristics (respectively).

3.2 Estimation Method

i) Ordered probit

The estimation technique used is the ordered probit. The method was adopted because the dependent variable "*self-reported health status",* is an ordered categorical variable – the response outcomes take on values from 0 to 4, if the person's heath status is poor, fair, good, very good, or excellent in that order. In this instance, the traditional linear regression model cannot be used because the error term has been shown to be heteroscedastic (Woodridge 2002; Greene 2000).[7] [8] Further, the resulting probability estimations may lie outside the 0-1 bounds, thereby violating the underlying probability theory (Woodridge 2002). These weaknesses may lead to misleading interpretations. To circumvent these problems, the ordered probit model which follows a standard normal probability distribution function and has a white noise error term is

[7] Heteroscedasticity violates the assumptions of the classical linear regression model and results in inconsistent estimations.

[8] Greene shows that from the familiar linear regression model F(X,B)=X'β, the variance of the error term is heteroscedastic as it depends on **β** values i.e Var[ε|**x**] =**x' β**(1- **x' β**) (Greene 2000; 665).

often used as a preferred discrete choice model (Woodridge 2002; Green 2000). The probit model constrains the resulting probabilities within the theoretically correct 0-1 probability range (Woodridge 2002).

Although an appropriate model for discrete choice outcomes, authors of econometric models observe that the underlying assumptions of probit estimation make coefficient interpretation difficult (Long 1997).[9] The adopted practice is to interpret the sign of the coefficient (+/-) in terms of an increase or decrease in the associated outcome probabilities without too much concern on the magnitudes. However, techniques to transform and interpret the magnitudes of coefficients have been formulated and now available in standard software. These techniques involve the conversion of coefficients into standardized beta coefficients associated with the latent variable or the computation of marginal effects on predicted probabilities thereby making interpretation of coefficients meaningful (Long 2001; Long and Freese 2006).[10]

ii) Econometric specification

Let *y* be a discrete variable defined as y = {0,1,2,3,4} if a respondent perceives their health as poor, fair, good, very good or excellent respectively. Following Wooldridge (2002), the ordered probit model for y conditional upon x, can be derived from a latent variable model. Assume that a latent variable *y** is determined by;

$$y^* = \mathbf{x\beta} + e, \qquad e|\mathbf{x} \sim \mathbf{N}(0,1).$$

The health *condition* is unobserved (latent), but an individual report of self-reported health status is taken as an index of measure. This outcome is defined as;

$$\begin{array}{lll} y = 0 & \text{If} & y^* \leq \alpha_1 \\ y = 1 & \text{If} & \alpha_1 < y^* \leq \alpha_2 \\ y = 2 & \text{If} & \alpha_2 < y^* \leq \alpha_3 \\ y = 3 & \text{If} & \alpha_3 < y^* \leq \alpha_4 \\ y = 4 & \text{If} & y^* > \alpha_4 \end{array}$$

[9] The dependent variable is assumed to be an unobserved latent variable whereas the coefficient estimates are based on the actual observable data. So, unlike linear models where both the regressand and regressors are observable, direct coefficient interpretation in oprobit is meaningless or unclear because the latent variable is unobservable (Long 2001). The difficulty is also compounded by the assumption that the model follows a non linear cumulative standard normal distribution function which implies that unlike linear models where coefficients are constant, ordered probit coefficients are constantly changing thereby making exact interpretations hard to make (Long & Freese, 2006).

[10] The underlying econometric theory and estimation computation is involved. Interested readers may consult Long and Freese (2006, Chapter 5) and Greene (2000, Chapter 21) for a detailed discussion and further references.

were $\alpha_1 < \alpha_2 \cdots < \alpha_4$ ($\alpha_i > 0$, for all i) are the unknown cut off points (or threshold parameters) which are estimated together with the betas. Given the standard normal assumption of the error term, the probabilities of observing each response category are obtained as;

$$\begin{aligned}
\text{Prob}(y=0|x) &= \Phi(\alpha_1 - X\,\beta) \\
\text{Prob}(y=1|x) &= \Phi(\alpha_2 - X\,\beta) - \Phi(\alpha_1 - X\,\beta) \\
\text{Prob}(y=2|x) &= \Phi(\alpha_3 - X\,\beta) - \Phi(\alpha_2 - X\,\beta) \\
\text{Prob}(y=3|x) &= \Phi(\alpha_4 - X\,\beta) - \Phi(\alpha_3 - X\,\beta) \\
\text{Prob}(y=4|x) &= 1 - \Phi(\alpha_4 - X\,\beta)
\end{aligned}$$

For each i^{th} response, the parameters of the model, α and β, can now be estimated using maximum likelihood estimation from the probit log likelihood function[11] using the statistical software STATA.

4.0 DATA DESCRIPTION AND DEFINITION OF VARIABLES

4.1 DATA

Data for this study came from the 2008 NIDS survey (wave 1) – South Africa's first nationally representative panel survey whose objective is to understand the dynamic structure of households and document changes in the general well-being of individuals and households in South Africa. The stratified sample consisted of 28,255 respondents from 7,305 households. This study used the adult survey, a subset of NIDS that had 16,885 observations.

4.1.1 Sampling design

A stratified, two-stage cluster sample design was employed in sampling the households to be included in the base wave. The NIDS sample is based on Statistics SA's 2003 master sample of 3000 PSU[12]. The explicit strata in the Master Sample are the 53 district councils (DCs). The sample was proportionally allocated to the strata based on the master sample DC PSU allocation and 400 PSUs were randomly selected within strata. Thereafter, an initial 9600 dwelling units were drawn from the various PSUs and all households living at selected dwelling units were interviewed (NIDS Metadata, 2009).

[11] $\ell i\,(\alpha,\beta) = 1[y_i=0]\log[\Phi\,(\alpha_1 - X_i\beta)] + 1[y_i=1]\log[\Phi\,(\alpha_2 - X_i\beta) - \Phi\,(\alpha_1 - X_i\beta)] + 1[y_i=2]\log[\Phi\,(\alpha_3 - X_i\beta) - \Phi\,(\alpha_2 - X_i\beta)] + 1[y_i=4]\log[1 - \Phi\,(\alpha_2 - X_i\beta)]$, were Φ is the standard normal density function and each i represents the i^{th} response variable (Greene, 2000).

[12] A PSU is defined as a geographical area that consists of at least one Enumeration Area (EA) or several EAs from the 2001 Census (-when the originally selected EA was found to have less than 74 households).

4.1.2 Survey weights

The NIDS weights were derived in two stages. In the first, the design weights were calculated as the inverse of the probability of inclusion. In the second, the weights were calibrated to the 2008 mid-year population estimates produced by Statistics SA. Design weights took into account the probability of a PSU being selected from the master sample and the probability of interviewing a household. The second set of weights - the post-stratification weights - adjusted the design weights such that the age-sex-race marginal totals in the NIDS data matched the population estimates produced by Stats SA for the mid-year population estimates for 2008 (NIDS Metadata, 2009).

4.2 DEFINITION OF VARIABLES

i) Dependent variable

The dependent variable, *health status*, was measured by a simple question that asked an adult respondent, "how would you describe your health at present…would you say it is excellent, very good, good, fair or poor?" The original survey provided five response options, but these were collapsed to three categories; i) very good or excellent; ii) good and; iii) fair or poor. We collapsed the categories to enable an uncluttered reporting of regression estimates. Empirical studies have found a positive correlation between self-rated health and physicians' medical examinations, and that self-rated health is a better predictor of mortality (Doctor 2001). In addition, subjective health measures are easier, cheaper and quicker to conduct compared to formal medical examinations. Further still, many health studies have shown a high degree of internal consistency - people with chronic illnesses or with illness symptoms tend to report poorer health status as well (Doctor 2001). Although self-reports of health have been found to be reliable, they are associated with problems such as reporting bias (Doctor 2001).[13]

ii) Explanatory variables

The explanatory variables were defined in a manner appropriate for a parsimonious model. Age was treated as a continuous variable with the a priori expectation that increases in age will be correlated with poor health reports. Broad age groups (15-49 and above 50) were included to capture the life-cycle effects of adult ill health. We introduced the square of age (age^2) to test for the quadratic effects of the Grossman

[13] Survey participants may give biased responses due to misperceptions and misunderstandings. For example, the health question may be perceived as an attempt to collect HIV/AIDS related data and this may result in respondents giving biased health scores.

(1972) hypothesis that health stock depreciates at an increasing rate as age increases. Gender was introduced, with male as the categorical reference dummy. A priori, we expect that being female will be associated with more reports of poor health relative to being male, in line with findings from other studies (i.e. Strauss et al 1993; Doctor 2001; Ramirez et al 2004). Our model incorporates race effects. We expect that race will influence self-reported health through more intermediary variables such as education and income rather than race by itself. Given that education and income will be controlled for, we expect that race will be insignificant unless other race specific factors outside this model, such as cultural factors or other, exert their influence through race. The race dummy *African* was used as the base category because Africans made up the largest racial group.

Education was measured by two variables. Firstly, education was measured by years of schooling- to test the linear effect of education on health outcomes. Secondly, education was measured as a set of categorical dummy variables to test for non-linear effects (i.e. no schooling, primary, secondary, tertiary and university level education). In so far as education increases the awareness of risky behavioural factors (e.g. knowledge that high cholesterol diets could lead to chronic illness), we expect that higher education will reduce the probability of reporting ill health (Steyn et al, 2006). Household income was measured by per capita expenditure (pce) in thousands of Rands. This is in line with other studies such as Handa (1997) which argue that expenditure is a better measure of long-run income. A priori, we expect that higher per capita expenditure will be associated with better self-reported health in line with theory and empirical studies (Grossman 1972; Strauss et al 1993). Residential dummies are included in the regressions on the basis that residential effects may work their way through more intermediary factors such as access to better medical or sanitation services. We collapsed residential location into three nominal categories: informal urban, formal urban and rural/traditional authority areas. A priori, we expect that adults resident in formal urban areas will report better health status relative to those living in rural or urban informal settlements. This expectation accords with empirical studies in Colombia and Jamaica (Ramirez 2004; Strauss 1993).

Marital status and adult's employment status were introduced as control variables. We expect that married adults would be less likely to engage in risky social behaviours such as excessive smoking and drinking (Umberton 1987), and would therefore be more likely to report better health outcomes relative to unmarried adults. We introduced an employment dummy that merely indicated whether an adult was employed (regular or casual) or not. Our expectation is that employed adults would have better access to health augmenting employment related benefits such as medical insurance and counselling which the unemployed would not typically have access to. We therefore expect that employed adults will report better health status relative to their unemployed counterparts. Table 1 below presents the summary of the variable definitions;

Table 1: Definition of variables

Variable	Definition
Dependent Variable	
Health Status	Categorical variable classified into 3 health ordinal outcomes; 2=poor or fair; 1 = good; 0 =very good or excellent.
Demographic variables	
Age	Adults' age in number of years
Age squared	Square of age
Gender	Dummy variables; male (1=male, 0=Otherwise)
Race	Race dummies African (0 or 1); Coloured (0 or 1), White (0 or 1) and Asian/Indian (0 or 1)
Socioeconomic variables	
Education level	• Educational dummies; No schooling (0 or 1); Primary (0 or 1) ; Secondary (0 or 1); Higher education (0 or 1) • Years of schooling
Per capita expenditure	Total household expenditure divided by the total number of people living in the household in thousands of Rands ('000).
Residential dummies	Residential dummies: rural/traditional authority (0 or 1); urbanformal (0 or 1); urban informal (0 or 1)
Control Variables	
Marital status	Whether an adult is married or not (0 or 1)
Employment dummy	Whether an adult is employed or not (0 or 1)[14]

5.0 RESULTS AND DISCUSSION OF RESULTS

5.1 DESCRIPTION OF THE DATA

We analysed the raw data to understand the nature and characteristics of the variables in our models. Of the total number of adults who revealed their perceived health status, 58% reported their health as very good or excellent, 23% reported good health while 19% reported ill health. About 25% of the adults living in informal urban areas reported being in ill health compared to 16% of those living in formal urban areas and 20 % of the adults in rural/tribal areas. Respondent's age was invariably related to ill health. About 54% of the adults

[14] Under NIDS, **unemployed** specifically means searching for work whereas **not employed** also includes the not economically active.

in the oldest age group (70 years and older) reported being in poor health compared to 39% in the 50-69 year age group and only 11% in the youngest age group (14-49 years). Furthermore, about 21% of the females reported being in ill health compared to about 15% of the males. Racial patterns in health exist – 22% of the Indians/Asians reported ill health status, compared to 20 % of the Africans, 18% of the Coloured and 13% of the Whites. Married adults reported proportionately more cases of ill health compared to unmarried ones, and employment on the other hand showed that being employed is correlated with lower reports of ill health relative to being not employed.

The table below shows the weighted descriptive statistics of the variables used in the probit regressions

Table 2: Descriptive statistics

	Male		Female	
Variable	Mean	Std. Dev.	Mean	Std. Dev.
Health Status;				
In very good/excellent health	0.638	0.481	0.533	0.499
In good health	0.209	0.407	0.250	0.433
In Poor/fair	0.153	0.360	0.216	0.412
Age	36.301	17.076	38.311	17.583
African	0.816	0.388	0.811	0.392
Coloured	0.076	0.265	0.086	0.281
Asian/Indian	0.015	0.123	0.016	0.127
White	0.092	0.290	0.087	0.281
Level of education;				
No Schooling	0.091	0.287	0.120	0.325
Primary	0.217	0.412	0.193	0.395
Secondary	0.564	0.496	0.569	0.495
Higher education	0.129	0.335	0.117	0.321
Yrs of Schooling	8.793	4.102	8.574	4.283
Percapita expenditure	1.629	3.036	1.421	3.176
Location of residency;				
Urban (formal)	0.521	0.500	0.488	0.500
Rural (Tribal)	0.381	0.486	0.404	0.491
Urban (informal)	0.098	0.297	0.107	0.310
Married	0.345	0.475	0.290	0.454
Employment	0.505	0.500	0.330	0.470

5.2 MAXIMUM LIKELIHOOD PROBIT RESULTS

We estimated maximum likelihood ordered probit regressions of the probability of reporting adult ill health by gender, and then by broader age groups. Estimation by gender allows us to test the gender differences in health, while estimation by age groups allows us to analyse the life-cycle affects. Tables 3 and 4 below present the ordered probit and marginal effects by gender.

Table 3.Ordered probit estimated coefficients

	Male			Female		
	Coef.	Std. Error	t-statistic	Coef.	Std. Error	t-statistic
age	0.058	0.007	7.820	0.050	0.005	9.550
age squared	-0.0003	0.000	-3.800	-0.0003	0.000	-4.710
Coloured	-0.010	0.087	-0.110	-0.152	0.102	-1.490
Asian/Indian	0.399	0.100	3.990	0.356	0.194	1.830
White	-0.152	0.126	-1.210	-0.209	0.156	-1.340
Schooling	-0.046	0.007	-6.950	-0.049	0.007	-7.180
Per capita expenditure	-0.006	0.010	-0.610	-0.017	0.013	-1.280
Rural area	-0.238	0.105	-2.260	-0.384	0.081	-4.730
Formal urban area	-0.120	0.116	-1.030	-0.330	0.095	-3.470
Married	-0.209	0.062	-3.380	-0.175	0.040	-4.340
Employed	-0.275	0.051	-5.390	-0.106	0.050	-2.140
P-values for χ^2 tests for gender differences by:						
age	*(1 df)*	0.406				
Schooling	*(1 df)*	0.754				
employed	*(1 df)*	0.018				
N	6,117			N	9,137	
F(11,336) =	60.08			F(11,334)=	80.43	
Prob > F =	0.0000			Prob > F =	0.0000	

Table 4. Probit estimated marginal effects

	Male			Female		
	dy/dx	Std. Error	t-statistic	dy/dx	Std. Error	t-statistic
age	0.011	0.002	7.500	0.012	0.001	9.290
Age squared	0.000	0.000	-3.700	0.000	0.000	-4.650
Coloured	-0.002	0.017	-0.110	-0.037	0.025	-1.480
Asian/Indian	0.078	0.019	4.090	0.087	0.047	1.840
White	-0.030	0.025	-1.200	-0.051	0.038	-1.340
Schooling	-0.009	0.001	-6.880	-0.012	0.002	-7.210
Percapita expenditure	-0.001	0.002	-0.610	-0.004	0.003	-1.280
Rural area	-0.047	0.021	-2.260	-0.093	0.020	-4.740
Formal urban areas	-0.023	0.023	-1.030	-0.080	0.023	-3.520

married	-0.041	0.012	-3.390	-0.042	0.010	-4.340
employed	-0.054	0.010	-5.450	-0.026	0.012	-2.140

Note: Marginal effects (dy/dx) computed at their means

5.2.1 Life-cycle effects on adult illness

Probit coefficients in table 3 above show that an increase in age is associated with an increase in reports of ill health for both men and women; and the results are statistically significant at 1% in both models. More specifically, marginal effects (table 4) show that, South African women are more likely to report ill health compared to their male counterparts, controlling for all socio-economic variables. A one year increase in a women's age is associated with a 1.2 percentage point increase in the probability of reporting ill health, compared with 1.1 percentage point increase for men. But the null hypothesis for gender differences in this effect cannot be rejected according to the chi-square test statistic in table 3. That the effect of age is significant and positive is consistent with theoretical expectations and empirical evidence from studies in Jamaica, Malawi and Colombia (Strauss 1993; Handa 1998; Doctor 2001; and Ramirez et al 2004). The coefficient of age squared is statistically significant – indicating that age may impact ill health in non-linear ways. Figure 1 below shows the non-linear patterns in which age may impact the probabilities of reporting ill health by gender. The simulated probabilities of reporting ill health were computed separately for each sub-sample, holding all independent variables at their means.

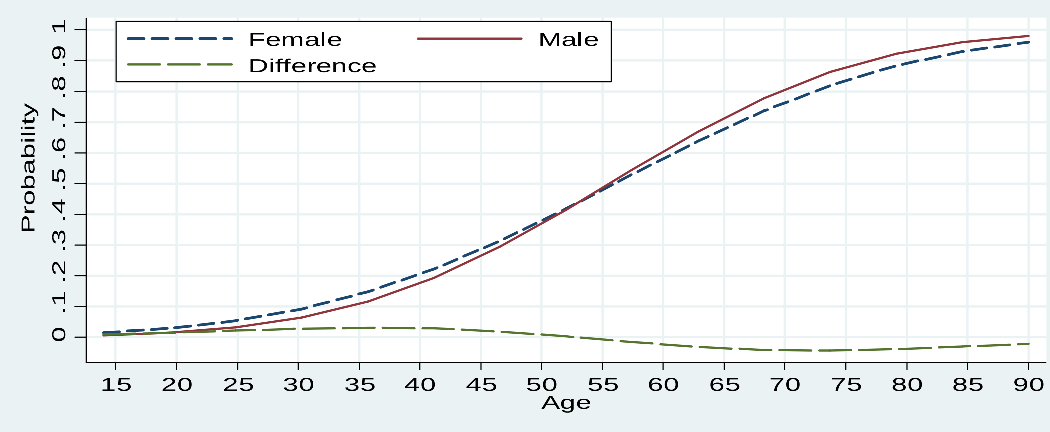


Fig 1. Simulated probabilities of ill health by age and gender

Note: Predicted probabilities based on the separate models as in Table 3 above

As can be seen in Figure 1 above, for both women and men, the probability of reporting ill health increases as age advances, but the increase in reporting incidence is slightly higher for women than men between the ages of 14 and 55 . As age approaches 55 years, differences in reporting probabilities between men and women converge, and thereafter, we observe a reversal in reporting patterns - with men having higher probabilities of reporting ill health compared to women. The reporting differentials between men and women above 55 years tend to increase steadily till about age 75 and thereafter begin to decline until about 90 years when the reporting probabilities tend to converge again. The narrow probability gap between men and women as observed in figure 1 supports the hypothesis that the differential impact of age on health between the genders is insignificant.

5.2.2 Education effects

Probit coefficients in table 3 indicate that education reduces the probability of reporting ill health for both males and females; and the results are robust. The marginal effects in table 4 show that the impact of education in reducing the probability of reporting ill health is higher for women than men. But as in studies by Handa (1998) and Strauss et al (1993), we fail to reject the null hypothesis of gender differences in the effect of schooling according to the chi-square test statistic (see table 3). Therefore, the reported difference in the incidence of ill health cannot be explained by the effect of education.

To gain a better understanding of the gender and life-cycle differentials in the probabilities of reporting ill health, we estimated probit models stratified by gender and by broader age groups; and then we simulated probabilities of reporting ill health at varying levels of education. The results of these estimates and simulations (respectively) are presented in tables 5 and 6 below.

Table 5: Oprobit and marginal coefficients for impact of schooling on ill health; by gender and age groups

	Males				**Females**			
	14-49		**50 years +**		**14-49**		**50 years +**	
Probit coefficients								
	Coef.	*t-stat*	*Coef.*	*t-stat*	*Coef.*	*t-stat*	*Coef.*	*t-stat*
Schooling	-0.038	-4.250	-0.060	-4.570	-0.053	-6.550	-0.036	-3.870
	(0.009)		(0.013)		(0.008)		(0.009)	
Marginal effects								
	dy/dx	*z-stat*	*dy/dx*	*z-stat*	*dy/dx*	*z-stat*	*dy/dx*	*z-stat*
Schooling	-0.006	-4.150	-0.020	-4.660	-0.010	-6.300	-0.013	-3.920
	(0.001)		(0.004)		(0.002)		(0.003)	

P-values for χ^2 tests for

age group differences;				
Schooling (1df) 0.1681			Schooling (1df) 0.1726	
	N=4689	N = 1428	N = 6577	N = 2560
	F(11,332) =14.59	F(11,311) =10.35	F(11,332) =25.01	F(11,312) =13.27
	P>F =0.000	P>F =0.000	P>F =0.000	P>F =0.000

Notes: Full oprobit and marginal effects estimates in appendix A; Marginal effects computed at the mean values; Standard errors in parenthesis.

Table 6. Predicted probability of reporting ill health: by level of schooling

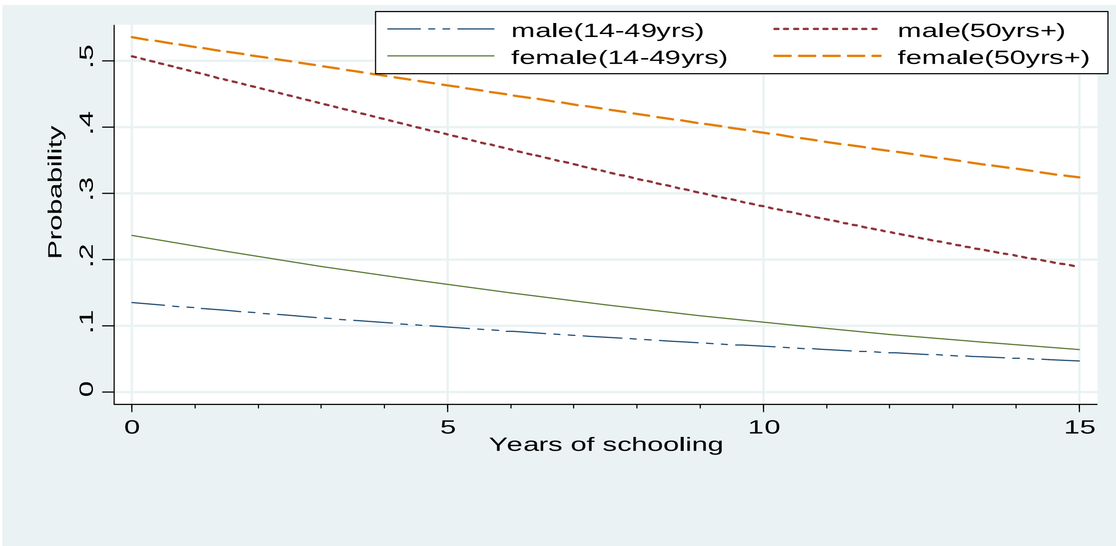
	Males		Females	
	14-49	50yrs+	14-49	50yrs+
No schooling	13.5%	50.6%	23.6%	53.5%
Primary	8.5%	34.4%	13.7%	43.4%
Secondary	5.9%	24.1%	8.7%	36.4%
Diploma	4.7%	18.9%	6.4%	32.3%

Note: Table based on estimated models in table 5

The ordered probit and marginal effect coefficients in table 5 reveal interesting patterns in the impact of schooling on the probability of reporting ill health. Specifically we observe that for the younger age group (14-49 yrs), the differential education effects are greater for females compared to males. Furthermore, we note that the health differentials tend to narrow out steadily as education increases in this age group. The simulated probabilities of ill health in table 6 (columns 1 and 3) present a vivid illustration of this trend. For a female, going from no schooling to secondary school completion reduces the probability of reporting ill health by 14.9 percentage points compared to males who only experience a 7.6 percentage point improvement. A visual representation of the narrowing of the reporting differentials among younger adults as education increases is depicted in figure 2 below. The steeper health gradient for young females in figure 2 also confirms that the education effects for females are higher than for males in the younger age group. Given that schooling effects cannot explain the gender differences among younger adults in South Africa (according to the chi-2 test), other factors outside our model may be at play.

The results in the older group (50 yrs and above) are reversed - simulations show that the differential impact of education on ill health is greater for males compared to their female counterparts. We can see this from the simulated probabilities in table 6 (columns 2 and 4), were for example, an improvement from no schooling to secondary education reduces the probability of reporting an illness by a higher magnitude for men (i.e 26.5 percentage points) relative to females (i.e 17.1 percentage points). This result can be seen in figure 2, where older males have a steeper gradient relative to older women. Thus in contrast with the results for younger adults, we find that the differential effects of education for older adults widens with more schooling.

Fig 2: Simulated probabilities of ill health against schooling: by age group and gender



vi) Income effects; and the impact of other variables

An important consideration in the literature on adult health and its socio-economic covariates is whether the impact of education on health primarily represents an income effect (Strauss et al, 1993; Handa 1998). In this section, we employ instrumental variables regressions to test whether income has an independent effect on health, and whether there is reverse causality between health and income. As mentioned in section 4, we use percapita household expenditure (pce) as a measure of long run or permanent income. Strauss et al (1993) and Handa (1998) argue that households may attempt to smooth consumption in the face of transitory shocks to income by saving in good times and dis-saving in bad times. Percapita expenditure was predicted using 6 variables (describing asset ownership and housing conditions) as identifying instruments. These are; i) whether a dwelling has a telephone, ii) cellphone, iii) electricity, iv) the number of rooms, v) type of wall material, and iv) whether the neighbourhood has street lighting. The instruments were chosen in to ensure instrument validity and in line with the logic of other studies (eg Strauss et al, 1993). For instance, whether a dwelling owns a telephone is unlikely to directly impact adult health; is likely correlated with income (pce); and is unlikely to be correlated with the error term in this model. The validity of our instruments is supported by the Amemiya-Lee-Newey (A-L-N) overidentification tests reported in table 7 below. The full sample instrumental variables probit regressions are presented in table 7 below.

Table 7. IV probit coefficients by gender

	Male			Female		
	Coef.	Std. Error	z-stat	Coef.	Std. Error	z-stat
Pce*	-0.048	0.047	-1.000	-0.028	0.043	-0.650
age	0.056	0.006	9.780	0.039	0.005	8.770
Age squared	0.000	0.000	-4.910	0.000	0.000	-2.610
Coloured	-0.098	0.054	-1.830	-0.205	0.042	-4.840
Asian/Indian	0.498	0.188	2.650	0.243	0.161	1.510
White	0.103	0.237	0.430	-0.200	0.220	-0.910
Schooling	-0.042	0.008	-5.340	-0.042	0.006	-7.110
Rural area	-0.065	0.074	-0.870	-0.241	0.058	-4.170
Formal urban area	0.047	0.078	0.610	-0.160	0.061	-2.640
married	-0.169	0.047	-3.580	-0.150	0.036	-4.170
employed	-0.181	0.043	-4.160	-0.053	0.035	-1.530
_cons	-1.283	0.145	-8.840	-0.645	0.114	-5.640
Wald test of exogeneity:						
X-2(1)=0.68 P-value =0.4109				χ-2 (1)=0.24 P-value =0.6261		
Amemiya-Lee-Newey minimum chi-sq statistic:						
X-2(5)=24.26 P-value =0.0002				χ-2 (5)=18.71 P-value =0.0022		
N	6,117			N	9,137	

Note: pce* (percapita expenditure) treated as endogenous; the wald test statistic is part of STATA; The Amemiya-Lee- Newey test for overidentification is a user written command (see Baum et al, 2006).

As can be seen in table 7 above, the wald test of exogeneity cannot be rejected, therefore we conclude that percapita expenditure is not endogenous in this model. This implies that our original uninstrumented models (table 3) are valid. The A-L-N test for overidentification further suggests that although our model is not endogenous, our choice of instruments is exogenous and therefore still valid. The results further show that the impact of schooling on adult health is still significant and independent of expenditure (pce). This finding is consistent with studies by Handa (1998) and Strauss et al (1993) who show the same results. The impact of percapita expenditure is insignificant in both the ordered probit and IV models.

We reviewed the effects of other variables on adult health in the model. *Being married* reduces the probability of reporting ill health and the coefficients were statistically significant at 1% for both females and males in the main regressions (table 3 and 4). Marginal effects indicate that marriage reduces the probability of reporting ill health by about the same percentage for both men and women. According to Umberton (1987), married adults are less likely to engage in risky social behaviours such as excessive smoking compared to unmarried ones. The effect of *employment* was found to be negative and significant (at the 1% for male and 5% for females), with employment effects being twice as higher for males than females. The chi-square test (p-value is 0.018) for gender differences in this effect shows that the differential impact can be explained by the influence of employment. These results may suggest that males may have better access to health augmenting employment related benefits such as better medical aid compared to their female counterparts. We found residential effects to be significant for both men and women. For women, we found that relative to living in informal townships, living in urban planned residences or rural areas was associated with a lower probability of reporting ill health. For men, residential effects were found to be only marginally significant if living in rural areas.

With regard to racial effects, being Coloured or White relative to being African was not statistically significant while being Asian/Indian was significant at 1% level for males and at 5% for females. Having controlled for other factors such as education, employment and income effects – the intermediary variables through which race may exert its effect on health status - we expected all race coefficients to be insignificant. The Coloured and White coefficients are indeed insignificant, implying that our model has successfully explained the racial differentials through the underlying variables in our model. For the Asian/Indian race, other factors outside this model may be affecting their self-perceived health status.

6. 0 CONCLUSION

This paper has attempted to understand the correlates of adult ill health in South Africa by gender and across the life cycle. Using data from the first wave of the 2008 National Income Dynamics Study, results show that in South Africa, as in other developing countries such as Jamaica (Strauss et al, 1993), schooling has a significant impact on the probability of reporting ill health. Among young adults (14-49 years), the impact of schooling on ill health is greater for women than for men but as years of education increase, reporting the differentials in health by gender tend to narrow out steadily. This finding may suggest that for younger adults, the benefit of education is greatest at lower levels of education. For older adults (50 years and above), we find a reversal in the effect of schooling – the differential impact of education on health by gender is greater for males than for females; and as education increases, we note that reporting differentials widen steadily. The impact of schooling persists even after controlling for income and other socioeconomic variables.

Results further show that employment reduces the probability of reporting ill health, with employment effects greatest among men relative to women. The gender differences in this effect are significance, suggesting that males may have better access to health augmenting employment related benefits such as better medical aid compared to their female counterparts. Location of residency matters. Adults living in informal townships are more likely to report ill health relative to those living in urban planned settlements or in rural areas. This finding may suggest that poor water and sanitation and a lack of health services in informal areas may be the intermediary variables through which residential differentials affect health status.

Finally, as in other countries, we find that South African women in general report higher cases of ill health compared to their male counterparts, although gender differences in this effect could not be attributed to the impact of age alone. An important area for future research would be to understand the intermediary mechanism through which employment and location of residency affects the probability of reporting ill health among South African adults.

Acknowledgements – I am sincerely grateful to SALDRU and NIDS for the award of the Masters Departmental scholarship. Thanks to Prof Ingrid D Woolard for kind supervision and to Prof Murray Leibbrandt for encouragements. All errors are my responsibility.

REFERENCES

Baum, C.F., Schaffer, M.E., Stillman, S., & V. Wiggins. 2006. Overid: Stata module to calculate tests of overidentifying restrictions after ivreg, ivreg2, ivprobit, ivtobit, reg3. Available at http://ideas.repec.org/c/boc/bocode/s396802.html

Doctor, H. 2001. Determinants of Self-reported health status in rural Malawi. A Paper Presented at the Annual Meeting of the Population Association of America. Washington DC, USA, 23-25 March 2001 (Unpublished).

Greene, W. 2000. *Econometric Analysis*. 5th ed. New York: Prentice Hall

Grossman, M. 1972. On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy*. 80(2): 223-255.

Gujarati, D. 1995. *Basic Econometrics.* 3rd ed. New York: McGraw-Hill

Handa, S. 1998. Gender and Life-Cycle Differecnes in the Impact of Schooling on Chronic Disease in Jamaica. Economics of Education Review. 17(3): 325-336.

Millunpalo, S., Vuori, I., Pekka, O., Panasen, M., & H. Uponen. 1997. Self-Rated Health Status as a Health Measure: The Predictive Value of Self-Reported Health Status on the use of Physician Services and on Mortality in the Working-Age population. *Journal of Clinical Epidemiology*. 50(5): 517-528.

Power C., Hertzman C., Matthews S. & O. Manor. 1997. Social Differences in Health: Life-Cycle Effects between Ages 23 and 33 in the 1958 British Birth Cohort. *American Journal of Public Health*. 87(9):1499-1503

Ramirez, M., Gallego, J. & Sepulveda, C. 2004. The Determinants of Health Status in a Developing Country: Results from the Colombian case. Economia Series Working Paper Number 41. University of Del Rosario.

Scott, L. 1997. *Regression Models for Categorical and Limited Dependent Variables.* California: Sage Publications.

Scott, L. & Freese, J. 2006. *Regression Models for Categorical Dependent Variables Using Stata*. Texas: Stata Press.

Strauss, J., Gertler, P., Rahman, O. & Kristin, F. 1993. Gender and Life-Cycle Differentials in the Patterns and Determinants of Adult Health. *The Journal of Human Resources.* 28(4): 791-837.

Steyn K., Fourie J. & N. Temple (eds). 2006. Chronic Diseases of Lifestyle in South Africa: 1995-2005. Technical Report. South African Medical Research Council.

Umberson, D. 1987. Family Status and Health Behaviors: Social Control as a Dimension of Social Integration. *Journal of Health and Social Behavior*. 28(3): 306-319.

United States Bureau of the Census. 2010. International Data Base. Available at http://www.census.gov/ipc/www/idb/groups.php

Woodridge, J. 2002. *Econometric analysis of Cross Section and Panel Data*. London: MIT press

World Health Organisation. 2008. *Global burden of Disease*. Geneva: WHO Press

World Health Organisation. 2001. *Macroeconomics and Health: Investing in Health for Economic Development*. Geneva: WHO Press

APPENDIX

Appendix A: Full probit and marginal effects estimates

i) Probit coefficients

	Males						Females					
	14-49			50 years +			14-49			50 years +		
	Coef.	S.e	t-stat	Coef.	S.e	t-stat	Coef.	S.e	t-stat	Coef.	S.e	t-stat
Schooling	-0.038	0.009	-4.250	-0.060	0.013	-4.570	-0.053	0.008	-6.550	-0.036	0.009	-3.870
age	0.060	0.022	2.760	0.006	0.059	0.100	0.059	0.017	3.580	-0.013	0.036	-0.370
age_squared	0.000	0.000	-1.040	0.000	0.000	0.070	0.000	0.000	-1.700	0.000	0.000	0.640
Coloured	-0.100	0.099	-1.010	0.174	0.139	1.250	-0.141	0.107	-1.320	-0.200	0.140	-1.430
Asian_Indian	0.302	0.167	1.810	0.560	0.238	2.350	0.357	0.198	1.800	0.324	0.446	0.730
White	-0.095	0.144	-0.660	-0.106	0.209	-0.510	-0.182	0.177	-1.030	-0.289	0.205	-1.410
pce	0.002	0.011	0.190	-0.017	0.015	-1.120	-0.014	0.020	-0.680	-0.022	0.013	-1.720
Rural(tribal)	-0.091	0.096	-0.950	-0.846	0.264	-3.200	-0.429	0.074	-5.800	-0.201	0.156	-1.290
Urban(formal)	-0.004	0.111	-0.030	-0.664	0.269	-2.470	-0.355	0.086	-4.110	-0.211	0.181	-1.160
married	-0.246	0.090	-2.740	-0.089	0.095	-0.940	-0.230	0.052	-4.400	-0.092	0.070	-1.310
employed	-0.244	0.067	-3.650	-0.389	0.092	-4.230	-0.052	0.052	-1.010	-0.275	0.087	-3.170
/cut1	1.416	0.298	4.750	-1.259	1.915	-0.660	0.690	0.261	2.640	-1.403	1.211	-1.160
/cut2	2.242	0.299	7.500	-0.477	1.921	-0.250	1.569	0.259	6.050	-0.632	1.215	-0.520

ii) Marginal effects

	Males						Females					
	14-49			50 years +			14-49			50 years +		
	dy/dx	S.e	z	dy/dx	S.e	z	dy/dx	S.e	z	dy/dx	S.e	z
Schooling	-0.006	0.001	-4.150	-0.020	0.004	-4.660	-0.010	0.002	-6.300	-0.013	0.003	-3.920
age	0.009	0.003	2.700	0.002	0.020	0.100	0.012	0.003	3.490	-0.005	0.013	-0.370
age^2	0.000	0.000	-1.040	0.000	0.000	0.070	0.000	0.000	-1.690	0.000	0.000	0.640
Coloured	-0.015	0.015	-1.010	0.059	0.047	1.260	-0.027	0.021	-1.310	-0.074	0.051	-1.430
Asian	0.044	0.024	1.820	0.188	0.079	2.380	0.069	0.038	1.800	0.119	0.163	0.730
White	-0.014	0.021	-0.660	-0.036	0.070	-0.510	-0.035	0.034	-1.030	-0.106	0.074	-1.430
pce	0.000	0.002	0.190	-0.006	0.005	-1.120	-0.003	0.004	-0.680	-0.008	0.005	-1.730
Rural(tribal)	-0.013	0.014	-0.960	-0.285	0.087	-3.290	-0.083	0.014	-5.760	-0.074	0.057	-1.300
Urban(formal)	-0.001	0.016	-0.030	-0.224	0.089	-2.520	-0.069	0.016	-4.210	-0.078	0.066	-1.170
married	-0.036	0.013	-2.730	-0.030	0.032	-0.940	-0.045	0.010	-4.370	-0.034	0.026	-1.310
employed	-0.036	0.010	-3.600	-0.131	0.030	-4.340	-0.010	0.010	-1.010	-0.101	0.032	-3.130

Appendix B: STATA Do file

```
*Author_Mashekwa M
*Title: Mini-thesis do file
*Gender and Life-cycle differentials in the correlates of adult ill health in South Africa
*Supervisor: Professor Ingrid D. Woolard
*___________________________________________________________________________
clear
set mem 500m
version 11.1
set more off
global IN "C:\Documents and Settings\Owner\Desktop\NOV 2010 RELEASE"
global OUT "C:\Documents and Settings\Owner\Desktop\NOV 2010 RELEASE\OUTPUT"
*___________________________________________________________________________

*Part 1: Merging files
*___________________________________________________________________________

use "$IN\indderived_Anon_V3.0.dta", clear
merge 1:1 hhid pid using "$IN\Adult_Anon_V3.0.dta"
tab _merge
keep if _m==3
drop _merge
merge m:1 hhid using "$IN\hhderived_Anon_V3.0.dta"
tab _merge
keep if _m==3
drop _merge
merge m:1 hhid using "$IN\HouseholdQ_Anon_V3.0.dta"
tab _merge
keep if _m==3
keep w1_a_hldes w1_a_best_age_yrs w1_a_gen w1_best_race w1_best_edu w1_hhincome w1_hhgeo w1_h_expenditure /*
*/ w1_hhsizer w1_a_marstt w1_empl_stat w1_h_dwltyp w1_h_dwlrms w1_h_dwlmatroof w1_h_dwlmatrwll w1_dwgt w1_hhdc w1_hhcluster /*
*/ w1_h_ownd w1_h_watsrc w1_h_toi w1_h_toishr w1_h_enrgelec w1_h_tellnd w1_h_telcel w1_h_strlght

*___________________________________________________________________________
```

```
*Part 2: Data cleaning & variable definition
*___________________________________________________________________________________


*1_health status                    /*recoding in line with section 4 definition*/
gen healthstatus=w1_a_hldes
drop if healthstatus==. | healthstatus==-9 | healthstatus==-8 | healthstatus==-3
recode healthstatus 5=1 4=2 3=3 2=4 1=5
label define healthstatus 1"Poor" 2"Fair" 3"Good" 4"Verg Good" 5"Excellent"
label value healthstatus healthstatus
tab healthstatus
gen h_status=healthstatus
recode h_status 1/2=2 3=1 4/5=0
label define h_status 2 "Poor & fair" 1 "Good" 0 "very good & better"
*creating health dummies
label value h_status h_status
gen poor_health=0 if  h_status!=.
replace poor_health=1 if h_status==2
replace poor_health=. if h_status==.
gen good_health=0 if  h_status!=.
replace good_health=1 if h_status==1
replace good_health=. if h_status==.
gen excellent_health=0 if h_status!=.
replace excellent_health=1 if h_status==0
replace excellent_health=. if h_status==.
*.........................................................................................
*2_age & age dummies
*.........................................................................................
gen age=w1_a_best_age_yrs
drop if age ==.
drop if age==-9 | age==-8 | age==-3
gen age_squared=(age)^2
gen ag_1=0 if age!=.
replace ag_1=1 if age<=49
replace ag_1=. if age==.
gen ag_2=0 if age!=.
replace ag_2=1 if age>=50 & age!=.
replace ag_2=. if age==.
gen ag_spline=.
replace ag_spline=1 if ag_1
replace ag_spline=2 if ag_2
*.........................................................................................
*3_Gender
*.........................................................................................

gen gender= w1_a_gen
gen male =0 if w1_a_gen!=.
replace male=1 if w1_a_gen==1
replace male=. if w1_a_gen==.
tab male, m
gen female =0 if w1_a_gen!=.
replace female=1 if w1_a_gen==2
replace female=. if w1_a_gen==.
tab female
*.........................................................................................
*4_Race
*.........................................................................................

gen race= w1_best_race
drop if race==.
label define race 1"African" 2"Coloured" 3"Asian_Indian" 4"White"
label value race race
gen African=0 if race!=.
replace African=1 if race==1
replace African=. if race==.
gen Coloured=0 if race!=.
replace Coloured=1 if race==2
replace Coloured=. if race==.
gen Asian_Indian=0 if race!=.
replace Asian_Indian=1 if race==3
replace Asian_Indian=. if race==.
gen White=0 if race!=.
replace White=1 if race==4
replace White=. if race==.
```

```
*..........................................................................................
*5_Education - defined by 2 variables
*..........................................................................................
*linear years of schooling
gen educ_years=w1_best_edu
drop if educ_years==. | educ_years==-9 |educ_years==-5 |educ_years==-3 |educ_years==-8
recode educ_years 25=0 24=13 15=14 16/17=13 18/19=14 20=15 21/22=16 23=17
tab educ_years

* educational dummies
gen edu_dummy=educ_years
recode edu_dummy 0=0 1/7=1 8/12=2 13/23=3
label define edu_dummy 0"No_schooling" 1"Primary_educ" 2"Secondary_educ" 3"higher_educ"
label value edu_dummy edu_dummy
tab edu_dummy
gen No_Schooling=0 if edu_dummy!=.
replace No_Schooling=1 if edu_dummy==0
replace No_Schooling=. if edu_dummy==.
gen Primary_educ=0 if edu_dummy!=.
replace Primary_educ=1 if edu_dummy==1
replace Primary_educ=. if edu_dummy==.
gen Secondary_educ=0 if edu_dummy!=.
replace Secondary_educ=1 if edu_dummy==2
replace Secondary_educ=. if edu_dummy==.
gen higher_educ=0 if edu_dummy!=.
replace higher_educ=1 if edu_dummy==3
replace higher_educ=. if edu_dummy==.
*..........................................................................................
*6_ per capita expenditure
*..........................................................................................
gen pce=(w1_h_expenditure/w1_hhsizer)/1000 /* pc monthly hh expenditure - full imputations*/

*..........................................................................................
*7_Residential variables/geo dummies
*..........................................................................................
gen location=w1_hhgeo
recode location 1/2=1 3=2 4=3
label define location 1 "tribal or formal rural" 2"Urban formal" 3"Urban Informal"
label value location location
tab location
gen tribal_rural=0 if location!=.
replace tribal_rural=1 if location==1
replace tribal_rural=. if location==.
gen Urban_formal=0 if location!=.
replace Urban_formal=1 if location==2
replace Urban_formal=. if location==.
gen Informal_urban=0 if location!=.
replace Informal_urban=1 if location==3
replace Informal_urban=. if location==.
*..........................................................................................
*8_marital status
*..........................................................................................
gen marital_status=w1_a_marstt
drop if marital_status==-3
gen married=0 if marital_status!=.
replace married=1 if marital_status==1
replace married=. if marital_status==.
gen not_married=0 if marital_status!=.
replace not_married=1 if marital_status==2 |marital_status==4|marital_status==4|marital_status==5
tab married not_married
*..........................................................................................
*9_employment
*..........................................................................................
drop if w1_empl_stat==.
gen employed=0 if w1_empl_stat!=.
replace employed=1 if w1_empl_stat==3
replace employed=. if w1_empl_stat==.
gen un_employed=0 if w1_empl_stat!=.
replace un_employed=1 if w1_empl_stat==0|w1_empl_stat==1|w1_empl_stat==2



*______________________________________________________________________________________________

*Part 3: Weighted descriptives, and o-probit estimations
```

```
*______________________________________________________________________________
*3.1 declare svy
gen weight=int(w1_dwgt)
svyset [pw=w1_dwgt], strata(w1_hhdc) psu(w1_hhcluster)

* some descriptives discussed in section 5.1
svy: tab h_status
svy: tab poor_health location, column
svy: tab poor_health ag_spline, column
svy: tab  poor_health female, column
svy: tab  poor_health race, column
svy: tab  poor_health married, column
svy: tab  poor_health employed, column

* Weighted descriptive statistics
* male sample
sum excellent_health good_health poor_health age African Coloured Asian_Indian White /*
*/ No_Schooling Primary_educ Secondary_educ higher_educ educ_years  /*
*/ pce Urban_formal tribal_rural Informal_urban married employed [w=w1_dwgt] if male==1
* female sample
sum excellent_health good_health poor_health age African Coloured Asian_Indian White /*
*/ No_Schooling Primary_educ Secondary_educ higher_educ educ_years  /*
*/ pce Urban_formal tribal_rural Informal_urban married employed [w=w1_dwgt] if male==0

*..........................................................................................
*3.2 Gender regressions
*..........................................................................................

* Tests for gender differences
preserve
parmby "svy:oprobit h_status age age_squared Coloured Asian_Indian White educ_years pce tribal_rural Urban_formal married employed", by(female) norestore
escal(N) ren(es_1 N)
estparmtest estimate stderr if parm=="age", obs(N)
estparmtest estimate stderr if parm=="educ_years", obs(N)
estparmtest estimate stderr if parm=="employed", obs(N)
restore

*male sample

svy: oprobit h_status age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed if female==0


prgen age, f(14) t(90) generate(m_age) ncases(15) /*creating predicted probs for male age gps*/

margins, dydx(age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed) predict(outcome(2))

*female sample

svy: oprobit h_status age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed if female==1

prgen age, f(14) t(90) generate(f_age) ncases(15) /*creating predicted probs for female age gps*/

margins, dydx(age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed) predict(outcome(2))


*..........................................................................................
* Graphing the predicted probabilities
*..........................................................................................

*1_probabilities - age and gender

gen diff= f_agep2- m_agep2
twoway line  f_agep2 m_agep2 diff m_agex, sort clwidth(thick . .)  /*fig 1 output*/

*______________________________________________________________________________
*Cohort Regressions - education and gender models
*______________________________________________________________________________

*2_probabilities - education and gender
*preparing regs by gender & by broad age grps & testing for schooling difference in all 6 models
```

```
*male groups
preserve
parmby "svy:oprobit h_status age age_squared Coloured Asian_Indian White educ_years pce tribal_rural Urban_formal married employed if female==0", by(ag_spline)
norestore escal(N) ren(es_1 N)
estparmtest estimate stderr if parm=="educ_years", obs(N)
restore
*female groups
preserve
parmby "svy:oprobit h_status age age_squared Coloured Asian_Indian White educ_years pce tribal_rural Urban_formal married employed if female==1", by(ag_spline)
norestore escal(N) ren(es_1 N)
estparmtest estimate stderr if parm=="educ_years", obs(N)
restore

* Marginal effects
*a)males in 14-49year age group
svy: oprobit h_status age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed if female==0 & ag_spline==1

prvalue, x( educ_years=0) rest(mean) /*no schooling*/
prvalue, x( educ_years=7) rest(mean) /*primary sch completion*/
prvalue, x( educ_years=12) rest(mean) /*secondary sch completion*/
prvalue, x( educ_years=15) rest(mean) /*higher education*/

prgen educ_years, f(0) t(15) generate(m_young)

margins, dydx(age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed) predict(outcome(2))

*b)males in 50yrs+ age group
svy: oprobit h_status age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed if female==0 & ag_spline==2

prvalue, x( educ_years=0) rest(mean) /*no schooling*/
prvalue, x( educ_years=7) rest(mean) /*primary sch completion*/
prvalue, x( educ_years=12) rest(mean) /*secondary sch completion*/
prvalue, x( educ_years=15) rest(mean) /*higher education*/


prgen educ_years, f(0) t(15) generate(m_old)

margins, dydx(age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed) predict(outcome(2))

*c)females in 14-49year age group
svy: oprobit h_status age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed if female==1 & ag_spline==1

prvalue, x( educ_years=0) rest(mean) /*no schooling*/
prvalue, x( educ_years=7) rest(mean) /*primary sch completion*/
prvalue, x( educ_years=12) rest(mean) /*secondary sch completion*/
prvalue, x( educ_years=15) rest(mean) /*higher education*/

prgen educ_years, f(0) t(15) generate(f_young)


margins, dydx(age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed) predict(outcome(2))

*d)females in 50yrs+ age group
svy: oprobit h_status age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed if female==1 & ag_spline==2

prvalue, x( educ_years=0) rest(mean) /*no schooling*/
prvalue, x( educ_years=7) rest(mean) /*primary sch completion*/
prvalue, x( educ_years=12) rest(mean) /*secondary sch completion*/
prvalue, x( educ_years=15) rest(mean) /*higher education*/

prgen educ_years, f(0) t(15) generate(f_old)

margins, dydx(age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed) predict(outcome(2))

twoway line m_youngp2 m_oldp2 f_youngp2 f_oldp2 f_oldx, sort clwidth (thin thick thin thick) msymbol(Oh S Oh Dh)
```

```
*______________________________________________________________________________________________
*IV Regressions & tests for endogeneity and overidentification of instruments
*______________________________________________________________________________________________
*Male IV estimates
svy: oprobit h_status age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed if female==0

ivprobit h_status age age_squared Coloured Asian_Indian White educ_years tribal_rural Urban_formal /*
*/ married employed (pce= w1_h_dwlrms w1_h_dwlmatrwll w1_h_enrgelec w1_h_tellnd w1_h_telcel w1_h_strlght) if female==0, twostep

overid /*user written ado file for test of overidentification*/

*Female IV estimates
svy: oprobit h_status age age_squared Coloured Asian_Indian White educ_years /*
*/ pce tribal_rural Urban_formal married employed if female==1

ivprobit h_status age age_squared Coloured Asian_Indian White educ_years tribal_rural Urban_formal /*
*/ married employed (pce= w1_h_dwlrms w1_h_dwlmatrwll w1_h_enrgelec w1_h_tellnd w1_h_telcel w1_h_strlght) if female==1, twostep

overid
cap linktest
cap testnl _b[age] =_b[age_squared] = _b[Coloured] =_b[Asian_Indian] /*
*/ =_b[White]= _b[Primary_educ] = _b[Secondary_educ] =_b[Tertiary_educ] /*
*/ =_b[University_educ]=_b[percapita]= _b[tribal_rural]=_b[Urban_formal]=_b[hhsize]= _b[married] =_b[employed]

* Declaration: This work is my own; and all errors are mine (Mashekwa).
```