Essays in Health Economics

by

Negar Razavilar

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Economics
University of Alberta

Abstract

The Effect of Leisure Time Physical Activity on Labour Market Earnings: Evidence from the Canadian National Population Health Survey

Using 6 cycles of data from the Canadian National Population Health Survey I estimate the effect of leisure time physical activity on labour market earnings among working age adults. The longitudinal nature of the data allows for measuring time spent on leisure time physical activity at different points in time thus capturing changes in the amount of time spent on physical activity and its impact on future labour market outcomes. Estimates for the male subsample indicate that increased lagged participation in daily leisure time physical activity has a positive and significant impact on both hourly wages and annual income. These estimates are robust to a variety of panel data estimation techniques, including dynamic panel data estimation methods, and to the inclusion of additional explanatory variables controlling for Body Mass Index (BMI) and self-rated health status. However, increased participation in leisure time physical activity does not have a significant impact on either hourly wages or annual income among women.

The Causal Effect of Unemployment on Smoking: Evidence from the Canadian Community Health Survey

In this study I estimate the causal effect of individual unemployment on individual smoking behaviors using data from one cycle (year 2012) of the Canadian Community Health Survey. Two separate instrumental variables (IV) approaches

(Two-Stage Residual Inclusion (2SRI) and Two-Stage Predictor Substitution (2SPS)) using provincial level unemployment rates by age cohort as the instrumental variable are used to identify causal effects. IV estimates from two-part models, using a probit model for smoking status and a negative binomial model for smoking intensity, indicate that individual unemployment status does not have a significant impact on the probability of being a smoker but has a negative and significant impact on the number of cigarettes smoked per day conditional on being a smoker. The IV estimates are robust but sensitive to the type of the IV approach used.

Decomposition of the Income Gap in Body Mass Index (BMI) in Canada

The aim of this study is to examine the demographic, socio-economic, and behavioral variables explaining the income gap in Body Mass Index (BMI) among Canadian adults using one cycle (year 2012) of the Canadian Community Health Survey (CCHS). Two different grouping strategies; 1) low-income cut-offs in year 2012 from Statistics Canada, and 2) the top and bottom three income deciles from the CCHS are used for defining the high- and low-income groups. Using the Oaxaca-Blinder decomposition method I decompose the difference in mean BMI among the high- and low-income male and female into a part explained by differences in the observed characteristics between the two groups, and a share attributable to the differences in the returns to those characteristics between the two groups. It follows that high-income men have higher average BMIs than low-income men, while the opposite is observed among women. In the male sample, the highest contributions to the explained gap belong to employment status and being single (as opposed to being

married) which both contribute to the low-income having lower mean BMIs than the high-income. Among women, the highest contribution to the explained gap belongs to average daily energy expenditure which contributes to the low-income women having higher mean BMIs than the high-income women.

Dedication

I dedicate this thesis to my beloved parents, Jila and Vadood, and to my most precious sister, Negin, for their endless love, support, and encouragement and for giving me the strength to carry on.

Acknowledgements

I would like to express my most sincere gratitude to my supervisor, Dr. Christopher McCabe, for his continuous support and guidance, and for providing me the unique opportunity to join his research team.

I am extremely grateful to my dissertation committee members, Dr. Xuejuan Su and Dr. Tilman Klumpp, for their useful discussions and insightful comments on my thesis, Dr. Dana Andersen for accepting to be my arm's length examiner, Dr. Sandy Tubeuf for being my external reviewer, and Dr. Li Zhou for accepting to be the chair of my examining committee. I would also like to thank Dr. Jane Ruseski for her help and her feedback on my work.

Many other people helped me throughout this long and tough journey. I sincerely thank my colleagues Philip Akude, Samprita Chakraborty, Benoit Kudinga, Klemens Wallner, Dr. Michael Paulden, Max Sties, Dr. Waleem Alausa, Dr. Jie Yang, Ning Chao, and other members of the Department of Economics and the Department of Emergency Medicine at the University of Alberta, especially Audrey Jackson, Christina Smith, Brenda Carrier, Charlene Hill, Olga Krol, and Stephanie Griffin.

The estimations for the first chapter of this dissertation were conducted at the Research Data Center (RDC), University of Alberta. The RDCs are initiated by

Statistics Canada, the Social Sciences and Humanities Research Council (SSHRC), and university consortia with additional support from the Canada Foundation for Innovation (CFI) and the Canadian Institutes for Health Research (CIHR). My sincere gratitude goes to Irene Wong at the University of Alberta's RDC for her kind help and support.

I would also like to take the opportunity to express my gratitude to my professors during my Masters and PhD program at York University and University of Alberta especially, Dr. Andrea Podhorsky, Dr. Sam Bucovetsky, Dr. Tasso Adamopoulos, Dr. Andrei Semenov, and Dr. Joseph Marchand.

Last but not least I would like to thank my friends Nazanin Akhavan, Andia and Mandana Rezapourian, Maryam and Leila Zargarzadeh, Leily Mohammadi, Alaleh Rouhi, Leslie and Paul Precht, Kerry and Kim Precht, and my dear grandmother, aunts, uncles, and cousins for their love, warmth, and support.

Table of Contents

List of Tables

Nomenclature

LTPA	Leisure Time Physical Activity
OLS	Ordinary Least Squares
GMM	Generalized Method of Moments
FD-GMM	First Differences Generalized Method of Moments
FE	Fixed Effects
RE	Random Effects
BMI	Body Mass Index
2SRI	Two Stage Residual Inclusion
2SPS	Two Stage Predictor Substitution
AR	Auto Regressive Model

Introduction

The majority of preventable morbidities and mortalities in developed countries are caused by chronic conditions (for example type 2 diabetes, obesity, high blood pressure, cardiovascular diseases, and several types of cancers) rather than infectious diseases. Particularly, unhealthy behaviors such as physical inactivity, smoking, excess drinking, and poor diet, are mostly responsible for causing these chronic conditions[1]. The World Health Organization (2009) published a ranking of the modifiable risk factors associated with mortality and morbidity (measured in terms of disability adjusted life years (DALSYs)) in high-income countries (countries with 2004 dollars per capita income greater than $10,065). The ranking indicates that smoking is number one on the list accounting for 18% of deaths and 11% of DALYs, excess body weight is third and accounts for 8% of deaths and 7% of DALYs, and physical inactivity comes fourth accounting for 8% of deaths and 4% of DALYs. The second, fifth and sixth risk factors are high blood pressure, blood glucose, and cholesterol respectively which are in turn the consequences of risky health behaviors such as poor diet, physical inactivity, and smoking. The consequences of risky health behaviours induce substantial economic burden to the healthcare system and to the society as a whole. In fact, smoking, physical inactivity, and obesity (which can

[1] Kenkel (2000)

potentially be the result of physical inactivity and/or a combination of health behaviors) are now three of the biggest public health concerns in Canada (for example see Katzmarzyk and Janssen 2004, Rehm et al. 2006, and Janssen 2009). For instance, in year 2001, the estimated economic burden associated with physical inactivity in Canada was $5.3 billion dollars, consisting of $1.6 billion in healthcare costs, and $3.7 billion in indirect costs as a result of output lost due to illnesses, injury-related work disability, or premature mortality[2]. The total economic burden of tobacco consumption was $17.7 billion (based on 2002 data), which included health care costs and productivity losses attributable to premature death and disability resulting from tobacco related diseases[3]. In 2005, the total economic burden of adult obesity in Canada was estimated as $3.42 billion consisting of $1.62 billion in direct costs and $1.80 billion in productivity losses[4].

Although there exists substantial evidence (Cawley and Ruhm 2011) that health behaviors are often strongly associated with an individual's position in the society (based on occupational, economic, and educational criteria or a combination of these factors), the extent to which these associations are causal, i.e. whether the correlations are running from socioeconomic status to health behaviors or vice versa, is still an empirical question which has important policy implications when aiming at promoting healthy behaviors and/or targeting specific subpopulations in doing so. Estimating correlations rather than casual effects can potentially reflect two factors:

[2] Katzmarzyk and Janssen (2004)
[3] Rehm et al. (2006)
[4] Janssen (2009)

1) reverse causality, i.e. the impact of poor economic outcomes on health behaviors, or the impact of health behaviors on poor economic outcomes (in other words, the health behavior could potentially be causing the economic outcome (such as unemployment/employment or lower/higher wages), or it might be the fact that the economic phenomenon is causing the health behavior), or 2) the impact of unobservable confounding variables that simultaneously affect the health behavior and the economic cause/outcome in question. In addition, the extent to which the socioeconomic gaps in health/health behaviors could be eliminated or reduced once certain factors contributing to the gaps are eliminated, is another important policy question that deserves attention particularly due to the fact that eliminating gaps in health is a goal of every developed country.

In this dissertation I focus on the economic causes and/or consequences of three of the most important health behaviors i.e. physical activity, smoking, and Body Mass Index (Body Mass Index is not a health behavior in itself but is used to determine overweight or obesity which are potentially caused by a number of unhealthy behaviors such as physical inactivity and poor diet). More specifically, in chapter 1 I examine the impact of participation in leisure time physical activity (LTPA) on labour market earnings using 6 cycles of the longitudinal National Population Health Survey (NPHS). Evidence on the direct effect of LTPA on labour market outcomes are scarce (especially when people are in their prime working years/ages) despite the fact that the positive impact of physical activity on determinants of labor market success has been well established in the literature. The majority of Canadian adults do not meet the minimum required level of physical

activity to maintain good health, a statistic which has motivated the Canadian Sport Policy (2012) to aim at increasing the number and diversity of physically active Canadians by year 2022. Knowing if and how participation or increasing participation in LTPA affects labor market outcomes has a number of important implications: 1) it can serve as an additional motivation for individuals to increase their physical activity levels, 2) it can increase the productivity of the labour force which is the goal of every developed country in order to strengthen the position of the economy and improve welfare, 3) it can help policy makers in designing effective policies to promote physical activity and sports. I take advantage of the longitudinal nature of the data in order to identify causal effects, account for the persistence of labor market earnings when people are in their prime working years, and to measure LTPA at multiple points in time (since people have transitions in and out of LTPA). I find that increasing time spent in daily LTPA leads to a positive and significant increase in hourly wages and annual income of men, and the point estimates are robust to the inclusion of Body Mass Index (BMI) and a measure of self-rated health. On the contrary, I do not find a significant impact of increasing daily amount of time spent in LTPA on labor market earnings among women.

In chapter 2 I focus on the causal impact of unemployment on smoking intensity using a wave of the Canadian Community Health Survey (CCHS). As unemployment substantially increased over the last decades in almost all Western (developed) countries, the interest in the association between unemployment and substance use has also increased. Smoking is the most commonly reported unhealthy behavior among the unemployed (Henkel 2011), however it is not certain whether

this association is due to the causal impact of unemployment on smoking or vice versa, or whether there are some unobserved individual specific factors that are driving both smoking behaviors and the probability that an individual is unemployed. The causal effect of unemployment on substance use, such as smoking, continues to be an important question in the economic literature today (Henkel 2011). Unemployment can affect smoking behaviors through a number of different channels. First of all, the financial distress and social isolation resulting from unemployment can potentially induce individuals to smoke or smoke more often. Second of all, unemployment can cause a loss of income which can potentially decrease smoking levels in order to reduce spending on cigarettes and save money. Finally, employed individuals may suffer from work-related stress and may have less time to invest in healthy ways of stress relief (such as physical activity), and therefore might be more likely to smoke or smoke more often than the unemployed. Given the many substantial adverse health and economic outcomes associated with smoking, the answer to this question should be of great value to the policy makers in targeting specific populations. The literature on employment status and smoking is mixed and inconclusive, either because of ignoring the potential endogeneity of unemployment, or because of using methods/instrumental variables that are not known to be robust in appropriately addressing the endogeneity issue. My study extends on the previous works by implementing an instrumental variable (IV) method via Two Stage Residual Inclusion (2SRI) which has been shown to produce more consistent estimates in a nonlinear setting. I find that after accounting for the endogeneity of unemployment in the smoking equation, it does not have a significant impact on the

probability of being a smoker but has a negative and significant impact on the average number of cigarettes smoked per day conditional on being a smoker.

Finally in chapter 3 I examine the factors behind the income gap in BMI among Canadian adults using a wave of the CCHS dataset. The Canadian Population Health Initiative (2008) has released a statement that there exists socioeconomic gaps in health and health behaviors in Canada, and they are aiming at reducing these gaps. In terms of policy implications, interventions should be targeted at those with lower-SES whenever the gap in a particular health indicator is wide and favoring the higher-SES group. On the other hand, when the SES gap in health is relatively narrow, a more universal approach that meets the needs of all SES subpopulations should be implemented (The Canadian Population Health Initiative 2004). I use the Oaxaca-Blinder decomposition method in order to decompose the gap in mean BMI between high- and low-income Canadians into a part that is explained by differences in observed characteristics between the high- and low-income groups, and another part attributable to differences in the returns to those characteristics. Knowing if and how much different SES and behavioral factors contribute to the income gap in mean BMI should be of importance to policy makers that aim at reducing this gap. The results indicate that the income gap in mean BMI is significant among men but insignificant among women when using the Low Income Cut-Offs (LICO) in grouping individuals into high- and low-income categories. However when using the top and bottom three income deciles in grouping individuals into high- and low-income categories respectively, the income gap in mean BMI among women also becomes significant. The decomposition of the gap in the male subsample indicates that age, being

Canadian born, average daily energy expenditure, being single (as opposed to being married), being employed, presence of young kids in the household, average daily fruit and vegetable consumption, and average daily cigarette consumption contribute to the explained gap. Among women, the highest contribution to the explained gap in mean BMI belongs to the difference in average daily energy expenditure which contributes to the low-income women having higher mean BMIs than the high-income women. Policies aimed at reducing the income gap in BMI/obesity should particularly focus on reducing differences in health behaviors and on specific populations (for example the employed) within an income group.

1. The Effect of Leisure Time Physical Activity on Labour Market Earnings: Evidence from the Canadian National Population Health Survey

1.1. Introduction

Physical inactivity is considered one of the greatest threats to public health in Canada (Katzmarzyk and Janssen 2004). Many types of illnesses, disabilities and chronic conditions have consistently been associated with physical inactivity resulting in substantial healthcare costs and economic losses. In year 2001, the estimated economic burden associated with physical inactivity in Canada was $5.3 billion dollars, consisting of $1.6 billion in healthcare costs, and $3.7 billion in indirect costs as a result of output lost due to illnesses, injury-related work disability, or premature mortality (Katzmarzyk and Janssen 2004). The Canadian Society for Exercise Physiology (CSEP) recommends a daily energy expenditure of at least 1.5 kilocalories per kilogram of body weight (kcal/kg) from all leisure time physical activities (Lechner and Sari 2015). The Centers for Disease Control and Prevention (CDC) and the American College of Sports Medicine (ACSM) recommend that

individuals should engage in at least 30 minutes of moderate-intensity physical activity on most days of the week (Pate et al. 1995). There has also been an update on the 1995 recommendation that all adults aged 18 to 65 need a minimum of 30 minutes of moderate-intensity aerobic activity five days a week (Haskell et al. 2007). According to Gilmour (2007), the percentage of Canadians who reported being at least moderately physically active during their leisure time rose from 43% in 1996-1997 (based on data from the 1996-1997 National Population Health Survey) to 52% in 2005 (based on data from the 2005 Canadian Community Health Survey). Although these self-reported data indicate a very modest improvement in physical activity participation rates over recent years, Colley et al. (2011) argued that only 15% of adult Canadians over age 20 meet the minimum required level of physical activity based on objective accelerometer data. Based on these statistics, Canadian Sport Policy has recently become motivated to aim at increasing the number and diversity of Canadians participating in sport over 2012-2022 ("Canadian Sport Policy 2012").

In an attempt to encourage more people to participate in leisure time physical activity (or increase their level of participation), some economists have argued that one step towards better understanding the benefits of regular exercise and healthy lifestyle choices is to examine whether individuals can gain financial benefits, in addition to health benefits, from participation in physical activity (PA) (Kosteas 2012). One important policy goal of developed countries is to improve the productivity of the labour force which will ultimately help to improve the position of the economy, increase welfare and reduce unemployment. The labour market

impact of some of these policies such as schooling, vocational training, and public employment services for the unemployed have been extensively examined by many researchers, however little attention has been paid to the labour market effects of participation in PA. Although there is a consensus that more physical activity (within limits) is always better for the health (Warburton et al. 2006) it is not certain whether these health effects translate one-to-one into earnings gains. Knowing that participation in PA would lead to labour market gains can potentially motivate more individuals to actively engage in PA and will further help policy makers in promoting PA.

From a theoretical point of view it is uncertain how PA can affect labour market performance. The main theoretical frameworks that explain the association between participation in leisure time physical activity (LTPA) and labour market outcomes are based on the effect of LTPA on human capital and time allocation. The human capital theory proposed by Becker (1965) and Mincer (1958) focuses on the role of cognitive skills in improving human capital. These theories suggest that individuals improve their cognitive skills by investing time and other resources in education that will positively affect their employment and earnings through signaling (Spence 1973). Consistent with the household production model of Becker (1965) individuals make decisions about how to allocate their time and resources (subject to budget and time constraints) to the production of the "final" goods in order to maximize their utility. Based on this latter theory and the human capital theory, participation in LTPA and investment in human capital are two mutually exclusive alternatives, therefore allocating time and resources towards LTPA will crowd out

investment in human capital which will have a negative impact on labour market outcomes. On the other hand a number of theories have challenged the negative hypothesised relationship between LTPA and labour market outcomes based on refinements of the human capital theory. Heckman and Rubinstein (2001) stress the importance of non-cognitive skills in the development of human capital and labour market outcomes. Particularly, participation in PA and sports can provide non-cognitive skills such as social networking skills (Lechner 2009), self-discipline, and tenacity that complement the mainly cognitive skills provided by education (Cabane and Lechner 2014). Sports participation can also serve as a signal to the employers that the individual has high motivation, discipline and dedication at work (Lechner 2009). Finally in Grossman's health production model (Grossman 1972) health is viewed as an investment good (as well as a consumption good) which produces a stock of healthy time which can be dedicated to activities such as labour market activities. Therefore investment in health through participation in PA can directly affect labour market outcomes by increasing the productive quality of time and also indirectly by serving as a signal of health and future productivity (Lechner 2009, Rooth 2011). Healthy individuals are more productive at work hence more likely to earn higher wages (Wellman and Friedberg 2002). The higher productivity levels that individuals benefit from could be due to higher energy levels as a result of engaging in regular exercise (Puetz et al. 2006), or less days absent from work due to illnesses and morbidities associated with physical inactivity. Participation in PA can also lead to improved mental function (Etnier et al. 1997), all of which affect individuals' productivity at work.

The aim of this study is to examine whether participation in leisure time PA has a causal impact on labour market outcomes, in terms of wages and annual income, using data from 6 waves of the National Population Health Survey (NPHS); a longitudinal survey of the Canadian population. The research question of interest is: How does the average number of hours of participation in leisure time PA per day affect hourly wages and annual income?

There are a limited number of studies in the literature that focus on the effect of participation in leisure time sports and/or PA on labour market outcomes, when individuals are in their prime working years. These studies suffer from a number of shortcomings mostly due to either lack of detailed information (including frequency, time spent, and intensity of participation) on leisure time PA, and/or lack of a dataset capturing changes in both labour market outcomes and participation in PA over a sufficiently long period of time. There is also a wide difference in research design used in these studies. One important factor to be considered is that participation in PA is episodic and individuals have movements in and out of PA meaning that they may be physically active in some periods and not active (or less active) in others. It is therefore important of capture Individuals' PA at different points in time. Another important issue often neglected (or poorly addressed) in previous studies is the potential endogeneity of PA. It has been shown that wages/income affect the decision to participate as well as time spent on physical activity (see for example Humphreys and Ruseski 2011), causing a potential endogeneity in estimating the effect of PA on wages. Unobserved individual characteristics such as time preference, and omitted variables that affect both wages and time spent in LTPA are other possible sources

of endogeneity. Failing to account for these econometric issues in estimating the causal effect of PA on wages is likely to produce inconsistent estimates. Existing studies in the literature fail to account for one or a number of the above mentioned factors.

This study extends on existing studies in a number of ways. First of all, detailed information on participation (frequency and time spent) in PA in the NPHS dataset allows us to calculate the number of hours spent on PA per day. The physical activity variable is constructed in a way that reflects both frequency and duration (amount of time spent) of participation not just participation per se. In most previous studies, the indicators used for participation in PA only capture participation as a dummy variable or capture frequency of participation at best, while the PA variable used in this study contains more detailed information on PA thus creating more variation in terms of participation in PA across the sample and providing more detailed information on PA habits of individuals. In addition, the longitudinal nature of the data allows for capturing transitions in and out of PA (rather than measuring PA at a single point in time) thus capturing how these transitions/changes ultimately affect labour market outcomes. Finally, I add to the literature by using a methodology that accounts for the endogeneity issues in estimating the casual effect of PA on wages. One challenge in estimating the effect of participation in LTPA on wages is the potential endogeneity of the physical activity variable. By taking advantage of the longitudinal nature of the data, I estimate the effect of PA on labour market earnings using a variety of panel data estimation techniques including static and dynamic models. I find that increasing lagged PA has a positive and significant

impact on hourly wages and annual income among men, and the point estimates are robust to a variety of estimation approaches, and to the inclusion of addition controls for self-rated health status and Body Mass Index (BMI). However, increasing lagged PA does not have a significant impact on either hourly wages or annual income among women.

This paper is organized as follows. In section 1.2 I discuss the relevant literature, in section 1.3 I explain about the methodology used to identify the model, in section 1.4 the data used for this study and some characteristics of the study sample are explained, in section 1.5 the results of the baseline model are presented and discussed, in section 1.6 I check for the robustness of the baseline results, and 1.7 I provide concluding remarks.

1.2. Literature Review

Empirical studies on the direct effect of physical activity on wages are scarce (Lechner 2015). There is a group of studies in the literature that examines the effect of a single measure of sports participation among young people, on educational attainment and future earnings. Some of these studies particularly focus on sports participation within the educational establishments, such as high school athletic participation, and most of them do not necessarily focus on causal effects. Ewing (1998) examined how high school athletic participation affects labour market outcomes in terms of performance-based pay, union member and number of workers supervised. He found that athletes are more likely to be in jobs that are associated

with better labour market outcomes. Also, Ewing (2007) found that former athletes have better outcomes in terms of wages and fringe benefits than their non-athlete counterparts. (Long and Caudill 1991) found an annual income premium of %4 for males who participated in intercollegiate athletics, and also suggested that athletic participation can result in increased levels of discipline, confidence, motivation, a competitive spirit and other factors that influence success. Barron et al. (2000) implemented a simple allocation of time model to and used an instrumental variables approach to explain individuals' choice to participate in high school athletics and how this choice affects their educational attainment and wages. The instruments they used include school size, and other school characteristics as well as the parents' income and the students' health, height and weight. They found a direct link between athletic participation and wages; although most of this association seems to be related to the individuals' ability and how they value leisure (see also Henderson et al. 2006, Persico and Postlewaite 2004, and Stevenson 2010). I is clear that some of the instruments used by Barron et al. 2000 such as health, height and weight are weak as they can be affected by athletic participation and may also directly affect the outcomes examined in their study. On the contrary, Maloney and Mccormick (2016) found that participation in college athletics reduced scholarly success, however they argued that the results may be due to sample selection effects since athlete entrants to high school had lower overall standardised test score. Eide and Ronan (2001) used height of students as an instrumental variable in order to examine the effect of high school-sponsored sports participation on academic success among different gender and race subgroups. They found that while black males and white females who

participated in sports had higher academic success, sports participation among white male had a negative impact on educational attainment. High school sports participation was also associated with higher earnings among black male student athletes. While height is a relatively stronger instrument than health or weight, it does not provide a high degree of exogenous variation in the treatment variable to allow for identification of causal effects. The fixed effects method along with information on the joining and quitting of clubs by individuals were used in this study to distinguish between selection and causality, while it can be argued that parental choice might confound the relationships between participation and club activity. Rees and Sabia (2010) also used students' height as an instrumental variable to estimate the impact of sports participation on a set of academic indicators such as grade-point average, paying attention in class and college aspirations. Stevenson (2010) used legislative change as an instrumental variable for sport participation to examine how high school girls' sports participation affects their college and labor force participation. He concluded that sports participation has a positive impact on labor force participation but its impact on wages is unstable. Using distance from sports facilities as an instrumental variable, and lagged variables to account for revers causality, Felfe et al. (2011) found that participation in club sports positively affects children's cognitive and non-cognitive development. Overall, most of these studies are suggestive a positive effect of sports participation on educational attainment and earnings, however the majority of them have used questionable methods and/or instruments in an attempt to identify causal relationships.

Another group of studies examine the association between participation in leisure time physical activity and/or sports on earnings and labour market outcomes. Using the German Socioeconomic Panel, Cabane (2014) found that men who reported participation in sports at least weekly earned 5% more than those who did not. Lechner (2009) estimated the long-run effect of sports participation on labour market outcomes using a semi-parametric matching estimation on the GSOEP and found a significant long run impact of sports participation on earnings and wages. Clark and Cabane (2013) concluded that lagged participation in sports is associated with higher wages. Rooth (2011) used a siblings fixed effects model on Swedish males and found that being physically fit increases earnings by 4% without controlling for cognitive skills, and 1.7% after controlling for cognitive skills. Kosteas (2012) estimated the causal relationship between participation in LTPA and wages in the US using a fixed effects method and a propensity score matching method, and found a 6% to 10% wage increase associated with regular exercise. Lechner and Downward (2013) estimated the effect of sports participation on annual household income among men and women aged 26 to 45. They found that men and women (in the mentioned age range) who participated in different sports earned respectively 4300 to 6500 GBP, and 3400 to 5300 GBP more in terms of annual household income than those who did not (the amount gained differs by the type of sport). Hyytinen and Lahtonen (2013) also used a sample of male twins to control for unobserved genetic confounding factors in estimating the long run effect of physical activity on income over a fifteen year period. Their within twin estimates suggest that physical activity has a positive effect on long term income. Lechner and

Sari (2015) used the Canadian National Population Health Survey (NPHS) in order to analyse the effect of a change in PA level form inactive to moderate and from moderate to active levels of participation. They used the same method as Lechner (2009) but with more informative data and found that only a change from moderate to active level had a significant impact (10% to 20%) on earning in the long-run (8 to 12 years).

This study makes a number of contributions to the literature. First of all I measure LTPA in a more detailed fashion using rich information on both frequency and time spent which gives a great deal of variation and heterogeneity among individuals with respect to their LTPA. This will allow me to examine whether there exists a "dose response" to an increase in the amount of time spent on LTPA. Most of the previous studies focus on team sports participation in high school or college when individuals are much younger, however age has been shown to be an important determinant of participation in sports and PA and normally participation decreases as people age (for example see Humphreys and Ruseski 2011). The only other study in the literature that examines a dose response effect to an increase in PA is the study by Lechner and Sari (2015) which only focuses on a change in the intensity (energy expenditure) of the activity level. However, intensities are important when only considering the health benefits of PA and its potential impact on earnings (individuals can reduce the amount of time spent on PA but increase the intensity of it), while using the amount of time spent on leisure time PA (implicitly) accounts for both the health and non-health benefits of participation in PA. I expect that the non-cognitive skills that individuals gain by participation in PA are mostly determined

by the amount of time they spend on these activates (especially on group sports) during their leisure time. The second contribution of this study is that I use methods beyond those implemented in previous studies in order to account for the endogeneity of LTPA, changes in the amount of time spent on LTPA over time (I measure LTPA at multiple time points in order to account for the episodic nature of LTPA and the fact that individuals have transitions in and out of LTPA), and the persistence of wages when people are in their prime working years/ages.

1.3. Methodology

In general, a panel data wage equation will take the following form:

$$\ln(W_{it}) = X_{it}\beta_1 + \beta_2 PA_{i,t\text{-}1} + v_{it} \qquad (1\text{-}1)$$

where $\ln(W_{it})$ is the log of hourly wages of individual i at time t, X_{it} is a vector of covariates affecting labour market earnings, $PA_{i,t\text{-}1}$ is a variable that represents the average number of hours of participation in physical activity per day of individual i at time t-1 (PA is in lag form since the impact of PA on labour market outcomes is not immediate), and $v_{it} = c_i + e_{it}$ is a composite error term which consists of a normally distributed error term, e_{it}, and the individual unobserved effect, c_i, also known as the latent variable or unobserved heterogeneity. The coefficient of interest in equation (1-1) is β_2 which shows the effect of an increase in lagged PA on the log of hourly wages today.

Assuming zero correlation between all the right-hand-side variables (X_{it} and $PA_{i,t\text{-}1}$) and the unobserved individual effects, equation (1-1) could be consistently

estimated by pooled OLS (Ordinary Least Squares) or by Random Effects (RE). However there is reason to expect that the unobservable individual factors that enter into the composite error term and affect wages, also affect the explanatory variables, particularly participation in PA. Some of the most important unobserved factors in our model are discipline, discount rate and ability. For example, Individuals with higher disciplines are more likely to perform better in the labour market and earn higher wages as a result. They are also more likely to engage in leisure time PA. Therefore the unobserved component in equation (1-1) is potentially correlated with the PA variable, which implies that the correlation between v_{it} and physical activity is no longer restricted to be zero. In this case, pooled OLS and random effects estimator are no longer capable of producing consistent estimates.

Assuming that the unobserved individual characteristics/omitted variables are constant over time, one possible panel data approach to eliminate the bias due to unobserved heterogeneity is to use a Fixed Effects (FE) estimation approach. FE relies on within individual variation in the covariates of the model hence omitting, from the estimation, the effect of any time invariant variable (Wooldridge 2002). However, in order for FE to produce consistent estimates we need to assume that the covariates in the model are exogenous once the unobserved time-invariant individual effect is eliminated ((E ($e_{it} \mid c_i, PA_{i,t-1}, X_{it}$) = 0). While implementing FE is a common approach in dealing with time-invariant individual unobserved effects, it does not account for potential endogeneity arising from time-variant omitted variables. The conventional approach in this case would be to use an instrument or a set of instruments that predict exogenous variation in PA but are uncorrelated with the error

term. Finding valid instruments in health related studies comes with challenge and this study is no exception. I therefore exploit the panel nature of the data in order to identify causal effects.

A natural extension to equation 1-1 is to add a lagged dependent variable (LDV) to the right hand side of the equation. Adding LDV can account for multiple important factors in determining wages: 1) The presence of LDV accounts for time-variant omitted variable that potentially bias the estimates due to endogeneity (Wooldridge 2010), 2) past wages are expected to have a strong impact on current wages especially due to the fact that individuals are in their prime working years and their wages are likely to be persistent, and 3) an important missing information in the data is the number of years of work experience of the respondents, therefore adding lagged wages can serve as a proxy for accumulated years of work experience and human capital. Incorporating the dynamics of wages in equation (1-1) yields the following dynamic wage equation:

$$\ln(W_{it}) = X_{it}\beta_1 + \beta_2 \ln(W_{i,t\text{-}1}) + \beta_3 PA_{i,t\text{-}1} + v_{it} \qquad (1\text{-}2)$$

where $\ln(W_{i,t\text{-}1})$ indicates log of hourly wages in the previous period (one cycle before). The coefficient of interest in equation (1-2) would be β_3 which captures the effect of an increase in lagged PA on current log of hourly wages. Equation (1-2) can be consistently estimated using pooled OLS assuming that there is no time-invariant unobserved component in the error term. Using FE to eliminate time-invariant unobserved heterogeneity in estimating equation (1-2) will result in dynamic panel bias given the small number of time-points (small T) in the sample (Nickel 1981). Holtz-Eakin et al. (1988) suggested removing the unobserved

individual component by first differencing. As a result of first differencing equation (1-2) becomes the following:

$$\Delta(\ln(W_{it})) = \Delta X_{it}\beta_1 + \beta_2\Delta(\ln(W_{i,t\text{-}1})) + \beta_3\Delta PA_{i,t\text{-}1} + \Delta e_{it} \quad (1\text{-}3)$$

Note that $\Delta(\ln(W_{i,t\text{-}1}))$ is still correlated with the error term since $\ln(W_{i,t\text{-}1})$ is correlated with $e_{i,t\text{-}1}$. Holtz-Eakin et al. (1988) argue that if there is no serial correlation between the error terms, further lags of the endogenous variables will be highly correlated with the endogenous differenced variables but uncorrelated with the error terms. Therefore $\ln(W_{i,t\text{-}2})$ can be used as an instrument for $\Delta(\ln(W_{i,t\text{-}1}))$. Similarly, assuming that $\Delta PA_{i,t\text{-}1}$ is also endogenous, it can be instrumented using $PA_{i,t\text{-}2}$. Arellano et al. 1991 argue that if the above approach is performed in a GMM (Generalized Method of Moments) context, it produces more efficient estimates (GMM is an approach that was first introduced by Hansen (1982)). This method, known as the First Differenced-GMM (FD-GMM) approach, has become increasingly popular in estimating dynamic panel data models when dealing with small T and large N panel dataset (small time period and large sample sizes), independent variables that are not strictly exogenous, and the presence of fixed effects (unobserved heterogeneity) (David Roodman 2009).

In the AB approach, there is a system of equations implicitly, one for each time period, where in each equation a different set of instruments is used. For example at time $t = 3$, the wage variable at time $t = 1$ is used as an instrument for $\Delta(\ln(W_{i2}))$. At $t = 4$, both wage variables at $t = 1$ and $t = 2$ are used as instruments for $\Delta(\ln(W_{i3}))$ and so on. Using these lagged variables as instruments implies that $E(\ln(W_{i,t\text{-}s})\,\Delta e_{it}) = 0$ for all i and t where $s = 1 \ldots \infty$, meaning that at each time period,

the differenced error term is uncorrelated with the instrument(s) used in that time period. All other variables in the model are treated as strictly exogenous and therefore act as their own instruments in the estimation.

Equation (1-3) is AR (1) (first-order autocorrelation of the residuals) by design since both Δe_{it} and $\Delta e_{i,t\text{-}1}$ have the component of $e_{it\text{-}1}$ in common which makes the transformed error terms correlated at time t and t-1. This in fact justifies using the second and later lags of the endogenous variables as instruments in the difference-GMM estimation. However, in the existence of serial correlation between the error terms, equation (1-3) will be at least AR (2) since $e_{i,t\text{-}1}$ in Δe_{it} will be correlated with $e_{i,t\text{-}2}$ in $\Delta e_{i,t\text{-}2}$. In this case the instruments need to be taken from one further lag behind. Therefore, at $t = 4$ instead of using both wages at $t = 1$ and $t = 2$ as instruments, we can only use wage at $t = 1$ as an instrument (at any given t use $\ln(W_{i,t\text{-}3})$ and further lags if available instead of $\ln(W_{i,t\text{-}2})$ as instruments) and so on. In the next section I discuss the results of the autocorrelation test as well as the Hansen J test for over identifying restrictions for joint validity of the moment conditions. The validity of the instruments is tested using Hansen J test which is robust to serial correlation and heteroscedasticity (Roodman 2009).

A more efficient alternative to FD-GMM is System-GMM, first designed by Blundell & Bond (1998), which undertakes the same procedure as in FD-GMM but with one additional assumption that the lagged differences in the dependent variable (and other potential endogenous variables) greater than or equal to 1 are valid instrument for the lagged level of the variable in the levels equation. This additional moment condition implies that $E(\Delta(\ln(W_{i,t\text{-}s}))\, e_{it}) = 0$ for all i and t where $s = 1, \ldots,$

∞. As the System-GMM method uses the differenced version of the dynamic model as well as the levels version, it allows for the estimation of the effect of time-invariant characteristics on the outcome variable as well (while these variables are differenced out in the FD-GMM approach). As the large number of instruments becomes an issue when implementing a System-GMM approach, I collapse the instrument matrix in order to limit instrument count whenever a System-GMM is used. In all GMM estimations (which I discuss below), both wages and PA are instrumented. A potential drawback of the GMM estimators is that the instruments might be weak. In order to partially overcome the issue of identification by weak instruments, I use only one available lag (instead of all available lags) as instruments (in both FD-GMM and System-GMM estimations) in order to restrict the number of instruments. Also, for the purpose of checking the robustness of the estimates, I compare the results of the GMM estimators with the alternative static and dynamic panel data models.

All estimations are carried out in STATA 13. For the dynamic panel model estimation, the XTABOND2 command introduced by Roodman (2009) is used with the two-step FD-GMM or System-GMM option which produces robust standard errors with the Windmeijer (2005) finite sample correction.

1.4. Data

The data used in this study are taken from cycles 4 to 9 of the Canadian National Population Health Survey (NPHS) household component. The longitudinal data collection started in year 1994/1995 (cycle 1) with 17,276 respondents participating in the survey. The youngest respondent was 12 years old in the first cycle. The same

individuals were interviewed every two years until year 2010/2011 (cycle 9). The survey provides rich information about the respondents' health status and socioeconomic factors. It also provides information on frequency and intensity of participation in physical activity in the last three months. The data is a representative data set of the Canadian population in year 1994. The NPHS is based on a two-stage, stratified, cluster design. Statistics Canada has provided sample weights to account for the complex survey design. The initial survey weights represent the inverse probability of selection into the survey which are then adjusted for factors such as non-response and the longitudinal sample due to attrition. Final adjustment consists of post-stratification within each province to ensure consistency with population estimates based on the 1996 Canadian Census (Statistics Canada 2010). The regressions in this paper are weighted but the summary statistics are un-weighted.

In this paper I use data from cycles 4, 6, 7, 8 and 9 (years 2000, 2002, 2004, 2006, 2008, and 2010) of the survey as the income measure used to construct the wage variable is reported starting from cycle 4. The final sample is restricted to individuals who are aged greater than or equal to 22 in cycle 4 and less than or equal to 67 years in the last cycle. The lower age limit is to ensure that respondents are out of school/college when they enter into the analysis and the upper age limit is to ensure that they are not in the retirement age at the end of the analysis time period. The sample used in this study includes respondents with complete response patterns across all cycles.

The dependent variable of interest is the log of hourly wages of the respondents. Three pieces of information were used to construct the wage variables; annual income, number of weeks worked in the year in which income is reported, and total usual hours worked per week in the given year. Respondents were asked about their personal income before taxes from all sources in the past 12 months (12 months prior to the interview time). The responses range from zero to $500,000 per year. Real income is calculated by converting each monetary value to 2002 dollars using information on the consumer price index 2009 basket. The second piece of information used to construct the wage variables is the number of weeks the respondent worked in the same year income was reported and the number of hours worked per week. It is important to mention that the annual income variable reported in the NPHS includes income from all source not just wages and salaries. Therefore the annual income might potentially reflect income from sources other than labour market earnings. In order to overcome this issue, I restricted the sample in a way to ensure that most of the annual income is coming from labour market activities (this is discussed in more detail below). In addition, the NPHS question that is asked from the respondents about the number of weeks and hours worked includes hours spent on any paid leave (such as paid vacation, paid sick leave, etc.). The total amount of income earned each year is divided by the number of weeks worked during that year to get the average amount of income earned in a week for a given year (cycle). The third piece of information used to construct hourly wages is the total usual number of hours the respondent worked per week (including paid vacation, paid sick leave, maternity leave, etc.). Hourly wage is estimated by dividing the average weekly

income by the total number of hours worked per week. Following Kosteas (2012) cycles in which the respondent worked less than 500 hours or more than 3,500 hours per year have been deleted from the sample in order to drop observations in which the individual had a weak labour market attachment or in cases where there has been an error in reporting the hours worked. Observations with weekly income of less than 100 dollars have also been dropped. Hourly wages below 1 dollar has been recoded to 1. Finally hourly wages enters the model in natural log form. The final sample is split by gender. Implementing all the restrictions to the sample, dropping observations with missing information on the dependant variable or any of the explanatory variables, and keeping individuals with at least three consecutive cycles of information results in an unbalanced panel (but with no gaps). The male and female subsamples each include approximately 1,000 individuals. The male sample includes 5,394 observation-years and the female sample includes 5,211 observation-years.

The explanatory variable of interest is the average number of hours spent on physical activity per day which is based on the response to the question on the frequency of participation in physical activity in the last three months and the amount of time spent on each occasion. Activities include walking, gardening, jogging, different sports such as basketball, volleyball and etc. Multiplying the "number of times participated in the last three months" by "the amount of time spent on each occasion", gives us the total amount of time (in minutes) spent on physical activity in the past three months. This number is divided by 3 which gives us the average minutes spent on physical activity in one month, then divided by 30 which gives the

average minutes per day, and finally divided by 60 which gives the average number of hours spent on all leisure time physical activities per day.

The demographic and socioeconomic variables used in this study include age, highest level of education, marital status, binary variables for whether the respondent is Canadian born, number of young children (under the age of 12) in the household, provincial dummies (based on the province of residence), and occupation indicators. The education variable consists of 4 binary variables indicating the highest level of education completed by the respondent: “less than high school”, “high school graduate”, “some college studies” and “college graduate”. The first group serves as the reference group. Marital status is captured by three binary variables for whether the individual is “married/living/ in a common law relationship/living with a partner”, “single”, or “widowed/divorced/separated”. The reference category is “married”. The occupation indicators/categories include management occupations (reference category), occupations related to business and finance, health, sales and services, natural and applied sciences, social sciences/education/government/religion, culture/recreation/sports, trades/transportation/equipment, and occupations unique to primary industry and unique to production/manufacturing/utility.

1.4.1. Descriptive Statistics

Table 1.1and Table 1.3 show the un-weighted descriptive statistics for the male subsample pooled across all 6 cycles and by each cycle respectively. The average

male in the sample is 45 years old and spends about 0.62 hours on LTPA per day. Average annual income is $60,341 (10.83 in log form). Average number of weeks worked per year is 50 weeks and an average of 45 hours is spent on labour market work per week. Mean wage is $28 per hour (3.17 in log form). The majority of men are Canadian born, married, and have graduated from college. By looking at the summary statistics by cycle we can observe that real annual income and real hourly wages have constantly increased over time while the average number of weeks worked per year and hours worked per week have remained fairly constant over time. Average number of hours spent on LTPA increase initially till year 2002, decreased from 2002 to 2006 and then increased again until the end of the observation period.

Table 1.2 and Table 1.4 show the un-weighted descriptive statistics for the female subsample pooled across all 6 cycles and by each cycle respectively. Real annual income of an average female is significantly lower than that of an average male ($39,416 per year). Average number of weeks worked is 50 weeks while average number of hours worked per week is 39 hours which is lower than that of an average male. The mean log of hourly wages among women is 2.90 which translates into an hourly wage of about $21. An average woman spends about 0.61 hours on LTPA per day. The majority of women are Canadian born, married, and are college graduates. Similar to the pattern observed in the male subsample, average number of hours spent on LTPA increased from 2000 to 2002, but then decreased for a while, and again increased from 2006 to 2010.

Table 1.1 Summary Statistics for the Male Subsample

Variable	Mean	Standard Deviation
Real Annual Income	60341	43796
Log of Real Annual Income	10.83	0.58
Real Hourly Wage	28.23	20.96
Log of Real Hourly Wage	3.17	0.56
Weeks Worked per Year	49.65	6.86
Hours Worked per Week	44.46	8.98
PA_Average Daily Hours	0.62	0.55
Age	44.65	9.54
Canadian Born	0.90	0.30
Married	0.79	0.40
Single	0.13	0.33
Wid/Div/Sep	0.08	0.27
No High School	0.10	0.30
High School Graduate	0.13	0.33
Some College	0.26	0.44
College Graduate	0.51	0.50
Kids	0.54	0.91
Management	0.14	0.34
Bus/Finance/Admin	0.13	0.34
Natural and Applied Sci	0.13	0.34
Health	0.02	0.15
Soc/Educ/Gov/Religion	0.08	0.28
Culture/Recreation/Sports	0.01	0.12
Sales/Services	0.13	0.33
Trades/Transport/Equip	0.24	0.43
Primary Industry	0.04	0.20
Produc/Manu/Util	0.07	0.26
Newfoundland and Labrador	0.07	0.25
Prince Edward Islands	0.05	0.23
Nova Scotia	0.05	0.22
New Brunswick	0.05	0.22
Quebec	0.24	0.43
Ontario	0.24	0.43
Manitoba	0.06	0.23
Saskatchewan	0.07	0.25
Alberta	0.09	0.29
British Columbia	0.08	0.27
Observations	5394	

Table 1.2 Summary Statistics for the Female Subsample

Variable	Mean	Standard Deviation
Real Annual Income	39416	23873
Log of Real Annual Income	10.42	0.59
Real Hourly Wage	20.98	13.29
Log of Real Hourly Wage	2.90	0.54
Weeks Worked per Year	49.79	6.35
Hours Worked per Week	38.79	9.14
PA_Average Daily Hours	0.61	0.51
Age	44.36	9.02
Canadian Born	0.90	0.30
Married	0.68	0.47
Single	0.15	0.36
Wid/Div/Sep	0.17	0.38
No High School	0.06	0.24
High School Graduate	0.13	0.33
Some College	0.25	0.44
College Graduate	0.56	0.50
Kids	0.41	0.77
Management	0.08	0.27
Bus/Finance/Admin	0.32	0.47
Natural and Applied Sci	0.03	0.18
Health	0.11	0.32
Soc/Educ/Gov/Religion	0.19	0.39
Culture/Recreation/Sports	0.03	0.17
Sales/Services	0.18	0.38
Trades/Transport/Equip	0.02	0.12
Primary Industry	0.01	0.10
Produc/Manu/Util	0.03	0.17
Newfoundland and Labrador	0.06	0.23
Prince Edward Islands	0.06	0.24
Nova Scotia	0.06	0.24
New Brunswick	0.05	0.22
Quebec	0.19	0.39
Ontario	0.23	0.42
Manitoba	0.07	0.26
Saskatchewan	0.07	0.25
Alberta	0.10	0.30
British Columbia	0.10	0.30
Observations	5211	

Table 1.3 Summary Statistics on Select Variables for the Male Subsample by Cycle

Variable	2000	2002	2004	2006	2008	2010
Real Annual Income	54865	56576	58338	62617	64896	65765
	(37428)	(37313)	(40764)	(46792)	(48594)	(50456)
Log of Real Annual Income	10.75	10.78	10.81	10.86	10.89	10.91
	(0.57)	(0.55)	(0.56)	(0.58)	(0.60)	(0.59)
Real Hourly Wage	25.49	26.39	27.27	29.13	30.24	31.45
	(19.33)	(16.55)	(19.68)	(20.46)	(23.45)	(25.78)
Log of Real Hourly Wage	3.08	3.13	3.14	3.20	3.22	3.26
	(0.55)	(0.53)	(0.55)	(0.57)	(0.60)	(0.58)
Weeks Worked per Year	49.65	49.75	49.73	49.48	49.68	49.63
	(7.15)	(6.82)	(6.89)	(6.93)	(6.70)	(6.68)
Hours Worked per Week	44.75	44.05	44.56	44.61	44.73	43.99
	(8.87)	(8.69)	(9.10)	(9.18)	(8.82)	(9.23)
PA_Average Daily Hours	0.56	0.66	0.62	0.54	0.67	0.71
	(0.51)	(0.59)	(0.52)	(0.52)	(0.56)	(0.57)
Age	40.29	42.16	44.14	45.85	47.34	48.80
	(9.23)	(9.23)	(9.25)	(9.13)	(8.99)	(8.78)
Canadian born	0.90	0.90	0.90	0.90	0.90	0.90
	(0.30)	(0.30)	(0.30)	(0.30)	(0.30)	(0.30)
Married	0.75	0.77	0.80	0.81	0.82	0.81
	(0.43)	(0.42)	(0.40)	(0.39)	(0.39)	(0.39)
Single	0.18	0.15	0.12	0.11	0.10	0.10
	(0.39)	(0.36)	(0.32)	(0.31)	(0.30)	(0.31)
Wid/Div/Sep	0.07	0.08	0.08	0.08	0.08	0.08
	(0.26)	(0.27)	(0.27)	(0.28)	(0.28)	(0.28)
No High School	0.10	0.10	0.10	0.10	0.09	0.09
	(0.29)	(0.30)	(0.30)	(0.30)	(0.29)	(0.28)
High School Graduate	0.14	0.13	0.13	0.12	0.12	0.12
	(0.34)	(0.34)	(0.33)	(0.33)	(0.33)	(0.32)
Some College	0.28	0.27	0.27	0.26	0.25	0.25
	(0.45)	(0.44)	(0.45)	(0.44)	(0.44)	(0.44)
College Graduate	0.49	0.50	0.50	0.52	0.53	0.54
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)

Standard deviations in parentheses.

Table 1.4 Summary Statistics on Select Variables for the Female Subsample by Cycle

Variable	2000	2002	2004	2006	2008	2010
Real Annual Income	35625	36154	37951	39430	43150	45647
	(19534)	(20839)	(22573)	(23938)	(26552)	(28269)
Log of Real Annual Income	10.34	10.34	10.39	10.41	10.51	10.56
	(0.55)	(0.59)	(0.57)	(0.61)	(0.59)	(0.59)
Real Hourly Wage	18.78	19.62	20.27	21.25	22.56	24.02
	(10.22)	(13.82)	(12.7)	(14.03)	(13.36)	(14.55)
Log of Real Hourly Wage	2.81	2.83	2.87	2.9	2.97	3.03
	(0.5)	(0.54)	(0.53)	(0.55)	(0.55)	(0.55)
Weeks Worked per Year	49.97	49.9	49.94	49.33	49.82	49.8
	(5.99)	(6.24)	(6.21)	(7.01)	(6.21)	(6.35)
Hours Worked per Week	38.66	38.21	38.58	38.84	39.37	39.22
	(8.71)	(9.07)	(9.16)	(9.42)	(9.07)	(9.39)
PA_Average Daily Hours	0.55	0.63	0.59	0.54	0.64	0.71
	(0.48)	(0.52)	(0.49)	(0.45)	(0.53)	(0.56)
Age	40.32	41.86	43.84	45.41	47	48.48
	(8.7)	(8.79)	(8.76)	(8.62)	(8.43)	(8.21)
Canadian born	0.9	0.9	0.89	0.9	0.9	0.91
	(0.3)	(0.3)	(0.31)	(0.3)	(0.3)	(0.29)
Married	0.64	0.65	0.68	0.68	0.7	0.7
	(0.48)	(0.48)	(0.47)	(0.47)	(0.46)	(0.46)
Single	0.19	0.17	0.15	0.14	0.13	0.13
	(0.39)	(0.38)	(0.35)	(0.35)	(0.33)	(0.33)
Wid/Div/Sep	0.17	0.17	0.17	0.18	0.17	0.17
	(0.38)	(0.38)	(0.38)	(0.38)	(0.38)	(0.38)
No High School	0.06	0.06	0.06	0.06	0.06	0.05
	(0.23)	(0.23)	(0.24)	(0.25)	(0.23)	(0.23)
High School Graduate	0.14	0.14	0.13	0.13	0.13	0.11
	(0.35)	(0.34)	(0.34)	(0.33)	(0.33)	(0.31)
Some College	0.26	0.26	0.26	0.25	0.24	0.25
	(0.44)	(0.44)	(0.44)	(0.44)	(0.43)	(0.43)
College Graduate	0.54	0.55	0.55	0.55	0.57	0.59
	(0.5)	(0.5)	(0.5)	(0.5)	(0.49)	(0.49)

Standard deviations in parentheses.

1.5. Results and Discussion

Table 1.5 shows the results of the static OLS, FE, and RE estimations for the male sample. Survey weights are used in all estimations except for the RE model for which weights were not allowed. The OLS estimation result for the male sample shows that, all else constant, an increase in lagged average daily hours of PA, is associated with an increase in hourly wages by 6%. Age consistently has a positive correlation with hourly wages. Widowed/divorced/separated men earn 7% less than married men. Graduating from high school, having some college studies, and being a college graduate are associated with increased hourly earnings by 7%, 19%, and 41% respectively compared to having no education. Having more kids under the age of 12 is associated with 4% increase in hourly wage. The FE estimation results for the male sample indicate that, after eliminating the effect of time-invariant unobserved variables, lagged PA still has a positive and significant impact on log of hourly wages (3%) but the magnitude is smaller compared to when pooled OLS is used. These results suggest that the OLS estimation approach, which does not account for the unobserved individual heterogeneity, producing upward biased estimates of this impact. The only other significant coefficient in the FE specification (other than the coefficient on the intercept) is the one on the number of young kids in the household which is still positive. The RE models results also show a positive and significant impact of increasing lagged PA on log of hourly wages of about 4%. In addition, the RE results indicate that being a high school graduate, having some college education, and being a college graduate all contribute positively to log of hourly wages by 3%,

20%, 34% respectively (compared to having no education completed). These numbers are smaller than that of the pooled OLS specification, with the exception of having some college studies which has the same magnitude of impact on earnings in both OLS and RE models. Unlike the OLS model where being single had no significant impact on log of hourly wages while being widowed/divorced/separated did, in the RE model being single decreases log of hourly wages by 9% (compared to being married) while being widowed/divorce/separated has no significant impact. Similar to the OLS and FE specification, the RE results also show a positive and significant effect of an increase in the number of young kids on log of hourly earnings. Overall the latter three specifications of the static model for men indicate that regardless of the assumption imposed on the unobserved individual time-invariant component of the error term, PA has a positive impact on log of hourly wages. The magnitude of the impact of PA on earnings is highest in the OLS specification and lowest in the FE model (The magnitude of the impact of PA on earnings in the RE model lies between that in the OLS and FE models).

Table 1.6 shows the static OLS, FE and RE estimations results for the female subsample. In all three cases, the coefficient on PA is insignificant. In the OLS specification, widowed/divorced/separated women earn 10% more than married ones. Being a high school graduate, having some college studies, and being a college graduate are associated with an hourly wage premium of 27%, 40%, and 57% respectively compared to having no high school education. An increase in the number of young kids in the household increases hourly earnings by 5% among women.

Table 1.5 Static OLS, FE, and RE Estimates for the Male Subsample

	OLS		FE		RE	
Variable	Coefficient	Std.Err	Coefficient	Std.Err	Coefficient	Std.Err
PA_Average Daily Hours	0.06***	0.02	0.03*	0.02	0.04***	0.01
Age	0.01***	0.00	−0.05	0.04	0.01***	0.00
Canadian Born	−0.04	0.04	.	.	−0.06	0.05
Single	−0.02	0.05	0.01	0.07	−0.09**	0.03
Wid/Div/Sep	−0.07**	0.03	0.10	0.06	−0.01	0.04
High School Graduate	0.07**	0.03	−0.03	0.20	0.03	0.05
Some College	0.19***	0.03	−0.03	0.21	0.20***	0.05
College Graduate	0.41***	0.03	0.06	0.21	0.34***	0.05
Kids	0.04***	0.01	0.05***	0.02	0.03***	0.01
Bus/Finance/Admin	−0.21***	0.05	0.03	0.06	−0.07*	0.04
Natural and Applied Sci	−0.11***	0.04	0.08	0.07	0.02	0.05
Health	0.02	0.09	0.00	0.14	0.15*	0.09
Soc/Educ/Gov/Religion	−0.23***	0.04	0.03	0.06	−0.09**	0.04
Culture/Recreation/Sports	−0.37***	0.06	0.05	0.12	−0.17	0.12
Sales/Services	−0.39***	0.03	0.02	0.05	−0.10**	0.04
Trades/Transport/Equip	−0.34***	0.03	0.07	0.05	−0.11***	0.03
Primary Industry	−0.54***	0.06	0.01	0.10	−0.15**	0.07
Produc/Manu/Util	−0.29***	0.04	0.11	0.08	−0.10**	0.05
Constant	2.73***	0.10	4.94***	1.79	2.76***	0.12
R-squared (overall)	0.24		0.01		0.19	
R-squared (within)			0.05		0.04	
R-squared (between)			0.02		0.24	
Observations	4401		4401		4401	

Dependent variable is log of hourly wage.
* p < 0.10, ** p < 0.05, *** p < 0.01
OLS: Ordinary Least Squares; FE: Fixed Effects; RE: Random Effects

Table 1.6 Static OLS, FE, and RE Estimates for the Female Subsample

	OLS		FE		RE	
Variable	Coefficient	Std.Err	Coefficient	Std.Err	Coefficient	Std.Err
PA_Average Daily Hours	0.01	0.02	−0.01	0.02	0.01	0.01
Age	0.01***	0.00	0.00	0.04	0.01***	0.00
Canadian Born	0.07**	0.03	.	.	0.04	0.05
Single	0.01***	0.03	−0.07	0.05	−0.01	0.03
Wid/Div/Sep	0.09***	0.03	−0.04	0.05	0.01	0.03
High School Graduate	0.27***	0.05	0.01	0.06	0.25***	0.06
Some College	0.40***	0.05	0.20	0.13	0.42***	0.06
College Graduate	0.57***	0.04	0.21*	0.12	0.56***	0.06
Kids	0.05***	0.01	0.01	0.02	0.03**	0.01
Bus/Finance/Admin	−0.23***	0.03	−0.06	0.05	−0.10***	0.04
Natural and Applied Sci	−0.02	0.05	−0.10	0.07	0.00	0.05
Health	−0.13***	0.04	0.03	0.07	0.03	0.04
Soc/Educ/Gov/Religion	−0.14***	0.03	0.04	0.08	0.02	0.04
Culture/Recreation/Sports	−0.08	0.06	−0.02	0.09	−0.04	0.07
Sales/Services	−0.57***	0.04	−0.07	0.06	−0.26***	0.04
Trades/Transport/Equip	−0.33***	0.08	0.06	0.14	−0.10	0.07
Primary Industry	−0.16	0.13	−0.14	0.17	−0.08	0.14
Produc/Manu/Util	−0.24***	0.05	0.10	0.09	−0.04	0.05
Constant	2.36***	0.09	2.74*	1.61	2.30***	0.13
R-squared (overall)	0.28		0.10		0.25	
R-squared (within)			0.08		0.06	
R-squared (between)			0.11		0.31	
Observations	4216		4216		4216	

Dependent variable is log of hourly wage.

* p < 0.10, ** p < 0.05, *** p < 0.01

OLS: Ordinary Least Squares; FE: Fixed Effects; RE: Random Effects

Table 1.7 shows the results of the dynamic models for the male subsample. The dynamic panel models are estimated using OLS, RE, and System-GMM. All estimations are performed using survey weights except for the RE model for which weights were not allowed. In order to restrict the number of instruments, the System-GMM is estimated using only one available lag (instead of all available lags at each time point) of PA and wages as IVs with the collapse option. The coefficient on lagged wages is positive and significant in all three specifications. The point estimates indicate that a one hour increase in daily LTPA results in a 2% increase in hourly wages in the OLS model, and results in a 3% increase in hourly wages in the RE and System-GMM models. In the OLS specification, having some college studies and being a college graduate increase hourly wages by 7% and 16% respectively compared to being a high school dropout. In addition, in the RE model, being single decreases hourly wages by 7% compared to being married. Having some college studies and graduating from college each increase hourly wages by 6% and 13% respectively compared to being a high school dropout. A comparison of the results of the static and dynamic models indicates that after adding a lagged dependent variable to the OLS and RE specifications, PA still has a positive and significant effect on earnings but the magnitude is lower compared to the case where (static OLS and static RE) the dynamic nature of wages in not accounted for.

Table 1.8 shows the above dynamic estimation results for females. The model for the female subsample is estimated using a FD-GMM (First Differenced-Generalized Method of Moments) approach since the System-GMM estimation approach did not produce estimates that would support the joint validity of the

moment conditions. Again, the FD-GMM model is estimated using only one available lag as IVs for PA and wages. The coefficient on lagged wages is positive and significant in the OLS and RE models but insignificant in the FD-GMM model. The coefficient on PA is insignificant in all three models. In the OLS and RE models, having some college studies and being a college graduate consistently have a positive impact on hourly wages compared to being a high school dropout. In addition, being a high school graduate and increasing the number of young kids in the household also contribute positively to earnings.

As mentioned earlier, the dynamic panel estimation is based on the assumption of no serial correlation between the error terms and the validity of the over-identifying restrictions given that multiple instruments are used in the estimation process. The AR (2) test is based on the null hypothesis that the residuals have no autocorrelation of degree 2, so rejection of the null hypothesis means rejecting the presence of AR (2). In both cases, the corresponding p-values of the AR (2) test indicates that the null hypothesis cannot be rejected at any conventional level of significance. The Hansen J test for over identifying restrictions is based on the null hypothesis that all the instruments are jointly valid. The corresponding p-values for this test in both male and female estimations indicate that the null hypothesis cannot be rejected and thus indicating the joint validity of the moment conditions.

Overall, the results of different specifications of the model indicate that PA consistently has a positive and significant impact on hourly wages, and the point

estimates are sensitive to controlling for lagged wages. On the other hand, PA has no insignificant impact on the wages of women.

Although a variety of estimation techniques are used above for examining the effect of PA on earnings, I further check for the robustness of the estimates by a) using an alternative measure of earnings, and by b) adding control variables to the model.

Table 1.7 Dynamic OLS, RE, and System-GMM Estimates for the Male Subsample

	OLS		RE		System-GMM	
Variable	Coefficient	Std.Err	Coefficient	Std.Err	Coefficient	Std.Err
Lagged Wage	0.72***	0.02	0.72***	0.02	0.24***	0.06
PA_Average Daily Hours	0.02*	0.01	0.03**	0.01	0.03*	0.02
Age	0.00**	0	0.00***	0	0.01***	0
Canadian Born	−0.03	0.03	−0.02	0.02	−0.03	0.04
Single	−0.01	0.03	−0.07***	0.02	−0.03	0.07
Wid/Div/Sep	0	0.02	−0.02	0.02	−0.04	0.03
High School Graduate	0.03	0.02	0.01	0.02	0.06	0.04
Some College	0.07***	0.02	0.06***	0.02	0.16***	0.04
College Graduate	0.16***	0.02	0.13***	0.02	0.32***	0.04
Kids	0.01	0.01	0.01	0.01	0.02*	0.01
Bus/Finance/Admin	−0.04	0.03	−0.05**	0.02	−0.14**	0.06
Natural and Applied Sci	−0.05**	0.02	−0.03	0.02	−0.04	0.04
Health	0	0.08	0.01	0.05	0.03	0.07
Soc/Educ/Gov/Religion	−0.07**	0.03	−0.07***	0.02	−0.12***	0.05
Culture/Recreation/Sports	−0.08*	0.05	−0.05	0.04	−0.17**	0.07
Sales/Services	−0.12***	0.02	−0.09***	0.02	−0.25***	0.04
Trades/Transport/Equip	−0.09***	0.02	−0.09***	0.02	−0.20***	0.04
Primary Industry	−0.20***	0.04	−0.15***	0.04	−0.35***	0.07
Produc/Manu/Util	−0.08***	0.03	−0.09***	0.02	−0.18***	0.05
Constant	0.86***	0.1	0.90***	0.08	2.14***	0.23
R-squared (overall)	0.61		0.60			
R-squared (within)			0.003			
R-squared (between)			0.91			
Observations	4401		4401		4401	
Number of Instruments					35	
AR (1)					z = −6.17	p > z = 0.00
AR (2)					z = 1.57	p > z = 0.12
Hansen J test					chi2 = 3	p > chi2 = 0.22

Dependent variable is log of hourly wage.

* p < 0.10, ** p < 0.05, *** p < 0.01.

OLS: Ordinary Least Squares; RE: Random Effects; System-GMM: System-Generalized Method of Moments; AR (.): Test for the presence of autocorrelation of first and second degree; Hansen test: test for the joint validity of the instruments.

Table 1.8 Dynamic OLS, RE, and FD-GMM Estimates for the Female Subsample

	OLS		RE		FD-GMM	
Variable	Coefficient	Std.Err	Coefficient	Std.Err	Coefficient	Std.Err
Lagged Wage	0.63***	0.02	0.64***	0.02	0.10	0.05
PA_Average Daily Hours	0.00	0.01	0.01	0.01	−0.04	0.03
Age	0.00**	0.00	0.00***	0.00	−0.01	0.04
Canadian Born	0.04*	0.02	0.02	0.02	.	.
Single	0.00	0.02	0.00	0.02	0.03	0.06
Wid/Div/Sep	0.03	0.02	0.02	0.01	0.05	0.05
High School Graduate	0.09**	0.04	0.06**	0.03	−0.17	0.06
Some College	0.13***	0.04	0.09***	0.03	−0.06	0.17
College Graduate	0.21***	0.04	0.16***	0.03	−0.09	0.16
Kids	0.02**	0.01	0.02**	0.01	0.01	0.03
Bus/Finance/Admin	−0.11***	0.03	−0.11***	0.02	−0.03	0.05
Natural and Applied Sci	−0.05	0.03	−0.02	0.03	−0.04	0.07
Health	−0.06*	0.03	−0.04	0.03	0.10	0.07
Soc/Educ/Gov/Religion	−0.05*	0.03	−0.04	0.02	0.14	0.10
Culture/Recreation/Sports	−0.03	0.04	−0.06*	0.04	−0.04	0.08
Sales/Services	−0.23***	0.03	−0.22***	0.03	0.02	0.07
Trades/Transport/Equip	−0.16***	0.05	−0.13**	0.05	0.03	0.11
Primary Industry	−0.08	0.11	−0.09	0.09	−0.11	0.13
Produc/Manu/Util	−0.11***	0.04	−0.10***	0.03	0.14	0.11
Constant	0.96***	0.09	1.01***	0.08	.	.
R-squared (overall)	0.57		0.57			
R-squared (within)			0.003			
R-squared (between)			0.90			
Observations	4216		4216		3221	
Number of Instruments					28	
AR (1)					z = −7.35	p > z = 0.00
AR (2)					z = 1.33	p > z = 0.18
Hansen J test					chi2 =4.12	p > chi2 = 0.66

Dependent variable is log of hourly wage.
* p < 0.10, ** p < 0.05, *** p < 0.01.
OLS: Ordinary Least Squares; RE: Random Effects; FD-GMM: First Differenced-Generalized Method of Moments; AR (.): Test for the presence of autocorrelation of first and second degree; Hansen test: test for the joint validity of the instruments.

1.6. Robustness Checks

1.6.1. Alternative Measure of Earnings

The above estimates have been performed using an alternative measure of earnings; annual income. As mentioned in the data section, this variable is based on income of the respondent in the year prior to the interview date. Using annual income instead of hourly wages, allows for a more comprehensive estimation of the impact of PA on earnings. The hourly wage variable adjusts for the number of hours worked and thus eliminates the potential impact of the difference in working hours (between the respondents) on the estimates, whereas the annual income variable implicitly incorporates differences in the hours worked. Therefore, I expect hourly wage to provide a better measure of the respondents' actual labour market earnings, while differences in the annual income might potentially reflect differences in the amount of time spent on labour market activities. The equations take the same form as equation (1-1) and (1-2), however the dependent variable of interest in this case is log of annual income. The dynamic model also includes lagged log of annual income (instead of lagged log of hourly wages) on the right hand side. The model with log of annual income as dependent variable is estimated using all the specifications as in the baseline analysis. Since the Difference/System-GMM post-estimation results for the female subsample indicate the presence of AR (2), they are not reported as I cannot draw inferences based on those results (a potential solution to overcome the problem of AR (2) would be to add a second lag of income to the right-hand side of the model, but this approach would substantially reduce sample size).

Table 1.11 and Table 1.12 indicate the estimates using log of annual income for the male and female subsamples. In the male subsample, the results of the static OLS and RE models indicate that an increase of one hour in PA is associated with a 4% and 2% increase in annual income respectively. The dynamic RE model also shows a positive and significant impact of PA on annual income (about 2%). The coefficient on PA is insignificant in the static FE and dynamic OLS. Finally using a system GMM approach (for the estimation of the dynamic model) with one available lag of income and PA as IVs, I find that PA has a positive and significant impact on annual income with a magnitude of 0.04. The test statistics for the presence of AR (2) and the joint validity of the instruments indicate that there is no second degree serial correlation between the errors and that the moment conditions are jointly valid.

Among women, similar to the results found in the previous section, PA does not have a significant impact on annual income. This finding holds using all different specifications of the model.

Table 1.9 Static OLS, FE, and RE Estimates for the Male Subsample

	OLS		FE		RE	
Variable	Coefficient	Std.Err	Coefficient	Std.Err	Coefficient	Std.Err
PA_Average Daily Hours	0.04***	0.02	0.02	0.01	0.02**	0.01
Age	0.01***	0.00	0.01	0.04	0.01***	0.00
Canadian Born	−0.04	0.03	.	.	−0.08	0.05
Single	−0.11**	0.05	−0.05	0.04	−0.10***	0.03
Wid/Div/Sep	−0.08***	0.03	0.04	0.04	−0.01	0.03
High School Graduate	0.11***	0.04	−0.40**	0.20	0.03	0.06
Some College	0.22***	0.03	−0.43**	0.20	0.17***	0.06
College Graduate	0.43***	0.03	−0.32	0.20	0.33***	0.06
Kids	0.04***	0.01	0.03*	0.02	0.03***	0.01
Bus/Finance/Admin	−0.37***	0.05	−0.05	0.04	−0.12***	0.04
Natural and Applied Sci	−0.27***	0.04	0.02	0.05	−0.02	0.04
Health	−0.10	0.10	−0.06	0.14	0.04	0.09
Soc/Educ/Gov/Religion	−0.34***	0.04	0.00	0.05	−0.10***	0.04
Culture/Recreation/Sports	−0.54***	0.06	0.00	0.08	−0.21**	0.10
Sales/Services	−0.51***	0.04	−0.01	0.05	−0.12***	0.04
Trades/Transport/Equip	−0.47***	0.03	0.07	0.04	−0.10***	0.03
Primary Industry	−0.75***	0.06	−0.10	0.08	−0.27***	0.06
Produc/Manu/Util	−0.44***	0.04	0.10**	0.05	−0.10***	0.04
Constant	10.68***	0.09	10.65***	1.57	10.72***	0.14
R-squared (overall)	0.27		0.0001		0.21	
R-squared (within)			0.05		0.04	
R-squared (between)			0.0007		0.24	
Observations	4401		4401		4401	

Dependent variable is log of annual income

* p < 0.10, ** p < 0.05, *** p < 0.01

OLS: Ordinary Least Squares; FE: Fixed Effects; RE: Random Effects

Table 1.10 Static OLS, FE, and RE Estimates for the Female Subsample

	OLS		FE		RE	
Variable	Coefficient	Std.Err	Coefficient	Std.Err	Coefficient	Std.Err
PA_Average Daily Hours	−0.01	0.02	−0.03	0.02	−0.01	0.01
Age	0.01***	0.00	0.01	0.03	0.00**	0.00
Canadian Born	0.07**	0.03	.	.	0.04	0.05
Single	0.04	0.03	−0.02	0.06	0.02	0.03
Wid/Div/Sep	0.13***	0.03	−0.01	0.04	0.02	0.03
High School Graduate	0.27***	0.05	−0.12*	0.07	0.29***	0.07
Some College	0.37***	0.05	−0.13	0.15	0.46***	0.06
College Graduate	0.56***	0.05	−0.11	0.15	0.58***	0.06
Kids	0.02	0.02	−0.02	0.02	0.00	0.01
Bus/Finance/Admin	−0.41***	0.03	−0.15***	0.06	−0.19***	0.04
Natural and Applied Sci	−0.09**	0.05	−0.17**	0.09	−0.08	0.06
Health	−0.33***	0.04	−0.06	0.11	−0.09	0.06
Soc/Educ/Gov/Religion	−0.21***	0.04	−0.09	0.09	−0.02	0.05
Culture/Recreation/Sports	−0.27***	0.07	−0.12	0.09	−0.14***	0.06
Sales/Services	−0.76***	0.04	−0.22***	0.07	−0.34***	0.05
Trades/Transport/Equip	−0.38***	0.09	0.10	0.14	−0.15*	0.08
Primary Industry	−0.26**	0.13	−0.16	0.12	−0.20*	0.10
Produc/Manu/Util	−0.37***	0.05	−0.04	0.07	−0.12**	0.05
Constant	10.17***	0.10	10.25***	1.32	10.05***	0.14
R-squared (overall)	0.28		0.04		0.27	
R-squared (within)			0.10		0.10	
R-squared (between)			0.02		0.31	
Observations	4216		4216		4216	

Dependent variable is log of annual income

* p < 0.10, ** p < 0.05, *** p < 0.01

OLS: Ordinary Least Squares; FE: Fixed Effects; RE: Random Effects

Table 1.11 Dynamic OLS, RE, and System-GMM Estimates for the Male Subsample

	OLS		RE		System-GMM	
Variable	Coefficient	Std.Err	Coefficient	Std.Err	Coefficient	Std.Err
Lagged Annual Income	0.79***	0.02	0.80***	0.02	0.38***	0.07
PA_Average Daily Hours	0.01	0.01	0.02**	0.01	0.04**	0.02
Age	0.00**	0	0.00***	0	0.00**	0
Canadian Born	−0.02	0.02	−0.03*	0.02	−0.03	0.04
Single	−0.04	0.03	−0.07***	0.02	−0.08	0.05
Wid/Div/Sep	−0.01	0.02	−0.01	0.02	−0.03	0.03
High School Graduate	0.03	0.02	0.01	0.02	0.07	0.05
Some College	0.04**	0.02	0.04***	0.02	0.13***	0.04
College Graduate	0.11***	0.02	0.09***	0.02	0.28***	0.04
Kids	0	0.01	0	0.01	0.02	0.01
Bus/Finance/Admin	−0.06**	0.02	−0.07***	0.02	−0.14**	0.06
Natural and Applied Sci	−0.08***	0.02	−0.05***	0.02	−0.12***	0.04
Health	−0.03	0.08	−0.04	0.05	−0.01	0.09
Soc/Educ/Gov/Religion	−0.07***	0.03	−0.06***	0.02	−0.14***	0.04
Culture/Recreation/Sports	−0.09**	0.04	−0.06	0.04	−0.21***	0.08
Sales/Services	−0.11***	0.02	−0.09***	0.02	−0.24***	0.05
Trades/Transport/Equip	−0.10***	0.02	−0.09***	0.02	−0.21***	0.04
Primary Industry	−0.20***	0.03	−0.17***	0.03	−0.39***	0.07
Produc/Manu/Util	−0.11***	0.02	−0.10***	0.02	−0.20***	0.05
Constant	2.48***	0.24	2.40***	0.21	6.67***	0.69
R-squared (overall)	0.71		0.70			
R-squared (within)			0.02			
R-squared (between)			0.94			
Observations	4401		4401		4401	
Number of Instruments					35	
AR (1)					z = −5.72	p > z = 0.00
AR (2)					z = 1.23	p > z = 0.22
Hansen J test					chi2 =1.15	p > chi2 = 0.56

Dependent variable is log of annual income

* p < 0.10, ** p < 0.05, *** p < 0.01

OLS: Ordinary Least Squares; RE: Random Effects; System-GMM: System-Generalized Method of Moments; AR(.): Test for the presence of autocorrelation of first and second degree; Hansen test: test for the joint validity of the instruments

Table 1.12 Dynamic OLS and RE for the Female Subsample

	OLS		RE	
Variable	Coefficient	Std.Err	Coefficient	Std.Err
Lagged Annual Income	0.74***	0.02	0.76***	0.02
PA_Average Daily Hours	−0.02	0.02	0	0.01
Age	0.00*	0	0.00***	0
Canadian Born	0.03	0.02	0.01	0.02
Single	0	0.02	0.01	0.01
Wid/Div/Sep	0.03**	0.02	0.02	0.01
High School Graduate	0.06	0.04	0.05*	0.02
Some College	0.08**	0.04	0.06**	0.02
College Graduate	0.15***	0.04	0.12***	0.02
Kids	0	0.01	0	0.01
Bus/Finance/Admin	−0.15***	0.02	−0.13***	0.02
Natural and Applied Sci	−0.09***	0.03	−0.06**	0.03
Health	−0.10***	0.03	−0.09***	0.02
Soc/Educ/Gov/Religion	−0.08***	0.02	−0.05***	0.02
Culture/Recreation/Sports	−0.11***	0.04	−0.11***	0.03
Sales/Services	−0.24***	0.03	−0.21***	0.03
Trades/Transport/Equip	−0.11**	0.05	−0.14***	0.05
Primary Industry	−0.10	0.11	−0.16**	0.07
Produc/Manu/Util	−0.16***	0.04	−0.15***	0.03
Constant	2.82***	0.22	2.69***	0.17
R-squared (overall)	0.70		0.71	
R-squared (within)			0.03	
R-squared (between)			0.94	
Observations	4216		4216	

Dependent variable is log of annual income
* p < 0.10, ** p < 0.05, *** p < 0.01
OLS: Ordinary Least Squares; RE: Random Effects

1.6.2. Added Controls

One way of testing for the robustness of the baseline estimates is to examine whether the estimates are sensitive to the inclusion of additional control variables. For this

purpose I added two explanatory variables to the baseline model; Body Mass Index (BMI) and a measure of self-rated health which is reported in 5 categories in the NPHS. It is established in the literature that physically active individuals have lower BMIs and that there is a wage penalty associated with increased BMI (for example see Chou et al. 2004). In addition, Mcleod and Ruseski (2015) used the NPHS data in order to examine the longitudinal relationship between participation in PA and a set of health indicators. They found that participation in PA significantly reduces the likelihood of being in fair/poor health (as opposed to being in good/very good/excellent health). The literature also shows a strong association between health and labour market earnings/wages. These additional control variables have been excluded from the baseline analysis since they are likely to be endogenous in a wage equation. Nevertheless, it is expected that they add explanatory power to the equation and allow for examining whether PA still has an impact on earnings even after controlling for two possible channels through which the effect is likely to occur. I have added these two explanatory variables to the models. In this case, the models take the same form as equation (1-1) and (1-2) but with two additional (contemporaneous) control variables. BMI is a continuous variables indicating body mass index (weight in kilograms divided by height in meters squared). The NPHS asks a question about the respondents' general health based on their own judgement, the responses are coded as 5 binary indicators: "poor", "fair", "good", "very good", and "excellent" health. I have used the same 5 binary variables as a measure of health status in the model (with "poor" health being the reference category).

Table 1.13 and Table 1.14 show the results of the robustness checks (using log of hourly wages as the dependent variable) after adding BMI and health indicators to the baseline models (for the male subsample). In all three static models, the coefficient on PA is still positive and significant with a magnitude identical to those in the baseline static models (except for the OLS specification where there is a very slight difference in the magnitude of the point estimates in the baseline OLS model and the OLS model with added controls). In the dynamic models with added controls, the coefficient on PA is significant in the RE model (with a magnitude slightly lower than that in the baseline dynamic RE model) but not in the OLS and system-GMM models.

Overall, these results indicate that the point estimates are robust to the inclusion of health status and BMI in the majority of model specifications used.

Table 1.13 Static OLS, FE, and RE Estimates with Added Controls for the Male Subsample

	OLS		FE		RE	
Variable	Coefficient	Std.Err	Coefficient	Std.Err	Coefficient	Std.Err
PA_Average Daily Hours	0.05***	0.02	0.03*	0.02	0.04***	0.01
Fair Health	0.29	0.21	0.20	0.21	0.10	0.15
Good Health	0.36*	0.21	0.18	0.21	0.10	0.14
Very Good Health	0.37*	0.21	0.18	0.21	0.13	0.14
Excellent Health	0.52**	0.21	0.14	0.21	0.14	0.14
BMI	0.00*	0.00	0.00	0.00	0.00	0.00
Age	0.01***	0.00	−0.05	0.04	0.01***	0.00
Canadian Born	−0.06*	0.04	.	.	−0.06	0.05
Single	0.02	0.05	0.01	0.07	−0.09***	0.03
Wid/Div/Sep	−0.07**	0.03	0.10	0.06	−0.01	0.04
High School Graduate	0.07**	0.03	−0.03	0.20	0.04	0.05
Some College	0.18***	0.03	−0.02	0.21	0.20***	0.05
College Graduate	0.39***	0.03	0.06	0.21	0.34***	0.05
Kids	0.04***	0.01	0.05***	0.02	0.03***	0.01
Bus/Finance/Admin	−0.21***	0.05	0.03	0.06	−0.08*	0.04
Natural and Applied Sci	−0.11***	0.03	0.08	0.07	0.02	0.05
Health	0.03	0.09	0.00	0.13	0.16*	0.09
Soc/Educ/Gov/Religion	−0.22***	0.04	0.03	0.06	−0.09**	0.04
Culture/Recreation/Sports	−0.37***	0.06	0.05	0.12	−0.17	0.12
Sales/Services	−0.40***	0.03	0.02	0.05	−0.10***	0.04
Trades/Transport/Equip	−0.33***	0.03	0.07	0.05	−0.11***	0.03
Primary Industry	−0.53***	0.06	0.00	0.10	−0.15**	0.07
Produc/Manu/Util	−0.29***	0.04	0.11	0.08	−0.10**	0.05
Constant	2.28***	0.23	4.85	1.79	2.62***	0.20

Dependent variable is log of hourly wage.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

OLS: Ordinary Least Squares; FE: Fixed Effects; RE: Random Effects.

Table 1.14 Dynamic OLS, RE, and System-GMM Estimates with Added Controls for the Male Subsample

	OLS		RE		System-GMM	
Variable	Coefficient	Std.Err	Coefficient	Std.Err	Coefficient	Std.Err
Lagged Hourly Wage	0.71***	0.02	0.71***	0.02	0.24***	0.06
PA_Average Daily Hours	0.02	0.01	0.02**	0.01	0.03	0.02
Fair Health	0.16	0.15	0.06	0.13	0.23	0.15
Good Health	0.14	0.14	0.07	0.12	0.19	0.15
Very Good Health	0.16	0.14	0.09	0.12	0.20	0.15
Excellent Health	0.18	0.14	0.11	0.12	0.24*	0.15
BMI	0.00	0.00	0.00	0.00	0.00	0.00
Age	0.00**	0.00	0.00***	0.00	0.01***	0.00
Canadian Born	−0.03	0.03	−0.02	0.02	−0.04	0.04
Single	−0.01	0.03	−0.06***	0.02	−0.03	0.06
Wid/Div/Sep	0.00	0.02	−0.02	0.02	−0.04	0.03
High School Graduate	0.03	0.02	0.01	0.02	0.06	0.04
Some College	0.06***	0.02	0.06***	0.02	0.16***	0.04
College Graduate	0.16***	0.02	0.13***	0.02	0.31***	0.04
Kids	0.01	0.01	0.01	0.01	0.02*	0.01
Bus/Finance/Admin	−0.04	0.03	−0.05**	0.02	−0.13**	0.06
Natural and Applied Sci	−0.05**	0.02	−0.03	0.02	−0.04	0.04
Health	0.00	0.08	0.01	0.05	0.03	0.07
Soc/Educ/Gov/Religion	−0.07***	0.03	−0.07***	0.02	−0.11***	0.04
Culture/Recreation/Sports	−0.08*	0.05	−0.06	0.04	−0.17**	0.07
Sales/Services	−0.12***	0.02	−0.10***	0.02	−0.25***	0.04
Trades/Transport/Equip	−0.09***	0.02	−0.09***	0.02	−0.20***	0.04
Primary Industry	−0.20***	0.04	−0.15***	0.04	−0.35***	0.07
Produc/Manu/Util	−0.08***	0.02	−0.09***	0.02	−0.17***	0.05
Constant	0.70***	0.17	0.80***	0.14	1.84***	0.26
Number of Instruments					40	
AR(1)					z = −6.12	p > z = 0.00
AR(2)					z = 1.54	p > z = 0.13
Hansen J Test					chi2 =3.19	p > chi2 =0.20

Dependent variable is log of hourly wage.

* p < 0.10, ** p < 0.05, *** p < 0.01

OLS: Ordinary Least Squares; RE: Random Effects, System-GMM: System-Generalized Method of Moments; AR(.): Test for the presence of autocorrelation of first and second degree; Hansen test: test for the joint validity of the instruments

1.7. Conclusion

In this study I estimated the effect of an increase in leisure time physical activity on labour market wages in a static and dynamic framework using 6 cycles of the longitudinal NPHS with detailed information on frequency and time spent in LTPA. The results of different model specifications show that, all else constant, an increase of one hour in lagged LTPA has a positive effect on log of hourly wages and log of annual income among men. These point estimates are robust to the inclusion of additional explanatory variables controlling for self-rated health and BMI. However, among women, increasing LTPA does not have a significant impact on either log of hourly wages or log of annual income. Overall, my findings complement the previously found effects of PA and/or sports participation on labour market outcomes among men. This study has a number of shortcomings which are either due to data limitations or methodology. The NPHS data does not collect information on wage rates but rather reports income from all sources. Although I tried to overcome this data limitation by imposing restrictions to the selected sample, having more detailed information on wages and salaries (with larger samples) could improve the precision of the estimates. Another shortcoming of the study is in the methodology used. As mentioned earlier, the instruments produced in the GMM estimations might be weak. This is expected to be less of an issue in this study since I restricted the number of instruments used in the analysis, which reduces the probability that the estimates are biased by weak instruments. Nevertheless, it would

be preferable to use a more valid exclusion restriction which was a challenge to find given the information available in the data.

The results of this study suggest that increasing public awareness about the labour market benefits of increased LTPA can motivate more people to engage in PA and/or increase their level of participation. It is also apparent that increasing the general level of PA will increase employees' productivity and labour market outcomes, therefore policy interventions aimed at expanding PA levels should be considered (Lechner 2015). For example, there are a number of studies that investigate the impact of employer-sponsored wellness promotion programs in the work place with the motivation of reducing insurance costs. One of the elements of these programs is providing free access to sports/PA clubs and/or work place exercise interventions. Workers who committed to the program showed a reduced prevalence of a number of health risk factors (for example see Chung et al. 2009 and Berry et al. 2010). Evidence suggests that these employer-sponsored health promotion programs may have benefits beyond reduced healthcare expenditures through increased job satisfaction as a result of increased exercise (Thogersen et al. 2005), and increased productivity of the workers (Mangione nd Quinn 1975) as a result of increased job satisfaction. Future research can focus on the effectiveness of these programs in raising the productivity of the labour force and increasing firms profit (Kosteas 2012).

2. The Causal Effect of Unemployment on Smoking: Evidence from the Canadian Community Health Survey

2.1. Introduction

It is well-established in the literature that the unemployed are more likely to engage in risky health behaviours than the employed. According to a systematic literature review by (Henkel 2011), smoking is the most commonly reported unhealthy behaviour among the unemployed. Tobacco use has been associated with substantial adverse health outcomes and economic losses. According to the World Health Organization (WHO), tobacco consumption is considered to be the first leading risk factor for disability and disease in high income countries. Smoking is linked to many types of chronic conditions, cancers, cardiovascular diseases, and a number of other health conditions, and is also known to be the strongest risk factor associated with premature mortality (World Health Organization 2009). There is also significant economic burden associated with tobacco consumption in Canada. Rehm et al. (2006) estimated the total economic burden of tobacco consumption to be $ 17.7 billion (based on 2002 data), which included health care costs and productivity losses attributable to premature death and disability resulting from tobacco related diseases.

These observations stress the importance of examining if and how unemployment affects individuals' health and lifestyle choices as the association between these two factors remains under researched in the economics literature. Understanding the nature of this relationship has important policy implications from a public health perspective in preventing smoking among more vulnerable subpopulations.

The theoretical framework for analyzing health behaviors in the literature is mainly motivated by Grossman's health production model (Grossman 1972). Based on this theory, health is viewed as a "consumption good" since individuals gain utility from being healthy, and an "investment good" which produces a stock of healthy time for other activities. Individuals make choices on how to allocate their time and resources to the production of different consumption commodities subject to time and monetary budget constraints. For example, investment in health is produced by household production functions with inputs such as time, medical care, health behaviours, etc. Cigarettes can be viewed as consumption goods for some individuals which yield immediate utility as a result of anxiety relieve, appetite suppression, etc. (Aubin et al. 2012). Therefore, smoking is considered a disinvestment in health in return for a short-term increase in utility (Cawley and Ruhm 2011). The adverse psychological consequences of unemployment such as fear of not finding a job and increased risk of further job losses, loss of the non-financial benefits of work such as respect and identity, decreased life satisfaction (for example see Charles and Stephens 2004, Akerlof and Kranton 2000), abandonment of the future, loss of hope (for example see Drydakis 2014, Karsten and Moser 2009) can potentially result in a shift from preference towards the future to preference

towards the present time (Fryer and Stambe 2015) and thus induce individuals to find unhealthy ways, such as smoking, excess drinking and overeating, to cope with their increased stress levels (for example see Hill and Angel 2005, and Kassel et al. 2003). On the contrary, unemployment can result in decreased consumption of substances due to two reasons. First of all, unemployment results in less income and although a reduction in income can cause psychological distress it can also potentially lead to a reduction in smoking as there is less money available to the agent to purchase tobacco. The latter argument holds as long as tobacco is considered a normal good for which consumption decreases as income decreases. Second, the unemployed do not suffer work related stress which is also a potential risk factor for risky health behaviours (Ruhm 1995, Azagba and Sharaf 2011). A third possible argument would be that unemployed benefit from more leisure time which has an uncertain effect on health depending on how agents invest in their leisure time. Some individuals may find healthy ways, such as increased levels of physical activity, in order to cope with stress. Empirical studies are needed to tackle this relationship and the possible direction of causality. The focus of this study is on answering the following question: Does individual unemployment have a causal effect on smoking behaviors among individuals?

While cross-sectional studies have established that unemployed are more likely to smoke than the employed, these results cannot be interpreted as casual as there might be a selection from poor health into unemployment or vice versa. It may also be the case that unemployment and health behaviours are jointly determined by some unobserved factors such as genes, lifestyle and culture (Gathergood 2013).

Therefore, conducting studies using individual level data and improved methods accounting for the endogeneity of unemployment can help us better understand the nature of this relationship. Previous studies that have tried to find a causal (rather than correlative) effect of unemployment on smoking in different countries, have used methods that are not robust in accounting for all potential sources of endogeneity. Besides, only a very limited number of these studies focus on smoking intensity in addition to smoking status. According to Karim et al. (2010) one reason for the divergence behind the findings of previous studies (other than the difference in the data used and methods implemented) could be difference in country-specific effects and welfare state arrangements. Therefore generalizing the results from one country to another would be highly problematic (Henkel 2011). This study, to the best of our knowledge, is the first study that addresses the causal effect of individual unemployment on smoking in Canada.

I aim to contribute to the literature by implementing an instrumental variable approach to estimate the causal effect of unemployment on smoking status and smoking intensity conditional on being a smoker using data from the Canadian Community Health Survey (CCHS).

This study is organized as follows: in section 2 I discuss the conceptual framework for the relationship between unemployment and health behaviours, in section 3 I provide a summary discussion of the literature, in section 4 I discuss the data, sample selection and the descriptive statistics of the study sample, in section 5 I provide a discussion of the econometric methods implemented in the study, section 6 the regression results are discussed, and in section 7 I provide conclusive remarks.

2.2. Literature

The literature on the effect of unemployment on smoking is mixed and inconclusive. The findings vary significantly depending on the type of data used and the methods implemented. On the one hand, for example Mathers and Schofield (1998) found that the unemployed are more likely to suffer from chronic illnesses, be in poor mental health, and have higher levels of smoking, drinking and a poor diet. De Vogli and Santinello (2005) studied the role of psychosocial factors as a mediator between unemployment and smoking status using logistic regressions on a cross sectional sample of Italian adults. They found that the relationship between unemployment and smoking weakened once the role of psychosocial factors were accounted for in the analysis. (Schunck and Rogge 2010) used one wave of the German Micro-census data in order to investigate the effect of unemployment on smoking and BMI using a multivariate regression analysis and found that unemployment was associated with a 46% higher probability of smoking. Khlat et al. (2004) documented the relationship between unemployment and health problems (including smoking) using data from cycle 1991-1992 of the French National Health Survey. They found that unemployment among men was associated with a higher prevalence of smoking. Fergusson et al. (1997) examined the impact of exposure to unemployment after school leaving on a number of health conditions including nicotine dependence. They followed a birth cohort of New Zealand young people up to the age of 18 in order to examine how duration of unemployment since age 16 affected these health conditions and found positive significant association between duration of

unemployment and nicotine dependence. (Merline et al. 2004, Montgomery et al. 1998, Hammarström and Janlert 2003, Falba et al. 2005, and Okechukwu et al. 2012 established that early unemployment, accumulated years of unemployment and involuntary job loss have a positive impact on either becoming a smoker, or smoking relapse. Schunck and Rogge (2012) used longitudinal data from the German Socio-Economic Panel to estimate the casual effect of unemployment on smoking take up, relapse, quitting, and intensity. They used a zero-inflated negative binomial model and compared the results with those obtained from a fixed effects method in order to eliminate unobserved individual effects that potentially cause endogeneity in the first approach. They found that although there was a cross-sectional association between unemployment duration and the probability to smoking, the fixed effects model showed no causal impact of unemployment on smoking take-up, quitting and relapse. With regards to the methods used in Schunck and Rogge (2012) the authors argue that implementation of fixed effects (which is used to eliminate the effect of unobserved individual heterogeneity and relies on within individual variation in the key variables) comes at price in that only few individuals in their sample changed their status in both unemployment and smoking behaviours. Besides their fixed effects method does not estimate the effect of unemployment on smoking intensity. Goel (2008) did not find a significant effect of unemployment on the demand for cigarettes, Morris et al. (1992) found that loss of unemployment is not significantly associated with an increase in smoking, and Ruhm (2005) showed that smoking decreases during temporary economic downturns. Bolton and Rodriguez (2009) findings stressed the importance of unemployment benefit programs as a protective

factor determining the health effects following periods of unemployment. Latif (2014) used longitudinal data from Canada in order to estimate the impact of recession on smoking and drinking, and found that unemployment rate (in macro level) has no effect on the probability of being a smoker but has a positive impact on the number of cigarettes smoked by daily smokers. Novo et al. (2000) studied the relationship between unemployment and smoking during the recession and boom. They found that during the recession daily smoking was less intense compared to the boom. They also found that unemployment was associated with smoking especially among women during the boom.

Overall, different studies in the literature suggest that the direction of the association between unemployment and smoking behaviours is still not clearly known. Studies that have tried to find a causal impact of unemployment on smoking have either focused on very specific populations/samples, or have not used robust methods to account for all possible sources of endogeneity. The current study is one of the very few studies in the economic literature that tackles endogeneity of unemployment in a smoking equation, using more robust estimation techniques than those used in previous studies. I contribute to the literature by implementing an instrumental variable approach to Canadian data in order to estimate how individual level unemployment affects smoking behaviours. Specifically I estimate the causal impact of individual unemployment not only on individual smoking status but also on smoking intensity conditional on being a smoker.

2.3. Econometric Methods

The econometric model to be estimated takes the following general form:

$$Y_{nsi} = \alpha Y_{ei} + X_i\beta + u_i \quad (2\text{-}1)$$

where Y_{nsi} is the count dependent variable representing the average number of cigarettes smoked per day (this variable includes zeros for non-smokers and positive values for smokers), Y_{ei} is a binary variable representing unemployment status, X_i represents the covariates affecting smoking intensity, and u_i is the error term.

There are two main econometric challenges that come with estimating the above equation. The first challenge when estimating the effect of a "treatment" on a specific outcome in observational studies is the issue of selection bias which is mainly caused by the fact that the objects of the study are not randomly assigned to the treatment but rather receiving the treatment is based on some variables and/or individual-specific characteristics that also affect the outcome. Therefore, it is important to account for confounders that affect both the treatment (here unemployment) and the outcome. However this can only be feasible by including observed confounders in the analysis model, while there might be a number of other confounders that are not observable and/or not quantifiable to be included in the model. In this study, the unobserved heterogeneity which is due to individual specific characteristics not known to the researcher, and the presence of some omitted variables that affect both unemployment and smoking (status and intensity) cause a potential source of endogeneity in the model. In addition, smoking behaviors can also affect the probability of being unemployed, thus causing another source of

endogeneity through reverse causality. As endogeneity of unemployment causes the error terms to be correlated with unemployment in the above equation, the question cannot be addressed by using standard regression models as they will not provide consistent estimates under the presence of endogeneity. In addition, the nature of the dependent variable requires using a non-linear regression approach as linear regressions yield inconsistent, inefficient, and biased estimates (Long 1997). In order to address these econometric issues I implement an instrumental variable (IV) approach which relies on finding a variable that produces exogenous variation in individual level unemployment but is not correlated with the error term in the equation. This is ideally a variable that affects the probability of an individual becoming/being unemployed but does not directly affect their smoking habits. In the data I explain, in more detail, the choice of the instrumental variable and the justification behind choosing it.

Two of the most common IV approaches used in health economics studies when dealing with endogeneity in a non-linear model are two-stage predictor substitution (2SPS) and two-stage residual inclusion (2SRI) which are both instrumental variable approaches. Terza 2008 compared the performance of these two methods and found that, in a generic parametric framework and a nonlinear model, 2SRI is consistent while 2SPS is not (Humphreys et al. 2011). I therefore implement the 2SRI approach in this study to estimate the effect of unemployment on smoking.

Another challenge in estimating the model is the presence of a large number of zeros in the dependent variable which may arise for three main reasons 1)

Infrequency of purchase/consumption due to short recording periods in the survey, 2) Cigarettes may not be a good for some individuals because they are non-smokers, 3) Some individuals might potentially be smokers but cannot currently afford to purchase cigarettes at current prices and income, in which case a corner solution of zero consumption is the utility maximizing decision for these individuals. In the CCHS survey used in this study individuals were first asked if they are smokers and then based on the answer to the first question, they were asked a follow up question about the amount of their cigarette consumption, therefore we are looking at a typical consumption pattern rather than recorded consumption. This implies that the observed zeros in the outcome variable are "genuine zeros", as discussed by Jones (2000), which are the result of utility maximization decisions as opposed to censoring. The general framework for dealing with "genuine zeros" in a model is the full double-hurdle approach (Jones 2000) where the individual is faced with two decisions: the participation decision, and the consumption decision. In this approach, individuals must pass two hurdles before they are observed with a positive consumption. According to (Madden 2008), there are two main questions that need to be answered in the double hurdle framework in order to determine the precise form of the model specification that will be used. The first question is whether the participation decision dominates the consumption decision, a phenomenon known as first hurdle dominance. As mentioned above, given the nature of the questions asked in the CCHS survey about smoking status and intensity, it is reasonable to assume that the observed zero consumption represents a discrete choice as opposed to a corner solution. In other words, the participation and consumption equations are

independent/not related and that the participation decision dominates the consumption decision.

The second question is whether the error terms in the participation and consumption equations are correlated, i.e. whether the unobserved factors affecting participation are correlated with the unobserved factors that affect consumption. Assuming dependence between the error terms calls for Heckman (1979) selection model whereas if the error terms are assumed to be independent, a two-part model is more appropriate. There is a well-established debate on whether to use a Heckman sample selection model or a two-part model in health econometrics when there is no obvious variable(s) that affects the participation decision but does not affect intensity decision i.e. a variable that can be included in the participation equation but is excluded from the intensity equation. Madden (2008) used data on female smoking and drinking in order to revisit this debate and found that generally a two-part model is favoured, however the comparison should be carried out on a case-by-case basis. As I do not find a variable (or a set of variables) that affects participation but does not affect intensity of consumption within the context of this study, I use a two-part model approach (with instrumental variables to account for endogeneity of unemployment) in order to carry out the analysis.

The two-part model is estimated in two separate steps. The first step involves a binary outcome model which captures the data generating process governing the zeros in the smoking intensity variable. This step estimates the probability of an individual being a smoker. The second step involves a count model for the average number of cigarettes smoked per day given that individual is a current smoker. In

order to implement the IV approach via 2SRI, the endogenous treatment variable (unemployment) is first regressed on all other covariates of the model plus the instrument in the first stage and the residuals of this regression are obtained. Angrist and Krueger (2001) suggest estimating the first stage binary outcome model as a linear probability model using Ordinary Least Squares (OLS) rather than using a non-linear probit approach since in order for the second stage of the two-stage IV estimates to be consistently estimated, the functional form of the first stage does not necessarily have to be correctly specified. If the functional form of the first stage model is not correctly specified, it will result in the second stage estimates to be inconsistent. Therefore, the first stage model is treated as a linear probability model and is estimated via Ordinary Least Squares (Angrist & Krueger 2001) as follows:

$$Y_{ei} = X_{ei}\beta_e + u_{ei} \tag{2-2}$$

where Y_{ei} is a binary variable for unemployment, X_{ei} is the set of covariates affecting unemployment plus the instrumental variable, and u_{ei} are the error terms.

The second stage of the 2SRI approach involves estimation of the two-part model discussed above; one for the binary outcome (probability of being a smoker) and another for the truncated-at-zero count variable (average number of cigarettes smoked per day conditional on being a smoker). The binary outcome model takes the following form:

$$Y_{si} = \alpha Y_{ei} + X_i\beta_s + y_u'\mu + u_{si} \tag{2-3}$$

where Y_{si} is the binary smoking status variable which takes the value of 1 if the individual is a current smoker and zero otherwise, Y_{ei} is the binary unemployment

variable, X_i is the set of covariates affecting smoking status, and y_u' are the residuals obtained from equation (2-2). Equation (2-3) is a probit model and is estimated via maximum likelihood.

The truncated at zero count model is the following:

$$Y_{nsi}\ (Y_{nsi} > 0) = \alpha Y_{ei} + X_i\beta_{ns} + y_u'\mu + u_{nsi} \qquad (2\text{-}4)$$

where Y_{nsi} is the count dependent variable representing the average number of cigarettes smoked per day conditional on being a current smoker, Y_{ei} is the binary unemployment variable, X_{nsi} is the set of covariates affecting smoking intensity, and y_u' are the residuals obtained from equation (2-2). Equation (2-4) is a count model and is also estimated via maximum likelihood.

The t-test on the estimated coefficient of the residuals, μ, can tell us whether unemployment is exogenous in the model. If this coefficient is statistically different from zero then unemployment is endogenous.

One potential problem with using Poisson regression is that if there is over-dispersion in the count variable meaning that the variance is greater than the mean, then Poisson model may not be an appropriate option and a negative binomial count model will be preferred. The negative binomial regression estimation produces the test result for the presence of over-dispersion. It tests the null hypothesis of the over-dispersion parameter being equal to zero, if the null is not rejected then the Poisson and negative binomial regressions are equivalent. If the null is rejected the negative binomial regression is the preferred specification. The likelihood ratio test of the null hypothesis (the over-dispersion being parameter being statistically equal to zero) is

rejected at all conventional significance levels here, indicating that there is over-dispersion and that the negative binomial is preferred over the Poisson model. The detailed coefficient estimation method for the two-part model is provided in appendix A.

2.3.1. The Instrumental Variable

As mentioned earlier, the instrumental variable used in the analysis must predict exogenous variations in unemployment but must not be correlated with the error term. Choosing a proper instrument in health related studies comes with challenge as any individual level data could potentially be correlated with both the treatment (unemployment in this study) and the outcome, mainly through unobserved heterogeneity.

Generally the best practice to estimate a causal relationship would be through setting up a Randomized Controlled Trial (RCT) where subjects of the study are randomly assigned to a particular treatment. However this approach might be unethical or impractical especially in health-related studies. As an alternative, using a "naturally varying phenomenon" (Rassen et al. 2009) as an instrumental variable that predicts getting the treatment but does not affect the outcome is recommended. The frequency by which the instrument can predict receiving the treatment define instrument strength.

Rassen et al. (2009) justified the use of geographical and regional variables as instruments in studies focusing on health outcomes. They also mention that the instrument (here exclusion restriction) must predict receiving the treatment

(unemployment), must not directly affect the outcome, and must not affect the outcome through common causes of the instrument and the outcome. Rather, any effect of the instrument on the outcome must take place through the effect of the instrument on the treatment received by each individual. In this study I use provincial unemployment rates across age groups as an instrument for individual level unemployment. I expect that aggregate unemployment rates would affect the probability of being/becoming unemployed but do not directly affect individuals' smoking behaviours.

In order to overcome the potential impact of unobserved year-specific heterogeneity on the estimates, it would have been ideal to use repeated cross-sectional data from the CCHS with a time-varying instrumental variable (proposed by Moffitt (1993) for limited dependent variables). The IV used in this study is in fact time-varying which makes it an ideal candidate in identifying the parameter on the endogenous variable (conditional on the IV being valid and strong). However, only one cycle of cross-sectional data (the latest cycle available) is used in this study due to the earlier cycles not differentiating between zeros and missing values which might potentially bias the estimates given that the most important variables of interest are binary. Therefore, I am only relying on the variation of the IV across provinces (and age cohorts) by using one cycle of the data in order to identify the model.

An instrument must be valid and strong. A valid instrument is one that is not correlated with the error terms in the equation in question, i.e. a valid instrument must be exogenous. Validity of an instrument also implies that it must not have a

direct effect on the dependent variable. In the case of a single endogenous variable and a single instrument, as in this study, there is no econometric test for checking the validity of the instrument. The logical justification for the validity of the instrument is that individuals' smoking status/intensity is not affected by the provincial unemployment rates directly but rather any potential impact of the provincial unemployment rates on smoking behaviours should take place through the impact of aggregate unemployment rates on individual's unemployment status. In addition, it is assumed that smoking and provincial unemployment rates are not simultaneously determined. The latter two assumptions, however, might fail in some cases. For example some unemployed individuals might be seeking a job by moving to provinces with a more robust economy and lower unemployment rates. In addition, there might be provincial legislations on smoking bans or tobacco prices that can potentially affect individuals smoking behaviours.

A strong instrument must be capable of predicting exogenous variations in the endogenous variable. The strength of an instrument can be tested empirically. The rule of thumb for testing the strength of an instrument is by looking at the partial F-statistic of the instruments in the first stage linear probability regression of unemployment on all covariate plus the instrument (Staiger and Stock 1997). An F-statistic above 10 confirms instrument strength. In the case of a single endogenous variable and a single instrument the t-value for the instrument should be at least 3.2 or its p-value should be below 0.0016. The results of the first stage regression in this study show that the corresponding t-value for the instrument (aggregate unemployment rates) is 9 with a p-value of 0.000, indicating that I do not have a

weak instrument. The partial F-statistic on the instrument in the first stage regression is 81.01, also indicating that I do not have a weak instrument.

2.4. Data

The data used for this study are taken from the Public Use Micro data Files (PUMF) of the Canadian Community Health Survey (CCHS). CCHS is a cross-sectional and nationally representative dataset which contains information on the health status of the respondents and their households as well as information on their demographic and socio-economic statuses. The CCHS data is collected through a random digit dial telephone survey and includes all Canadians over the age of 12 except those living in First Nations reserves, institutions, and those serving in the armed forces. The latest available cycle belongs to year 2012. For this study I use the one year PUMF produced for year 2012 which contains 61,707 observations (individuals). I only use one cycle of the CCHS data since the earlier cycles of the data do not differentiate between zeros and missing values which might potentially affect the estimates given that the most important explanatory variable of interest (unemployment) is a binary variable which takes the value of zero or one. Statistics Canada assigns a survey weight to each respondent included in the final sample which corresponds to the number of persons that the respondent represents in the sample. The survey weights are used in the estimations whereas the summary statistics are un-weighted.

Theory suggests that there are a number of impact factors, other than labour force participation, that affect individuals' health and health related behaviors. Some

of the most important factors are household income, education and socioeconomic status. Better educated people have higher health knowledge and are expected to know more about the adverse effects of risky health behaviors. For instance, Kenkel (1995) found that education has a negative impact on smoking and drinking. Education is also highly correlated with labour income leading to high opportunity costs of illness. Moreover, since better educated people allocate medical services more efficiently and have better knowledge on how to use them, education is expected to be positively correlated with the efficiency of the health production function (Kenkel 1995). There are a number of standard demographic variables used in the literature such as age, gender, marital status and geographical locations. In CCHS, starting from age 20, age is reported in 5 year intervals, so I define midpoints in coding the age variable. For example the age category of 20 to 24 years is coded as 22 years, therefore respondents whose age falls within this category are reported as being 22 years old. In this study I use individuals with an age (midpoint) between 22 to 62 years old. Marital status is divided into three categories. "Married" refers to respondents who are married or in a common-law relationship, "single" refers to respondents who are single and have never been married, "widowed/divorced/separated" refers to those who have been either divorced, separated or widowed. The latter group is the reference group in the estimation. A dummy variable for country of origin is included which takes the value of one if the respondent is an immigrant as opposed to being Canadian born. The socio-economic variables included in the model are education, household income and an indicator for whether the respondent or a household member owns the home. The education

variable includes four categories: less than high school (reference category), high school graduate, some college studies and college graduate. In the CCHS (year 2012), household income is reported in 5 categories: less than $20k a year, $20k-$39,999k per year, $40k-$59,999k per year, $60k-$79,999k per year and greater than $80k per year, I use the first category (< $20k) as the reference category. I also add a variable indicating whether there are kids under the age of 12 in household.

The independent variable of interest is a binary variable which takes the value of one if the respondent is unemployed. This variable is constructed based on the answer to the question: "Are you an employee or self-employed?" This question was asked form respondents who were employed at the time the survey data was collected (Both employees and self-employed responses were coded as being employed).

The dependent variable of interest is smoking intensity which captures the average number of cigarettes smoked per day. This variable in constructed based on the response to two separate questions in the survey. The first question is: "At the present time, do you smoke cigarettes daily, occasionally or not at all?" This question allows us to identify individuals with a current cigarette consumption of zero. The second question, which is asked from respondents who smoke either daily or occasionally, is: "How many cigarettes do you smoke each day now?" The latter question includes only positive values for those who are current smokers. The information provided by the responses to these two separate questions are combined in order to construct the dependent variable for the average number of cigarettes smoked per day which includes zeros for non-smokers and positive values for current smokers.

Another variable used in my analysis which serves as an instrument is aggregate unemployment rates in year 2012 for different age groups across the 9 provinces included in the model (see table of descriptive statistics for the list of provinces included in the study). These data are taken from Statistics Canada, Table 282-0002 Labour Force Survey Estimates by sex and detailed age groups. The table was downloaded in June 2014 and is based on population data from the 2006 Census. For each age group, the unemployment rate is the number of unemployed in that age group as a percentage of the labour force in the same group (Statistics Canada 2014). Age groups in this table are categorised the same way as in the master CCHS file (4 year intervals) and coded the same way. Finally data from this table are merged to the master CCHS file based on province of residence and age midpoints of respondents. The detailed information on provincial level unemployment rates by age groups in year 2012 are provided in the appendix (Table A 1). The final sample includes 29,385 individuals after restricting the age range and dropping students, respondents who are permanently unable to work, retired respondents, and individuals with missing values on the variables used in this study.

2.4.1. Descriptive Statistics

Table 2.1 and Table 2.2 show the un-weighted descriptive statistics for the entire sample (descriptive statistics for the smoking variables are shown separately in Table 2.2). The majority are Canadian born and almost half of the sample are male/female. About 21% of the sample are unemployed. The average unemployment rate across the entire pooled sample is 6.45%. It is worth mentioning that there is a substantial

difference between the unemployment rate and the percentage of unemployed individuals in the selected sample. This might partly be explained by the fact that Statistics Canada differentiates between employed, unemployed, and not in the labour force when releasing data on unemployment rates, whereas in the CCHS we do not know which individuals are actively in the labour force. Therefore, any individual who is not in the labour force in the selected sample will automatically be coded as being "unemployed". Another reason for this observation might be the fact that the descriptive statistics of the selected sample here are not weighted. The sample consists of 24% current smokers with 19% being daily smokers and 5% being occasional smokers making. The descriptive statistics for the unemployed and employed subsamples (Table 2.3) show very similar statistics for the smoking variables, with only slightly higher amounts for the unemployed sample compared to the employed (except for the percent of occasional smokers). Among the unemployed, 23% are daily smokers and 4% are occasional smokers, thus making 27% of the unemployed "current smokers". In the employed subsample, 19% are daily smokers and 5% are occasional smokers which makes 24% of this subsample "current smokers". The average number of cigarettes smoked per day (conditional on being a smoker) among the unemployed and employed is about 14.65 and 12.71 cigarettes respectively. These statistics indicate that there is not much difference, in terms of smoking status and intensity statistics, between the employed and unemployed in the study sample.

Table 2.4 also shows descriptive statistics for the "smokers" subsample which includes both daily and occasional smokers. Only 23% of smokers are

unemployed and about 68% of them are college graduates. The majority of the smokers sample consists of daily smokers (80%), and the average number of cigarettes smoked per day across both type of smokers is about 13 cigarettes per day.

Table 2.1 Summary Statistics for Socioeconomic Variables- Full Sample

Variable	Mean	Std. Dev.
Male	0.45	0.5
Age	45.43	12.53
Canadian	0.86	0.35
Less than High School	0.06	0.23
High School Graduate	0.13	0.33
Some College	0.03	0.18
College Graduate	0.78	0.41
Kids	0 .25	0.44
Unemployed	0.21	0.41
HH Income $0-$19,999k	0.07	0.26
HH Income $20-$39,999k	0.15	0.36
HH Income $40-$59,999k	0.18	0.39
HH Income $60-$79,999k	0.17	0.37
HH Income > $80k	0.43	0.5
Single	0.22	0.42
Wid/Div/Sep	0.14	0.35
Married	0.64	0.48
Unemployment Rate	0.06	0.02
Observations	29385	

Table 2.2 Summary Statistics for Smoking Variables- Full Sample

Variable	Mean	Std.Dev
Smoke_Daily	0.19	0.4
Smoke_Occasionally	0.049	0.22
Current Smoker	0.24	0.43
Average # of Cigarettes Smoked per day	3.2	7.24
Observations	29385	

Table 2.3 Summary Statistics For Smoking Variables by Employment Status

Variable	Mean	Std.Dev
(Unemployed = 1)		
Smoke_Daily	0.23	0.42
Smoke_Occasionally	0.04	0.20
Current Smoker	0.27	0.44
Average # of Cigarettes Smoked per day (if smoker = 1)	14.65	10.25
Observations	6199	
(Employed = 1)		
Smoke_Daily	0.19	0.40
Smoke_Occasionally	0.05	0.22
Current Smoker	0.24	0.42
Average # of Cigarettes Smoked per day (if smoker = 1)	12.71	8.80
Observations	23186	

Table 2.4 Summary Statistics for the "Current Smoker" Subsample

Variable	Mean	Std.Dev
Male	0.51	0.50
Daily Smoker	0.80	0.40
Age	43.96	12.71
Canadian	0.91	0.28
Less than High School	0.10	0.30
High School Graduate	0.17	0.38
Some College	0.05	0.21
College Graduate	0.68	0.47
Occasional Smoker	0.20	0.40
Unemployed	0.23	0.42
HH Income $0-$19,999k	0.12	0.33
HH Income $20-$39,999k	0.20	0.40
HH Income $40-$59,999k	0.20	0.40
HH Income $60-$79,999k	0.16	0.37
HH Income >$80k	0.31	0.46
Average # of Cigarettes Smoked per day	13.16	9.19
Single	0.30	0.46
Married	0.50	0.50
Wid/div/sep	0.19	0.39
Observations	7134	

2.5. Regression Results

2.5.1. Exogeneity Tests

The exogeneity test for the 2SRI models can be performed by looking at the coefficient on y_u' in equations (2-3) and (2-4). Recall that y_u' are the residuals obtained from equation (2-2). If the null hypothesis is rejected, i.e. μ is statistically different from zero, unemployment is not exogenous. The results in Table 2.9 and

Table 2.10 show that the coefficient on the residuals in the probit model is not significant while it is significant in the zero-truncated negative binomial model for smoking intensity. This finding implies that unemployment is endogenous in the smoking intensity equation but not in the smoking status equation.

2.5.2. Coefficients and Marginal Effects Estimate

Table 2.5 to Table 2.10 show the coefficient estimates and marginal effects for the two separate specifications of the two-part model; one where unemployment is treated as exogenous and the other one where unemployment is treated as endogenous. As the main focus is on the marginal impacts, I first indicate the estimates of the marginal effects in Table 2.5, followed by the coefficient estimates in Table 2.6 to Table 2.10. Table 2.6 and Table 2.7 show the coefficient estimates of the probit and negative binomial models where unemployment is treated as exogenous, whereas Table 2.9 and Table 2.10 show the coefficient estimates of the IV method (using 2SRI) when unemployment is treated as endogenous. Each specification includes a probit model for the probability of being a smoker, and a count model for the average number of cigarettes smoked per day conditional on being a smoker. The marginal effects are reported only for the key variable (unemployment) and are calculated for a discrete change in unemployment from zero to one while all other explanatory variables are held at their means. Since the full sample is large, in regressions where the full sample is used I considered a p-value threshold of 5% instead of 10% in order to avoid over interpreting the estimates. Both probit models for smoking status indicate that unemployment does not have a

significant effect on the probability of being a smoker regardless of the endogeneity/exogeneity assumption imposed on unemployment. As for the effect of unemployment on smoking intensity, a different pattern is observed. While the simple negative binomial (where unemployment is treated as exogeneous) shows a positive but insignificant impact of unemployment on smoking intensity, the marginal effects estimation of the 2SRI negative binomial indicates that being unemployed decreases the average number of cigarettes smoked per day by about 11 cigarettes. The coefficient on the residuals in the latter model is significant indicating that unemployment is an endogenous regressor in the smoking intensity equation. Therefore the negative binomial model estimated via 2SRI is the preferred specification for smoking intensity. Overall, the results of the analyses imply that unemployment has a causal effect on the intensity of cigarette consumption among smokers but not on the probability of being a smoker.

Another interesting observation is that factors affecting smoking status and smoking intensity do not necessarily have the same signs. By looking at the simple probit model for smoking status and the negative binomial (with endogenous treatment) for smoking intensity we can observe that while the probability of being a smoker decreases with age, intensity of consumption increases among smokers as they age. Single and widowed/divorced/separated individuals are more likely to smoke than married ones, but there is no significant association between marital status and smoking intensity conditional on being a smoker. Education consistently has a negative effect on the probability of being a smoker as well as on smoking

intensity. Individuals with education levels of high school or above are less likely to smoke than high school drop outs, and smoke less conditional on being a smoker.

2.5.3. Sensitivity and Robustness Check

An alternative approach to the 2SRI is the 2SPS method which is an IV method but instead of adding the residuals from the first stage regression (of unemployment on all variables plus the instrument) to the second stage equation, the predicted values obtained from the first stage are substituted for the observed values of the unemployment variable in the second stage. Therefore the second stage equations for the count model becomes the following:

$$Y_{nsi} = \hat{y}_{ei}\alpha + X_{nsi}\beta_{ns} + u_{nsi} \quad (2\text{-}5)$$

where $\hat{y}_{ei}$ are the predicted values obtained from the first stage linear probability model of unemployment on all variables plus the instrument. The marginal effects from the 2SPS estimation show that unemployment decreases the average number of cigarettes smoked per day by about 17 units.

The estimation results for both smoking status and smoking intensity are sensitive to the assumption imposed on exogeneity of unemployment. The marginal effect of unemployment on smoking status shifts from being positive in the simple probit model where unemployment is treated as exogenous to negative in the 2SRI approach where unemployment is treated as endogenous (though the effect is not significant in either specification). The negative binomial estimates show that unemployment has a positive but insignificant effect on smoking intensity when

unemployment is treated as exogenous while this effect becomes negative and significant once potential endogeneity of unemployment is accounted for. As mentioned earlier the exogeneity tests indicate that unemployment should be treated as an endogenous regressor in a smoking intensity equation, therefore I would argue that the IV approach used for smoking intensity performs better in producing consistent estimates compared to the simple negative binomial.

The IV estimates of the marginal effect of unemployment on smoking intensity are robust in that both IV methods (2SPS and 2SRI) consistently show a negative and significant marginal effect of unemployment on the average number of cigarettes smoked per day. However the IV results are sensitive to the type of the estimation approach in that the 2SPS method yields larger effects of unemployment on smoking intensity than the 2SRI approach. Terza (2008) argues that in a non-linear framework 2SRI is consistent while 2SPS in not, therefore assuming that 2SPS produces biased estimates compared to 2SRI, the 2SPS estimation results in our study are upward biased.

Table 2.5 Marginal Effects for Main Estimates and Robustness Checks

Alternative Models	Unemployed	Std.Err
Smoking Status		
Single Equation Probit	0.003	0.01
Probit_2SRI	−0.16	0.08
Smoking Intensity		
ZT Negative Binomial	0.67	0.51
ZT Negative Binomial_2SRI	−11.41 ***	2.15
ZT Negative Binomial_2SPS	−16.63***	4.28

*** $p < 0.01$. ZT=Zero Truncated, 2SRI: Two-Stage Residual Inclusion, 2SPS: Two-Stage Predictor Substitution. Main estimates include the Single Equation Probit, ZT Negative Binomial, Probit_2SRI, and ZT Negative Binomial_2SRI.

Table 2.6 Probit Model Coefficient Estimates_Smoking Status

Variable	Current Smoker	Std. Err.
Unemployed	0.01	0.05
Newfoundland	0.40***	0.09
Prince Edward Islands	0.30**	0.11
Nova Scotia	0.30***	0.09
New Brunswick	0.32***	0.08
Quebec	0.40***	0.06
Ontario	0.23***	0.05
Manitoba	0.21**	0.09
Saskatchewan	0.21***	0.07
Alberta	0.33***	0.07
Male	0.22***	0.03
Single	0.17***	0.04
Wid/Div/Sep	0.20***	0.05
Age	−0.005	0.002
Canadian	0.38***	0.05
High School Graduate	−0.25***	0.08
Some College	−0.24**	0.11
College Graduate	−0.53***	0.08
Kids	0	0.04
HH Income $20-$39,999k	−0.16**	0.07
HH Income $40-$59,999k	−0.23***	0.07
HH Income $60-$79,999k	−0.40***	0.03
HH Income > $80k	−0.47***	0.07
Constant	−0.40***	0.13
Pseudo R-Squared	0.06	
Observations	29385	

** $p < 0.05$, *** $p < 0.01$

Table 2.7 Zero-Truncated Negative Binomial Model Coefficient Estimates- Smoking Intensity

Variable	Smoking Intensity (y>0)	Std. Err.
Unemployed	0.06	0.04
Newfoundland	0.1	0.1
Prince Edward Islands	0.01	0.12
Nova Scotia	−0.01	0.09
New Brunswick	0.14	0.08
Quebec	0.07	0.07
Ontario	0.02	0.07
Manitoba	0.03	0.09
Saskatchewan	0.07	0.08
Alberta	0.14	0.08
Male	0.33***	0.03
Single	−0.03	0.04
Wid/Div/Sep	0.04	0.05
Age	0.013***	0.001
Canadian	0.42***	0.06
High School Graduate	−0.06	0.07
Some College	0.04	0.08
College Graduate	−0.22***	0.06
Kids	−0.03	0.04
HH Income $20-$39,999k	−0.11*	0.06
HH Income $40-$59,999k	−0.02	0.06
HH Income $60-$79,999k	−0.15**	0.07
HH Income > $80k	−0.17***	0.06
Constant	1.56***	0.14
Wald chi2	399.15	p>chi2=0
Observations	7134	

* p < 0.10, ** p < 0.05, *** p < 0.01

Table 2.8 First Stage Linear Probability Coefficient Estimates- Unemployment Equation

Variable	Unemployed	Std. Err.
Unemployment Rate	2.62***	0.29
Newfoundland	−0.11***	0.03
Prince Edward Islands	−0.19***	0.03
Nova Scotia	−.05***	0.02
New Brunswick	−0.06***	0.02
Quebec	−0.07***	0.01
Ontario	−0.02	0.01
Manitoba	0.03	0.02
Saskatchewan	0.01	0.02
Alberta	0.05***	0.02
Male	−0.08***	0.01
Single	−0.02	0.01
Widow/Sep/Div	−0.07***	0.01
Age	0.01***	0
Canadian	−0.02**	0.01
High School Graduate	−0.11***	0.03
Some College	−0.10***	0.03
College Graduate	−0.14***	0.03
Kids	0.05***	0.01
HH Income $20-$39,999k	−0.23***	0.03
HH Income $40-$59,999k	−0.34***	0.02
HH Income $60-$79,999k	−0.37***	0.03
HH Income > $80k	−0.42***	0.02
Constant	0.30***	0.05
R-squared	0.13	
Observations	29385	

** $p < 0.05$, *** $p < 0.01$

Table 2.9 Second Stage IV Probit Model Coefficient Estimates- Smoking Status

Variable	Current Smoker	Std. Err.
Unemployed	−0.62	0.4
Residual	0.63	0.41
Newfoundland	0.41***	0.09
Prince Edward Islands	0.23**	0.12
Nova Scotia	0.29***	0.09
New Brunswick	0.34***	0.08
Quebec	0.33***	0.06
Ontario	0.23***	0.05
Manitoba	0.20**	0.09
Saskatchewan	0.19***	0.07
Alberta	0.33***	0.07
Male	0.17***	0.04
Single	0.16***	0.04
Widow/Div/Sep	0.15***	0.06
Age	−0.003	0.002
Canadian	0.37***	0.05
High School Graduate	−0.32***	0.1
Some College	−0.30***	0.11
College Graduate	−0.62***	0.1
Kids	0.02	0.04
HH Income $20-$39,999k	−0.30***	0.11
HH Income $40-$59,999k	−0.44***	0.15
HH Income $60-$79,999k	−0.56***	0.16
HH Income > $80k	−0.74***	0.18
Constant	−0.06	0.25
Pseudo R-squared	0.06	
Observations	29385	

$p < 0.05$, *** $p < 0.01$. Residual: residuals obtained from the first stage linear probability regression of unemployment on all covariates plus the instrument.

Table 2.10 Second Stage IV Negative Binomial Model Coefficient Estimates-Smoking Intensity

Variable	Smoking Intensity(y>0)	Std. Err.
Unemployed	−1.42***	0.37
Residual	1.49***	0.37
Newfoundland	0.17	0.1
Prince Edward Islands	−0.11	0.12
Nova Scotia	−0.01	0.09
New Brunswick	0.19**	0.09
Quebec	−0.01	0.07
Ontario	0.02	0.07
Manitoba	0.01	0.09
Saskatchewan	0.01	0.08
Alberta	0.13	0.08
Male	0.21***	0.04
Single	−0.05	0.04
Widow/Div/Sep	−0.06	0.06
Age	0.02***	0.002
Canadian	0.38***	0.06
High School Graduate	−0.22***	0.08
Some College	−0.12	0.09
College Graduate	−0.43***	0.09
Kids	0.02	0.05
HH Income $20-$39,999k	−0.45***	0.1
HH Income $40-$59,999k	−0.54***	0.14
HH Income $60-$79,999k	−0.70***	0.15
HH Income > $80k	−0.80***	0.17
Constant	2.38***	0.25
Wald chi2	422.48	p>chi2=0
Observations	7134	

* p < 0.10, ** p < 0.05, *** p < 0.01. Residual: residuals obtained from the first stage linear probability regression of unemployment on all covariates plus the instrument.

2.6. Discussion and Conclusion

In this study I estimated the casual effect of unemployment on smoking using micro data from the Canadian Community Health Survey. I contribute to the literature by implementing an instrumental variable approach to tackle endogeneity of unemployment. Our results suggest that unemployment has a negative impact on the average number of cigarettes smoked per day once potential endogeneity of unemployment is accounted for. These findings are in contradiction with the assumption that the risky health behaviours that many unemployed engage in are actually produced by the experience of unemployment (Schunck and Rogge 2012). The results of this study can be explained in a number of different ways. First of all, it is not known from the data what portion of the unemployed in the sample are in the labour force and looking for a job. I am expecting that individuals who are subject to involuntary unemployment/job loss would suffer more from the distress/adverse health effects of being unemployed than those who are not looking for a job. Prochaska et al. 2013 showed that the job-seeking unemployed are more likely to smoke than the non-job seeking unemployed. The time period used in this study might also be another source of explanation behind the findings, in the sense that the data belong to year 2012 following the economic recession of 2008. The impacts of recession are expected to persist, and though the recession is now over, the unemployment rate is still above its rate before the recession (Latif 2014). Ruhm (2005) used micro data in order to estimate the effect of a reduction in employment rate on a number of individual health behaviours (such as physical activity, smoking

and obesity) and found that smoking decreases and physical activity increases as the employment rate decreases. Changes in income and the amount of leisure time were examined as two possible channels through which these findings could be explained. Ruhm argued that a decrease in working hours, which potentially increases non-market time available for investments in health behaviours (and potentially decreases work-related stress), is one possible explanation for the findings. Particularly Ruhm found that a decline in the number of hours worked was associated with a decline in smoking and an increase in physical activity. Smoking can be viewed as a form of self-medication (which is not time consuming) for individuals who are suffering from employment-related stress since healthy methods of stress relief (such as physical activity) are time consuming. On the contrary, he finds that reduction in income does not play an important role in mediating the effect of the decline in employment rate on health behaviours. Another reason for the divergence between the findings of this study and most previous studies might be in the data and methods used. To the best of my knowledge, the data and methods used in this study have not been used in previous studies to examine the causal effect of individual unemployment on smoking behaviours. While the methods implemented in this study are improvements to the ones previously used, they suffer a number of shortcomings. First of all, the validity assumption behind the instrumental variable used in this study might potentially fail due to the reasons discussed in the IV section of the paper. In addition, the data used in this study are cross-sectional while more informative longitudinal data are needed in order to account for the duration of

unemployment, changes in employment status, and whether the individual is in the labour force.

Finally the results of this study do not imply that unemployment is not a public health concern. Nonetheless, the results indicate that once the effect of unobserved factors driving both smoking behaviours and unemployment are accounted for in the sampled Canadian population, unemployment does not seem to deepen social inequalities in smoking behaviours. An extension to the current study would be to examine whether there is a differential impact of unemployment on smoking among different socioeconomic subpopulations since the unemployed group is likely to be a heterogeneous group with respect to socioeconomic status.

3. Decomposition of the Income Gap in Body Mass Index in Canada

3.1. Introduction

Inequalities in health among different socioeconomic segments (SES) of the population remain a public health concern in Canada as in many other developed countries where the SES gaps in health are normally persistent despite the efforts of the health systems to eliminate inequalities in access to healthcare. The Canadan Population Health Initiative (2008) has demonstrated the existence of SES gaps, particularly income and education gaps, in health and has stated that "…in seeking to address gaps in health as a result of unequal SES, it is important to consider the individual-level factors and the broader social determinants of health that contribute to those gaps." There is an argument that the SES inequalities in health are normally exerted by their impact on health behaviors or living conditions. The mechanisms behind the socioeconomic inequalities in health are still an important question for many economic researchers (Costa-Font et al. 2014). Although there are a number of studies focusing on inequalities in health (For example see Marmot 2005, Kawachi et al. 2002, O'Donnell and Doorslaer 2008). SES inequalities in obesity have received far less attention despite the fact that the policy implications of the

studies on inequalities in health may not be directly applicable to inequalities in obesity (Hajizadeh et al. 2014), and despite the dramatic increase in obesity prevalence in Canada over the last few decades. There is epidemiological evidence that the rise in the mean BMI of the Canadian population is causing the increasing incidence of obesity (Raine 2004). The percentage of obese adults in Canada increased from 10 % in 1970 (Starky 2005) to 25 % in 2008 (CIHI and PHAC: Obesity in Canada 2011). Obesity is considered a major public health concern as it is strongly associated with many negative physical and mental health outcomes such as heart disease, type 2 diabetes, depression, cancer, dementia, arthritis, and hypertension (McLaren 2007, Rashad 2003, and Tjepkema 2004). The obesity epidemic, induces substantial economic burden to the society in terms of health care costs and productivity losses (Klarenbach et al. 2006). In 2005, the total economic burden of adult obesity in Canada was estimated as $3.42 billion consisting of $1.62 billion in direct costs and $1.80 billion in productivity losses (Janssen 2009).

There is evidence that there exists an income related gap in BMI/obesity prevalence among Canadian adults (Hajizadeh et al. 2014); the prevalence of obesity increases with income among men whereas it decreases with income among women. There are several socioeconomic and behavioral factors that differ by income status that can potentially lie behind the income gap in body weight. In other words, any factor that is unequally distributed by income and has an impact on BMI can contribute the income gap in BMI. There is an argument that the growth of fast foods has been one of the main factors underlying the rise in obesity prevalence which has disproportionately affected the poor. According to Drewnowski and Specter (2004),

the poor may choose diets that provide the maximum amount of calories at the least cost since they are not able to afford healthy food. In Grossman (1972) health production model, individuals allocate time and resources such as medical care and physical activity to produce a stock of health, therefore an increase in income can potentially promote investment in health through increased physical activity and improved eating habits. However, the opposite of the latter theory may also be true since individuals with higher incomes have a higher opportunity cost of time which can potentially cause them to spend less time on exercise. In addition, income differences can cause gap in BMI through labour force status. For example, employed individuals face a higher opportunity cost associated with sick days and might be more likely to invest in health compared to the unemployed. On the contrary however, the employed may have less time to invest in their health since many health behaviours such as healthy eating by preparing own meals and regular physical activity are time-intensive (Kpelitse et al. 2014).

To date, there is little known about the extent to which the income-related gap in body weight (obesity risk) could be eliminated or reduced if certain factors contributing to this gap could be eliminated. From a policy perspective it is important to know whether the observed gap in obesity prevalence among the high-income and the low-income is due to a differential distribution of variables (i.e. difference in the observed characteristics of the high-income and the low-income such as difference in education, participation in physical activity, etc.) or whether there are other factors that cause these observed variables/characteristics to manifest differently. The latter is commonly known as the difference between the returns to observed characteristics.

As the two effects take place simultaneously it is important to quantify and decompose them in order to predict the magnitude of the impact of possible interventions.

The objective of this study is to decompose income inequalities in body weight in Canada among men and women. The question is: how far can the observed gap in BMI between the "rich" and the "poor" be explained by the difference in the commonly cited determinants of BMI between these two groups, and what portion remains unexplained? What is the relative contribution of each factor in causing the BMI gap between the "rich" and the "poor"? For this purpose an Oaxaca-Blinder decomposition method is applied that allows for the decomposition in the income gap into an "explained" and an "unexplained" portion. The "unexplained" gap is due to the difference in the returns to the observed characteristics which will exist even if both groups were to attain the same characteristics. As mentioned earlier, decomposition of the income gap in BMI has important policy implications as it is an important step towards determining how far this gap is explained by observed differences between these two groups in terms of SES statutes and health behaviors. To the best of our knowledge, this is the first study that addresses the income gap in BMI in Canada using an Oaxaca-Blinder decomposition method.

The study is organized as follows: in section 2 some relevant literature is discussed, in section 3 the proposed methodology is discussed, in section 4 the data used to carry out the analysis as well as some descriptive statistics are presented, in

section 5 the results of the analysis are discussed, and finally in section 6 I provide a discussion of the results and conclusive remarks.

3.2. Literature Review

There are a limited number of studies in the literature that examine SES inequalities/differences in BMI/obesity prevalence. Volland (2012) augmented 12 waves of the Behavioral Risk Factor Surveillance System in order to examine how the distribution of income can determine variations in BMI and obesity across the US and how a change in the distribution of income have contributed to the increase on obesity prevalence. He found that income inequality in fact has an impact on weight outcomes. Jolliffe (2011) used NHANES data form 2003 to 2006 and found no statistically significant difference between the prevalence of overweight and obesity between the poor and non-poor. Babey et al. (2010) examined disparities in obesity among California adolescent using data from the California Health Interview Survey between 2001 and 2007. They found that the prevalence of obesity rose significantly over this period among lower-income (below the poverty line) adolescents but not among higher-income (at or above 300% of the federal poverty line) adolescents, and the disparities in obesity prevalence between these two income groups rose from a 7 percentage point difference in 2001 to a 15 percentage point difference in 2007. Using three decades of data from the National Health and Nutrition Examination Survey, Chang and Lauderdale (2005) examined income difference in BMI and the change in the prevalence of obesity among different income groups over time. They found that the prevalence of obesity increased at all

levels of income over the observation period, and the largest increases were not necessarily among the poor. They also found that income gradients in BMI exists among all races and genders, however among white women there is a consistent inverse income gradient in BMI throughout the observation period, while for black and Mexican women the income-gradient in BMI becomes positive in the later waves of the data. In a study including 21 developed countries, Pickett et al. (2005) found a positive correlation between income inequality and percentage of obese men and women. Zhang and Wang (2004) and Nikolaou and Nikolaou (2008) found that there is a negative relationship between SES and obesity among women in the US and 10 European countries. Examining the trends in SES inequalities in obesity prevalence (in the US) form 1971 to 2002, Zhang and Wang (2007) and Zhang and Wang (2015) also found that the association between obesity and SES decreased over time. Costa-Font and Gil (2008) found that SES inequalities have a significant impact on the probability of being obese in Spain. The decomposition of their concentration index indicates that education and demographic variables contribute mostly to the income-related inequality in obesity. Using three cycles of longitudinal data from Sweden and the corrected concentration index, Ljungvall and Gerdtham (2010) found that obesity is more prevalent among the poor however the inequality has decreased over time. Madden (2013) also used the concentration index to indicate that the SES inequality in obesity prevalence is higher among women than men in Ireland, and income and education mostly explain this observed inequality. Hajizadeh et al. (2014) used the concentration index to quantify the socioeconomic inequalities in obesity prevalence in Canada. Their estimated concentration index indicates that

obesity is concentrated among the more affluent men, with an increasing trend over time, while it is more concentrated among the less affluent women. The decomposition of their concentration index indicates that income, demographic variables, immigration status, education, and physical activity and drinking habit explain the income-inequality in obesity prevalence. Finally, Alaba and Chola (2014) also used the concentration index to measure SES inequalities in obesity risk in South Africa and found that men with higher income are more likely to be obese than poorer men. However women have similar patterns in obesity regardless of their SES.

3.3. Methods

The decomposition tool developed by (Oaxaca 1973) and (Blinder 1973) allows for the decomposition of the gap in an outcome variable between two groups into a part that can be explained by differences in the observed characteristics between the two groups, and a part that is attributable to the differences in the returns to those characteristics (Bauer and Sinning 2008). It is also used to identify and quantify the contribution of observed characteristics to the gap in the outcome variable of interest. For this purpose, I first need to estimate the conditional mean of BMI given a set of observable covariates. In other words, BMI should be regressed on a set of covariates, including demographic, socioeconomic, and behavioral covariates that affect individuals' BMI. The BMI equation for the high-income and low-income groups are given by the following two equations:

$$BMI_i^{rich} = X_i\beta^{rich} + \varepsilon_i^{rich} \quad (3\text{-}1)$$

$$BMI_i^{poor} = X_i\beta^{poor} + \varepsilon_i^{poor} \quad (3\text{-}2)$$

where BMI_i^j indicates the body mass index of individual *i* in group *j* (*j*=high-income, low-income), X_i is a vector of variables/characteristics affecting BMI, and ε_i is the error term. Consistent with the literature, I use BMI instead of log of BMI as the dependent variable (for example see Sen 2014 and Powell et al. 2012). Equations (3-1) and (3-2) are estimated using Ordinary Least Squares (OLS).

The gap in mean BMI between the high-income and low-income can be shown as:

$$BMI^{rich} - BMI^{poor} = X^{rich}\,\beta^{rich} - X^{poor}\beta^{poor} \quad (3\text{-}3)$$

where BMI^{rich} and BMI^{poor} are mean BMI's in the high-income and low-income groups respectively, and X^{rich} and X^{poor} are the variables at their means. The gap in the mean outcome can be decomposed into a component that is due to differences in the characteristics/explanatory variables (*X*) between the high-income and the low-income, and a component that is due to difference if the coefficients (β). The decomposition can be written as:

$$BMI^{rich} - BMI^{poor} = \Delta X\beta^{poor} - \Delta\beta X^{rich} \quad (3\text{-}4)$$

$$BMI^{rich} - BMI^{poor} = \Delta X\beta^{rich} - \Delta\beta X^{poor} \quad (3\text{-}5)$$

where $\Delta X = X^{rich} - X^{poor}$, and $\Delta\beta = \beta^{rich} - \beta^{poor}$. The Blinder-Oaxaca decomposition is a special case of a more comprehensive decomposition as follows (Rashad and Sharaf 2016):

$$BMI^{rich} - BMI^{poor} = \Delta X(D\beta^{rich} + (I - D)\beta^{poor}) + \Delta\beta(X^{poor}(I - D) + X^{rich}D) \quad (3\text{-}6)$$

where D is a matrix of weights and I is an identity matrix. Since there is no reason to believe that the income-related "discrimination" is strictly favouring one group over the other, and since the high-income and low-income groups in this study have very different sample sizes (there is substantially more observations in the high-income group than in the low-income group), I follow the suggestion by Reimers (1983) and Cotton (1988) of weighting the equation by the average mean ($\beta^* = \frac{1}{2}\beta^{rich} + \frac{1}{2}\beta^{poor}$) and relative sample sizes ($\beta^* = \frac{n_{rich}}{n_{rich}+n_{poor}}\beta^{rich} + \frac{n_{poor}}{n_{rich}+n_{poor}}\beta^{poor}$) respectively.

3.4. Data and Summary Statistics

The data used for this study are taken from one cycle of the Canadian Community Health Survey (CCHS). CCHS is a cross-sectional and nationally representative dataset which contains information on the health status of the respondents and their households as well as information on their demographic and socio-economic statuses. The CCHS data is collected through a random digit dial telephone survey and includes all Canadians over the age of 12 except those living in First Nations reserves, institutions, and those serving in the armed forces. I use year 2012 of the data which contains 61,707 observations (individuals). Statistics Canada assigns a survey weight to each respondent included in the final sample which corresponds to the number of persons that the respondent represents in the sample. The survey

weights are implemented in the estimations but in the summary statistics. In this study I use individuals who are 18.5 years and above. The final sample is split by gender.

The demographic and socioeconomic variables in the model include age, gender, marital status, household income, household arrangements, education, occupation, immigration status, behavioral variables, and geographical variables. Age is reported in 4 year intervals in the dataset therefore midpoints are used to indicate the age of the respondents. For marital status three binary variables are constructed: married=1 if the respondent is married, divorced=1 if the respondent is divorced or widowed, and single=1 if the respondent has never been married (reference group). Home ownership is defined as a binary variable which is equal to one if the home is owned by the respondent or a household member and zero otherwise. Education level is defined in 4 categories: “less than high school” which is equal to one if the respondent is a high school dropout and zero otherwise (reference group), “high school” which is equal to one if the respondent has finished high school and zero otherwise, “some college education” which is equal to one if the respondent has attended college and zero otherwise, and “college graduate” which is equal to one if the respondent has a college degree and zero otherwise. Employment status is captured by a binary variable, “employed”, which is equal to one if the respondent is currently employed and zero otherwise. A binary variable is constructed in order to indicate whether the respondent is Canadian born as opposed to an immigrant.

Behavioral variables include fruit and vegetable consumption, level of physical activity, and drinking and smoking habits. “Fruit and vegetables consumption” is a continuous variable indicating total servings of fruits and vegetables consumed per day. The physical activity variable is a continuous variable indicating average daily energy expenditure from all leisure time physical activities. Drinking habit is captured as a continuous variable indicating average daily alcohol consumption, i.e. average number of drinks consumed per day. Smoking is also captured in a continuous variable indicating average number of cigarettes smoked per day which is equal to zero for non-smokers and positive values for daily or occasional smokers. In addition provincial binary variables for all 10 provinces are constructed as indicators for province of residence.

The outcome variable of interest is a continuous variable indicating body mass index (BMI) of the individuals. This variable is provided in the data based on self-reported information on height and weight of the respondents. BMI is reported for all respondents except pregnant women.

In order to have a measure of poverty (low-income) I use HH income. The HH income categories in CCHS include: HH income less than $20,000, HH income between $20,000 and $39,999, HH income between $40,000 and $59,999, HH income between $60,000 and $79,999, and HH income of $80,000 or higher. In order to categorize the respondents into low- and high-income groups to carry out the analysis we need to have an income cut-off below which the respondent is considered “low-income”. Although there is no official poverty measure in Canada, statistics

Canada provides a few measures of poverty, the oldest and most common of which is the low Income cut-off (LIC). As there is no consensus as to which poverty measure should be used, I use the low income cut-offs reported by Statistics Canada for year 2012 based on HH size (Statistics Canada 2015). "They [low-income cut-offs] reflect a consistent and well-defined methodology that identifies those who are substantially worse off than the average. In the absence of an accepted definition of poverty, these statistics have been used by many analysts who wanted to study the characteristics of the relatively worse off families in Canada." (Statistics Canada 1999). In Statistics Canada there is a low income cut-off reported for each size of family unit (from 1 person to 7 or more persons) in each year for each community size. These cut-off are reported every year and represent the income level at (or below) which a family spends 20 percentage points more than an average family on basic necessities (foods, shelter, clothing). The cut-offs are estimated by Statistics Canada using the 1992 Family Expenditure Survey. As I only have information on province of residence for the respondents in our sample and do not have detailed information on their community size and/or census area of residence, I used the cut-offs for the biggest community size in Statistics Canada which belongs to a census metropolitan area of 500,000 inhabitants or more. Since HH income is reported in categories in the CCHS, in order to be able to compare the respondents' HH income with the cut offs defined in Statistics Canada, I constructed mid-points for each HH income category reported in the CCHS. Therefore the midpoint for the first income category (less than $20,000) is $10,000, the midpoint of the second category (between $20,000 and $39,999) is $30,000, the midpoint of the third category

(between $40,000 and $59,999) is $50,000, the midpoint of the fourth category (between $60,000 and $79,999) is $70,000, and finally the midpoint of the fifth HH income category (greater than $80,000) is $120,000. The CCHS provides information on HH size in 5 categories: 1 person, 2 persons, 3 persons, 4 persons, and 5 or more persons. For the HH size of 5 persons or more in the CCHS, the average LIC of the last three HH size categories from the Statistics Canada table (5 persons, 6 persons, and 7 or more persons) are used. Accounting for HH size, if the respondents HH income reported in CCHS falls below the cut-off, then the respondent is categorized as "low-income", and if HH income is above the cut-off they are categorized as "high-income". For example, if there is a respondent with an annual HH income of $50,000 per year and a HH size of 2 persons, the respondent is categorized as high-income since the corresponding LIC for a HH size of 2 (in year 2011) reported in Statistics Canada is $29,004 per year. The final sample includes individuals with no missing information on any of the variables described above.

Table 3.1 shows the un-weighted summary statistics for the male and female subsamples. The male subsample includes size is 20,027 observations. Mean age is 51 years old. 60% of men are married (or in a common law relationship) and the majority of non-married men are single as opposed to being widowed/separated/divorced. 74% are college graduates. About 16% of men have kids under the age of 12 in their households, 77% own their homes (or a member in their household owns the home, average HH size is 2.24 persons, and average HH income is 73,230 dollars per year. Average fruit and vegetable consumption is 4.27

grams per day, average daily energy expenditure is 2.36 units, average daily alcohol consumption is 0.51 units, and average daily cigarette consumption is 3.36 units. Mean BMI is 27.12 and about 22% of men are obese. 89% of men belong to the high income group. Among high income men, 64% are married, 23% are single, and 13% are widowed/separated/divorced. The majority are college graduates, employed, and own their homes. Only 16% have kids under 12 years of age. Average daily fruit and vegetable consumption, average daily alcohol consumption, and average daily cigarettes consumption among high income men is 4.33 units, 0.52 units, and 3.04 units respectively. Average daily energy expenditure is 2.39 units among high income men. Average BMI is 27.20 in the high income category and about 22% of men in this category are obese. Among low income men, 41% are single, 33% are married, and 26% are widowed/divorced/separated. A little over half of the sample are college graduates, 34% are employed, and 46% own their homes. Average daily fruit and vegetable consumption, average daily alcohol consumption, and average daily cigarette consumption is 3.83, 0.37 and 5.96 units respectively. Average daily energy expenditure is 2.11 units. Average BMI is 26.49 in the low income category and about 21% of men in this category are obese. Comparing the statistics of the high and low income men we can observe that the portion of low income men that are single or widowed/divorced/separated is almost twice as much as that of high income men while the portion of low income married men is almost half as much as the portion of high income married men. There are more college graduates among the high income than among the low income. The percentage of employed men in the high income group is twice as high as the percentage of the employed in the low

income category. While high income men have higher levels of average daily alcohol and fruit/vegetable consumption and average daily energy expenditure than low income men, the difference in these values in the high and low income categories is relatively small.

There are 25,406 women in the sample. An average woman is 53 years old and the majority of women are married (or in a common law relationship). Among non-married women, the majority are separated/widowed/divorced as opposed to being single (never married). 71% of women are college graduates. 75% own their homes (or belong to households where a member of the household owns the home), 18% have kids under the age of 12 years in the HH, average HH size is 2.16 persons, and the average household income is 64,279.7 dollars per year. Average daily number of cigarette consumption among women is 2.19 units, average daily alcohol consumption is 0.21 units, and average daily energy expenditure form leisure time physical activity is 2.05 units. Average BMI is 26.21 and about 21% of women are obese. An average woman in the sample consumes 5.10 grams of fruits and/or vegetables per day. 83% of women belong to the high-income group. Mean BMI is 26.62 among the low-income group and 25% of women in this income category are obese. Mean BMI is 26.13 among the high income group and 21% of women in this group are obese. Among high income women, 60% are married, 18% are single, and 22% are widowed/divorced/separated. The majority are college graduates and own homes. 17% of women in this category have kids under the age of 12 in their HHs. About 57% are employed. Average daily fruit and vegetable consumption, alcohol consumption, and cigarette consumption is 5.19, 0.22, and 1.94 units respectively.

Average daily energy expenditure among high income women is 2.13 units. Among low income women, 23% are married, 26% are single, and 51% are widowed/divorced/separated. Less than half of the subsample (45%) are college graduates. 23% have young kids, 26% are employed, and 44% live in HHs where a member owns the home. Average daily fruit/vegetable consumption is 4.61 units, average daily alcohol consumption is 0.13 units, and average daily number of cigarettes consumed is 3.43 units. An average low income woman expends 1.68 units on leisure time physical activities per day. Overall, the percentage of single (widowed/divorced/separated) women in the low income group is twice (over twice) as high as that of the income group, whereas the portion of married women in the low income category is less than half of the portion of married women in the high income category. An average high income woman has a higher daily energy expenditure level, consumes more alcohol and fruits and vegetables, and smokes less than an average low income woman.

Table 3.1 Summary Statistics by Subsample and Income Group

	Male				Female			
	High-Income		Low-Income		High-Income		Low-Income	
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Age	51.43	17.82	50.45	19.01	52.75	17.84	55.84	21.08
Canadian Born	0.85	0.36	0.82	0.39	0.86	0.35	0.83	0.38
No High School	0.08	0.27	0.26	0.44	0.08	0.27	0.34	0.47
High School Grad	0.13	0.33	0.16	0.37	0.12	0.33	0.16	0.36
Some College	0.03	0.18	0.06	0.24	0.03	0.17	0.06	0.23
College Grad	0.76	0.42	0.52	0.5	0.77	0.42	0.45	0.5
Kids	0.16	0.37	0.19	0.39	0.17	0.38	0.22	0.41
HH Size	2.26	1.09	2.11	1.41	2.2	1.1	1.94	1.33
HH Income	80241	36654	16591	11768	74160	37135	15153	10497
Owns Home	0.81	0.39	0.46	0.5	0.82	0.39	0.44	0.5
Avg Daily Fruit/Veg	4.33	2.43	3.83	2.52	5.19	2.57	4.61	2.58
Avg Daily Energy	2.39	2.38	2.11	2.54	2.13	2.09	1.68	1.91
Avg Daily Cigarettes	3.04	7.37	5.96	10.39	1.94	5.38	3.43	7.37
Avg Daily Alcohol	0.52	1.1	0.37	1.13	0.22	0.58	0.13	0.56
Single	0.23	0.42	0.41	0.49	0.18	0.38	0.26	0.44
Widow/Div/Sep	0.13	0.34	0.26	0.44	0.22	0.42	0.51	0.5
Married	0.64	0.48	0.33	0.47	0.6	0.49	0.23	0.42
Employed	0.66	0.47	0.34	0.47	0.57	0.5	0.26	0.44
BMI	27.2	4.65	26.49	5.21	26.13	5.54	26.62	6.31
Obese	0.22	0.42	0.21	0.41	0.21	0.4	0.25	0.43
Observations	17821		2206		21152		4254	

3.5. Results

3.5.1. Regression Results

Table 3.2 and Table 3.3 show the coefficient estimates from regressing BMI on the explanatory variables for the high-income and low-income groups separately. Since the full male and female samples as well as the high-income subsamples within each gender sample are large, I considered a p-value threshold of 5% instead of 10% in

order to avoid over interpreting the estimates in cases where these large samples are used.

Among both high- and low-income men, BMI increases with age. Having young kids in the HH is positively associated with BMI among both income groups while being single is negatively associated with BMI in both groups. Average daily fruit and vegetable consumption, average daily energy expenditure, and average daily cigarette consumption are all negatively correlated with BMI among the high-income male. Employed high-income men have higher average BMIs than unemployed high-income men. Among the coefficients on SES in the high-income male sample, the largest coefficients belong to being single and being employed (after being Canadian born). Among coefficients on health behaviours, average daily energy expenditure has the largest magnitude.

Among women, average daily fruit and vegetable consumption, average daily energy expenditure and average and average daily alcohol consumption are all negatively associated with BMI among both income groups. Being employed is positively associated with BMI among high-income women but not among the low-income. Being single, is negatively associated with BMI in both income groups, while being widowed/divorced/separated is negatively correlated with BMI only in the high-income group.

Table 3.2 OLS Estimates by Income Group for the Male Subsample

	High-Income		Low-Income	
Variable	Mean	Std. Err.	Mean	Std. Err.
Age	0.03***	0.00	0.03**	0.01
Canadian Born	1.52***	0.14	0.78*	0.43
High School Graduate	−0.02	0.29	0.36	0.56
Some College	0.21	0.45	−0.19	0.61
College Graduate	−0.14	0.24	0.15	0.49
Kids	0.37**	0.17	0.91*	0.49
Owns Home	0.30**	0.15	−0.28	0.38
Avg Daily Fruit/Veg	−0.09***	0.02	−0.01	0.07
Avg Daily Energy	−0.14***	0.02	−0.10	0.07
Avg Daily Cigarettes	−0.02**	0.01	0.00	0.02
Avg Daily Alcohol	−0.08	0.05	−0.15	0.14
Single	−0.74***	0.18	−0.78*	0.46
Widow/Div/Sep	−0.20	0.20	0.34	0.59
Employed	0.68***	0.14	−0.29	0.40
Constant	24.87***	0.52	24.45***	1.28
R-squared	0.06		0.08	
Observations	17821		2206	

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.3 OLS Estimates by Income Group for the Female Subsample

	High-Income		Low-Income	
Variable	Mean	Std. Error	Mean	Std. Error
Age	0.04***	0.01	0.04***	0.01
Canadian Born	1.16***	0.19	1.39***	0.44
High School Graduate	0.33	0.31	−0.06	0.78
Some College	−0.01	0.39	0.34	0.86
College Graduate	−0.02	0.25	−0.55	0.50
Kids	−0.24	0.20	0.75	0.48
Owns Home	−0.25	0.18	−0.81**	0.39
Avg Daily Fruit/Veg	−0.07**	0.03	−0.14*	0.07
Avg Daily Energy	−0.34***	0.03	−0.21**	0.08
Avg Daily Cigarettes	−0.02	0.01	−0.02	0.03
Avg Daily Alcohol	−0.43***	0.10	−0.77***	0.16
Single	−0.80***	0.21	−0.87*	0.48
Widow/Div/Sep	−0.44***	0.18	−0.56	0.58
Employed	0.42***	0.17	−0.63	0.44
Constant	25.32***	0.67	25.73***	1.42
R-squared	0.07		0.09	
Observations	21152		4254	

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.5.2. Decompositions Results

The first panel of Table 3.4 shows the mean BMI in the high- and low-income group, as well as the difference in mean BMI between the two income groups for the male and female samples separately. The predicted average BMI is 26.75 among high-income men, and 25.79 among low-income men, producing a predicted gap of 0.96 units which is highly significant. The predicted mean BMI among high-income women is 25.55, and 25.92 among low-income women, producing a predicted gap of 0.36 units. As the gap in female BMI is not significant, in the subsequent analysis I will only focus on the decomposition results of the male subsample.

The second panel of Table 3.4 shows the results of the decomposition (for the male sample) using two different weighting schemes. The first column indicates the results when weighting the gap by the average mean of the coefficients of the high- and low-income groups. The explained part of the gap is 0.37 units and significant which indicates that 0.37 units of the gap in mean BMI between the two income groups can be explained by differences in the magnitude of the determinants of BMI between the two groups. On the other hand, 0.59 units of the gap in mean BMI is unexplained, meaning that even if the high- and low-income men had the same average level of the observed control variables, 0.59 units of the gap would still exists. This portion of the gap is due to the difference in the coefficients i.e. the difference in the returns to the observed characteristics between the high- and low-income. In column 2, the coefficients of both groups weighted by the relative sample sizes of the groups are used for weighting the gap in mean X's. The explained portion of the gap is significant and accounts for most of the gap in mean BMI (0.59 units of the gap in average BMI is explained by differences in the determinants respectively).

Table 3.5 shows the contributions of each individual characteristic to the explained gap among men using different the two different weighting schemes. In both cases age, average daily energy expenditure, having kids under 12 years old in the HH, and being single (as opposed to being married) consistently contribute to the explained gap. Among other determinants that contribute to the explained gap in either specification used are being employed, average daily fruit and vegetable consumption, and average daily cigarette consumption. In both weighting schemes, 0.13 units of the gap is explained by being single (as opposed to being married)

meaning that being single contributes to the low-income men having lower average BMIs than high-income men. This result can be explained by the difference in the portion of single/married men between the high- and low-income groups. Among high-income men 64% are married and 23% are single whereas among the low-income 0.33% are married and 41% are single. In addition, the OLS estimates show that being single is associated with lower BMI (among both income groups) than being married, therefore the difference in the percentage of single/married people in the high- and low-income categories can, at least partly, explain this phenomenon. The difference in average daily energy expenditure explains 0.06 units of the gap which contributes to the low-income men having higher mean BMIs than high-income men. In the second column, employment status has the highest contribution (other than being Canadian born) towards the low-income having lower average BMIs than the high-income (0.18 units), this is potentially due to the fact that employed men in the high-income group are more likely to have white collar occupations whereas men in the low-income groups are more likely to have blue-collar occupations. As blue-collar occupations are more physically demanding and more labour intensive, they can potentially explain why employment status contributes to the low-income having lower mean BMI tan the high-income. In the same latter case, average daily fruit and vegetable consumption and average daily cigarette consumption contribute to the gap in opposite directions in that average daily cigarette consumption contributes to low-income men having lower average BMIs while average daily fruit and vegetable consumption contributes to low-income men having higher BMIs than the high-income. The latter results are likely

due to the fact that smoking is more prevalent among the low-income and since cigarette consumption is associated with a decreased appetite, it contributes to the low-income having lower BMIs. On the contrary, low-income men expend (slightly) less average energy on leisure time physical activities which contributes to them having higher average BMIs than the high-income, though this contribution is small compared to that of employment status which implies that the high-income are generally more sedentary (due to their employment status) than the low-income. Overall, although difference in health behaviours generally contribute to high-income men having lower average BMIs than low-income men (the high income have higher average daily energy expenditures and lower amounts of cigarette consumption which both have negative impacts on body weight), the magnitude of the contribution of the gap in these behaviours is not enough to offset the contributions that marital status and employment status have to high-income men having higher average BMIs than the low-income.

Table 3.4 Decomposition Results of the Income Gap in Mean BMI by Weighting Scheme

	Male	Female
High-Income	26.75***	25.55***
	(0.06)	(0.07)
Low-Income	25.79***	25.92***
	(0.18)	(0.20)
Difference	0.96***	0.36
	(0.19)	(0.21)
	Male	
	D = 0.5	D = 0.891
Unexplained (U){C+(1-D)*CE}:	0.59***	0.37
	(0.19)	(0.20)
Explained (V) {E+D*CE}:	0.37***	0.59***
	(0.13)	(0.10)

Standard errors in parentheses. *** $p < 0.01$

Table 3.5 Contributions to the Explained Income Gap in Mean BMI for Male by Weighting Scheme

	D=0.5		D=0.891	
Variable	Mean	Std. Error	Mean	Std. Error
Age	0.09***	0.03	0.09***	0.03
Canadian Born	0.18***	0.05	0.23***	0.04
High School Graduate	−0.011	0.02	0	0.02
Some College	0	0.01	0	0.01
College Graduate	0.001	0.06	−0.02	0.04
Kids	−0.06**	0.03	−0.04**	0.02
Owns Home	0.003	0.07	0.08	0.05
Avg Daily Fruit/Veg	−0.03	0.02	−0.04***	0.01
Avg Daily Energy	−0.06***	0.02	−0.06***	0.02
Avg Daily Cigarettes	0.02	0.02	0.03**	0.02
Avg Daily Alcohol	−0.03	0.02	−0.02	0.01
Single	0.13***	0.04	0.13***	0.03
Widow/Div/Sep	−0.003	0.01	0.01	0.01
Employed	0.06	0.07	0.18***	0.04
Total	0.37***	0.13	0.59***	0.1
Observations	20027		20027	

** $p < 0.05$, *** $p < 0.01$

3.6. Sensitivity Check

In the baseline approach, the sampling method (for categorizing the respondents into high- and low-income) puts the majority of the respondents into the high-income group. Although the use of the weighting schemes in the decomposition estimations (especially the weighting scheme in which the coefficients are weighted by the relative sample sizes) should, to a certain degree, address this problem, as a check for robustness/sensitivity I have consider an alternative sampling method. For this

purpose, I use information provided in the CCHS on the income deciles that the respondents belong to. Income deciles indicate the distribution of the respondents, in deciles, based on the adjusted ration of their HH income to the low income cut-off corresponding to their HH size and community size (Note that the community sizes that the HHs belong to are not reported in the data and are only used by Statistics Canada in order to calculate the adjusted ratio and to group respondents into the 10 income deciles reported in the shared data. Therefore, in the baseline analysis, I had to make an assumption bout the community size. Nevertheless, given the values of the low-income cut-offs corresponding to other community sizes (reported by Stats Canada) and given the HH income categories provided in the shared CCHS data set, the majority of individuals would still fall into the same income categories that they have in the baseline analysis of this study, regardless of their community sizes). Respondents in the bottom three income deciles are categorized as low-income, and the ones in the top three income deciles are categorized as high-income. This categorization is different from the one used in the baseline analysis in two different ways: 1) it produces high- and low-income groups that are more similar in size compared to the income groups used in the baseline analysis, and 2) it includes only individuals in the two extreme income categories (top three and bottom three of the income distribution). Using this sampling method, a number of individuals in the original sample (in the baseline analysis) are automatically dropped from analysis since they belong the middle income deciles (the analysis only includes a subset of the respondents that belong to either the bottom three or top three income deciles), therefore the overall sample size is smaller than in the baseline analysis. The male

subsample includes 11,468 individuals with 7,155 individuals belonging to the high-income group and 4,313 individuals belonging to the low-income group. The female subsample include 14,682 individuals with 7028 individuals belonging to the high-income group and 7654 individuals belonging to the low-income group. The summary statistics and OLS regression results are presented in Appendix A (Table A 2, Table A 3 and Table A 4).

Table 3.6 shows the decomposition results for the male and female subsamples using the alternative sampling strategy. The first panel of the table indicates that, similar to the results in the baseline analysis, mean BMI is higher in the high-income male group than in the low-income group, and the gap (1.33 units of BMI) is highly significant (and larger than in the baseline results). The income gap in mean BMI in the female subsample (0.47 units of BMI) also becomes highly significant (with a higher mean BMI in the low-income group) using this alternative grouping strategy. The second and third panels of the table show the decomposition results using the two different weighting schemes used in the baseline analysis. Among men, both the explained and unexplained portion of the income gap in mean BMI are significant, whereas among women, only the explained portion of the gap is significant and constitutes the majority of the gap. Among men, although a huge bulk of the gap (0.50 units of the gap when weighting by the average mean, and 0.52 units of the gap when weighting by relative sample sizes) is still explained by differences in the characteristics between the two income groups, the majority of the gap is unexplained and is due to differences in how those characteristics manifest differently among the high- and low-income groups.

Table 3.7 and Table 3.8 shows the contributions of the observed covariates to the explained portion of the gap. In the male subsample, regardless of the weighting scheme used, average daily energy expenditure and being single are highly significant in contributing to the explained gap. In addition, using Cotton's weighting scheme, average daily fruit/vegetable consumption, average daily alcohol consumption, and being employed also contribute to the explained gap in mean BMI among men. These results are similar to the baseline case with the exception of the contribution of alcohol consumption which was not significant in the baseline analysis but is significant using this alternative grouping strategy. The coefficient on average daily alcohol consumption indicates that if the low-income men were to have the same level of daily alcohol consumption as in high-income men, their mean BMI would be 0.02 units lower than the current mean level. The magnitudes of the contribution of the observed covariates are slightly different from that in the baseline case. For example, in the baseline analysis, being single had a higher contribution to the explained gap than average daily energy expenditure, whereas as this is the opposite in this current alternative analysis. Overall, the results are robust but sensitive to the income grouping strategy used in the male subsample.

In the female subsample where the income gap in mean BMI is now significant, age, average daily fruit/vegetable consumption, average daily alcohol consumption, average daily energy expenditure and being single all contribute to the explained gap. The signs on the coefficients of the health behaviors are negative (as expected) meaning that if the low-income women had the same average daily fruit/vegetable consumption, energy expenditure and alcohol consumption, their

mean BMI would be lower the current value. The difference in average daily energy expenditure between high- and low-income women (which is favoring the high-income women) has the highest contribution to the explained income gap in mean BMI.

Table 3.6 Decomposition Results of the Income Gap in Mean BMI by Weighting Scheme

	Male	Female
High-Income	27.11***	25.30***
	(0.09)	(0.11)
Low-Income	25.78***	25.77***
	(0.12)	(0.14)
Difference	1.33***	−0.47***
	(0.14)	(0.18)
	Male	
	D = 0.5	D = 602
Unexplained (U){C+(1-D)*CE}:	0.83***	0.81***
	(0.17)	(0.18)
Explained (V) {E+D*CE}:	0.50***	0.52***
	(0.13)	(0.13)
	Female	
	D = 0.5	D = 0.511
Unexplained (U){C+(1-D)*CE}:	−0.125	−0.121
	(0.24)	(0.24)
Explained (V) {E+D*CE}:	−0.34*	−0.35*
	(0.18)	(0.18)

Standard errors in parentheses. * p < 0.10, ** p < 0.05, *** p < 0.01

Table 3.7 Contributions to the Explained Income Gap in Mean BMI for Male by Weighting Scheme

	D=0.5		D=0.602	
Variable	Mean	Std. Error	Mean	Std. Error
Age	−0.03	0.02	−0.03	0.02
Canadian Born	0.28***	0.05	0.28***	0.05
High School Graduate	0.03	0.04	0.04	0.04
Some College	0.02	0.01	0.02	0.01
College Graduate	−0.10	0.08	−0.12	0.08
Kids	−0.01	0.01	−0.01	0.01
Owns Home	0.14**	0.07	0.16**	0.08
Avg Daily Fruit/Veg	−0.02	0.01	−0.02*	0.01
Avg Daily Energy	−0.10***	0.03	−0.11***	0.03
Avg Daily Cigarettes	0.02	0.02	0.02	0.02
Avg Daily Alcohol	−0.02*	0.02	−0.02	0.02
Single	0.08***	0.03	0.08***	0.03
Widow/Div/Sep	0	0.01	0	0.01
Employed	0.11	0.07	0.13*	0.07
Total	0.50***	0.13	0.52***	0.13
Observations	11468		11468	

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 3.8 Contributions to the Explained Income Gap in Mean BMI for Female by Weighting Scheme

	D=0.5		D=0.511	
Variable	Mean	Std. Error	Mean	Std. Error
Age	−0.16***	0.03	−0.16***	0.03
Canadian Born	0.19***	0.05	0.18***	0.05
High School Graduate	−0.03	0.06	−0.03	0.06
Some College	−0.02	0.04	−0.02	0.04
College Graduate	−0.11	0.17	−0.11	0.17
Kids	−0.01	0.02	−0.01	0.02
Owns Home	−0.12	0.1	−0.12	0.11
Avg Daily Fruit/Veg	−0.05*	0.03	−0.05*	0.03
Avg Daily Energy	−0.23***	0.03	−0.23***	0.03
Avg Daily Cigarettes	0.01	0.02	0.01	0.02
Avg Daily Alcohol	−0.06***	0.02	−0.06***	0.02
Single	0.07**	0.03	0.07**	0.03
Widow/Div/Sep	0.06	0.05	0.06	0.05
Employed	0.02	0.07	0.02	0.07
Total	−0.34*	0.18	−0.34*	0.18
Observations	14682		14682	

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

3.7. Discussion and Conclusion

In this study I used one wave of the Canadian Community Health Survey and a Blinder-Oaxaca decomposition tool in order to decompose the income gap in mean BMI among men and women. Low-Income Cut-offs by household size from Statistics Canada were used to define the two (high and low) income categories. I find that the BMI gap is highly significant among men but insignificant among women (in the baseline analysis). Using two weighting schemes (proposed in the

literature) for the decomposition analysis, the results indicate that a huge bulk of the gap in mean BMI between high- and low-income men can be explained by differences in the SES and behavioural characteristics between the two groups. More specifically, age, being single (as opposed to being married), employment status, and average daily cigarette consumption contribute to the low-income having lower mean BMIs than the high-income men since the average age of the low-income is lower, the proportion of single men in the low-income group is substantially higher, there are more unemployed men in the low-income group, and average daily cigarette consumption is higher among the low-income than among the high-income men, all of which negatively affect BMI. On the contrary, average daily energy expenditure on leisure time physical activities, average daily fruit and vegetable consumption, and having young kids (under 12 years of age) in the household all contribute to the low-income men having higher mean BMIs than the high-income since the low-income consume less fruits and vegetables daily and expend less energy on daily physical activity, both of which negatively affect BMI. The results are sensitive to the strategy implemented in grouping individuals into high- and low- income categories. When using an alternative strategy (top and bottom three income deciles) for grouping individuals into the high- and low-income groups, the income gap in mean BMI among women also becomes significant. The highest contribution to the explained gap in mean BMI among women belongs to the difference in average daily energy expenditure between the two groups.

An important limitation of this study is that the socioeconomic factors and health behaviours that explain differences in BMI do not necessarily reflect causal

effects and it is beyond the scope of this paper to establish causality. For example, as much as average daily energy expenditure has a negative effect on individuals' BMI, people with lower BMIs might be more likely to be physically active. Therefore the results of this paper should be interpreted as correlations/associations rather than causal effects. Another limitation of the study is that any unobserved determinant of BMI that is missing from the analysis gets absorbed into the unexplained portion of the gap. For example, detailed information on race/ethnicity is missing in the data which can potentially capture differences in BMI through different biological factors and eating habits between different ethnic and racial groups. The explanatory power of the OLS models are relatively low ranging from 0.06 to 0.09, therefore future research can benefit from more informative data to address these issues. Finally the robustness of the estimates can further be examined by using alternative strategies for grouping individuals into high- and low-income, or by considering a pairwise comparison between any of the two income categories provided in the dataset. Nevertheless it provides some insights that can help policy makers in designing effective policies aimed at reducing the explained income gap in BMI which is due to differences in the socioeconomic status of the individuals and their health behaviors. The results of this study suggest that policies aimed at reducing the income gap in obesity prevalence should particularly focus on high-income men and low-income women. For example, policies should aim at increasing leisure time physical activity among low-income women and among employed men in the high-income category. In addition, interventions aimed at increasing fruit and

vegetable consumption and reducing the prevalence of smoking among the low-income/poor can prove to be effective in reducing the income gap in BMI.

4. References

Alaba, Olufunke, and Lumbwe Chola. 2014. "Socioeconomic Inequalities in Adult Obesity Prevalence in South Africa: A Decomposition Analysis." *International Journal of Environmental Research and Public Health* 11 (3): 3387–3406.

Akerlof, George, and Rachel E Kranton. 2000. "Economics and Identity." *The Quarterly Journal of Economics* 115 (3).

Angrist, Joshua D, and Alan B Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments." *Journal of Economic Perspectives* 15 (4): 69–85.

Arellano, Manuel, and Stephen Bond. 1991. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations." *The Review of Economic Studies* 58 (2): 277–97.

Aubin, Henry-Jean, Amanda Farley, Deborah Lycett, Pierre Lahmek, and Paul Aveyard. 2012. "Weight Gain in Smokers after Quitting Cigarettes." *The British Medical Journal* 4439: 1–21.

Azagba, Sunday, and Mesbah F Sharaf. 2011. "The Effect of Job Stress on Smoking and Alcohol Consumption." *Health Economics Review* 1 (1): 1–15.

Babey, Susan H, Theresa A Hastert, Joelle Wolstein, and Allison L Diamant. 2010. "Income Disparities in Obesity Trends among California Adolescents."

American Journal of Public Health 100 (11): 2149–55.

Barron, John M, Bradley T Ewing, and Glen R Waddell. 2000. "The Effects of High School Athletic Participation on Education and Labor Market Outcomes." *The Review of Economics and Statistics* 82 (3): 409–21.

Bauer, Thomas K, and Mathias Sinning. 2008. "An Extension of the Blinder-Oaxaca Decomposition to Nonlinear Models." *Advances in Statistical Analysis* 92 (2): 197–206.

Becker, Gary S. 1965. "A Theory of the Allocation of Time." *The Economic Journal* 75 (299): 493–517.

Becker, Gary S. 1976. "The Economic Approach to Human Behavior." Chicago, IL: University of Chicago Press.

Berry, Leonard L, Ann M Mirabito, and William B Baun. 2010. "What's the hard return on employee wellness programs?" *Harvard Business Review*. December: 104–112

Blinder, Alan. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *Journal of Human Resources* 8 (4): 436–55.

Blundell, Richard, and Stephen Bond. 1998. "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models." *Journal of Econometric*s 87 (1): 115–143.

Bolton, Kelly L, and Eunice Rodriguez. 2009. "Smoking, Drinking and Body Weight after Re-Employment: Does Unemployment Experience and Compensation Make a Difference?" *BioMed Central Public Health* 9 (1): 77.

Cabane, C. 2014. "Unemployment duration and sport participation." *International Journal of Sport Finance* 9 (3): 261–280.

Cabane, Charlotte, and Michael Lechner. 2015. "Physical Activity of Adults : A Survey of Correlates , Determinants , and Effects." *Journal of Economics and Statistics* 235 (4): 367–402.

Cabane, Charlotte, and Andrew E Clark. 2015. "Childhood Sporting Activities and Adults Labour-Market Outcomes." *Annals of Economics and Statistics* 119: 123–148.

Canadian Sport Policy. 2012. Canadian Sport Policy 2012. Technical Report.

Cawley, John, and Donald S Kenkel. 2008. "Introduction." In: Cawley and Kenkel (Eds.). *The Economics of Health Behaviours.* Northampton, MA: Edward Elgar.

Cawley, John, and Christopher J Ruhm. 2011. "The Economics of Risky Health Behaviors." *NBER Working Paper No. 17081.*

Chang, Virginia W, and Diane S Lauderdale. 2005. "Income Disparities in Body Mass Index and Obesity in the United States, 1971-2002." *Archives of Internal Medicine* 165 (18): 2122–28.

Chou, Shin-Yi, Michael Grossman, and Henry Saffer. 2004. “An Economic Analysis of Adult Obesity: Results from the Behavioral Risk Factor Surveillance System.” *Journal of Health Economics* 23 (3): 565–87.

Chung, Martin, Peter Melnyk, Donald Blue, Donald Renaud, Marie-Claude Breton. 2009. "Worksite Health Promotion: the value of the tune up your heart program". *Population Health Management* 12(6): 297–304

Colley, Rachel C, Didier Garriguet, Ian Janssen, Cora L Craig, Janine Clarke, Mark S Tremblay. 2011. “Physical Activity of Canadian Children and Youth : Accelerometer Results from the 2007 to 2009 Canadian Health Measures Survey.” *Health Reports* 22 (1): 15–23

Costa-Font, Joan, and Joan Gil. 2008. “What Lies Behind Socio-Economic Inequalities in Obesity in Spain? A Decomposition Approach.” *Food Policy* 33 (1): 61–73.

Costa-Font, Joan, Quevedo C Hernandez, and Rubio D Jimenez. 2014. “Income Inequalities in Unhealthy Life Styles in England and Spain.” *Economics and Human Biology* 13 (1): 66–75.

Cotton, Jeremiah. 1988. “On the Decomposition of Wage Differentials.” *Review Of Economics & Statistics* 70 (2): 236–243.

De Vogli, Roberto, and Massimo Santinello. 2005. “Unemployment and Smoking: Does Psychosocial Stress Matter?” *Tobacco Control* 14 (6): 389–95.

Drewnowski, Adam, and SE Specter. 2004. “Special Article Poverty and Obesity : The Role of Energy Density and Energy Costs.” *American Journal of Clinical Nutrition* 79: 6–16.

Drydakis, Nick. 2014. “The Effect of Unemployment on Self-Reported Health and Mental Health in Greece from 2008 to 2013: A Longitudinal Study Before and During the Financial Crisis.” *IZA Discussion Paper No. 8742*.

Eide, Eric R, and Nick Ronan. 2001. “Is Participation in High School Athletics an Investment or a Consumption Good?” *Economics of Education Review* 20 (5): 431–42.

Etnier, Jennifer L, Walter Salazar, Daniel M Landers, Steven J Petruzzello, Myungwoo Han, and Priscilla Nowell. 1997. “The Influence of Physical Fitness and Exercise upon Cognitive Functioning: A Metaanalysis.” *Journal of Sport & Exercise Psychology* 19: 249–77.

Ewing, Bradley T. 2007. “The Labor Market Effects of High School Athletic Participation: Evidence From Wage and Fringe Benefit Differentials.” *Journal of Sports Economics* 8 (3): 255–65.

Ewing, Bradley T. 1998. “Athletes and Work.” *Economics Letters* 59 (1): 113–17.

Falba, Tracy, Hsun-mei Teng, Jody L Sindelar, and William T Gallo. 2005. “The Effect of Involuntary Job Loss on Smoking Intensity and Relapse.” *Addiction* 100 (9): 1330–39.

Felfe, Christina, Michael Lechner, and Andreas Steinmayr. 2016. "Sports and Child Development." *Public Library of Science ONE* 11(5).

Fergusson, David M, L John Horwood, and M T Lynskey. 1997. "The Effects of Unemployment on Psychiatric Illness during Young Adulthood." *Psychological Medicine* 27 (2): 371–81.

Fryer, David, and Rose Stambe. 2015. "Unemployment and Mental Health." *International Encyclopedia of Social and Behavoral Sciences* 2: 733–37.

Gathergood, John. 2013. "An Instrumental Variables Approach to Unemployment, Psychological Health and Social Norm Effects." *Health Economics* 22 (6): 643–54.

Gilmour, Heather. 2007. "Physically Active Canadians." *Health Reports* 18 (3): 45-65.

Goel, Rajeev K. 2008. "Unemployment, Insurance and Smoking." *Applied Economics* 40 (20): 2593–99.

Grossman, Michael. 1972. "On the Concept of Health Capital and the Demand for Health." *Journal of Political Economy* 80 (2): 223–49.

Hajizadeh, Mohammad, M Karen Campbell, and Sisira Sarma. 2014. "Socioeconomic Inequalities in Adult Obesity Risk in Canada: Trends and Decomposition Analyses." *European Journal of Health Economics* 15 (2): 203–21.

Hammarström, Anne, and Urban Janlert. 2003. “Unemployment -- an Important Predictor for Future Smoking: A 14-Year Follow-up Study of School Leavers.” *Scandinavian Journal of Public Health* 31 (3): 229–32.

Hansen, Lars P, and Kenneth J Singleton. 1982. “Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models” *Econometrica* 50 (5): 1269–1286.

Hanson, Margaret D, and Edith Chen. 2007. “Socioeconomic Status and Health Behaviors in Adolescence: A Review of the Literature.” *Journal of Behavioral Medicine* 30 (3): 263–85.

Haskell, William L, I Min Lee, Russell R Pate, Kenneth E Powell, Steven N Blair, Barry A Franklin, Caroline A Macera, Gregory W Heath, Paul D Thompson, and Adrian Bauman. 2007. “Physical Activity and Public Health: Updated Recommendation for Adults from the American College of Sports Medicine and the American Heart Association.” *Circulation* 116 (9): 1081–93.

Heckman, James J, and Yona Rubinstein. 2001. “The Importance of Non-Cognitive Skills: Lessons from the GED Testing Program.” *The American Economic Review* 91 (2): 145–49.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47: 53–161.

Henderson, Daniel J, Alexandre Olbrecht, and Solomon W Polachek. 2006. “Do

Former College Athletes Earn More at Work? A Nonparametric Assessment." *Journal of Human Resources* 41 (3): 558–77.

Henkel, Dieter. 2011. "Unemployment and Substance Use: A Review of the Literature (1990-2010)." *Current Drug Abuse Reviews* 4 (1): 4–27.

Hill, Terrence D, Catherine E Ross, and Ronald J Angel. 2005. "Neighborhood Disorder, Psychophysiological Distress, and Health." *Journal of Health and Social Behavior* 46 (2): 170–86.

Holtz-Eakin, Douglas, Whitney Newey, and Harvey S Rosen. 1988. "Estimating Vector Autoregressions with Panel Data." *Econometrica* 56 (6): 1371–95.

Humphreys, Brad R, John Nyman, and Jane E Ruseski. 2011. "The Effect of Gambling on Health : Evidence from Canada" *American Society of Health Economists (ASHEcon) Paper No. 2011–18.*

Humphreys, Brad R, and Jane E Ruseski. 2011. "An Economic Analysis of Participation and Time Spent in Physical Activity." *The B.E. Journal of Economic Analysis & Policy* 11 (1): 138–59.

Hyytinen, Ari, and Jukka Lahtonen. 2013. "The Effect of Physical Activity on Long-Term Income." *Social Science and Medicine* 96: 129–37.

Janssen, Ian, and A Diener. 2009. "Economic Burden of Obesity in Canada in 2005." Public Health Agency of Canada. Ottawa.

Jolliffe, Dean. 2011. “Overweight and Poor? On the Relationship between Income and the Body Mass Index.” *Economics and Human Biology* 9 (4): 342–55.

Jones, Andrew M. 2000. “Chapter 6 Health Econometrics.” *Handbook of Health Economics* 1: 265–344.

Karim, Syahirah A, Terje A Eikemo, and Clare Bambra. 2010. “Welfare State Regimes and Population Health” *Health Policy* 94 (1): 45–53.

Kassel, Jon D, Laura R Stroud, and Carol A Paronis. 2003. “Smoking, Stress, and Negative Affect: Correlation, Causation, and Context across Stages of Smoking.” *Psychological Bulletin* 129 (2): 270–304.

Katzmarzyk, Peter T, and Ian Janssen. 2004. “The Economic Costs Associated with Physical Inactivity and Obesity in Canada: An Update.” *Canadian Journal of Applied Physiology* 29 (1): 90–115.

Kawachi, I, S V Subramanian, and N Almeida-Filho. 2002. “A Glossary for Health Inequalities.” *Journal of Epidemiology and Community Health* 56 (9): 647–52.

Kenkel, Donald S. 1995. “Should you Eat Breakfast? Estimates from Health Production Functions” *Health Economics* 4 (1): 15–29.

Kenkel, Donald S. 2000. “Prevention.” In Anthony J Culyer and Joseph P Newhouse (Eds.) *Handbook of Health Economics* 1: 1675–1720.

Khlat, Myriam, Catherine Sermet, and Annick Le Pape. 2004. “Increased Prevalence

of Depression , Smoking , Heavy Drinking and Use of Psycho-Active Drugs among Unemployed Men in France." *European Journal of Epidemiology* 19 (5): 445–51.

Klarenbach, Scott, Raj Padwal, Anderson Chuck, and Philip Jacobs. 2006. "Population-Based Analysis of Obesity and Workforce Participation." *Obesity* 14 (5): 920–27.

Kofi, Kerwin, and Melvin Stephens. 2004. "Job Displacement , Disability , and Divorce." *Journal of Labor Economics* 22 (2): 489–522.

Kosteas, Vasilios D. 2012. "The Effect of Exercise on Earnings: Evidence from the NLSY." *Journal of Labor Research* 33 (2): 225–50.

Kpelitse, Koffi-Ahoto, Rose Anne Devlin, and Sisira Sarma. 2014. "The Effect of Income on Obesity among Canadian Adults." *Canadian Centre for Health Economics Working Paper No. 2014–C02.*

Latif, Ehsan. 2014. "The Impact of Recession on Drinking and Smoking Behaviours in Canada." *Economic Modelling* 42: 43–56.

Lechner, Michael. 2009. "Long-Run Labour Market and Health Effects of Individual Sports Activities." *Journal of Health Economics* 28 (4): 839–54.

Lechner, Michael. 2015. "Sports, Exercise, and Labor Market Outcomes." *IZA World of Labor Paper No. 126.*

Lechner, Michael, and Paul Downward. 2013. "Heterogeneous Sports Participation and Labour Market Outcomes in England." *IZA Discussion Paper No. 7690.*

Lechner, Michael, and Nazmi Sari. 2015. "Labor Market Effects of Sports and Exercise: Evidence from Canadian Panel Data." *Labour Economics* 35: 1–15.

Ljungvall, Åsa, and Ulf G Gerdtham. 2010. "More Equal but Heavier: A Longitudinal Analysis of Income-Related Obesity Inequalities in an Adult Swedish Cohort." *Social Science and Medicine* 70 (2): 221–31.

Long, James E, and Steven B Caudill. 1991. "The Impact of Participation in Intercollegiate Athletics on Income and Graduation." *The Review of Economics and Statistics* 73 (3): 525–31.

Long, J Scott. 1997. "Pseudo-R2's Based on R2 in LRM." In J Scott Long (Eds). *Regression Models for Categorical and Limited Dependent Variables* 104–6. Thousand Oaks, CA: SAGE Publications.

Madden, David. 2008. "Sample Selection versus Two-Part Models Revisited: The Case of Female Smoking and Drinking." *Journal of Health Economics* 27 (2): 300–307.

Madden, David. 2013. "The Socio-Economic Gradient of Obesity in Ireland." *The Economic and Social Review* 44 (2): 181–96.

Maloney, Michael T, and Robert E Mccormick. 2016. "An Examination of the Role That Intercollegiate Athletic Participation Plays in Academic Achievement :

Athletes ' Feats in the Classroom." *The Journal of Human Resources* 28 (3): 555–70.

Mangione, Thomas W, and Robert P Quinn. 1975. "Job satisfaction, counterproductive behavior and drug use at work." *Journal of Applied Psychology* 60:114–116

Marmot, Michael. 2005. "Public Health Social Determinants of Health Inequalities." *Lancet* 365: 1099–1104.

Mas-Colell, Andreu, Michael D Whinston, and Jerry R Green. 1995. "Microeconomic Theory". New York: Oxford University Press.

Mathers, Colin D, and Deborah J Schofield. 1998. "The Health Consequences of Unemployment: The Evidence." *The Medical Journal of Australia* 168 (4): 178–82.

McLaren, Lindsay. 2007. "Socioeconomic Status and Obesity." *Epidemiologic Reviews* 29 (1): 29–48.

Mcleod, Logan, and Jane E Ruseski. 2015. "Longitudinal Relationship between Participation in Physical Activity and Health." *Canadian Centre for Health Economics Working Paper No. 150002.*

Merline, Alicia C, Patrick M O'Malley, and John E Schulenberg. 2004. "Substance Use Among Adults 35 Years of Age: Prevalence, Adulthood Predictors, and Impact of Adolescent Substance Use." *American Journal of Public Health* 94

(1): 96–102.

Mincer, Jacob. 1958. “Investment in Human Capital and Personal Income Distribution.” *The Journal of Political Economy* 66 (4): 281–302.

Moffitt, Robert. 1993. “Identification and estimation of dynamic models with a time series of repeated cross-sections.” *Journal of Econometrics* 59: 99–123.

Montgomery, S M, Derek G Cook, Mel J Bartley, and Michael E J Wadsworth. 1998. “Unemployment, Cigarette Smoking, Alcohol Consumption and Body Weight in Young British Men.” *European Journal of Public Health* 8 (1): 21–27.

Morris, Joan K, Derek G Cook, and A Gerald Shaper. 1992. “Nonemployment and Changes in Smoking, Drinking, and Body-Weight.” *British Medical Journal* 304 (6826): 536–41.

Nickell, Stephen. 1981. “Biases in Dynamic Models with Fixed Effects." *Econometrica* 49 (6): 1417–1426.

Nikolaou, Agelike, and Dimitrios Nikolaou. 2008. “Income-Related Inequality in the Distribution of Obesity among Europeans.” *Journal of Public Health* 16 (6): 403–11.

Novo, Mehmed, Anne Hammarström, and Urban Janlert. 2000. “Smoking Habits-a Question of Trend or Unemployment? A Comparison of Young Men and Women between Boom and Recession.” *Public Health* 114 (6): 460–63.

O'Donnell, Owen, Eddy Van Doorslaer, Adam Wagstaff, and Magnus Lindelo. 2008. "Analyzing Health Equity Using Household Survey Data: A Guide to Techniques and Their Implementation." *The World Bank.*

Oaxaca, Ronald. 1973. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review* 14 (3): 693–709.

Okechukwu, Cassandra, Janine Bacic, Kai-wen Cheng, and Ralph Catalano. 2012. "Smoking among Construction Workers : The Nonlinear in Fl Uence of the Economy , Cigarette Prices , and Antismoking Sentiment." *Social Science and Medicine* 75 (8): 1379–86.

Pate, Russell R, Michael Pratt, Steven N Blair, William L Haskell, Caroline A Macera, Claude Bouchard, David Buchner, et al. 1995. "Physical Activity and Public Health. A Recommendation from the Centers for Disease Control and Prevention and the American College of Sports Medicine." *The Journal of the Amercian Medical Association* 273 (5): 402–7.

Paul, Karsten I, and Klaus Moser. 2009. "Unemployment Impairs Mental Health : Meta- Analyses." *Journal of Vocational Behavior* 74 (3): 264–282.

Persico, Nicola, Andrew Postlewaite, and Dan Silverman. 2004. "The Effect of Adolescent Experience on Labor Market Outcomes : The Case of Height." *Journal of Political Economy* 112 (5).

Pickett, Kate E, Shona Kelly, Eric Brunner, Tim Lobstein, and Richard G Wilkinson.

2005. “Wider Income Gaps, Wider Waistbands? An Ecological Study of Obesity and Income Inequality.” *Journal of Epidemiology and Community Health* 59 (8): 670–74.

Powell, Lisa M, Roy Wada, Ramona C Krauss, and Youfa Wang. 2012. "Ethnic disparities in adolescent body mass index in the United States: The role of parental socioeconomic status and economic contextual factors". *Social Sciences and Medicine* 75(3): 469–476

Prochaska, James O, Colleen A Redding, and Kerry E Evers. 2013. Transtheoretical Model of Behavior Change. *Encyclopedia of Behavioral Medicine*.

Public Health Agency of Canada. 2011. *Obesity in Canada: A Joint Report from the Public Health Agency of Canada and the Canadian Institute for Health Information.* Public Health Agency of Canada and Canadian Institute for Health Information.

Puetz, Timothy W, Patrick J O’Connor, and Rod K Dishman. 2006. “Effects of Chronic Exercise on Feelings of Energy and Fatigue: A Quantitative Synthesis.” *Psychological Bulletin* 132 (6): 866–76.

Raine, Kim D. 2004. “Overweight and Obesity in Canada: A Population Health Perspective.” *Canadian Population Health Initiative* and *Canadian Institute for Health Information.*

Rashad, Ahmed, and Mesbah Sharaf. 2016. “Regional Inequalities in Child

Malnutrition in Egypt , Jordan , and Yemen : A Blinder-Oaxaca Decomposition Analysis." *University of Alberta Working Paper No. 2016–03*

Rashad, Inas. 2003. "Assessing the Underlying Economic Causes and Consequences of Obesity." *Gender Issues* 21 (3): 17–29.

Rassen, Jeremy A, M Alan Brookhart, Robert J Glynn, Murray A Mittleman, and Sebastian Schneeweiss. 2009. "Instrumental Variables 1: Instrumental Variables Exploit Natural Variation in Nonexperimental Data to Estimate Causal Relationships." *Journal of Clinical Epidemiology* 62 (12): 1226–32.

Canadian Populatin Health Initiative. 2008. *Reducing Gaps in Health: A Focus on Socio-Economic Status in Urban Canada.* Canadian Population Health Initiative.

Rees, Daniel I, and Joseph J Sabia. 2010. "Economics of Education Review Sports Participation and Academic Performance : Evidence from the National Longitudinal Study of Adolescent Health." *Economics of Education Review* 29 (5): 751–59.

Rehm, Jürgen, D Baliunas, S Brochu, Benedikt Fischer, William Gnam, J Patra, S Popova, and B Taylor. 2006. "The Costs of Substance Abuse in Canada 2002: Highlights." Ottawa, ON: The Canadian Centre on Substance Abuse.

Reimers, Cordelia W. 1983. "Labor Market Discrimination Against Hispanic and Black Men." *The Review of Economics and Statistics* 65 (4): 570–579.

Roodman, David. 2009. “How to Do xtabond2: An Introduction to Difference and System GMM in Stata.” *The Stata Journal* 9 (1): 86–136.

Rooth, Dan-Olof. 2011. “Work Out or Out of Work--The Labor Market Return to Physical Fitness and Leisure Sports Activities.” *Labour Economics* 18 (3): 399–409.

Ruhm, Christopher J. 1995. “Economic Conditions and Alcohol Problems.” *Journal of Health Economics* 14 (5): 583–603.

Ruhm, Christopher J. 2005. “Healthy Living in Hard Times.” *Journal of Health Economics* 24 (2): 341–63.

Schunck, Reinhard, and Benedikt G Rogge. 2010. “Unemployment and Its Association with Health-Relevant Actions: Investigating the Role of Time Perspective with German Census Data.” *International Journal of Public Health* 55 (4): 271–78.

Schunck, Reinhard, and Benedikt G Rogge. 2012. “Unemployment and Smoking: Causation, Selection, or Common Cause? Evidence from Longitudinal Data.” *SOEP papers on Multidisciplinary Panel Data Research No. 491*.

Sen, Bisakha. 2014. "Using the Oaxaca-Blinder Decomposition as an Empirical Tool to Anlayize Racial Disparities in Obesity". *Obesity* 22: 1750–1755

Spence, Michael. 1973. “Job Market Signaling.” *The Quarterly Journal of Economics* 87 (3): 355–74.

Staiger, B Y Douglas, and James H Stock. 1997. “Instrumental Variables Regression with Weak Instruments.” *Econometrica* 65 (3): 557–86.

Starky, Sheena. 2005. “The Obesity Epidemic in Canada.” *Parliamentary Information and Research Service, Library of Parliament Canada.*

Statistics Canada. 1999. *Low Income After Tax (LICO-IAT 1992 Base and LIM-IAT) (13-592-X).* http://www.statcan.ca/bsolc/english/bsolc?catno=13-592-X.

Statistics Canada. 2014. *CANSIM-Table 282-0002-Labour Force Survey Estimates by Sex and Detailed Age Groups*. http://www5.statcan.gc.ca/cansim/a47.

Statistics Canada. 2015. *Low Income Cut-Offs (1992 Base) after Tax.* http://www.statcan.gc.ca/pub/75f0002m/2013002/tbl/tbl01-eng.htm.

Stevenson, Betsey. 2010. “Beyond the Classroom : Using Title IX to Measure the Return to High School Sports.” *NBER Working Paper No. 15728.*

Terza, Joseph V, Anirban Basu, and Paul J Rathouz. 2008. “Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling.” *Journal of Health Economics* 27:531–543.

Thogerson-Ntoumani, Cecilie, Kenneth R Fox, and Nikos Ntoumanis. 2005. "Relationships between exercise and three components of mental well-being in corporate employees". *Psychology of Sport and Exercise* 6: 609–627

Tjepkema, Michael. 2006. “Adult Obesity.” *Health Reports* 17 (3).

Volland, Benjamin. 2012. "The Effects of Income Inequality on BMI and Obesity: Evidence from the BRFSS." *Papers on Economics and Evolution No. 1210.*

Warburton, Darren E R, Crystal Whitney Nicol, and Shannon S D Bredin. 2006. "Health Benefits of Physical Activity: The Evidence." *Canadian Medical Association Journal* 174 (6): 801–9.

Wasserman, Jeffrey, Willard G Manning, Joseph P Newhouse, and John D Winkler. 1991. "The Effects of Excise Taxes and Regulations on Cigarette Smoking." *Journal of Health Economics* 10 (1): 43–64.

Wellman, Nancy S, and Barbara Friedberg. 2002. "Causes and Consequences of Adult Obesity : Health , Social and Economic Impacts in the United States" *Asia Pacific Journal of Clinical Nutrition* 11: 705–709.

Windmeijer, Frank. 2005. "A Finite Sample Correction for the Variance of Linear Efficient Two-Step GMM Estimators." *Journal of Econometrics* 126 (1): 25–51.

Wooldridge, Jeffry M. 2010. "Econometric Analysis of Cross Section and Panel Data." Cambridge, MA: MIT Press.

Wooldridge, Jeffrey M. 2002. "Econometric Analysis of Cross Section and Panel Data." Cambridge, MA: MIT Press.

World Health Organization. 2009. *Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks.* Geneva: WHO press.

Zhang, Qi, and Youfa Wang. 2004. "Socioeconomic Inequality of Obesity in the United States: Do Gender, Age, and Ethnicity Matter?" *Social Science and Medicine* 58 (6): 1171–80.

Zhang, Qi., Youfa Wang. 2007. "Using Concentration Index to Study Changes in Socio-Economic Inequality of Overweight among US Adolescents between 1971 and 2002." *International Journal of Epidemiology* 36 (4): 916–25.

Zhang, Qi, Youfa Wang. 2015. "Secular Trends in Socioeconomic Inequality of Obesity in the United States." *In Studies on Economic Well-Being: Essays in the Honor of John P. Formby* 12 (4): 481–99.

5. Appendix A

For the binary dependent variable of smoking status we have the following probability mass function:

$$Pr(Y = y) = \begin{cases} \pi, & y = 0 \\ 1 - \pi, & y = 1, 2, 3, \ldots \end{cases}$$

The zero truncated count model process has the following probability mass function:

$$\Pr(Y = y | Y \neq 0) = \begin{cases} \dfrac{\lambda^y}{(e^\lambda - 1)y!}, & y = 1, 2, 3, \ldots \\ 0, & \text{Otherwise} \end{cases}$$

The unconditional probability mass function for Y is:

$$\Pr(Y = y) = \begin{cases} \pi, & y = 0 \\ (1 - \pi)\dfrac{\lambda^y}{(e^\lambda - 1)y!}, & y = 1, 2, 3, \ldots \end{cases}$$

Assuming that the observations are independently and identically distributed, the log likelihood for the t^{th} observation is:

$$\ln L\,(\pi_i, y_i, \lambda_i) = \begin{cases} \ln(\pi_i), & y = 0 \\ \ln\left\{(1 - \pi_i)\left(\dfrac{\lambda_i^{y_i}}{(e^{\lambda_i} - 1)y_i!}\right)\right\}, & y = 1, 2, 3, \ldots \end{cases}$$

where:

$\pi_i = e^{-e^{x_i\beta_1}}$

and

$\lambda_i = e^{x_i\beta_2}$

where β_1 and β_2 are the parameters of the binary choice model and zero truncated count model respectively.

The log likelihood function of our two-part model can be written as:

$$\ln L = \ln\left\{\prod_{i\in\Omega_0}\left(e^{-e^{x_i\beta_1}}\right)\prod_{i\in\Omega_1}\left(1-e^{-e^{x_i\beta_1}}\right)\prod_{i\in\Omega_1}\left(\frac{e^{y_i x_i\beta_2}}{(e^{e^{y_i\beta_2}}-1)y_i!}\right)\right\} =$$

$$\sum_{i\in\Omega_0}\{-e^{x_i\beta_1}+\sum_{i\in\Omega_1}\ln(1-e^{-e^{x_i\beta_1}})\}+\{\sum_{i\in\Omega_1}y_i x_i\beta_2-\sum_{i\in\Omega_1}\ln(e^{e^{x_i\beta_2}}-1)-\sum_{i\in\Omega_1}\ln(y_i!)\}=\ln\{L_1(\beta_1)\}+\ln\{L_2(\beta_2)\}$$

where $\ln\{L_1(\beta_1)\}$ is the log likelihood of the binary outcome model and $\ln\{L_2(\beta_2)\}$ is the log likelihood of the zero-truncated count model. The above log likelihood function is the sum of log likelihoods of the binary choice model and the truncated at zero count mode, thus can be fitted in two steps (McDowell 2003).

Table A 1 Annual Provincial Unemployment Rates in year 2012 by Age Groups

	Age								
Province	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64
Newfoundland and Labrador	0.17	0.13	0.103	0.106	0.092	0.102	0.108	0.131	0.167
Prince Edward Island	0.169	0.086	0.1	0.086	0.071	0.087	0.105	0.106	0.153
Nova Scotia	0.166	0.119	0.074	0.069	0.076	0.061	0.052	0.058	0.098
New Brunswick	0.144	0.107	0.081	0.073	0.082	0.081	0.093	0.094	0.118
Quebec	0.103	0.078	0.071	0.063	0.06	0.066	0.062	0.066	0.081
Ontario	0.13	0.079	0.071	0.059	0.058	0.055	0.06	0.062	0.061
Manitoba	0.072	0.052	0.042	0.044	0.037	0.037	0.039	0.036	0.046
Saskatchewan	0.071	0.056	0.038	0.045	0.025	0.035	0.032	0.033	0.04
Alberta	0.064	0.042	0.034	0.037	0.037	0.039	0.036	0.042	0.041
British Columbia	0.099	0.073	0.061	0.047	0.062	0.048	0.048	0.058	0.061

Source: Statistics Canada. Table 282-0002 - Labour force survey estimates (LFS), by sex and detailed age group, annual

Table A 2 Summary Statistics by Subsample and Income Group

	Male				Female			
	High-Income		Low-Income		High-Income		Low-Income	
Variable	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
Age	48.70	15.67	54.39	19.72	48.26	15.22	58.42	20.09
Canadian Born	0.89	0.32	0.77	0.42	0.89	0.31	0.81	0.40
No High School	0.03	0.16	0.22	0.42	0.01	0.11	0.29	0.45
High School Grad	0.09	0.29	0.17	0.38	0.07	0.26	0.17	0.37
Some College	0.02	0.15	0.05	0.22	0.02	0.14	0.05	0.22
College Grad	0.86	0.35	0.55	0.50	0.90	0.30	0.49	0.50
Kids	0.17	0.38	0.13	0.34	0.19	0.39	0.16	0.37
HH Size	2.34	1.05	2.02	1.17	2.40	1.04	1.84	1.15
HH Income	112308	18770	25776	14344	111885	19148	23505	13499
Owns Home	0.91	0.29	0.55	0.50	0.93	0.25	0.53	0.50
Avg Daily Fruit/Veg	4.46	2.43	4.00	2.49	5.44	2.61	4.70	2.51
Avg Daily Energy	2.65	2.48	2.06	2.40	2.48	2.19	1.66	1.86
Avg Daily Cigarettes	2.55	6.79	4.75	9.35	1.63	4.89	2.90	6.76
Avg Daily Alcohol	0.61	1.12	0.39	1.07	0.28	0.61	0.14	0.54
Single	0.20	0.40	0.34	0.47	0.15	0.36	0.22	0.42
Widow/Div/Sep	0.10	0.30	0.22	0.42	0.11	0.32	0.46	0.50
Married	0.70	0.46	0.44	0.50	0.73	0.44	0.32	0.47
Employed	0.81	0.39	0.36	0.48	0.73	0.44	0.28	0.45
BMI	27.48	4.44	26.55	4.97	25.83	5.34	26.52	6.02
Obese	0.23	0.42	0.20	0.40	0.19	0.39	0.24	0.42
Observations	7155		4313		7028		7654	

Table A 3 OLS Estimates by Income Group for the Male Subsample

	High-Income		Low-Income	
Variable	Mean	Std. Err.	Mean	Std. Err.
Age	0.03***	0.01	0.02***	0.01
Canadian Born	1.13***	0.26	1.12***	0.26
High School Graduate	−0.82	0.58	0.22	0.39
Some College	−0.97	0.66	−0.05	0.52
College Graduate	−0.76	0.48	0.01	0.31
Kids	0.36	0.25	0.63*	0.34
Owns Home	0.55**	0.26	0.15	0.25
Avg Daily Fruit/Veg	−0.11***	0.04	−0.02	0.04
Avg Daily Energy	−0.16***	0.04	−0.07	0.05
Avg Daily Cigarettes	−0.02	0.01	−0.01	0.01
Avg Daily Alcohol	−0.08	0.09	−0.14	0.09
Single	−0.61**	0.28	−0.69**	0.31
Widow/Div/Sep	−0.35	0.29	0.43	0.39
Employed	0.51*	0.27	0.10	0.27
Constant	25.35***	0.84	23.48***	0.84
R-squared	0.07		0.07	
Observations	7155		4313	

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A 4 OLS Estimates by Income Group for the Female Subsample

	High-Income		Low-Income	
Variable	Mean	Std. Err.	Mean	Std. Err.
Age	0.05***	0.01	0.04***	0.01
Canadian Born	0.64	0.41	1.29***	0.32
High School Graduate	0.45	1.15	0.14	0.47
Some College	0.53	1.29	0.35	0.62
College Graduate	−0.24	1.07	−0.51	0.33
Kids	−0.52*	0.28	0.78**	0.39
Owns Home	−0.07	0.39	−0.50*	0.28
Avg Daily Fruit/Veg	−0.06	0.05	−0.08	0.05
Avg Daily Energy	−0.37***	0.04	−0.25***	0.06
Avg Daily Cigarettes	0.01	0.02	−0.02	0.02
Avg Daily Alcohol	−0.32*	0.17	−0.59***	0.13
Single	−0.52	0.32	−0.67*	0.38
Widow/Div/Sep	−0.31	0.38	−0.25	0.36
Employed	0.35	0.25	−0.25	0.33
Constant	23.01***	1.26	23.83***	0.89
R-squared	0.08		0.07	
Observations	7028		7654	

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$