

2013•2014
FACULTY OF SCIENCES
Master of Statistics

Master's thesis
Human biomonitoring: structural equation models

Promotor :
Mevrouw Liesbeth BRUCKERS

Lazaro Mwandigha
Thesis presented in fulfillment of the requirements for the degree of Master of Statistics

Transnational University Limburg is a unique collaboration of two universities in two countries:
the University of Hasselt and Maastricht University.



Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE-3500 Hasselt
Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE-3590 Diepenbeek



2013•2014
FACULTY OF SCIENCES
Master of Statistics

Master's thesis

Human biomonitoring: structural equation models

Promotor :
Mevrouw Liesbeth BRUCKERS

Lazaro Mwandigha

*Thesis presented in fulfillment of the requirements for the degree of Master of
Statistics*

Contents

1	Introduction	1
1.1	Statements of objectives	4
2	Data description	5
3	Methodology	7
3.1	Exploratory data analysis	7
3.2	Testing for mediation or indirect effect	7
3.2.1	Use of regression models	7
3.2.2	Use of structural equation models (SEM)	11
3.2.3	Path analysis	12
3.3	Parameter estimation	13
3.3.1	Choice of parameter estimator	13
3.3.2	Maximum likelihood	14
3.4	Model estimation	14
3.5	Goodness of fit	15
3.5.1	Descriptive goodness of fit measure	15
3.6	Variable selection	17
3.7	Software used	17
4	Results	18
4.1	Exploratory data analysis	18
4.2	Inferential statistics	20
4.2.1	Outcome SHBG	20
4.2.2	Outcome FT3	21
4.2.3	Outcome FT4	23
4.2.4	Outcome TSH	24
5	Discussion and conclusion	26
6	Limitations and recommendations	28
7	Acknowledgement	29
8	Appendix	
9	Software codes	

1 Introduction

According to the world health report commissioned by the World Health Organisation (WHO) in 2004, environmental risk factors contribute to disease burden in 85 of the 102 major disease groupings. The specific fraction of disease attributable to the environment vary widely across different disease conditions. Globally, an estimated 24% of the disease burden (healthy life years lost) and an estimated 23% of all deaths (premature mortality) is attributable to environmental factors. Large regional differences in the environmental contribution to various disease conditions exist due to differences in environmental exposures and access to health care across different regions[1].

Environmental exposure to pollutants such as cadmium and thallium has been the subject of many studies. **Cadmium** is a compound that is regarded as non-essential to human health. It may get into the human body through water and food contaminated with cadmium, inhaling cadmium contaminated air and smoking. Eating food or drinking water with high levels of cadmium can lead to severe irritation of the stomach, may cause vomiting and diarrhea. Breathing high doses of cadmium can irritate, damage the lungs and can cause death. However, the greatest concern is the exposure to lower doses of cadmium over a long period of time. The lower and long-term exposure to cadmium through air or through diet can cause kidney damage. Although the damage is not life-threatening, it can lead to the formation of kidney stones and affect the skeleton, which can be painful and debilitating. In such instances lung damage has also been observed. Cadmium may be detected in blood or urine. The presence of cadmium in blood and urine confirm recent acute and moderate longterm exposure respectively [2].

In the past, **thallium** was used as a therapeutic agent to treat syphilis, gonorrhoea, tuberculosis, ringworm, as a depilatory for excess hair and rodenticide. Currently, thallium is used in the manufacture of electronic components, optical lenses, semiconductor materials, alloys, gamma radiation detection equipment, imitation jewelry, artist's paints, low temperature thermometers, and green fireworks. Thallium exposure may occur at smelters in the maintenance and cleaning of ducts and flues and through contamination of cocaine, heroin, and herbal products. Thallium may also enter the human body through food and water contaminated with thallium, smoking, living near hazardous waste sites containing thallium, touching or, for children, eating soil contaminated with thallium and through inhalation. Criminal and unintentional thallium poisonings are reported in different parts of the world, some leading to death. The presence of thallium may be detected in blood or urine [3][4].

Different exposures affect different organs and glands in the human body, cadmium and thallium have been reported to affect the thyroid. The thyroid is a gland which is integral to regulation of development and metabolism and thus the impact of these metallic exposures may be assessed by the examination of the state and functionality of the thyroid [5]. In order to assess the status of the thyroid, typically levels of **Thyroid Stimulating Hormone (TSH)** is measured. The TSH is a measure of how much the hypothalamic-pituitary system is attempting to stimulate the thyroid gland. A "normal" TSH is indicative of a thyroid gland that is not failing or has not already failed.

On the other hand, in order to assess the functionality of the thyroid gland, the levels of **Free**

Triiodothyronine (FT3) and **Free Thyroxine (FT4)** are checked. These are hormones that are produced by the throid gland. Normal levels of TSH coupled with normal levels of FT3 and FT4 are indicative of active throid that is working at optimal levels and is free of disease(s) or conditions such as inflammation of the thyroid, goitre, graves disease, hashimoto thyroiditis among others [6].

Table 1 shows a summary of the different clinical conditions associated with different levels of TSH, FT3 and FT4. Normal serum TSH levels range between 0.5 and 4.0 $\mu\text{U/ml}$. Values of 0.1-0.45 $\mu\text{U/ml}$ and 4.5-10 $\mu\text{U/ml}$ although abnormal, have been associated with normal levels of thyroid hormones. FT3 and FT4 normal reference values are 2.3-4.2 pg/ml and 0.9-2.4 ng/dL respectively [7].

TSH	FT4	FT3	Interpretation
Normal	Normal	Normal	Normal thyroid
High	Normal	Normal	Mild (subclinical) hypothyroidism (underactive thyroid)
High	Low	Low or Normal	Hypothyroidism (underactive thyroid)
Low	Normal	Normal	Mild (subclinical) hyperthyroidism (overactive thyroid)
Low	High or Normal	High or Normal	Hyperthyroidism (overactive thyroid)
Low	Low or Normal	Low or Normal	Non-thyroidal illness; rare pituitary (secondary) hypothyroidism

Table 1: *Summary of conditions associated with abnormal levels of TSH, FT3 and FT4*

Polycyclic aromatic hydrocarbons (PAHs) are widely present in urban air pollution, cigarette smoking, food contamination or occupational exposure. The urinary **hydroxypyrene** is considered the main biological biomarker currently available to measure the exposure to PAH [8]. A limited number of studies have suggested that exposures to polycyclic aromatic hydrocarbon (PAH) metabolites such as hydroxypyrene may affect **sex hormone-binding globulin (SHBG)** [9]. SHBG is a protein produced primarily in the liver, although the testes, uterus, brain, and placenta also synthesize it. It serves as a transport carrier, shuttling estrogen and testosterone to sex hormone receptors throughout the body. SHBG also safeguards these vital hormones from degrading too rapidly and prevents their clearance from the body. It thus acts as the master regulator of sex hormone levels, maintaining the delicate balance between estrogen and testosterone critical to overall health in aging humans.

Studies have shown that SHBG steadily rises with age, in addition, some enviromental factors have been associated with elevated levels of SHBG. Elevated levels of SHBG may lead to testosterone being sequestered and thus functionally unavailable to healthy tissues. The ultimate result is gynecomastia (the development of fatty breast tissue in men), diminished libido and poor sexual performance, cognitive decline, and chronic fatigue [10].

Enviromental pollutants have been the subject of many researches which is understandable as their effects on human health is severe. The statistics documented in the preamble demonstrate this fact clearly. Interestingly, previous focus of research has taken one of two approaches. Researches have endeavoured to either determine the existence of possible effect of **covariate(s)** of interest and **enviromental exposure(s)** on **health outcome**, that is

$$Outcome = Exposure(s) + Covariate(s)$$

or the existence of possible effect of **Covariate(s)** of interest on **Environmental exposure**, that is

$$Exposure = Covariate(s)$$

The goal of this project was to determine whether the presence of exposures mediate the effect of covariates on the health outcome in the population residing in Flanders, Belgium. The concept of mediation or indirect effect is diagrammatically described in Figure 1 and Figure 2. Figure 1 shows the effect of a covariate on an outcome. This effect is called **total effect** and is labelled as path **c**.

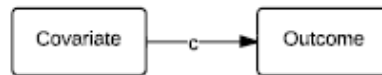


Figure 1: *Total effect of covariate on outcome*

Figure 2 shows a number of associations, first the effect of the covariate on the outcome that is not mediated by an exposure is shown, this is called **direct effect (also called partial effect)** and is signified by the path **c'**. There is also the effect of the covariate on exposure defined by the path **a** and the effect of exposure on the outcome defined by the path **b**. The role of the exposure variable is that of the mediator. The idea of mediation analysis will involve the assessment of the joint path **a * b**. The path **a * b** is also known as the **indirect effect** of the covariate on the outcome. The basic premise of mediation analysis is that the total effect can be decomposed into direct and indirect effect, that is.

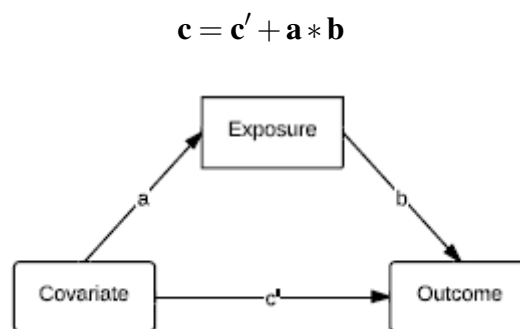


Figure 2: *Representation of mediating (indirect) effect and direct effect*

To assess the mediating effect of different exposures or indirect effect of selected covariate(s) on the outcome of interest, data from a study conducted by **vision on technology (VITO)** in the years 2008-2011 in Flanders, Belgium was analysed. The data was collected from 8 regions in Flanders with different environmental characteristics. The participants were recruited by a stratified clustered multi-stage design and belonged to 3 different age groups namely newborns and their mothers, 14-15 year adolescents and 50-65 years adults. In the study, exposures to various pollutants such as cadmium, lead, thallium, benzene among others were of interest. For this project, a subset of the data collected from adolescents aged between 14-15 years was analysed. Potentially useful covariates were pre-selected and the exposures considered were thallium, cadmium and hydroxy pyrene.

1.1 Statements of objectives

Determine the indirect effect of covariates of interest on human health using the four biomarkers TSH, SHBG, FT3 and FT4 as outcome variables.

2 Data description

The data provided consisted of 4 outcome with 606 observations. Table 2 shows information regarding the 4 outcomes while table 3 shows the information regarding variables associated with outcomes TSH, FT3 and FT4. Table 4 shows information regarding variables associated with outcome SHBG whereas table 5 gives information regarding the possible exposures associated with each outcome.

Variable	Description	Units of Measurements
TSH	Thyroid Stimulating Hormone	μ U/ml
FT3	Free Triiodothyronine	pg/ml
FT4	Free Thyroxine	ng/dl
SHBG	Sex Hormone Binding Globulin	nmol/l

Table 2: *Description of the four outcome variables*

Variable	Description	Levels
Persnr	Unique number identifying each adolescent	-
Par-edul	Highest education level of adolescent's parent	1= Lower secondary, 2= Higher secondary, 3= Tertiary
Adol-edul	Highest education level of the adolescent	1=ASO ,2=TSO ,3=BSO
Season	Season of year data was collected	1= Winter ,2= Spring ,3=Summer, 4= Autumn
Age	Age of the adolescent	Continuous
Gender	Sex of the adolescent	1=Male, 2=Female.
BMI	Adolescent's Body Mass Index	Continuous
Adol-brfd	Adolescent breastfed as a baby?	0=No, 1=Yes
Smoking	Adolescent smoking status	0=Never smoked, 1= Ex-smoker, 2= Daily

Table 3: *Data description of all the variables associated with TSH, FT3 and FT4 outcomes*

Variable	Description	Levels
Persnr	Unique number identifying each adolescent	-
Par-edul	Highest education level of adolescent's parent	1=Lower secondary, 2=Higher secondary, 3=Tertiary
Adol-edul	Highest education level of the adolescent	1=ASO ,2=TSO ,3=BSO
Season	Season of year data was collected	1=Winter, 2= Spring, 3=Summer, 4=Autumn
Adol-brfd	Adolescent breastfed as a baby?	0=No, 1=Yes
Sober	Adolescent sober when blood sample taken?	0=No , 1=Yes
Age	Age of the adolescent	Continous
BMI	Adolescent's Body Mass Index	Continuous
Smoking	Adolescent smoking status	0= Never smoked, 1= Ex-smoker, 2=Daily

Table 4: *Data description of all variables associated with SHBG outcome*

Outcome	Exposure	Description
TSH	BTL	Thallium in blood
	UTL-CRT	Thallium in urine
	HPYR-CRT	Hydroxy pyrene
FT3	BCD	Cadmium in blood
	UCD-CRT	Cadmium in urine
	UTL-CRT	Thallium in urine
FT4	UCD-CRT	Cadmium in urine
	HPYR-CRT	Hydroxy pyrene
SHBG	HPYR-CRT	Hydroxy pyrene

Table 5: *Data description of the possible doses associated with TSH, FT3, FT4 and SHBG outcomes*

3 Methodology

3.1 Exploratory data analysis

In order to obtain insight into the data, exploratory data analysis was conducted by use of tables and summary statistics.

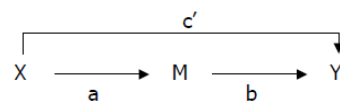
3.2 Testing for mediation or indirect effect

3.2.1 Use of regression models

Baron and Kenny approach

Baron and Kenny proposed the use of regression models to conduct mediation analysis. The regression models are fitted in 4 steps, in each step the significance of the coefficients are examined. To illustrate the steps, take 3 variables X, Y and M. Where X is the covariate of interest, Y is the outcome and M is the potential mediating variable. The steps taken under the Baron and Kenny approach are summarised in Figure 3. The purpose of Steps 1 to 3 is to establish that relationships among the variables exist. If one or more of these relationships are nonsignificant, one may usually conclude that mediation is not possible or likely. Assuming there are significant relationships from Steps 1 through 3, one proceeds to Step 4. In the Step 4 model, some form of mediation is supported if the effect of M (path b) remains significant after controlling for X. If X is no longer significant when M is controlled, the finding supports **full mediation**. If X is still significant (i.e., both X and M both significantly predict Y), the finding supports **partial mediation** [11].

The Baron and Kenny’s approach although popular often fails in testing for mediation for two reasons; first, the significance of the indirect path (i.e. that X affects Y through the compound pathway of **a** and **b**) is not tested.



	Analysis	Visual Depiction
Step 1	Conduct a simple regression analysis with X predicting Y to test for path c alone, $Y = B_0 + B_1X + e$	
Step 2	Conduct a simple regression analysis with X predicting M to test for path a, $M = B_0 + B_1X + e$.	
Step 3	Conduct a simple regression analysis with M predicting Y to test the significance of path b alone, $Y = B_0 + B_1M + e$.	
Step 4	Conduct a multiple regression analysis with X and M predicting Y, $Y = B_0 + B_1X + B_2M + e$	

Figure 3: Baron and Kenny’s approach to mediation analysis¹

¹Source:Newsom1USP 655 SEMWinter 2012

Second, mediation effect may be missed due to the insistence of significance of path **c** in Step 1 before mediation analysis is conducted. In more recent treatments of mediation analysis, it has been pointed out that in situations where one of the path coefficients is negative, there can be significant mediated effects even when there is no significant association between X and Y. For example, if one of the paths in the mediation model is negative, a form of **suppression** may occur such that positive direct and negative indirect effects tend to cancel each other out to yield a small and nonsignificant total effect. (If **a** is negative, while **b** and **c'** are positive, then when we combine a negative $a * b$ product with a positive **c'** coefficient to reconstitute the total effect **c**, the total effect **c** can be quite small even if the separate positive direct path and negative indirect paths are quite large). In cases where the mediator acts as a supressor variable, **inconsistent mediation** is said to occur. For these reasons, the Baron and Kenny approach has low statistical power [14].

Judd & Kenny difference of coefficients approach

Fortunately, there are alternative methods to Baron and Kenny's method of conducting mediation analysis, these methods do not suffer from the drawback of the Baron and Kenny approach as they involve the testing for the indirect effect. One of those methods was proposed by Judd and Kenny. Unlike the 4 regression model steps of Baron and Kenny, only 2 regression models are needed. The method is called the Judd and Kenny difference of coefficients approach, so called because it involves getting the difference of regression coefficients from 2 regression models.

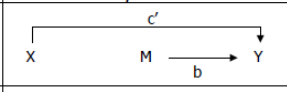
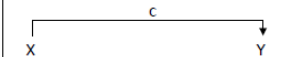
	<i>Analysis</i>	<i>Visual Depiction</i>
<i>Model 1</i>	$Y = B_0 + B_1X + B_2M + e$	
<i>Model 2</i>	$Y = B_0 + BX + e$	

Figure 4: *Judd & Kenny difference of coefficients approach*²

Figure 4 shows the 2 regression models fitted under this approach. The approach involves subtracting the partial regression coefficient, B_1 , obtained in Model 1, from the simple regression coefficient, B , obtained from Model 2. Note that both represent the effect of X on Y but that B is the coefficient from the simple regression of Y on X while B_1 is the partial regression coefficient from a multiple regression. The indirect effect is the difference between these two coefficients[11].

$$B_{indirect} = B - B_1 = c - c'$$

The next stage is the testing of the hypothesis $H_0 : c - c' = 0$. There are a number of ways of computing the standard error to be used in assessing significance of $c - c'$. One popular method derived by Freedman and Schatzkin is expressed as

$$SE_{c-c'} = \sqrt{s_c^2 + s_{c'}^2 - 2s_{cc'}\sqrt{1 - \rho_{XM}}}$$

where

- s_c^2 and $s_{c'}^2$ are the standard errors of raw regression coefficients of the effect of X on Y and X on Y while correcting for M, respectively.

²Source:Newsom1USP 655 SEMWinter 2012

- $s_{cc'}$ is the covariance between coefficient of path c and c'
- ρ_{XM} is the correlation coefficient between X and M

The Judd & Kenny difference of coefficients approach, $c - c'$, is easy to implement in simple mediation models. One of its major drawback is that it's not easily generalized to situations where there are multiple mediators between a covariate, X , and outcome, Y . [12][13]. As such, this approach is not widely used.

Sobel product of coefficients approach

This method is equivalent to the Judd & Kenny difference coefficient approach as it leads to the same conclusion as that of Judd & Kenny difference of coefficient approach when testing for mediation in simple models. Infact, for a model with simple model with a single X, Y and M the 2 approaches yield similar value of the indirect effect. This can be illustrated in the following equations. In the introduction section, It was established that

$$c = c' + a * b$$

which may be algebraically expressed as

$$c - c' = a * b$$

The term of the left side of the equation is the basis of the Judd & Kenny difference coefficient approach while the the term of the right is the basis of the Sobel product of coefficients approach. Both approaches involve fitting of two regression models. However, there is a slight difference between the 2 methods. Although the 2 methods involve fitting the same multiple regression model, the simple regression model fitted in the Sobel's product coefficient approach is different as it involves fitting the regression model of M on X rather than the regression model of Y on X .

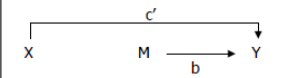
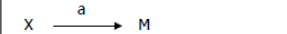
	<i>Analysis</i>	<i>Visual Depiction</i>
<i>Model 1</i>	$Y = B_0 + B_1X + B_2M + e$	
<i>Model 2</i>	$M = B_0 + BX + e$	

Figure 5: *Sobel product of coefficients approach*³

Figure 5 shows the 2 regression models fitted under this approach. After the fitting of the models, the product of coefficients is formed by multiplying two coefficients together, the partial regression effect, B_2 , for M predicting Y and the simple coefficient, B , for X predicting M [11].

$$B_{indirect} = B_2 * B = b * a$$

Thereafter, the significance of the product of coefficients becomes the most important task. This involves testing of the hypotheses

$$H_0 : b * a = 0$$

³Source:Newsom1USP 655 SEMWinter 2012

To set up a Z test statistic, an estimate of the standard error of this ab product, SE_{ab} , is needed. Sobel provided the following approximate estimate for the SE_{ab} :

$$SE_{ab} \approx \sqrt{b^2 s_a^2 + a^2 s_b^2}$$

where

- a and b are the (unstandardized) raw regression coefficients that represent the effect of X on M and that of M on Y corrected for X respectively.
- s_a^2 and s_b^2 are the standard errors for the raw regression coefficient a and b respectively.

The obtained estimate of SE_{ab} can then be used in the computation of $Z = a * b / SE_{ab}$. The Z statistic can then be used in testing for $H_0 : \mathbf{a} * \mathbf{b} = \mathbf{0}$. There has been a lot of controversy, however, on how to compute SE_{ab} as the Z tests for the significance of $a * b$ assume that values of the $a * b$ product are normally distributed across samples from the same population. Empirical data suggests that this assumption is incorrect for many values of a and b . Because of this, it is recommended that bootstrapping methods be used to obtain confidence intervals for estimates of $a * b$ [14].

Bootstrapping has become widely used in situations where the analytic formula for the standard error of a statistic is not known and/or there are violations of assumptions of normality. Bootstrapping involves drawing a sample from the population (with replacement). The values of a , b , and $b * a$ are calculated for this sample. This process is repeated many times (1,000 times, say). Subsequently, the value of $b * a$ is tabulated across these samples; this provides an empirical sampling distribution that can be used to derive a value for the standard error of ab . Results of such bootstrapping indicate that the distribution of $b * a$ values may be asymmetrical, this asymmetry therefore has been taken into account when setting up confidence interval (CI) of the estimates of $b * a$ [14].

The obtained CI provides a basis for evaluation of the single estimate of $b * a$ obtained from analysis of the entire data set. Bootstrapped CI's do not require that the values of $b * a$ have a normal distribution across all samples. If this CI does not include zero, then the conclusion may be made that there is statistically significant mediation or indirect effect. Currently, there are a number of SAS and SPSS macros that perform mediation analysis by using the Sobel product of coefficients approach with the corresponding bootstrap CI. The macros developed by A.F. Hayes like *indirect*, *sobel* and *process* can be procured online free of charge and can be used for this purpose [15].

It should be noted that the macros, although useful, have some serious limitations, for one there are restrictions on the use of these macros with categorical mediators and unfortunately categorical mediators occur frequently in practice. Moreover, although one can incorporate many continuous mediators, this can only be done with the use of one covariate, X, and one outcome, Y, at a single time when conducting test for mediation. This presents a problem for any (aspiring) mediation analyst in cases where multiple covariates, multiple (categorical) mediators and multiple outcomes are to be considered.

The solution to the problem is the use of **Structural Equation Modeling (SEM)** which allow

the modeling of multiple associations at the same time. By using SEM, one can fit models with multiple covariates and multiple outcomes. They have the added benefit of a number of single **Goodness of fit (GOF)** statistics upon which one may assess how well the model fits the data. In addition, software packages like *SPSS* and *MPLUS* provide tests of mediation using the sobel tests and their corresponding bootstrap CI's for any specified path(s) that one may be interested in. The possibility of being able to incorporate multiple covariates, multiple outcomes and testing of several indirect effects makes SEM an attractive approach for mediation analysis.

3.2.2 Use of structural equation models (SEM)

SEM has become one of the techniques of choice for researchers across disciplines and is increasingly a 'must' for researchers in the social sciences [16]. It is a powerful tool to explore and contrast hypotheses on causal relationships among variables. It studies causal relationships assuming the existence of linear relationships, although non-linear relationships can be modelled as well. These approaches do not actually prove causality, but they help to select relevant hypotheses, discarding those that are not backed by empirical evidence. Although the presence of a correlation between two variables does not necessarily imply the existence of a causal relationship between them, the existence of a causal relationship between two variables does imply the existence of a correlation between them. This is in essence the basis for the SEM approach. Structural equation modeling assumes that there is an underlying mechanism that leads to a theoretical covariance structure between a vector of random variables [17]. SEM is a very general and convenient framework for statistical analysis that includes several traditional multivariate procedures, for example regression analysis, factor analysis, discriminant analysis, and canonical correlation as special cases [18].

SEM has its roots in path analysis, which was invented by the geneticist Sewall Wright in 1921. It is customary to start a SEM analysis by drawing a path diagram. A path diagram consists of boxes and circles, which are connected by arrows. In Wright's notation, observed (measured) variables are represented by rectangle or square box, and latent (unmeasured) factors by circle or ellipse. Single headed arrow or 'path' are used to define causal relationships in the model, with the variable at the tail of the arrow causing the variable at the point. Double headed curved arrows indicate covariances or correlations, without a causal interpretation [18]. Arrows not originating from a variable represent residual unexplained variances which are unanalysed components of the diagram, which express current ignorance of the variables that determine them [17].

In SEM, the terms endogenous and exogenous variables are frequently used to describe variables in the model that are dependent on other variables and those that are not dependent on other variable in the model, respectively. Endogenous variables will have a directed arrow entering into them (i.e., prediction) both from the substantive predictors and a residual term that represents the variance not explained by the predictors. On the other hand, an independent (exogenous) variable is a variable that has causes that are assumed to be external to the model and not influenced by any other variable in the model. Exogenous variables can only have double headed arrows (i.e., correlation) going into them.

SEM is usually not conceived in terms of directly measurable, possibly well defined, hypothetical or theoretical constructs. It usually take into account potential errors of measurement in all observed variables, whether it is endogenous or exogenous variable, from which observed and unobserved(latent) variables come. The SEM framework consists of two types of models, **Path analysis**, which is a multivariate regression model, and **Measurement model**, which describes the relationship between set of observed variables and a set of unobserved (latent) variables. Path analysis focuses mainly on the structural relationships between observed variables. On the other hand, Measurement model focuses on the relationship between observed and unobserved variables. For this project, all variables were observed therefore only path analysis part of SEM was utilised.

3.2.3 Path analysis

Path analysis was originally developed by geneticist Sewall Wright to examine the effect of hypothesized models in phylogenetic studies. A few decades later, Path models were adopted by sociologists to describe human behaviour [19]. Path models can be perceived as an extension of multiple linear regression , with several multiple regression models or equations that are estimated simultaneously. Path analysis is used to explain causal relationship i.e how independent variables influence dependent variables. Structural Equations explains the relationship between endogenous and exogenous variable. However a variable can serve both as endogenous and exogenous. For simplicity, we can relate exogenous and endogenous variables to independent and dependent variables respectively in linear regression analysis. The exogenous are assumed to be measured without error. Due to complexity of the SEM, causal graphs called path diagrams comprising of flow diagrams with rectangles, circles and arrow joining the various components are often used to simplify it.

Indicator (observed) variable are represented by a rectangle or square while latent (unobserved) variables that are estimated by the indicators are represented by a circle or ellipse. Paths between two variables comprise of single or double headed arrows and they seek to explain the structural influence between the variables in question. There are four types of structural influence between two variables, say X and Y [20]:

- $X \rightarrow Y$: X structurally influences Y but not vice versa
- $X \leftarrow Y$: Y structurally influences X but not vice versa
- $X \leftrightarrow Y$: Y structurally influences X and X structurally influence Y
- XY: No structural relationship between X and Y

Figure 6 below is an example of a path diagram which describes the association between three endogenous variables (Y's) and two exogenous variables (X's).

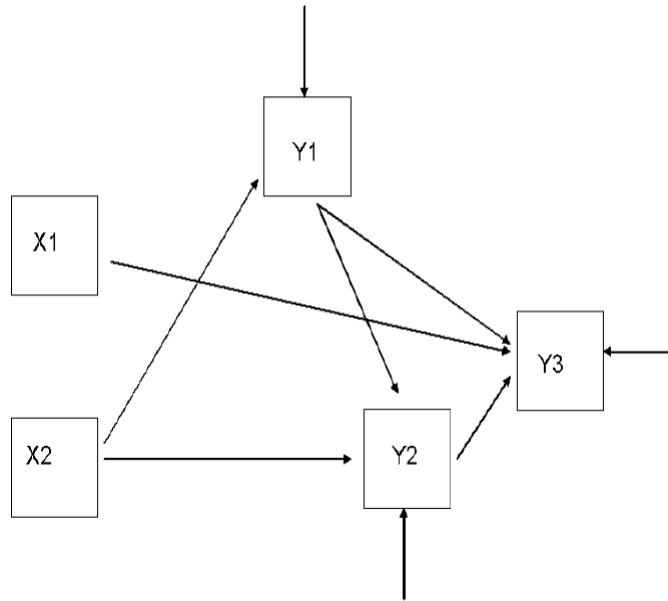


Figure 6: *Path diagram example*

We can visualize that, X_1 and X_2 are exogenous variables, Y_1 , Y_2 and Y_3 are endogenous variables. The arrows pointing to the endogenous variable represents error terms, which are not directly measured and reflect variability in the outcome. X_2 directly influences Y_1 , Y_2 and indirectly influences Y_3 through Y_2 [20][22].

3.3 Parameter estimation

3.3.1 Choice of parameter estimator

There are a number of parameter estimators available for fitting SEM. MPLUS offers 12 different types of estimators. With so much choice, one has to carefully consider what is appropriate to meet the required objective(s). The choice of the estimator depends on the type of data at hand and the scale of the endogenous variable(s). For example, **Weighted least squares (WLS)** does not require the normality assumption of the data to be satisfied. The estimates, standard errors and chi-square statistic are computed using a full weight matrix, \mathbf{W} , which is a consistent estimate of the asymptotic covariance matrix of the sample variances and covariances being analyzed. WLS typically requires large sample sizes, uses complete cases and is prone to convergence difficulty. Another type of estimator is the **Diagonally weighted least squares (DWLS)** called **WLSMV** in MPLUS. It employs the use of a diagonal matrix, \mathbf{W} , and requires that at least one of the endogenous variables be ordinal. **Maximum Likelihood (ML)** is the most widely used fitting function for structural equation models. ML relies on the assumption of multivariate normality of the data. In cases where multivariate normality assumption has been violated, several estimators are available in MPLUS such which allow for the computation of ML estimates with robust standard errors. Unfortunately, these methods are not usable together with the bootstrapping technique required when computing CI for the indirect effect. Therefore mediation analysis with these estimators is not feasible [21].

For this project, Maximum Likelihood (ML) estimator was chosen. That decision was motivated by several considerations. First, although ML has a limitation of the assumption of multivariate normality and violations of this distribution assumptions are common and often unavoidable in practice and can potentially lead to seriously misleading results, simulation studies have suggested that under conditions of severe nonnormality, ML parameter estimates are still consistent but not necessarily efficient. Second, when ML is used with bootstrapping, the distribution assumption may be relaxed and thus one can be confident that the inference made from the analysis is valid. Third, ML comes with formal statistical tests of overall model fit for overidentified models such as RMSEA and SRMR that are not available with many other estimators [24].

3.3.2 Maximum likelihood

Maximum likelihood or simply ML leads to estimates for the parameters which maximize the likelihood \mathbf{L} that the empirical covariance matrix \mathbf{S} is drawn from a population for which the model-implied covariance matrix $\Sigma(\theta)$ is valid. Maximising the log-likelihood function is equivalent to minimizing the fitting function,

$$F_{ML} = \log|\Sigma(\theta)| - \log\mathbf{S} + \text{tr}[\mathbf{S}\Sigma(\theta)^{-1}] - p$$

where

- $\Sigma(\theta)$ is the model implied covariance matrix and $|\Sigma(\theta)|$ it's determinant
- \mathbf{S} is the empirical covariance matrix
- θ is the vector of parameters
- tr is the trace of the matrix while p is the number of observed variables.

The ML estimator assumes that the variables in the model are multivariate normal, \mathbf{S} and $\Sigma(\theta)$ in the fitting function are positive definite (which implies that the matrices must be nonsingular). If the observed data stem from a multivariate normal distribution, if the model is specified correctly, and if the sample size is sufficiently large, ML provides parameter estimates and standard errors that are asymptotically unbiased, consistent, and efficient. Furthermore, with increasing sample size the distribution of the estimator approximates a normal distribution [24].

3.4 Model estimation

Due to flexibility in SEM formulation, a number of models can be conceived. However, not all models can be identified and estimated. A basic principle of identification is that a model cannot have a number of unknown parameters to be estimated than the number of unique pieces of information provided by the data [22].

When a model is **just identified** or **saturated**, it contains the same number of parameters as the number of elements in the observed covariance matrix. In this case every model parameter can be uniquely estimated. A model is said to be **over-identified** if it contains fewer parameters

to be estimated than the number of observed elements of the covariance matrix, in this case there might exist no set of parameter estimates that reproduces the observed covariance matrix exactly. Various statistical criteria, such as ML, may be used to choose parameter estimates that approximately reproduce the observed covariance matrix. If ML or WLS is estimation is used, statistical test of the goodness of fit of the model are computed. An **under-identified** model contains fewer number of observed elements of the covariance matrix than the number of parameters. In this case the model parameters can not be estimated [20].

The number of observed elements of the covariance matrix may be estimated by the following equation

$$p(p+1)/2$$

where p is the number of variables in the model. The number of parameters in the model is obtained as follows,

$$P + V + C + D$$

where

- P is the number of defined paths in the model
- V and C are the number of variances and covariances of exogeneous variables respectively in the model
- D is the number of disturbance (residual error) terms in the model

The degrees of freedom (d.f.) of the model is obtained by taking the number of elements in the covariance matrix less the number of parameters to be estimated from the model.

3.5 Goodness of fit

3.5.1 Descriptive goodness of fit measure

The coefficient of determination, R^2 , measures the goodness of fit of each equation in the model. However, since there are several equations in SEM, a summary measure that would assess goodness of fit for the model, as a whole, is necessary.

The formal statistical testing is done using the framework of traditional χ^2 goodness of fit test. In SEM, parameters are estimated by minimizing the fitting function, $F(\mathbf{S}, \Sigma(\hat{\theta}))$, which measures how close is the predicted covariance, $\Sigma(\hat{\theta})$, and the observed covariance, \mathbf{S} . The predicted covariance reflects the structure of the relationship proposed for the data, therefore, a good model would have a small value for the χ^2 statistic such that

$$\chi^2 = (p-1)F(\mathbf{S}, \Sigma(\hat{\theta})) \leq \chi_{p-1, \alpha}^2,$$

causing $H_0 : S = \Sigma(\theta)$, not to be rejected. In this case, the covariance structure fitted by the proposed model is not significantly different to the covariance structure derived from a saturated model, implying a good fit [20].

The tests mentioned above has several loopholes since they assume χ^2 -distributed test statistics. They may not be robust to assumption violations, and their values are considerably affected by

the sample size [26]. For a constant degree of freedom, smaller sample sizes lead to smaller χ^2 values, while larger sample sizes lead to larger χ^2 values [24]. Thus, large sample size would tend to reject the null hypothesis even though the difference is not important [26]. Interestingly, even model complexity could influence these test statistics. χ^2 value decreases as more parameters are added to the model such that you either do not reject the null hypothesis because of a correctly specified model or a highly overparameterized model [24]. Hence, several descriptive measures for goodness of fit such as RMSEA (Root Mean Square Error of Approximation) are recommended.

For RMSEA, is the square root of the estimated discrepancy due to approximation per degree of freedom,

$$\hat{\epsilon}_a = \sqrt{\max \left\{ \left(\frac{F(\mathbf{S}, \Sigma(\hat{\theta}))}{df} - \frac{1}{N-1} \right), 0 \right\}},$$

where

$F(\mathbf{S}, \Sigma(\hat{\theta})) =$ the minimum of the fitting function

$df =$ degree of freedom

$N =$ sample size,

the following criteria could be followed [25]

$RMSEA \leq .05$ – good fit

$.05 < RMSEA \leq .08$ – adequate fit

$.08 < RMSEA \leq .10$ – mediocre fit

$RMSEA > .10$ – not acceptable.

For RMR, is based on the square root of the mean of the fitted residuals,

$$RMR = \sqrt{\frac{\sum_{l=1}^p \sum_{j=1}^i (s_{ij} - \sigma_{ij})^2}{p(p+1)/2}},$$

where

$s_{ij} =$ an element of the empirical covariance matrix \mathbf{S}

$\sigma_{ij} =$ an element of the model – implied covariance matrix $\Sigma(\hat{\theta})$

$p =$ number of observed variables,

In principle, RMR values close to zero suggest a good fit. But as the elements of S and $\Sigma(\hat{\theta})$ are scale dependent, the RMR depends on the sizes of the variances and covariances of the observed. To overcome this problem, the Standardized Root Mean Square Residual (SRMR) which standardizes the scale of all variables prior to computing RMR is used. The rule of thumb is that the SRMR should be less than 0.05 for a good fit, whereas values smaller than 0.10 may be interpreted as acceptable [27].

3.6 Variable selection

The data provided a number of covariates of interest, in order to select variables that were to be included in modeling, **backward variable selection** was utilised for each set of outcome and associated exposure(s). Categorical variables pose a challenge when conducting variable selection as levels of the categorical variables are treated as distinct variables. This often unfortunately leads to some of the levels of the categorical variable being dropped in the process of variable selection.

One of the ways of solving this problem is to select the variable(s) for modeling based on their correlation with the outcome(s) of interest. A correlation matrix is computed between the variables and subsequently, instead of using the raw data as input data for *PROC REG* of SAS software, the computed correlation matrix is used.

Care has to be taken however when computing the correlation matrix as the nature of variables has to be taken into consideration. For ordinal variables, the popular **pearson product-moment correlation** is not advised as the resulting correlation is often artificially deflated, instead **polychoric correlations** should be computed [32]. Polychoric correlation is computed by assuming that two ordinal variables represent latent continuous normally distributed variables[31]. Polychoric correlations were computed using the SAS macro *POLYCHOR* which is available online free of charge[29]. Correlations between sets of continuous variables may be estimated by computing pearson product-moment correlation while correlations between ordinal and continuous variables may be estimated by computing **polyserial correlations**. Polyserial correlation is computed as follows; the values of the continuous variable are taken as they are whereas the ordinal variable is assumed to represent an underlying latent normally distributed variable which is used in computing the correlation[30]. Pearson product-moment and polyserial correlation were computed using *PROC CORR* of SAS software[32].

3.7 Software used

Data manipulation and exploratory data analysis was performed using SAS 9.3 and R 3.0.2. The SEM were fitted using MPLUS Version 5.0. All hypotheses were tested at $\alpha = 5\%$. The codes for data manipulation, variable selection and SEM fitted may be consulted in the Appendix.

4 Results

4.1 Exploratory data analysis

Table 6, 7, 8 and 9 summarize the missing data patterns for the 4 outcomes TSH, FT3, FT4 and SHBG respectively. It was observed that amount of missingness was highest for the data associated FT4 and TSH which had the same amount of missingness at 18.81% followed by FT3 whose associated amount of missingness stood at 18.32%. SHBG had the least amount of missingness at 8.33%.

The data associated with outcome FT4 and TSH had the largest number of missingness pattern at 11 followed by FT3 which had 9 missingness pattern whereas SHBG had the least number of missingness pattern which stood at 7.

Group	Variables												Count	%	
	TSH	Adol-edl	Season	Par-edl	Gender	Adol-brfid	BMI	Age	Smoking	BTL	UTL-CRT	HPYR-CRT			
Complete															
1	O	O	O	O	O	O	O	O	O	O	O	O	O	492	81.19
Missing															
2	O	O	O	O	O	O	O	O	O	O	O	O	M	3	0.50
3	O	O	O	O	O	O	O	O	O	O	O	M	O	69	11.39
4	O	O	O	O	O	O	O	O	O	O	O	M	M	1	0.17
5	O	O	O	O	O	O	O	O	M	O	O	O	O	4	0.66
6	O	O	O	O	O	M	O	O	O	O	O	O	O	7	1.16
7	O	O	O	M	O	O	O	O	O	O	O	O	O	2	0.33
8	O	O	O	M	O	O	O	O	O	O	O	O	O	9	1.49
9	O	O	O	M	O	O	O	O	O	O	M	O	O	1	0.17
10	O	M	O	O	O	O	O	O	O	O	O	O	O	11	1.82
11	M	O	O	O	O	O	O	O	O	O	O	O	O	7	1.16

Table 6: Missingness patterns for outcome TSH ('O' denotes observed and 'M' missing).

Group	Variables												Count	%	
	FT3	Adol-edl	Season	Par-edl	Gender	Adol-brfid	BMI	Age	Smoking	BCD	UCD-CRT	UTL-CRT			
Complete															
1	O	O	O	O	O	O	O	O	O	O	O	O	O	495	81.68
Missing															
2	O	O	O	O	O	O	O	O	O	O	M	M	M	70	11.55
3	O	O	O	O	O	O	O	O	M	O	O	O	O	4	0.66
4	O	O	O	O	O	M	O	O	O	O	O	O	O	7	1.16
5	O	O	O	O	O	M	O	O	O	O	M	M	M	2	0.33
6	O	O	O	M	O	O	O	O	O	O	O	O	O	9	1.49
7	O	O	O	M	O	O	O	O	O	O	M	M	M	1	0.17
8	O	M	O	O	O	O	O	O	O	O	O	O	O	11	1.82
9	M	O	O	O	O	O	O	O	O	O	O	O	O	7	1.16

Table 7: Missingness patterns for outcome FT3 ('O' denotes observed and 'M' missing).

Summary statistics of the 4 outcome variables, exposures and 2 continuous covariates, Age and BMI, may be consulted on table 10 . It was observed that exposure HPYR-CRT and outcome SHBG had the widest range of observed values. Frequencies of the potential covariates are displayed on table 11 . It was observed that the distribution of females and males was somewhat

evenly distributed. The variable Gender, Season and Sober had no missing values.

Group	Variables											Count	%	
	FT4	Adol-edl	Season	Par-edl	Gender	Adol-brfd	BMI	Age	Smoking	UCD-CRT	HPYR-CRT			
Complete														
1	O	O	O	O	O	O	O	O	O	O	O	O	492	81.19
Missing														
2	O	O	O	O	O	O	O	O	O	O	M	O	3	0.50
3	O	O	O	O	O	O	O	O	O	M	O	O	69	11.39
4	O	O	O	O	O	O	O	O	O	M	M	O	1	0.17
5	O	O	O	O	O	O	O	O	M	O	O	O	4	0.66
6	O	O	O	O	O	M	O	O	O	O	O	O	7	1.16
7	O	O	O	O	O	M	O	O	O	M	O	O	2	0.33
8	O	O	O	M	O	O	O	O	O	O	O	O	9	1.49
9	O	O	O	M	O	O	O	O	O	M	O	O	1	0.17
10	O	M	O	M	O	O	O	O	O	O	O	O	11	1.82
11	M	M	O	O	O	O	O	O	O	O	O	O	7	1.16

Table 8: Missingness patterns for outcome FT4 ('O' denotes observed and 'M' missing).

Group	Variables											Count	%	
	SHBG	Adol-edl	Season	Parl-edl	Adol-brfd	Sober	BMI	Age	Smoking	HPYR-CRT				
Complete														
1	O	O	O	O	O	O	O	O	O	O	O	O	297	91.67
Missing														
2	O	O	O	O	O	O	O	O	O	O	M	O	3	0.97
3	O	O	O	O	O	O	O	O	M	O	O	O	1	0.31
4	O	O	O	O	M	O	O	O	O	O	O	O	6	1.85
5	O	O	O	M	O	O	O	O	O	O	O	O	7	2.16
6	O	M	O	O	O	O	O	O	O	O	O	O	7	2.16
7	M	O	O	O	O	O	O	O	O	O	O	O	3	0.93

Table 9: Missingness patterns for outcome SHBG ('O' denotes observed and 'M' missing).

Variable	N	Min	Max	Mean	SD
TSH	599	0.350	11.920	2.363	1.120
FT3	599	2.880	5.900	4.151	0.519
FT4	599	0.770	2.020	1.242	1.167
SHBG	321	5.000	137.330	42.866	19.764
BTL	606	0.002	0.099	0.0312	0.009
UTL-CRT	533	0.050	0.748	0.178	0.083
HPYR-CRT	602	2.200	2588.20	157.658	146.387
BCD	606	0.030	22.878	0.3097	0.991
UCD-CRT	533	0.028	0.705	0.2271	0.0968
Age	606	13.624	17.029	14.92	0.665
BMI	606	15.137	34.256	20.398	3.081

Table 10: Summary statistics of outcome variables and doses

Variable	Level	Frequency	%
gender	1=Male	324	53.47
	2=Female	282	46.53
	Missing	-	-
adol-edl	1=	316	52.15
	2=	192	31.68
	3=	87	14.36
	Missing	11	1.82
adol-brfd	0=	433	71.45
	1=	164	27.06
	Missing	9	1.49
season	1=	246	40.59
	2=	232	38.28
	3=	16	2.64
	4=	112	18.48
	Missing	-	-

Variable	Level	Frequency	%
par-edl	1=	93	15.35
	2=	181	29.87
	3=	322	53.14
	Missing	10	1.65
sober	0=	551	90.92
	1=	55	9.08
	Missing	-	-
smoking	0=	558	92.08
	1=	22	3.63
	2=	22	3.63
	Missing	4	0.66

Table 11: *Frequency distribution of the potential covariates*

4.2 Inferential statistics

4.2.1 Outcome SHBG

Of particular interest, given the potential covariates, was to determine if **HPYR-CRT (hydroxy pyrene)** had any mediating effect on the outcome variable **sex hormone-binding globulin (SHBG)**. The covariates to be included in modeling were selected by backward selection procedure at $\alpha = 5\%$. First, the correlation matrix to be used as data input for *PROC REG* of SAS was computed.

Table 17 in the Appendix shows the obtained correlation of all the variables. The results of the backward selection are shown on table 18 in the Appendix. It was seen that the variables **Sober, BMI, Age, Adol-brd, Par-edl** and **HPYR-CRT** were associated with the outcome. **Smoking** was found to be associated with the exposure HPYR-CRT. Based on these findings, a path diagram was formulated and is shown on figure 7. No evidence from literature was found to suggest that parents education level might influence their children's SHBG level. It's association with the outcome SHBG may be due to the fact that it masks social economic status of the region in which the participants reside. Therefore the variable Par-edl was excluded from analysis.

The parameters to be estimated were 32 since there were 9 paths to be modelled, 6 variances and 15 covariances associated with exogenous variables and 2 endogenous variables. There were 8 observed variables which resulted in 36 elements in the covariance matrix. Since the number of parameters to be estimated was less than the number of elements in the covariance matrix, the model was said to be overidentified and therefore parameter estimation was indeed possible. The reported RMSEA and SRMR values were 0.000 and 0.005 respectively for the model indicating that it possessed a good fit to the data.

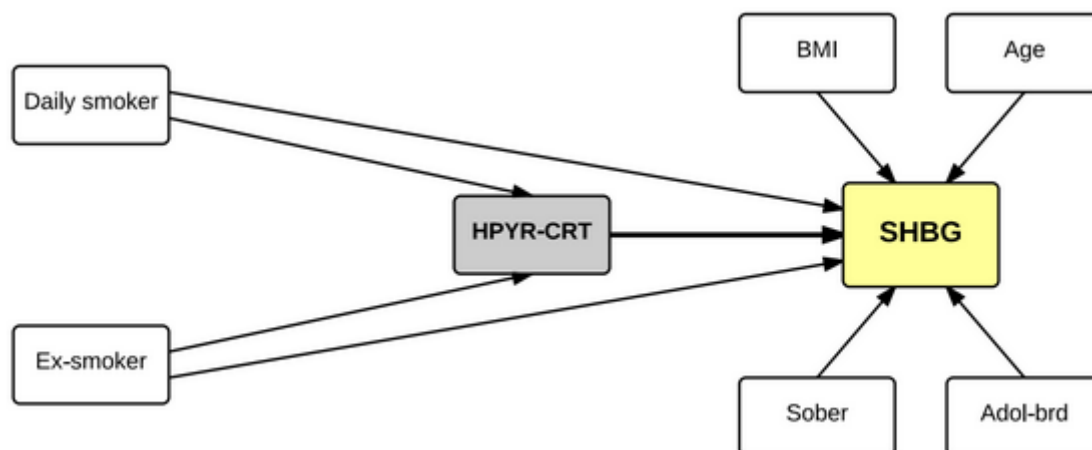


Figure 7: Proposed model for outcome SHBG

Effect: Ex smoker to SHBG		
Total effect	Indirect effect	Direct effect
0.297(-6.501, 8.615)	0.109(-0.260, 1.007)	-0.187 (-6.851, 8.776)

Table 12: Effect of covariate Ex smoker on SHBG

Effect: Daily smoker to SHBG		
Total effect	Indirect effect	Direct effect
8.592(0.701 , 16.577)	1.918 (-0.239, 4.851)	6.674 (-1.330 , 13.924)

Table 13: Effect of covariate Daily smoker on SHBG

Table 12 and 13 summarize the effect of not being a ex-smoker and daily-smoker (compared to never smoking) on outcome SHBG together with their corresponding 95% bootstrap CIs. It was observed that neither of the variables had an indirect effect on the outcome. Without controlling for the exposure, smokers have significantly higher SHBG levels than individuals who have never smoked. There was no discernable significant differences between SHBG levels of ex-smokers and those who have never smoked.

4.2.2 Outcome FT3

Given the outcome, FT3, the goal was to determine whether **BCD (cadmium in blood)**, **UCD-CRT (cadmium in urine)** and **UTL-CRT (thallium in urine)** mediated the effect of the covariates of interest. The covariates to be included in the model were selected by backward selection procedure. Like before, a correlation matrix was used as input data for *PROC REG* of SAS. Polychoric, polyserial and pearson correlations were computed appropriately for each set of variables as needed. The computed correlation matrix may be viewed on table 19 in the Appendix. Backward selection procedure conducted at $\alpha = 5\%$ revealed that **Age, BMI, Gender, Smoking, Season** and **UCD-CRT** were associated with outcome FT3. **Adol-edl, Par-edl, Gender, BMI** and **Smoking** were associated with BCD whereas **Age, BMI** and **Smoking** were found to be associated with UCD-CRT. Finally, **Age** and **Smoking** were found to be associated

with UTL-CRT. Table 20 in the Appendix contains details of the backward selection procedure. Figure 8 shows the diagrammatic representation of the proposed path model. Like in the case of outcome SHBG, the covariate Par-edl was excluded from analysis. Adol-edl was accorded the same treatment as Par-edl.

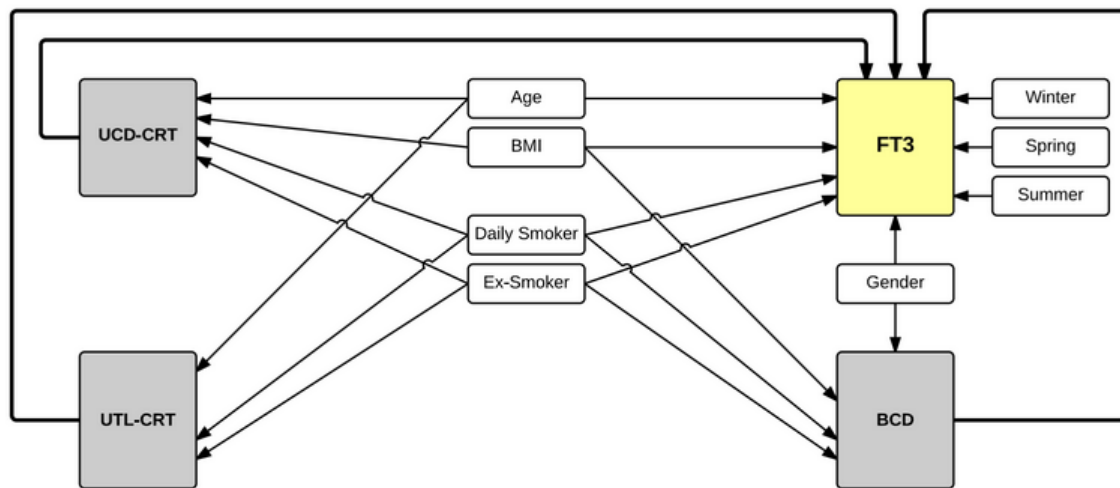


Figure 8: Proposed model for outcome FT3

The number of paths in the model was 22 with 8 variances, 28 covariances and 4 residual terms bringing the total number of parameters to 62. The model comprised of a total of 12 variables which resulted in 78 elements in the covariance matrix. Since the number of parameters to be estimated was less than 78, the model was said to be overidentified and thus parameters of the model could be estimated.

The model was fitted and yielded RMSEA and SRMR values of 0.057 and 0.031 respectively suggesting the overall fit of the proposed model to the data was adequate. The tests for mediation was conducted and the results of the estimates and corresponding 95% bootstrap CI may be consulted on table 14 . The results revealed that there was a total and direct effect of Age and Gender on FT3. There was an indirect effect of Smoking on FT3 through the UTL-CRT (thallium in urine). This suggests that, compared to those adolescents who have never smoked, adolescent smokers have 0.015 pg/ml higher levels of FT3 due to the mediation effect of thallium.

Effect	Total effect	Direct effect	Specific indirect path	Indirect effect
BMI to FT3	0.000 (-0.013, 0.010)	0.001 (-0.012, 0.012)	Through BCD Through UCD-CRT	0.000 (0.000, 0.001) -0.001(-0.004, 0.000)
Age to FT3	-0.156 (-0.213, -0.099)	-0.142 (-0.199, -0.086)	Through UCD-CRT Through UTL-CRT	-0.007 (-0.020, 0.000) -0.007 (-0.022, 0.000)
Ex-smoker to FT3	-0.003 (-0.190, 0.207)	0.013 (-0.158, 0.224)	Through BCD Through UCD-CRT Through UTL-CRT	-0.002 (-0.018, 0.003) -0.013 (-0.057, 0.002) -0.002 (-0.025, 0.024)
Daily smoker to FT3	0.055 (-0.045, 0.146)	0.057 (-0.053, 0.168)	Through BCD Through UCD-CRT Through UTL-CRT	-0.017 (-0.037, 0.055) 0.000 (-0.009, 0.000) 0.015 (0.001, 0.050)
Gender to FT3	-0.543 (-0.615, -0.468)	-0.545 (-0.616,- 0.471)	Through BCD	0.002 (-0.003, 0.009)

Table 14: Summary of results for tests of indirect effect with corresponding 95% bootstrap CIs for outcome FT3

4.2.3 Outcome FT4

The objective in this context was to investigate whether **UCD-CRT (cadmium in urine)** and **HPYR-CRT (hydro pyrene)** mediated the effect of the covariates on the outcome **FT4**. The covariates to be considered were selected by backward selection procedure. Like in the settings of outcome FT3 and SHBG, this was done by using a correlation matrix as input data for *PROC REG* of SAS. Similarly polychoric, polyserial and pearson correlation coefficients were appropriately computed as needed. The computed correlation values may be viewed on table 21 in the Appendix. The backward selection procedure conducted at $\alpha = 5\%$ showed that the variables **Age, Gender, Adol-brd** and **UCD-CRT** are associated with outcome FT4. The variables **Season** and **Smoking** are associated with HPYR-CRT whereas **Age, BMI** and **Smoking** were associated with variable UCD-CRT. Details of the results may be viewed on table 22 in the Appendix. Figure 9 shows the path diagram of the proposed model.

The number of parameters to be estimated was 68 compared to the number of elements in the covariance matrix which was 78. Therefore the model was said to be overidentified and all parameters of interest were estimable. The results from fitting the proposed model RMSEA and SRMR values of 0.069 and 0.026 respectively indicating that the fit of the model was adequate. The test for mediation was conducted and the results may be consulted on table 15. It was observed that there was a direct effect of Age on the outcome FT4. Adolescents whose measurements were taken in Spring also exhibited higher values of FT4 than in other seasons. No indirect effects were detected.

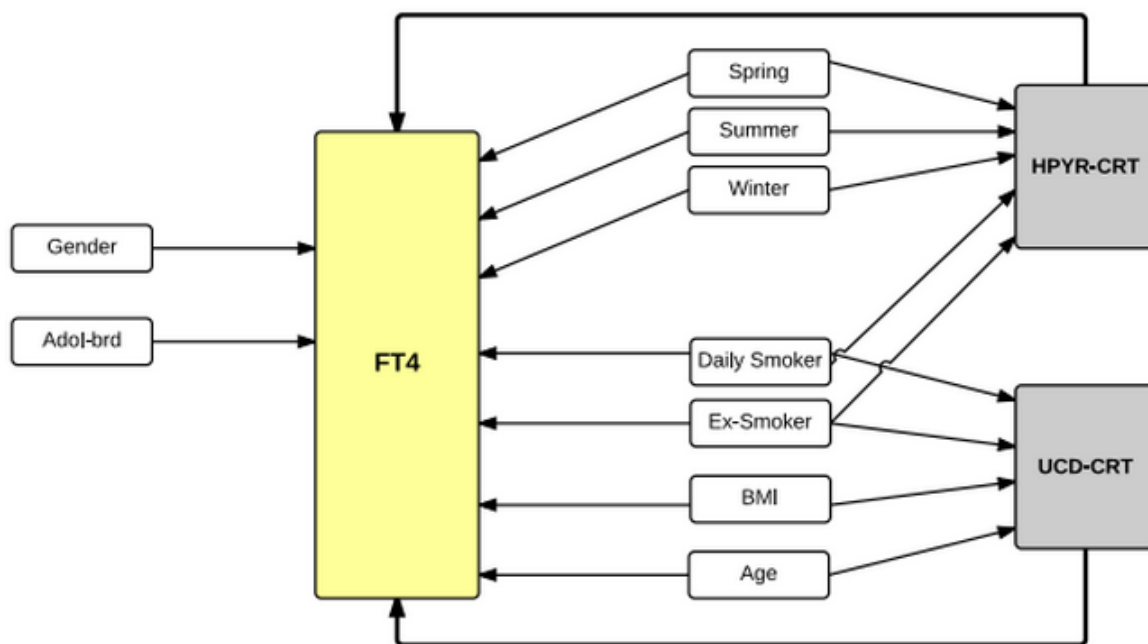


Figure 9: Proposed model for outcome FT4

Effect	Total effect	Direct effect	Specific indirect path	Indirect effect
BMI to FT4	-0.003 (-0.008, 0.000)	-0.004 (-0.009, 0.000)	Through UCD-CRT	0.001 (0.000, 0.002)
Age to FT4	0.034 (0.010, 0.058)	0.031 (0.006, 0.056)	Through UCD-CRT	0.003 (0.000, 0.007)
Ex-smoker to FT4	-0.005 (-0.085, 0.067)	-0.009 (-0.093, 0.065)	Through HPYR-CRT Through UCD-CRT	-0.001 (-0.010, 0.001) 0.005 (-0.002, 0.019)
Daily smoker to FT4	0.025(-0.048, 0.099)	0.031 (-0.044, 0.107)	Through UCD-CRT Through HPYR-CRT	-0.001 (-0.011, 0.005) -0.005 (-0.025, 0.001)
Winter to FT4	0.004 (-0.039,0.046)	0.004 (-0.049,0.046)	Through HYPR-CRT	0.000 (-0.003, 0.002)
Spring to FT4	0.066 (0.021, 0.107)	0.066 (0.021, 0.108)	Through HYPR-CRT	0.000 (-0.001,0.004)
Summer to FT4	0.013 (-0.081, 0.116)	0.014 (-0.084,0.113)	Through HYPR-CRT	-0.001 (-0.016, 0.002)

Table 15: Summary of results for tests of indirect effect with corresponding 95% bootstrap CIs for outcome FT4

4.2.4 Outcome TSH

As in the 3 previous outcomes. The same approach was taken for outcome TSH. The exposures that were considered to be potential mediators were **BTL (thallium in blood)**, **UTL-CRT (thallium in urine)** and **HPYR-CRT (hydro pyrene)**. The correlation matrix that was used as input data for *PROC REG* of SAS was computed as in the other three settings and is displayed on table 23. The results of the backward selection procedure at $\alpha = 5\%$ showed that **BMI**, **Gender** and **Smoking** were associated with TSH while **Adol-edl**, **Season**, **Adol-brd**, **Gender** and **Smoking** were associated with BTL. **Age** and **Smoke** were associated with UTL-CRT whereas **Season** and **Smoke** were associated with HPYR-CRT. Table 24 in the Appendix shows the details of the results obtained from the backward selection. Figure 10 shows the path diagram for the proposed model. Like before, the variable Adol-edl was not considered in the model. The number of parameters to be estimated were 76 while the number of observed elements of the covariance matrix was 91 implying that the model was over identified and thus estimable. The results of the fitted model yielded RMSEA and SRMR values of 0.094 and 0.066 respectively suggesting the fit of the proposed model was mediocre. This is not surprising as the backward selection procedure suggested that none of the exposures (potential mediators) considered were associated with the outcome TSH. Nevertheless, the model is still useful as for each test for indirect effect, specific paths as opposed to the entire model are considered. Furthermore, the modeling of the paths between exposure and outcome although not significant are still needed in order to be able to estimate the value and significance of the indirect effect.

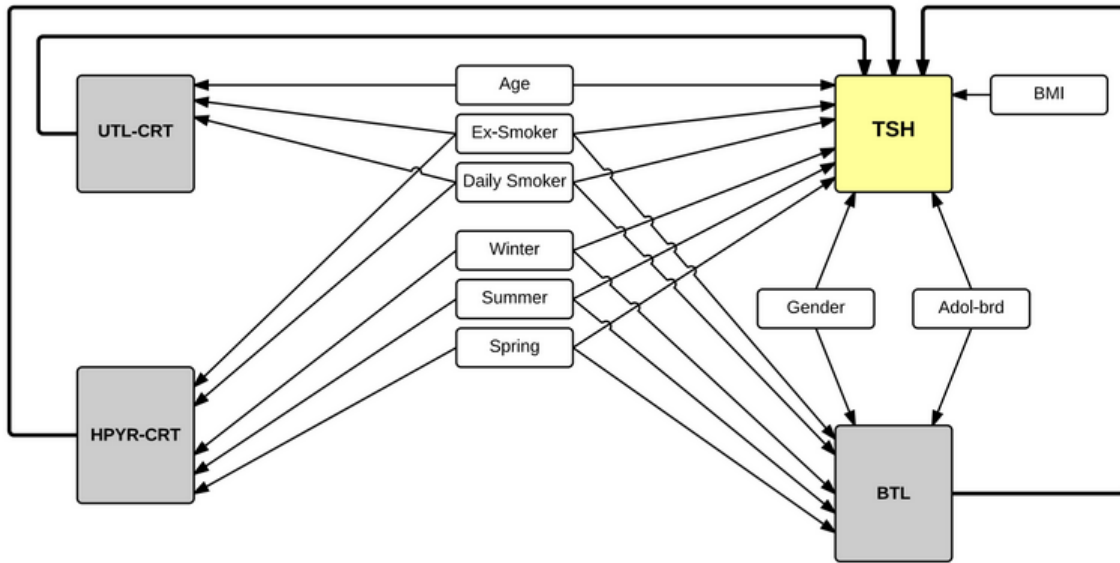


Figure 10: *Proposed model for outcome TSH*

Table 16 shows the results obtained from the analysis. An indirect effect of covariate Daily smoker on outcome TSH through HPYR-CRT was detected. There was also an indirect season effect of Winter on the outcome TSH through HPYR-CRT. Age had a direct effect on the outcome TSH while Gender had both a direct and total effect on the outcome TSH.

Effect	Total effect	Direct effect	Specific indirect path	Indirect effect
Ex smoker to TSH	-0.029 (-0.419, 0.390)	-0.038 (-0.434, 0.381)	Through BTL Through HPYR-CRT Through UTL-CRT	0.003 (-0.009, 0.038) 0.006 (-0.005, 0.033) -0.001 (-0.028, 0.022)
Daily smoker to TSH	-0.177(-0.383, 0.010)	-0.225 (-0.435, -0.036)	Through BTL Through HPYR-CRT Through UTL-CRT	0.014 (-0.010, 0.123) 0.032 (0.003, 0.105) 0.003 (-0.029, 0.104)
Winter to TSH	0.166 (-0.156, 0.423)	0.138 (-0.184, 0.384)	Through BTL Through HPYR-CRT	0.015 (-0.019, 0.066) 0.013 (0.001, 0.038)
Spring to TSH	-0.082 (-0.419, 0.187)	-0.096 (-0.441, 0.173)	Through BTL Through HPYR-CRT	0.007 (-0.008, 0.048) 0.006 (-0.001, 0.024)
Summer to TSH	0.251 (-0.281, 0.784)	0.244 (-0.282, 0.808)	Through BTL Through HPYR-CRT	-0.014 (-0.060, 0.021) 0.017 (-0.005, 0.068)
AGE to TSH	-0.113(-0.275, 0.035)	-0.112 (-0.267, -0.042)	Through UTL-CRT	-0.001 (-0.019, 0.016)
Gender to TSH	-0.210 (-0.365, -0.034)	-0.196 (-0.380, -0.010)	Through BTL	-0.014 (-0.060, 0.021)
Adol-brd to TSH	-0.129 (-0.353, 0.090)	-0.134 (-0.355, 0.087)	Through BTL	0.005 (-0.009, 0.034)

Table 16: *Summary of results for tests of indirect effect with corresponding 95% bootstrap CIs for outcome TSH*

5 Discussion and conclusion

This project was concerned with determining whether some or all of the covariates of interest had an indirect effect on some outcomes through apriori chosen exposure variables. The outcomes considered were 4, namely TSH, FT3, FT4 and SHBG. This necessitated the consideration of 4 distinct models. To test for mediation or indirect effect, Structural Equation Models (SEM) was chosen because unlike linear regression approach to mediation analysis, it allows the fitting of several paths at one go, single goodness of fit statistics are reported that enable one to assess the overall fit of the paths to the specified data. Moreover, they allow for testing for many indirect effects in at one go. All this properties make SEM approach quite an appealing tool for mediation analysis.

For each outcome, several exposures were considered as potential mediators. Maximum likelihood (ML) was chosen as the best estimation method in this setting. ML assumes that the data stem from a multivariate normal distribution. If the model is specified correctly, and if the sample size is sufficiently large, ML provides parameter estimates and standard errors that are asymptotically unbiased, consistent, and efficient. Furthermore, with increasing sample size the distribution of the estimator approximates a normal distribution. In the event that the assumption of multivariate normality is violated and this happens often and is unavoidable in practice, the results obtained can be seriously misleading. Fortunately, simulation studies have suggested that under conditions of severe nonnormality, ML parameter estimates are still consistent but not necessarily efficient. Moreover, when ML is used with bootstrapping, as was the case in this project, the normality assumption may be relaxed and thus one can be confident that the inference made from the analysis is valid[24].

ML also has some other features that make it useful in mediation analysis. ML comes with formal statistical tests of overall model fit for overidentified models such as RMSEA and SRMR that are not available with many other estimators. In the event of missing data, as was the case in this project, the missing data mechanism may be ignored if the missing data mechanism is deemed to be Missing At Random (MAR). Multiple imputation (MI) is another alternative that may be pursued to deal with the missing data especially in instances where non likelihood methods are to be used. The parameter estimates are averaged over the set of analyses and the standard errors are computed using the average of the standard errors over the set of analysis and the between analysis parameter estimate variation. Presently, MPLUS does not support mediation analysis with multiple imputed data [21].

The outcomes and exposures to be considered were determined apriori. The task therefore was to determine the covariates to be used for testing. This was done by use of backward selection procedure. All variables selected were associated with the exposure(s) under consideration at $\alpha = 5\%$. In testing the indirect effect, the Sobel product of coefficient method was used. In order not to rely on the likely incorrect assumption that the the size of indirect effect, $b * a$, are normally distributed across samples from the same population. Bootstrapping was used to obtain confidence intervals for estimates of $b * a$ for the purpose of making inference [14]. The results of indirect effect were documented under each of the outcome variables.

For outcome SHBG, no indirect effects were detected for covariates Daily smoker and Ex

smoker. Daily smoker was significantly associated with the outcome, 8.592 (0.701, 16.577), however this association ceased to be significant when the covariate outcome relationship was assessed while controlling for the exposure HPYR-CRT. Smoking has been shown to be a predictor of HPYR-CRT [8], however, this predictor effect does not carry over to the outcome through HPYR-CRT. Therefore HYPR-CRT (hydroxy pyrene) does not mediate the effect of smoking on SHBG.

For the case of outcome FT3, Age was found to have a significant negative total -0.156 (-0.213, -0.099) and direct effect -0.142 (-0.199, -0.086) on FT3. Indeed, this seems to be in agreement with what has been documented in literature from studies of reference intervals for thyroid hormone levels from birth to adulthood which suggest that FT3 levels decline with age [33]. The presence of cadmium (UCD-CRT) and thallium (UTL-CRT) in urine is indicative of moderate to longterm exposure which increases with age [2]. This is because as one ages, there is an accumulation of levels of cadmium and thallium in the body. Therefore a persons age may be a predictor of the level of exposure in the body, however this predictor effect can not be carried over to the outcome FT3 through the exposures of cadmium and thallium detected in urine. As such the exposures cadmium and thallium detected in urine can not be deemed to be mediating the effect of age on FT3. Daily smoking was found to have an indirect effect on FT3, 0.015 (0.001, 0.050), through thallium detected in urine. This suggests that, compared to those adolescents who have never smoked, adolescent smokers have 0.015 pg/ml higher levels of FT3 due to the mediation effect of thallium detected in urine. Research has shown that smoking is associated with exposure to thallium [4]. It was seen that smoking has an indirect effect on FT3 through thallium and therefore thallium mediates the effect of smoking on FT3.

For the outcome FT4, Age was found to have a significant total 0.034 (0.010, 0.058) and direct effect 0.031 (0.006, 0.056). The results obtained contradict what has been documented in literature from studies of reference intervals for thyroid hormone levels from birth to adulthood which suggest that FT4 levels decline with age [33]. There was also a significant total 0.066 (0.021, 0.107) and direct effect 0.066 (0.021, 0.108) of Spring on FT4 suggesting that FT4 levels were much higher for participants whose measurements were assessed in spring than in any other season. A review of literature did not yield any documentation suggesting that there might be seasonal variation associated with FT4 levels. There is need therefore to investigate further to determine whether the results obtained under outcome FT4 constitute chance finding.

For the outcome TSH, a significant direct effect of Age -0.112 (-0.267, -0.042) on TSH was detected. The results suggest that TSH levels decline with increase in age, this finding has been collaborated by most studies [33] [34]. There was a significant direct -0.210 (-0.365, -0.034) and indirect effect -0.196 (-0.380, -0.010) of Gender on TSH. Female adolescents were found to have lower levels of TSH than their male counterparts. This finding is also consistent with some studies which have been conducted in the past [35].

6 Limitations and recommendations

The data analyzed for this project comprised of missing data. The method used to estimate parameters was Maximum Likelihood (ML). The use of likelihood based estimators makes the missing data mechanism ignorable under the assumption of Missing At Random (MAR). Sensitivity analysis was not conducted to check the impact of this assumption on the results obtained.

The mediation model framework adopted for this project was simple mediation model in which a single mediating variable, M , was incorporated between the covariate, X , of interest and the outcome Y for every indirect path considered. It may be interesting to consider models in which multiple mediators are incorporated between the X and Y relationship for every indirect path considered. Also worth considering is the counterfactual framework in which interaction between a covariate and mediator may be assessed.

7 Acknowledgement

I would like to express my deep appreciation to VLIR (*Vlaamse Interuniversitaire Raad*) for funding both my residence and study in Belgium. Their support in the last two years has made it possible for me to pursue my dream of practicing statistics. I also do appreciate my lecturers for the training and guidance during the master program. What I have learnt has been useful on this project and I foresee that to be the case in my career too. I owe a debt of gratitude to my group mates *Jedelyn, Mohammed, Maria* and *Kevin*, all those days we laboured to complete and submit projects have taught me alot about research work. I am also grateful to VITO for providing the data analyzed on this project. Lastly but certantly not least, I am thankful to *Mevrouw Liesbeth Bruckers* whose availability for meetings, responding to emails and offering of much needed guidance has been a key factor in the completion of this project.

References

- [1] Preventing disease through healthy environments (2006). *Towards an estimate of the environmental burden of disease*. <http://www.who.int/quantifying-ehimpacts/publications/preventingdisease.pdf>. Accessed on 09 July 2014.
- [2] Environmental Protection Agency. *Cadmium*. <http://www.epa.gov/osw/hazard/wastemin/minimize/factsheets/cadmium.pdf>. Accessed on 22 July 2014.
- [3] Medscape. *Thallium toxicity*. <http://emedicine.medscape.com/article/821465-overview>. Accessed on 22 July 2014.
- [4] MedicineNet *Thallium: Bringing doctor's knowledge to you*. <http://www.medicinenet.com/thallium/article.htm>. Accessed on 06 September 2014.
- [5] Christensen, K. L. *Metals in Blood and Urine, and Thyroid Function Among Adults in the United States 2007-2008*. International Journal of Hygiene and Environmental Health. Urban & Fischer Verlag Jena, Jena, Germany, 216(6):624-632, (2013)
- [6] Lab tests online (2013): *a public resource on clinical lab testing from the laboratory professionals who do the testing* .<http://labtestsonline.org/understanding/analytes/t3/tab/test/>. Accessed on 26 Aug 2014.
- [7] Scoscia E., Baglioni S., *Low triiodothyronine (T3) state: a predictor of outcome in respiratory failure? Results of a clinical pilot study*. European Journal of Endocrinology. 151: 557-560,(2004)
- [8] A. Sancini, L. Montuori *Urinary hydroxypyrene and estradiol in an occupationally exposed outdoor population* .Ann Ig 2014; 26: 311-320 doi:10.7416/ai.2014.1991
- [9] PubMed: *Reproductive hormones in relation to polycyclic aromatic hydrocarbon (PAH) metabolites among non-occupational exposure of males* .<http://www.ncbi.nlm.nih.gov/pubmed/19942255>. Accessed on 26 Aug 2014.
- [10] Life Extension Magazine: *Do You Know Your Sex Hormone Status?* .<http://www.lef.org/magazine/mag2011/may2011-Do-You-Know-Your-Sex-Hormone-Status-01.htm>. Accessed on 26 Aug 2014.
- [11] Newsom1USP 655 SEMWinter 2012: *Testing Mediation with Regression Analysis* .<http://www.upa.pdx.edu/IOA/newsom/semclass/ho-indirect.pdf> Accessed on 26 July 2014.
- [12] David P. M., Chondra M. L., Jeanne M. H., Stephen G. W., & Virgil S. (2010): *A Comparison of Methods to Test Mediation and Other Intervening Variable Effects* .<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2819363/?report=classic>.Accessed on 06 Sept 2014.
- [13] Kristopher J. P. & Hayes A. F.: *Contemporary Approaches to Assessing Mediation in Communication Research*.<http://www.sagepub.com/upm-data/23657-Chapter2.pdf>. Accessed on 06 Sept 2014.
- [14] Sage Publications: *Mediation* .<http://www.sagepub.com/upm-data/47570-ch-16.pdf>. Accessed on 26 July 2014.

- [15] Statistical horizons: *Mediation and Moderation Analysis using SPSS and SAS*. <http://www.afhayes.com/> Accessed on 26 July 2014.
- [16] Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural Equation Modelling: Guidelines for Determining Model Fit. *Electronic Journal of Business Research Methods*, 6[1]: 53-60.
- [17] Iriondoa, J.M., Albert, M.J., & Escudero, A. (2003). Structural equation modelling: an alternative for assessing causal relationships in threatened plant populations. *Biological Conservation* 113: 367-377. <http://www.escet.urjc.es/biodiversos/publica/SEMBioCon.pdf>
- [18] Hox, J.J. & Bechger, T.M. (1998). An Introduction to Structural Equation Modeling. *Family Science Review*, 11: 354-373.
- [19] Lleras, Christy.(2005). *Path Analysis*. Encyclopedia of social measurement vol.3 The Pennsylvania State University
- [20] Shkedy, Z. (2004). Structural Equation Modeling An Introduction: Path Analysis and Confirmatory FactorAnalysis. Master in Biostatistics Centre for Statistics Limburgs Universitair Centrum
- [21] Muthen, L. K.& Muthen, B.O.(1998-2007). *Mplus User's Guide. Fifth Edition*. Los Angeles, C.A :Muthen & Muthen
- [22] Lei, P. & Wu, Q. (2007). An NCME Instructional Module on Introduction to Structural Equation Modeling: Issues and Practical Considerations .The Pennsylvania State University.
- [23] Sudano, J. & Perzynski, A. (2013). Applied Structural Equation Modeling for Dummies, by Dummies. www.indiana.edu/~wim/docs/2-22-2013Sudano-Perz-presentation.ppt. Accessed on 21 July 2014.
- [24] Schermelleh-Engel, K. & Moosbrugger, H. (2003). Evaluating the Fit of Structural Equation Models: Tests of Significance and Descriptive Goodness-of-Fit Measures. *Methods of Psychological Research Online* 2003 8(2), 23-74. <http://www.mpr-online.de>. Accessed on 10 December, 2013.
- [25] Browne, M. & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.
- [26] Bentler, P. (1990). Comparative Fit Indexes in Structural Models. *Psychological Bulletin* 107(2), 238-246.
- [27] Hu, L. & Bentler, P. (1998). Fit indices in covariance structure analysis: Sensitivity to underparameterized model misspecification. *Psychological Methods*, 3, 424-453.
- [28] Mulaik, S., James, L., Van Alstine, J., Bennett, N., Lind, S. & Stilwell, C. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin*, 105, 430-445.
- [29] SAS Support (2010). *Polychor Macro*. <http://support.sas.com/kb/25/addl/fusion25010-1-polychor.sas.txt>. Accessed on 22 July 2014.

- [30] Base SAS(R) 9.3 Procedures Guide:Statistical Procedures (2014). *The CORR procedure*. <http://support.sas.com/documentation/cdl/en/procstat/63963/HTML/default/viewer.htm#procstat-corr-sect019.htm>. Accessed on 06 September 2014
- [31] Introduction to SAS. UCLA (2006): *Statistical Consulting Group*. <http://www.ats.ucla.edu/stat/sas/notes2>. Accessed on 24 November 2007.
- [32] SAS Global Forum (2010). *Let SAS Do the Work: Correlation Crossroads*. <http://support.sas.com/resources/papers/proceedings10/256-2010.pdf>. Accessed on 22 July 2014.
- [33] Kapelari K., Kirchlechner C., Hogler W., Schweitzer K., Virgolini I. & Moncayo R.(2008). *Pediatric reference intervals for thyroid hormone levels from birth to adulthood: a retrospective study*. <http://www.biomedcentral.com/1472-6823/8/15>. Accessed on 09 September 2014.
- [34] Robin P. Peeters (2008). *Thyroid hormones and aging*. <http://www.hormones.gr/205/article/article.html>. Accessed on 06 September 2014.
- [35] Hadlow NC, Rothacker KM, Wardrop R, Brown SJ, Lim EM, Walsh JP (2013). *The relationship between TSH and free T4 in a large population is complex, non-linear and differs by age and gender*. <http://www.thyroid.org/thyroid-physicians-professionals/thyroid-disease-information/clinical-thyroidology/july-2013-volume-25-issue-7/clin-thyroidol-201325156-157/>. Accessed on 06 September 2014

8 Appendix

	Age	BMI	HPYR	SHBG	Adol-edl	Adol-brd	Season	Par-edl	Smoking	Sober
Age	1									
BMI	0.0125	1								
HPYR	0.005	-0.017	1							
SHBG	-0.194	-0.383	0.121	1						
Adol-edl	0.090	0.158	-0.004	-0.029	1					
Adol-brd	0.134	0.116	-0.023	-0.222	0.018	1				
Season	-0.282	0.021	-0.097	0.105	0.141	-0.259	1			
Par-edl	0.002	-0.147	-0.001	0.108	-0.534	-0.068	0.048	1		
Smoking	0.325	-0.090	0.137	0.055	0.078	0.141	-0.083	-0.035	1	
Sober	-0.033	0.017	0.025	0.115	0.163	-0.096	-0.008	-0.360	0.154	1

Table 17: Polychoric, polyserial and pearson product moment correlation for outcome SHBG

Outcome	Covariate	P-Value
SHBG	Par-edl	0.0157
	Age	0.0005
	Adol-brd	0.0095
	BMI	< 0.0001
	Sober	0.0108
	HPYR-CRT	0.0284
HPYR-CRT	Smoking	0.0146

Table 18: Variables selected for SHBG outcome with P-value less than 5%

	Age	BCD	BMI	FT3	UCD	UTL	Gender	Adol-edl	Adol-brd	Season	Par-edl	Smoking
Age	1											
BCD	0.061	1										
BMI	0.0317	-0.011	1									
FT3	-0.144	-0.033	-0.057	1								
UCD	-0.144	0.082	-0.092	0.115	1							
UTL	-0.086	-0.010	-0.020	0.140	0.221	1						
Gender	-0.044	-0.035	0.093	-0.637	-0.0235	-0.05577	1					
Adol-edl	0.109	0.490	0.194	0.049	-0.005	-0.021	-0.137	1				
Adol-brd	0.083	-0.014	0.095	-0.109	-0.087	-0.074	0.118	0.121	1			
Season	-0.186	0.027	0.037	0.043	0.005	0.021	-0.167	0.0919	-0.218	1		
Par-edl	-0.067	-0.017	-0.201	0.074	0.043	-0.037	-0.099	-0.489	-0.102	0.037	1	
Smoking	0.326	0.202	0.020	-0.002	-0.130	0.122	-0.012	0.245	0.045	0.027	-0.167	1

Table 19: Polychoric, polyserial and pearson product moment correlation for outcome FT3

Outcome	Covariate	P-Value
FT3	Season	0.0017
	Age	< 0.0001
	Gender	< 0.0001
	BMI	< 0.0001
	Smoking	0.0471
	UCD-CRT	0.0332
BCD	Adol-edl	< 0.0001
	Par-edl	< 0.0001
	Gender	0.0062
	BMI	0.0282
	Smoking	0.0077
UTL-CRT	Age	0.0019
	Smoking	0.0002
UCD-CRT	Age	0.0141
	Smoking	0.0421
	BMI	0.0421

Table 20: Variables selected for FT3 outcome with P-value less than 5%

	Age	BMI	FT4	HPYR	UCD	Gender	Adol-edl	Adol-brd	Season	Par-edl	Smoking
Age	1										
BMI	0.032	1									
FT4	0.191	-0.047	1								
HPYR	0.047	0.007	-0.061	1							
UCD	-0.144	-0.092	-0.117	0.227	1						
Gender	-0.044	0.093	-0.086	0.043	-0.023	1					
Adol-edl	0.109	0.194	-0.038	0.035	-0.005	-0.137	1				
Adol-brd	0.083	0.095	0.085	0.071	-0.087	0.118	0.121	1			
Season	-0.186	0.037	0.022	-0.081	0.005	-0.167	0.092	-0.218	1		
Par-edl	-0.067	-0.201	0.044	-0.093	0.043	-0.099	-0.489	-0.102	0.037	1	
Smoking	0.326	0.020	0.071	0.175	-0.130	-0.012	0.245	0.045	0.027	-0.167	1

Table 21: Polychoric, polyserial and pearson product moment correlation for outcome FT4

Outcome	Covariate	P-Value
FT4	Age	< 0.0001
	Gender	0.0184
	Adol-brd	0.0495
	UCD-CRT	0.0418
HPYR-CRT	Season	0.0453
	Smoking	< 0.0001
UCD-CRT	Age	0.0141
	Smoking	0.0421
	BMI	0.0421

Table 22: Variables selected for FT4 outcome with P-value less than 5%

	HPYR	Age	BMI	BTL	TSH	UTL	Gender	Adol-edl	Adol-brd	Season	Par-edl	Smoking
HPYR-CRT	1											
Age	0.047	1										
BMI	0.007	0.031	1									
BTL	0.005	0.112	-0.003	1								
TSH	0.037	-0.078	0.084	0.037	1							
UTL	0.076	-0.086	-0.020	0.429	0.030	1						
Gender	0.043	-0.044	0.093	-0.187	-0.108	-0.056	1					
Adol-edl	0.035	0.109	0.194	-0.023	-0.025	-0.021	-0.137	1				
Adol-brd	0.071	0.083	0.095	0.102	-0.046	-0.074	0.118	0.121	1			
Season	-0.081	-0.186	0.038	-0.147	-0.040	0.021	-0.167	0.092	-0.218	1		
Par-edl	-0.093	-0.067	-0.201	-0.022	-0.017	-0.037	-0.099	-0.489	-0.102	0.037	1	
Smoking	0.175	0.326	0.020	0.164	-0.156	0.122	-0.012	0.245	0.045	0.027	-0.167	1

Table 23: Polychoric, polyserial and pearson product moment correlation for outcome TSH

Outcome	Covariate	P-Value
TSH	Gender	0.0043
	BMI	0.0210
	Smoking	< 0.0001
BTL	Adol-edl	0.0253
	Season	0.0002
	Gender	< 0.0001
	Smoking	< 0.0001
	Adol-brd	0.0231
UTL-CRT	Age	0.0019
	Smoking	0.0002
HYPT-CRT	Smoking	0.0453
	Season	< 0.0001

Table 24: Variables selected for TSH outcome with P-value less than 5%

9 Software codes

```
/*-----*/
/*STARTING OVER !!!                                     */
/*-----*/

/*IMPORTING DATA WITH NEW VARIABLE NAMES */

libname last "D:\Year 2\sem2\Thesis_new_approach" ;
proc print data=last.adol_sem;run;

PROC IMPORT OUT= last.med_data
            DATAFILE="D:\Year 2\sem2\Thesis_new_approach\adol_sem_new.csv"
            DBMS=CSV REPLACE;
            GETNAMES=YES;
            DATAROW=2;
RUN;

/* Creating 5 Datasets for each response */

title "TSH DATA";
data last.tsh;
    set last.med_data(keep=persnr TSH  adol_edl season par_edl gender
        adol_brd bmi_cat bmi_con age_cat age_con smoking BTL UTL_CRT
        HPYR_CRT);
run;

title "FT4 DATA";
data last.ft4;
    set last.med_data(keep=persnr FT4  adol_edl season par_edl gender
        adol_brd bmi_cat bmi_con age_cat age_con smoking UCD_CRT HPYR_CRT);
run;

title "FT3 DATA";
data last.ft3;
    set last.med_data(keep=persnr FT3  adol_edl season par_edl gender
        adol_brd bmi_cat bmi_con age_cat age_con smoking
        BCD UCD_CRT UTL_CRT);
run;

title "SHBG DATA";
data last.shbg;
    set last.med_data(keep=persnr SHBG gender adol_edl season par_edl
        adol_brd sober bmi_cat bmi_con age_cat age_con smoking HPYR_CRT);
    if gender=1;
run;

/*----- */
/* MISSING DATA PATTERNS                               */
/*----- */

* TSH DATA ;
data tsh_pat;
set last.tsh(keep=TSH  adol_edl season par_edl
gender adol_brd  bmi_con  age_con smoking BTL UTL_CRT
HPYR_CRT);
run;

proc mi data=tsh_pat nimpute=0;
var TSH  adol_edl season par_edl gender adol_brd
bmi_con  age_con smoking BTL UTL_CRT HPYR_CRT;
run;

* FT3 DATA ;
data ft3_pat;
set last.ft3(keep=FT3  adol_edl season par_edl
gender adol_brd  bmi_con  age_con smoking BCD UCD_CRT UTL_CRT);
run;
```



```

proc mi data=ft3_pat nimpute=0;
var FT3  adol_edl season par_edl gender adol_brd
bmi_con  age_con smoking BCD UCD_CRT UTL_CRT;
run;

* FT4 DATA ;
data ft4_pat;
set last.ft4(keep=FT4  adol_edl season par_edl
gender adol_brd  bmi_con  age_con smoking UCD_CRT HPYR_CRT);
run;

proc mi data=ft4_pat nimpute=0;
var FT4  adol_edl season par_edl gender adol_brd
bmi_con  age_con smoking UCD_CRT HPYR_CRT;
run;

* SHBG DATA ;
data shbg_pat;
set last.shbg(keep=SHBG adol_edl season par_edl
adol_brd sober  bmi_con  age_con smoking HPYR_CRT);
run;

proc mi data=shbg_pat nimpute=0;
var SHBG adol_edl season par_edl adol_brd
sober  bmi_con  age_con smoking HPYR_CRT;
run;

* FREQ;
proc means data=last.med_data;
var age_con bmi_con;
run;

/*-----
/*      VARIABLE SELECTION SHBG
/*-----
*/

*--- STEP 1-----:Compute polyserial and polychoric correlation ;
%inc"D:\Year 2\sem2\Thesis_new_approach\polychor.sas" ;

%polychor(data=last.shbg, var= adol_edl adol_brd
season par_edl  smoking sober, out=last.tetcorb_shbg, type=corr);

proc corr data=last.shbg polyserial;
with adol_edl adol_brd season par_edl  smoking sober ;*ordinal;
var age_con bmi_con  HPYR_CRT SHBG  ;*continuous ;
run;

proc corr data=last.shbg pearson;
var SHBG HPYR_CRT bmi_con age_con;
run;

*Import assembled corr into SAS;

PROC IMPORT OUT= last.corrm_shbg
DATAFILE="D:\Year 2\sem2\Thesis_new_approach\corr_m_shbg.csv"
DBMS=CSV REPLACE;
GETNAMES=YES;
DATAROW=2;
RUN;

*--- STEP 2-----:Backward Selection of Covariates using LS;

proc reg data=last.corrm_shbg (type=corr);* using correlation
takes a/c of variable type;
model_outcome: model SHBG= par_edl adol_edl

```

*/

```

season age_con adol_brd bmi_con smoking sober HPYR_CRT /selection=backward;
model_Exposure: model HPYR_CRT= par_ed1 adol_ed1
season age_con adol_brd bmi_con smoking sober /selection=backward;
model_bmi:model bmi_con=smoking age_con /selection=backward;
model_smoke:model smoking=sober age_con/selection=backward;
run;quit;

```

```

/*-----
/* M PLUS PREPARATION
/*-----

```

*/

```

*----- SHBG-----;

```

```

*Step 1: Get descriptive statistics for the file ;

```

```

proc corr data=last.shbg nomiss;run;

```

```

*Step 2 : Convert all of the missing values to a single missing value code;

```

```

data shbg_mplus;
  set last.shbg(keep= smoking sober adol_brd
  age_con bmi_con SHBG HPYR_CRT);
  array allvars _numeric_ ;

```

```

if smoking=1 then smoking_1= 1;
else smoking_1=0;
if smoking=2 then smoking_2= 1;
else smoking_2=0;

```

```

  do over allvars;
    if missing(allvars) then allvars = 9 ;
  end;
run;

```

```

proc print data=shbg_mplus;run;

```

```

*Step 3 : Export data for use in MPLUS;

```

```

proc export data=shbg_mplus outfile="D:\Year 2\sem2\Thesis_new_approach
\desperate_method\shbg.txt" dbms=dml replace ;
run;

```

```

! MPLUS CODE FOR OUTCOME SHBG

```

```

Title: Path Analysis with Indirect effects

```

```

Data:

```

```

  File is "D:\Year 2\sem2\Thesis_new_approach\desperate_method\shbg.txt";
  FORMAT is FREE;

```

```

Variable:

```

```

  Names are

```

```

    adol_brd age bmi smoking sober SHBG HPYR smoke_1 smoke_2;

```

```

  Missing are all(9) ;

```

```

  Usevariables are adol_brd age bmi sober SHBG HPYR smoke_1 smoke_2 ;

```

```

  !Categorical is ;! Categorical Outcome Variables

```

```

  !Note WLSMV can't be used with Nominal Variables

```

```

ANALYSIS:

```

```

  bootstrap = 1000;

```

```

  estimator = ML;

```

```

  iterations=50000;

```

```

MODEL:

```

```

  SHBG on HPYR smoke_1 smoke_2 adol_brd age bmi sober ;

```

```

  HPYR on smoke_1 smoke_2 ;

```

```

MODEL INDIRECT:

```

```

  SHBG ind smoke_1;

```

```

  SHBG ind smoke_2;

```

```

OUTPUT:

```

```

  cinterval(bcbootstrap);

```

```

/*-----
/*   VARIABLE SELECTION FT3
/*-----
*/

*--- STEP 1-----:Compute polyserial and polychoric correlation ;
%inc"D:\Year 2\sem2\Thesis_new_approach\polychor.sas" ;

%polychor(data=last.ft3, var=gender adol_edl adol_brd season par_edl
smoking , out=last.tetcorb_ft3, type=corr);

proc corr data=last.ft3 polyserial;
  with adol_edl season par_edl gender adol_brd smoking ;*ordinal;
  var age_con BCD bmi_con FT3 UCD_CRT UTL_CRT ;*continuous ;
run;

proc corr data=last.ft3 pearson;
var age_con BCD bmi_con FT3 UCD_CRT UTL_CRT;
run;

*Import assembled corr into SAS;

PROC IMPORT OUT= last.corrm_ft3
  DATAFILE="D:\Year 2\sem2\Thesis_new_approach\corr_m_ft3.csv"
  DBMS=CSV REPLACE;
  GETNAMES=YES;
  DATAROW=2;
RUN;

*--- STEP 2-----:Backward Selection of Covariates using LS;

proc reg data=last.corrm_ft3 (type=corr);* using correlation takes a/c
of variable type;
modelb: model FT3=adol_edl season par_edl age_con gender bmi_con adol_brd
smoking BCD UCD_CRT UTL_CRT /selection=backward;
modelc1:model FT3= BCD /selection=backward;
modelc2:model FT3= UCD_CRT /selection=backward;
modelc3:model FT3= UTL_CRT /selection=backward;
model_BCD:model BCD= adol_edl season par_edl age_con gender bmi_con
adol_brd smoking /selection=backward;
model_UCD:model UCD_CRT=adol_edl season par_edl age_con gender bmi_con
adol_brd smoking /selection=backward;
model_UTL:model UTL_CRT=adol_edl season par_edl age_con gender bmi_con
adol_brd smoking /selection=backward;
run;quit;

/*-----
/* M PLUS PREPARATION-FT3
/*-----
*/

*Step 1:Get descriptive statistics for the file ;

proc corr data=last.ft3 nomiss;run;

*Step 2 :Convert all of the missing values to a single missing value code;

data ft3_mplus;
  set last.dumvarft3(keep=FT3 season gender smoking age_con bmi_con
  BCD UTL_CRT UCD_CRT);
  array allvars _numeric_ ;

if season=1 then season_1=1;
else season_1=0;
if season=2 then season_2=1;
else season_2=0;
if season=3 then season_3=1;
else season_3=0;

if smoking=1 then smoking_1= 1;
else smoking_1=0;
if smoking=2 then smoking_2= 2;
else smoking_2=0;

```

```

do over allvars;
  if missing(allvars) then allvars = 9 ;
end;
run;

proc print data=ft3_mplus;run;

*Step 3 :Confirming if N=324;

proc means data=ft3_mplus;
run;

*Step 4 : Export data for use in MPLUS;
proc export data=ft3_mplus outfile="D:\Year 2\sem2\Thesis_new_approach
\desperate_method\ft3.txt" dbms=dlm replace ;
run;
/*-----
/*   VARIABLE SELECTION FT4
/*----- */

*--- STEP 1-----:Compute polyserial and polychoric correlation ;
%inc"D:\Year 2\sem2\Thesis_new_approach\polychor.sas" ;

%polychor(data=last.ft4, var=gender adol_edl adol_brd season par_edl
smoking ,out=last.tetcorb_ft4, type=corr);

proc corr data=last.ft4 polyserial;
  with adol_edl season par_edl gender adol_brd smoking ;*ordinal;
  var age_con bmi_con FT4 HPYR_CRT UCD_CRT ;*continuous ;
run;

proc corr data=last.ft4 pearson;
var age_con bmi_con FT4 HPYR_CRT UCD_CRT;
run;

*Import assembled corr into SAS;

PROC IMPORT OUT= last.corrm_ft4
  DATAFILE="D:\Year 2\sem2\Thesis_new_approach\corr_m_ft4.csv"
  DBMS=CSV REPLACE;
  GETNAMES=YES;
  DATAROW=2;
RUN;

*--- STEP 2-----:Backward Selection of Covariates using LS;

proc reg data=last.corrm_ft4 (type=corr);* using correlation
takes a/c of variable type;
modelb: model FT4= age_con season par_edl adol_edl bmi_con gender adol_brd
smoking HPYR_CRT UCD_CRT /selection=backward;
model_con_HPYR_CRT: model HPYR_CRT=age_con season par_edl adol_edl
bmi_con gender adol_brd smoking /selection=backward;
model_con_UCD_CRT: model UCD_CRT= age_con season par_edl adol_edl
bmi_con gender adol_brd smoking /selection=backward;
modelc1:model FT4= HPYR_CRT /selection=backward;
modelc2:model FT4= UCD_CRT /selection=backward;
run;quit;
/*-----
/* M PLUS PREPARATION-FT4
/*----- */

*Step 1:Get descriptive statistics for the file ;

proc corr data=last.ft4 nomiss;run;

*Step 2 :Convert all of the missing values to a single missing value code;

```

```

data ft4_mplus;
  set last.dumvarft4(keep=FT4 gender smoking bmi_con age_con  adol_brd
  season HPYR_CRT UCD_CRT);
  array allvars _numeric_ ;

if season=1 then season_1=1;
else season_1=0;
if season=2 then season_2=1;
else season_2=0;
if season=3 then season_3=1;
else season_3=0;

if smoking=1 then smoking_1= 1;
else smoking_1=0;
if smoking=2 then smoking_2= 1;
else smoking_2=0;

  do over allvars;
    if missing(allvars) then allvars = 9 ;
  end;
run;

proc print data=ft4_mplus;run;

proc freq data=ft4_mplus;tables smoking smoking_1 smoking_2;
run;

```

*Step 3 :Confirming if N=324;

```

proc means data=ft4_mplus;
run;

```

*Step 4 : Export data for use in MPLUS;

```

proc export data=ft4_mplus outfile="D:\Year 2\sem2\Thesis_new_approach\desperate_method\ft4.txt" dbms=dlm rep;
run;

```

```

!-----
! MPLUS code for outcome FT4
!-----

```

Title: Path Analysis with Indirect effects

Data:

```

File is "D:\Year 2\sem2\Thesis_new_approach\desperate_method\ft4.txt";
FORMAT is FREE;
listwise=on;

```

Variable:

Names are

```

gender adol_brd age s bmi smoking FT4 HPYR UCD s_1 s_2 s_3 smoke_1
smoke_2;

```

Missing are all(9) ;

```

Usevariables are gender adol_brd age bmi FT4 HPYR UCD s_1 s_2 s_3
smoke_1 smoke_2;

```

!Categorical is adol_edl;! Categorical Outcome Variables

!Note WLSMV can't be used with Nominal Variables

ANALYSIS:

```

bootstrap = 1000;
estimator = ML;
iterations= 100000;
!parameterization=theta;

```

MODEL:

```

FT4 on age gender adol_brd UCD HPYR s_1 s_2 s_3 smoke_1 smoke_2;
UCD on age bmi smoke_1 smoke_2;
HPYR on s_1 s_2 s_3 smoke_1 smoke_2;

```

MODEL INDIRECT:

```

FT4 ind bmi;
FT4 ind age;
FT4 ind smoke_1;

```

```

        FT4 ind smoke_2;
        FT4 ind s_1;
        FT4 ind s_2;
        FT4 ind s_3;
OUTPUT:
        cinterval(bcbootstrap);
        TECH1;

/*-----
/*  VARIABLE SELECTION TSH
/*-----
*/

*--- STEP 1-----:Compute polyserial and polychoric correlation ;
%inc"D:\Year 2\sem2\Thesis_new_approach\polychor.sas" ;

%polychor(data=last.tsh, var=gender adol_edl adol_brd  season par_edl
smoking, out=last.tetcorb_tsh, type=corr);

proc corr data=last.tsh polyserial;
    with adol_brd adol_edl  gender par_edl season smoking ;*ordinal;
    var age_con bmi_con  BTL TSH  UTL_CRT HPYR_CRT ;*continuous ;
run;

proc corr data=last.tsh pearson;
var age_con bmi_con  BTL TSH HPYR_CRT UTL_CRT;
run;

*Import assembled corr into SAS;

PROC IMPORT OUT= last.corrm_tsh
    DATAFILE="D:\Year 2\sem2\Thesis_new_approach\corr_m_tsh.csv"
    DBMS=CSV REPLACE;
    GETNAMES=YES;
    DATAROW=2;
RUN;

*--- STEP 2-----:Backward Selection of Covariates using LS;

proc reg data=last.corrm_tsh (type=corr);* using correlation takes a/c of
variable type;
modelb: model TSH= par_edl adol_edl season age_con bmi_con adol_brd
gender smoking BTL UTL_CRT HPYR_CRT /selection=backward;
modelc1:model TSH= BTL /selection=backward;
modelc2:model TSH= UTL_CRT /selection=backward;
modelc2:model TSH= HPYR_CRT /selection=backward;
modelcon_BTL:model BTL= par_edl adol_edl season age_con bmi_con
adol_brd  gender smoking/selection=backward;
modelcon_UTL_CRT:model UTL_CRT= par_edl adol_edl season age_con bmi_con
adol_brd  gender smoking/selection=backward;
modelcon_HPYR_CRT:model HPYR_CRT= par_edl adol_edl season age_con bmi_con
adol_brd  gender smoking/selection=backward;
run;quit;

/*-----
/*  M PLUS PREPARATION-TSH
/*-----
*/

*Step 1:Get descriptive statistics for the file ;

proc corr data=last.tsh nomiss;run;

*Step 2 :Convert all of the missing values to a single missing value code;

data tsh_mplus;
    set last.tsh(keep= TSH bmi_con smoking season  gender adol_brd age_con
HPYR_CRT UTL_CRT BTL);
    array allvars _numeric_ ;

if smoking=1 then smoking_1= 1;

```

```

else smoking_1=0;
if smoking=2 then smoking_2= 2;
else smoking_2=0;

if season=1 then season_1=1;
else season_1=0;
if season=2 then season_2=1;
else season_2=0;
if season=3 then season_3=1;
else season_3=0;

do over allvars;
  if missing(allvars) then allvars = 9 ;
end;
run;

proc print data=tsh_mplus;run;

*Step 3 :Confirming if N=324;

proc means data=tsh_mplus;
run;

*Step 4 : Export data for use in MPLUS;

proc export data=tsh_mplus outfile="D:\Year 2\sem2\Thesis_new_approach\
desperate_method
\tsh.txt" dbms=dlm replace ;
run;

!-----
! MPLUS Code for Outcome TSH
!-----
Title: Path Analysis with Indirect effects
Data:
File is "D:\Year 2\sem2\Thesis_new_approach\desperate_method\tsh.txt";
FORMAT is FREE;
Variable:
Names are
gender adol_brd age s bmi smoking TSH BTL HPYR UTL smoke_1 smoke_2
s_1 s_2 s_3;
Missing are all(9) ;
Usevariables are gender adol_brd age bmi TSH BTL HPYR UTL
smoke_1 smoke_2 s_1 s_2 s_3;
!Note WLSMV can't be used with Nominal Variables

ANALYSIS:
bootstrap = 1000;
estimator = ML;
iterations=100000;

MODEL:
TSH on bmi gender smoke_1 smoke_2 s_1 s_2 s_3 adol_brd age BTL HPYR
UTL ;
BTL on smoke_1 smoke_2 s_1 s_2 s_3 adol_brd gender;
HPYR on smoke_1 smoke_2 s_1 s_2 s_3;
UTL on age smoke_1 smoke_2;

MODEL INDIRECT:
TSH ind smoke_1;
TSH ind smoke_2;
TSH ind s_1;
TSH ind s_2;
TSH ind s_3;
TSH ind age;
TSH ind gender;
TSH ind adol_brd;

OUTPUT:
cinterval(bcbootstrap);
TECH1;

```


Auteursrechtelijke overeenkomst

Ik/wij verlenen het wereldwijde auteursrecht voor de ingediende eindverhandeling:

Human biomonitoring: structural equation models

Richting: **Master of Statistics-Biostatistics**

Jaar: **2014**

in alle mogelijke mediaformaten, - bestaande en in de toekomst te ontwikkelen - , aan de Universiteit Hasselt.

Niet tegenstaand deze toekenning van het auteursrecht aan de Universiteit Hasselt behoud ik als auteur het recht om de eindverhandeling, - in zijn geheel of gedeeltelijk -, vrij te reproduceren, (her)publiceren of distribueren zonder de toelating te moeten verkrijgen van de Universiteit Hasselt.

Ik bevestig dat de eindverhandeling mijn origineel werk is, en dat ik het recht heb om de rechten te verlenen die in deze overeenkomst worden beschreven. Ik verklaar tevens dat de eindverhandeling, naar mijn weten, het auteursrecht van anderen niet overtreedt.

Ik verklaar tevens dat ik voor het materiaal in de eindverhandeling dat beschermd wordt door het auteursrecht, de nodige toelatingen heb verkregen zodat ik deze ook aan de Universiteit Hasselt kan overdragen en dat dit duidelijk in de tekst en inhoud van de eindverhandeling werd genotificeerd.

Universiteit Hasselt zal mij als auteur(s) van de eindverhandeling identificeren en zal geen wijzigingen aanbrengen aan de eindverhandeling, uitgezonderd deze toegelaten door deze overeenkomst.

Voor akkoord,

Mwandigha, Lazaro

Datum: **10/09/2014**