VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

MODELLING AND DATA ANALYSIS MASTER’S STUDY PROGRAMME

Master's thesis

Modeling Human Capital

Žmogiškojo Kapitalo Modeliavimas

Tomas Jaškevičius

Supervisor prof. habil. dr. Vydas Čekanavičius

Vilnius, 2021

VILNIUS UNIVERSITY

FACULTY OF MATHEMATICS AND INFORMATICS

MODELLING AND DATA ANALYSIS MASTER’S STUDY PROGRAMME

Darbo vadovas prof. habil. dr. Vydas Čekanavičius

Darbas apgintas (įrašoma data)

Registravimo NR.

Modeling Human Capital

Abstract

The purpose of this research is to investigate human capital. Reviewing various analysis techniques and methods, delving into the models used, examining world-popular and published reports and articles. Having enough literature, to prepare country level based human capital model, relying on previous papers, using structural equation modeling for better latent variables expresions. After creating the final form of the model, to compare most important variables obtained at the country level model with the results discussed in the more global version. The idea is to find new, as yet unused components that are not addressed, but having a significant impact on the evaluation of human capital. One of the conclusions of my modeling is precisely about the use of new technologies in society and the availability of those technologies in everyday life, this attribute has a strong influence on the estimate of human capital. Since there are only few works created on the analysis of human capital based on Lithuania data, this thesis could become a very good introduction to further, more complex, and more significant discoveries in the future.

Key words : Human capital, structural equation, modeling, development.

Žmogiškojo Kapitalo Modeliavimas

Santrauka

Šio darbo tikslas yra išnagrinėti žmogiškąjį kapitalą. Apžvelgiant įvairias analizės technikas bei metodus, įsigilinant į naudojamus modelius, išnagrinėjant pasaulyje populiarius ir publikuojamus raportus bei straipsnius. Siekiamybė yra sukurti šalies lygio modelį, besiremiant prieš tai publikuotais darbais šia tema ir panaudojant struktūrinių lygčių modelį, kuris leistų lengviau išreikšti latentinius kintamuosius. Sumodeliavus galutinę versiją, palyginti gautus šalies lygio rezultatus su globalesnėje versijoje aptariamais rezultatais. Pagrindinė šio darbo idėja, atrasti naujų, dar nenaudotų komponentų, į kuriuos neatkreipiamas dėmesys, bet jie daro reikšmingą įtaką žmogiškojo kapitalo vertinimui. Viena iš gautų mano modeliavimo išvadų yra būtent apie naujų technologijų naudojimą visuomenėje ir tų technologijų pasiekiamumas kasdieniniame gyvenime, šis matmuo daro stiprią įtaką žmogiškojo kapitalo įverčiui. Kadangi yra sukurta vos keletas darbų apie žmogiškojo kapitalo analizę remiantis Lietuvos duomenimis, manau šis mano darbas gali tapti labai gera įžanga į tolimesnius, sudėtingesnius, ir reikšmingesnius atradimus ateityje.

Raktiniai žodžiai : Žmogiškasis kapitalas, struktūrinės lygtys, modeliavimas, vystymasis.

Content

Introduction

Human capital, and things related to it, becoming more and more important in nowadays life, especially in countries level. While it was discovered and started to be explained not very long ago, also it is not straight forward measurable variable, exist lots of interpretations and methods to evaluate it. There are already created world level reports and models for doing that, but separated country cases are with high interest, they let to understand better what could help to improve economic growth, which parts of human capital still lack clarity, need more attention or improvements. Identifying those things will help not only increase human capital score on global level, but make greater conditions for society and to prepare improved world for tomorrow. As there are no only one strict method and explaining human capital could consist of different attributes, it is becoming quite a challenge to find significant insights. Since human capital feature is latent it must be explained through some other measurable variables, structural equation modeling is one of the handy techniques that could be applied and help to find out relevant components for explaining the same country human capital. The main idea of this thesis is to create a structural equation model based on Lithuania data, while doing that find out attributes which explain human capital and in final part understand whether those variables are different from general approach of human capital explanation, if so, then give insights how they differ. In literature overview part explained human capital in more details and should give a good understanding of what it is and why it is needed. In second chapter more technical overview of used methods and mathematical operations. Finally, in the last section described model building phase and final results, what was achieved, and what still waiting in the future works.

1. Analytical part

1.1. Human capital in general

Human capital is the economic value of the abilities and qualities of labor that influence productivity. Investing in these qualities produces greater economic output. Although there are no very strict rule how to define human capital, everyone could have a little bit of freedom for own interpretetions, but those should not go far away from the previous mentioned definition. Human capital recognizes the intangible assets and qualities that improve worker performance and benefit the economy. These qualities cannot be separated from the people who receive or possess them.

The starting point of the human capital idea can be followed back to crafted by Adam Smith in the 18th century. Smith underlined the importance of "the acquired and useful abilities of all the inhabitants or members of the society"[11]. The useful implications of the thought were not broadly perceived until the 1964, where Nobel Prize winners and University of Chicago financial specialists Gary Becker and Theodore Schultz made the theory of human capital. Becker realized the investment in workers was the same as investing in capital equipment, which is another factor of production. Both are resources that yield income and other outputs[9].

1.1.1. Human capital in real life

Human capital includes any human quality or value that can improve economic output and productivity. Because these are intangible assets cannot be separated from individual workers, quantifying them can be difficult. However, they consistently lead to increased economic performance. Human capital can include qualities like education, technical or on-the-job training, health, mental and emotional well-being, punctuality, problem-solving, people management, communication skills and much more. Interest in these characteristics improves the capacities of the workforce. The outcome is more noteworthy yield for the economy and higher pay for the person. As economies become more information based and globalized, the financial significance of human capital to both individual's competitive advantage and to countries' economic achievement become more critical than any time in recent memory. Likewise, human capital investment delivers many other non-economic benefits as well, such as improved health status, enhanced personal well-being and greater social cohesion. These more extensive advantages are viewed by many authors as being as important as, if not bigger than, the economic benefits in the form of higher earnings and economic growth[11].

Measuring the stock of human capital can help to better understand what drives economic growth, to assess the long-term sustainability of a country's development path, and to estimate the output and productivity performance of the educational sector. While all these perspectives emphasize the importance of measuring the total stock of human capital, later conversations have prompted developing consideration being paid to the distribution of human capital across households and individuals, and on the non-monetary benefits stemming from it. Maximizing current income and consumption in a context of limited resources will not guarantee the supportability of a nation's improvement way. Practical turn of events, in its inter-generational dimension, is usually perceived as necessitating that an unchanged stock of capital per capita be passed on to the next generation. To produce meaningful measures of the complete capital stock of every country, measures of every one of its parts are required. Not only the total stock of human capital but also its evolution over the long run gives important information for monitoring sustainability. For instance, measures of changes in human capital due to demographic factors such as population ageing, may provide an early warning of the danger that the amassing of human capital may not be sustainable over time. This would permit pre-emptive strategies aimed at encouraging alternative forms of investment, to offset the decline in the total capital stock due to ageing. The idea of individuals' prosperity stretches beyond its material side, to encompass a variety of non-financial measurements which, together, characterize individuals' personal satisfaction. This broader perspective has implications for the measurement of human capital as it highlights that, in addition to its economic returns, interest in human capital can generate other benefits that will improve individuals' well-being. These non-economic benefits can incorporate the improved health conditions that are for the most part related to advanced education and which may upgrade not just an individual's productivity and earnings but also his/her subjective well-being. Furthermore, these non-economic benefits are not restricted to individuals, but can stretch out to the general public on the loose. For example, education may lead to better-informed residents, more tolerant of social and cultural diversity and more willing to actively participate in a modern democratic society. While some of these non-economic benefits of education are captured through the monetary measures of human capital (e.g. the longer life expectancy of more educated individuals), this is not the case for most other benefits. Besides, the formation of human capital itself may be influenced by activities that enhance health conditions as well as family and community well-being. This, once more, also has implications for human capital estimation.

The concept of human capital has relatively more importance in labour-surplus countries. These countries are naturally endowed with more of labour because of high birth rate under the given climatic conditions. The surplus labour in these countries is the human asset accessible in more abundance than the tangible capital resource. This human resource can be transformed into human capital with compelling contributions of schooling, wellbeing and virtues. The transformation of raw human resource into exceptionally profitable human resource with these sources is the process of human capital formation. The problem of scarcity of tangible capital in the the work overflow countries can be settled by quickening the pace of human capital formation with both private and public interest in education and health sectors of their national economies. The tangible financial capital is an effective instrument of promoting economic growth of the country. The intangible human capital, on the other hand, is an instrument of advancing exhaustive improvement of the nation since human capital is straightforwardly identified with human development, and when there is human development, the qualitative and quantitative progress of the country is inescapable[7]. This importance of human capital is explicit in the changed approach of United Nations towards comparative evaluation of economic development of different nations in the world economy. The United Nations publishes the Human Development Report on human development in various countries with the target of evaluating the rate of human capital formation in these countries.

The statistical indicator of estimating human development in each nation is Human Development Index (HDI). It is the combination of "Life Expectancy Index", "Education Index" and "Income Index". The life expectancy index uncovers the standard of health of the population in the country; the education index reveals the educational standard and the literacy ratio of the population; and the income index reveals the standard of living of the population. If all these indices have a rising pattern throughout a significant stretch of time, it is reflected in a rising trend in HDI. Human capital is measured by health, education and quality of standard of living. Hence, the components of HDI, Life Expectancy Index, Education Index and Income Index, are directly related to human capital formation within the nation. HDI is indicator of positive correlation between human capital formation and economic development. If HDI increases, there is a higher rate of human capital arrangement in response to a better quality of education and health. Similarly, if HDI increases, per capita income of the nation likewise increases. Implicitly, HDI uncovers that the higher is human capital formation due to good levels of health and education, the higher is the per capita income of the country. This process of human development is the solid establishment of a ceaseless cycle of economic development of the nation for a long period of time. This significance of the concept of human capital in generating long-term economic development of the nation cannot be neglected. It is expected that the macroeconomic strategies of the relative multitude of countries are engaged towards advancement of human development and subsequently economic development.

Human capital is the backbone of human development and economic development in every nation. Mahroum (2007) suggested that at the macro-level, human capital management is about three key capacities: the capacity to develop talent, the capacity to deploy talent, and the capacity to draw talent from elsewhere. Collectively, these three capacities form the backbone of any country's human capital competitiveness. Recent USA research shows that geographic regions that invest in the human capital and economic advancement of immigrants who are already living in their jurisdictions help boost their short-term and long-term economic growth[3]. There is also strong evidence that organizations that possess and cultivate their human capital outperform other organizations lacking human capital.

1.1.2. Human capital and its parameters

Since 2012 the World Economic Forum has annually published its Global Human Capital Report, which includes the Global Human Capital Index (GHCI). In the 2017 edition, 130 countries are ranked according to the quality of their investments in human capital. In October 2018, the World Bank published the Human Capital Index (HCI) as a measurement of economic success. The Index ranks countries according to how much is invested in education and health care for young people. The World Bank's 2019 World Development Report on The Changing Nature of Work showcases the Index and explains its importance given the impact of technology on labor markets and the future of work.

Because of many various variables are possible, it is quite hard to write down only few important groups of indicators, which used to model human capital. Still most of the reasearchers and organizations stick to such three pillars: survival rate, education level, health level. While some works includes income, crime value or other specifics, those not so very popular and assumption could be made, that not important and do not give significant value to the final human capital index. On the other hand, for extraordinary countries or areas, one of those not significant variable could be very required and could show valuable impact, so they should not be forgotten when trying to create country specific model.

The HCI captures key phases of a kid's direction from birth to adulthood. In the poorest countries on the planet, there is a critical danger that a youngster won't make due to fifth birthday celebration. Even if child does reach school age, there is a further risk that it will not start school, not to mention total the full pattern of 14 years of studying, from preschool to grade 12, which is the standard in rich countries. The time child does spend in school may make an interpretation of unevenly into learning, depending on a variety of factors including the quality of of instructors and schools that it encounters. At the point when it turns 18, it carries the enduring impacts of chronic frailty and poor nutrition during youth that limit physical and cognitive abilities as child develop into adulthood. In the case of survival, the overall profitability understanding is stark: children who do not survive childhood never become productive adults. As a result, expected efficiency as a future specialist of a child born today is decreased by a factor equivalent to the survival rate, relative to the benchmark where all children survive[12].

The health and education components of human capital have intrinsic value that is undeniably important yet hard to evaluate. This in turn makes it challenging to combine the various components into a single index. In the case of health, the relative productivity interpretation depends on the empirical literature estimating the financial returns to better health at the individual level. The key challenge is that there is no exceptional, straightforwardly estimated outline pointer of the different parts of health that matter for productivity. Microeconometric literature often uses proxy indicators for health[12].

Education is the process and result of mastering systematic knowledge and skills, a fundamental condition for setting up an individual for life and work. The motivation of education is the formation of character, personality, which is able to adjust to life through such characteristics as independence, activity, creativity, and so on. The education system is designed to meet the requirements of society in the socialization of young people, in the development of socially endorsed personal conduct standards, in the development of a certain institutionalized value system by people. Various sciences study education from their own perspective. For instance, in philosophy the concept of "education" is utilized in the importance of the overall otherworldly cycle of human formation and the result of this process – the profound picture of an individual. Education is investigated as a cultural and historical phenomenon, a methods for protecting, moving and duplicating the gatherings of the otherworldly culture of humankind, people groups, and countries. The achievement of a person in education is determined, first, by how well he has figured out how to assimilate the dominant culture, and secondly, by the cultural capital controlled by the dominant group[1]. Education plays a significant role in the human capital of young people. Numerous conventional individuals comprehend education as a passive and formal process of accumulation of knowledge, because of which you can get a recognition or certificate. The process of education in the modern world ought not be perceived as a detached cycle of gathering of information. In the learning process, our personality is formed by motivational capacities, resolution to accomplish certain objectives, entrepreneurial skills, communication abilities, dedication, innovativeness in tackling different issues, the development of competitive quality that implements in full. In this day and age, to be educated or a certified specialist means to be competitive in all senses. Finally, while education is perhaps the main components of human capital, important to call attention to that the expense of training included time just as cash. Pursuing an education means that students lost the occasion to work, travel, or have children. People only pursued an education if the potential income gain was greater than the cost.

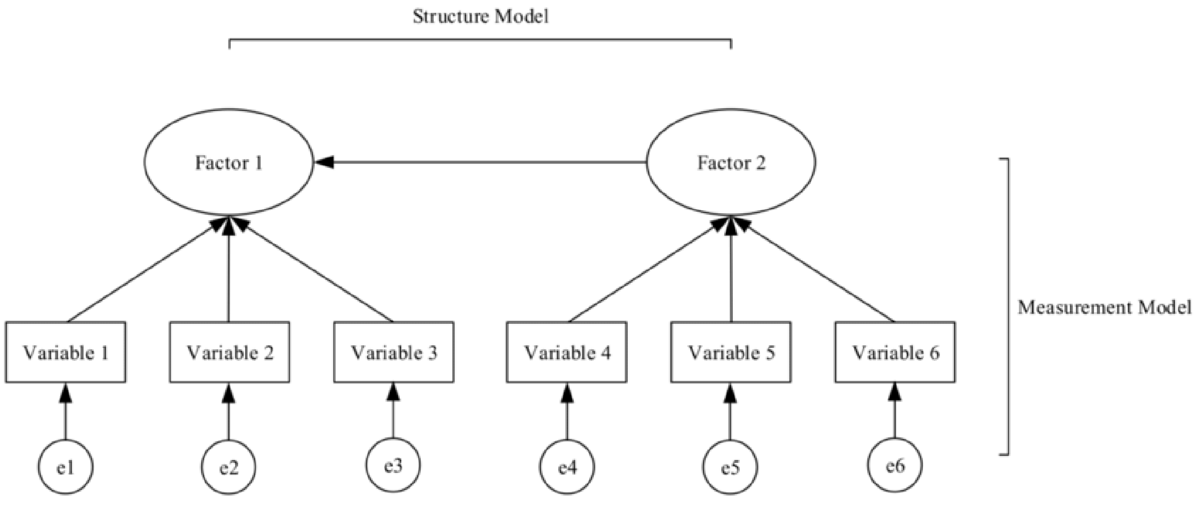
2. Methodological part

2.1. Model for human capital

Lots of the previous researches used structural equation modeling (SEM) as a kind of technique to explain a human capital and explain how it could be interpreted. Writing my work, I relied on such papers, and while in my work I will be using SEM too, so I will try to explain more wide about it in this part of the thesis. It should be well known analyses technique for most, which really helps in defining or studying social sciences.

2.1.1. Structural equation modeling

Structural equation modeling is a statistical technique for building and testing statistical models, which are often causal models. It is a hybrid technique that encompasses aspects of confirmatory factor analysis, path analysis and regression, which can be seen as special cases of SEM. SEM encourages confirmatory, rather than exploratory, modeling, thus, it is suited to theory testing, rather than theory development. It usually starts with a hypothesis, represents it as a model, operationalises the constructs of interest with a measurement instrument and tests the model. With an accepted theory or otherwise confirmed model, one can also use SEM inductively by specifying a model and using data to estimate the values of free parameters. One of its strengths is the ability to model constructs as unobservable latent variables — variables which are not measured directly, but are estimated in the model from measured variables. This allows the modeller to explicitly capture unreliability of measurement in the model, in theory allowing the structural relations between latent variables to be accurately estimated. SEM is an extension of the general linear model that simultaneously estimates relationships between multiple independent, dependent and latent variables.



1 pic. Structural equation model.

The structure of SEM is not complicated, althought few symbols might be in the figure:

- Ovals represent factors, also known as latent variables, unobserved variables or unmeasured variables in SEM. These are theoretical concepts which can be inferred but not directly measured.
- Rectangles are used to represent attributes, also called measured variables, observed variables, or manifest variables.
- Single-headed arrows pointing from one latent variable to another depict hypothetical causal relationships. These can be likened to regression coefficients. The single-headed arrows running from the latent variables to the attributes are equivalent to loadings in Factor Analysis.
- The double-headed arrow is the correlation between the latent exogenous (independent) variables or covariance between errors.
- The numbers adjacent the arrows are the regression coefficients, correlation coefficients and factor loadings. In SEM, regression coefficients are normally smaller than correlations and loadings.
- Error shown as a small circle, pointing out an arrow to specific attribute.

Before starting to build a model, initial dataset should be checked for few requirements fulfilment. First one — no missing value in the data, there should not be such things, cause they will distort the results. Second — number of records. From the theory it is known, that number of records should be from 10 to 20 times more than attributes. While each case is separate, final number should rely on the separate task and desired results. Even the lower or higher ratio could be acceptable in some extraordinary model, but most of the times this recommended interval is the best of choice, at least to start, and then in further investigation can be adjusted if needed. Third thing — stationary data. This one is not strict rule "must be", but rather nice to have feature. While some say it is needed to have a stationary data for acceptable SEM, there are lots of already proved cases when a good model was created using non-stationary data. From one point of view, it will be definitely easier to build a better model, when data is stationary, but as practice showing it is possible to do it and without stationarity. Next requirement — data should have normal distribution. It is like the previous one, not very strict rule, but preferable for better model creation. If variables are non-normal, then data analytic techniques for non-normal continuous variables should be used. There are three popular strategies used to accommodate non-normal data in SEM: Satorra-Bentler scaled chi-square and robust standard errors — for non-normal variables and bigger samples; Yuan-Bentler chi-square for non-normal variables and smaller samples; Diagonally weighted least squares estimation, for non-normal categorical variables. Last thing — no multicollinearity. If correlations within data reach close to the value of one, it might distort the dependencies between variables.

After creatting, it is important to examine the "fit" of an estimated model to determine how well it models the data. This is a basic task in SEM modeling, forming the basis for accepting or rejecting models and, more usually, accepting one competing model over another. The output of SEM programs includes matrices of the estimated relationships between variables in the model. Assessment of fit essentially calculates how similar the predicted data are to matrices containing the relationships in the actual data. Formal statistical tests and fit indices have been developed for these purposes. SEM model tests are based on the assumption that the correct and complete relevant data have been modeled. There are differing approaches to assessing fit. Traditional approaches to modeling start from a null hypothesis, rewarding more parsimonious models, to other such as Akaike information criterion (AIC) that focus on how little the fitted values deviate from a saturated model, taking into account the number of free parameters used. Because different measures of fit capture different elements of the fit of the model, it is appropriate to report a selection of different fit measures[6]. Some of the more commonly used measures of fit:

- Chi-squared (χ^2) — conceptually it is a function of the sample size and the difference between the observed covariance matrix and the model covariance matrix, could be understood as a "badness-of-fit" index, smaller values indicate better fit.
- Akaike information criterion — a test of relative model fit, the preferred model is the one with the lowest AIC value:

  $$AIC = 2k - 2\ln(L),$$

  where k is the number of parameters in the statistical model, and L is the maximized value of the likelihood of the model.
- Root Mean Square Error of Approximation (RMSEA) — fit index where a value of zero indicates the best fit. RMSEA scales F_0 - the population minimum of the Maximum-Likelihood fitting function by the model degrees of freedom. F_0 defined as:

  $$F_0 = \log|S| - \log|\hat{\Sigma}| + tr(\hat{\Sigma}S) - p,$$

  here S is the $p \times p$ population covariance matrix, $\hat{\Sigma}$ - $p \times p$ model-implied covariance matrix and p - the number of observed variables[2]. Then RMSEA:

  $$RMSEA = \sqrt{(F_0/df)}.$$

- Standardized Root Mean Residual (SRMR) — a popular absolute fit indicator, it is a measure of the average difference between the standardized model-implied and population covariance matrix. Hu and Bentler suggested 0.08 or smaller as a guideline for good fit.
- Comparative Fit Index (CFI) — in examining baseline comparisons, the CFI depends in large part on the average size of the correlations in the data. If the average correlation between variables is not high, then the CFI will not be very high.

$$CFI = \frac{1 - max\left[\left({\chi^2}_t - df_t\right), 0\right]}{max\left[\left({\chi^2}_t - df_t\right), \left({\chi^2}_n - df_n\right), 0\right]},$$

here ${\chi^2}_t$ – the chi-square value of the specified and estimated theoretical model, ${\chi^2}_n$ - the chi-square value of the baseline model, df_t - the degrees of freedom of the specified and estimated theoretical model, df_n - the degrees of freedom of the baseline model[8].

For each measure of fit, a decision as to what represents a good-enough fit between the model and the data must reflect other contextual factors such as sample size, the ratio of indicators to factors, and the overall complexity of the model.

2.1.2. Data preparation for SEM

2.1.2.1. Denton-Cholette data disaggregation

Before starting to build a SEM one of the requirements is correct number of records. While using real life data, sometimes it is hard to achieve that, precisely in this case — human capital modeling — because most of the data is collected only on yearly basis, and to have more than 100 records requires quite a history. Also it might be not very adequate to use hundred years of history, it is just simply too big interval. But to use last 15-20 years is completely acceptable for such analyse. This kind of interval should be enough to review all possible variables and take into consideration the most important ones. To attain the correct number of records data disaggregation should be applied. Of course lots of different techniques possible for that result, but I will review more widely the one which I applied in my work — Denton-Cholette temporal disaggregation of time series.

Denton (Denton, 1971) and Denton-Cholette (Dagum and Cholette, 2006) are fundamentally concerned with movement preservation, generating a series that is comparative to the indicator series whether or not the indicator is correlated with the low frequency series. Alternatively, these strategies can disaggregate a series without an indicator. The point of temporal disaggregation is to discover an unknown high frequency series y, whose averages, sums, first or last values are consistent with a known low frequency series y_1 (subscript l denotes low frequency variables). In order to estimate y, possible to use one or more other high frequency indicator variables. Collection of these high frequency series is a matrix X. Terms annual and quarterly will be used instead of low frequency and high frequency hereafter. The diversity of temporal disaggregation methods can be narrowed by creating a two-step framework for them: first, a preliminary quarterly series p has to be determined; second, the differences between the annual values of the preliminary series and the annual values of the observed series have to be distributed among the preliminary quarterly series. The sum of the preliminary quarterly series and the distributed annual residuals yields the final estimation of the quarterly series, $\hat{y}$. Formally

$$\hat{y} = p + Du_1.$$

D is a $n \times n_1$ distribution matrix, with n and n_1 denoting the number of quarterly and annual observations, respectively. u_1 is a vector of length n_1 and contains the differences between the annualized values of p and the actual annual values, y_1:

$$u_1 = y_1 - Cp.$$

Multiplying the $n_1 \times n$ conversion matrix, C, with a quarterly series performs annualization. With two years and eight quarters, and annual values representing the sum of the quarterly values, the conversion matrix is constructed in the following way:

$$C = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The methods of Denton and Denton-Cholette use a single indicator as their preliminary series:

$$p = X,$$

where X is a $n \times 1$ matrix. As a special case, a constant can be embodied as an indicator, allowing for temporal disaggregation without high frequency indicator series. For Denton method, the distribution matrix of temporal disaggregation is a function of the variance-covariance matrix, Σ:

$$D = \Sigma C'(C\Sigma \mathrm{C}')^{-1}.$$

The Denton methods minimize the squared absolute or relative deviations from an indicator series, where the parameter h defines the degree of differencing. For the additive Denton methods and for $h = 0$, the sum of the squared absolute deviations between the indicator and the final series is minimized. For $h = 1$, the deviations of first differences are minimized, for $h = 2$, the deviations of the differences of the first differences, and so on. For the proportional Denton methods, deviations are measured in relative terms. For the additive Denton method with $h = 1$, the variance-covariance matrix has the following structure:

$$\Sigma_D = (\Delta'\Delta)^{-1} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 2 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & \cdots & n \end{bmatrix},$$

where Δ is a $n \times n$ difference matrix with 1 on its main diagonal, −1 on its first subdiagonal and 0 elsewhere. For h = 2, $\Delta'\Delta$ is multiplied by Δ' from the left and Δ from the right side. For h = 0, it is the identity matrix of size n. Denton-Cholette is a modification of the original approach and removes the spurious transient movement at the beginning of the resulting series[5].

2.1.2.2. Normality and stationarity

Data normality and stationarity — also preferable requirements for SEM building. Those should be checked before starting to create a model, for better understanding what type of actions in later stages be needed and get better overview of the used data. While I mentioned, that nowadays in lots of researches and works described models built under non-normal distributed data and without stationarity — it is completely possible to achieve that, and sometimes real world data is not perfect as it should be in theory, still I will describe in this chapter a bit wider about these two requirements.

Stationarity means that the statistical properties (parameters such as mean and variance) of a process generating a time series do not change over time. Strong stationarity requires the shift-invariance (in time) of the finite-dimensional distributions of a stochastic process. This means that the distribution of a finite sub-sequence of random variables of the stochastic process remains the same as we shift it along the time index axis. For example, all independent identically distributed stochastic processes are stationary. The discrete stochastic process $X = \{x_i; i \in \mathbb{Z}\}$ is stationary if:

$$F_X(x_{t_{1+\tau}}, \dots, x_{t_{n+\tau}}) = F_X(x_{t_1}, \dots, x_{t_n}),$$

for $T \subset \mathbb{Z}$ with $n \in \mathbb{N}$ and any $\tau \in \mathbb{Z}$. For continuous stochastic processes the condition is similar, with $T \subset \mathbb{R}$, $n \in \mathbb{N}$ and any $\tau \in \mathbb{R}$ instead[10].

Weak stationarity only requires the shift-invariance (in time) of the first moment and the cross moment (the auto-covariance). This means the process has the same mean at all time points, and that the covariance between the values at any two time points, t and $t - k$, depend only on k, the difference between the two times, and not on the location of the points along the time axis. The process $X = \{x_i; i \in \mathbb{Z}\}$ is weakly stationary if:

- The first moment of x_i is constant; $\forall t, E[x_i] = \mu$
- The second moment of x_i is finite for all t; $\forall t, E[{x_i}^2] < \infty$ (which also implies of course that variance is finite for all t; $E[(x_i - \mu)^2] < \infty$)

- The auto-covariance depends only on the difference u - v; $\forall u, v, a, cov(x_u, x_v) = cov(x_{u+a}, x_{v+a})$

Few most popular tests in practical applications to check whether data is stationary or not — Augmented Dickey–Fuller (ADF) t-statistic test; Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test; Phillips-Perron (PP) test.

The Dickey-Fuller test is testing if $\phi = 0$ in this model of the data:

$$y_t = \alpha + \beta t + \phi y_{t-1} + e_t,$$

which is written as:

$$\Delta y_t = y_t - y_{t-1} = \alpha + \beta t + \gamma y_{t-1} + e_t,$$

where y_t is data. It is written this way that it will be possible to do a linear regression of Δy_t against t and y_{t-1} and test if γ is different from 0. If $\gamma = 0$, then it is a random walk process. If not and $-1 < 1 + \gamma < 1$, then it is a stationary process. The Augmented Dickey-Fuller test allows for higher-order autoregressive processes by including y_{t-p} in the model. But test is still if $\gamma = 0$.

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \dots,$$

the null hypothesis for both tests is that the data are non-stationary. To reject the null hypothesis for this test, a p value should be <0.05.

The KPSS test figures out if a time series is stationary around a mean or linear trend, or is non-stationary due to a unit root. The null hypothesis for the test is that the data is stationary, and alternative hypothesis for the test is that the data is not stationary. The KPSS test is based on linear regression. It breaks up a series into three parts: a deterministic trend (βt), a random walk (r_t), and a stationary error (ε_t), with the regression equation:

$$x_t = r_t + \beta t + \varepsilon_1.$$

If the data is stationary, it will have a fixed element for an intercept or the series will be stationary around a fixed level.

PP test is a unit root test, it is used in time series analysis to test the null hypothesis that a time series is non-stationary. It builds on the Dickey–Fuller test of the null hypothesis, like the Augmented Dickey–Fuller test, the Phillips–Perron test addresses the issue that the process generating data might have a higher order of autocorrelation than is admitted in the test equation. Whilst the ADF test addresses this issue by introducing lags of regressors in the test equation, the Phillips–Perron test makes a non-parametric correction to the t-test statistic. The test is robust with respect to unspecified autocorrelation and heteroscedasticity in the disturbance process of the test equation.

A normal (or Gaussian) distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its probability density function is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},$$

parameter μ is the mean or expectation of the distribution, while the parameter σ is its standard deviation. The variance of the distribution is σ^2. A random variable with a Gaussian distribution is said to be normally distributed, and is called a normal deviate.

Continuous distributions are typically described by their mean (central tendency), variance (spread), skew (asymmetry), and kurtosis (thickness of tails). A normal distribution assumes a skew and kurtosis of zero, but truly normal distributions are rare in practice. Unfortunately, the fitting of standard SEMs to non-normal data can result in inflated model test statistics (leading models to be rejected more often than they should). The assumption of normality is a characteristic of the estimator and not the model itself, structural equation model does not assume normality, but the widely-used normal-theory maximum likelihood estimator does. The assumption of normality applies to the residuals and only relevant for dependent variables, independent variables can take any distributional form. If normality is in doubt, remedial steps can be taken to help solve problems associated with violating this assumption. One option is to apply non-linear transformations to the problem variables. Although these can sometimes help sample data better approximate a normal distribution, non-linear transformations also alter the relationships between variables and can impede substantive interpretation. A second and often better option is to use a method of estimation that is less impacted by the deleterious effects of non-normality like robust maximum likelihood[4].

To test whether data is normaly distributed, some simple things could be done firstly. QQ (quantile-quantile) plot, it compares two different distributions. If the two sets of data came from the same distribution, the points will fall on a 45 degree reference line. Boxplot, if data comes from a normal distribution, the box will be symmetrical with the mean and median in the center. The normal probability plot, it was designed to test for the assumption of normality. If data has a normal distribution, the points on the graph forming a line. Histogram can give a good insight about what kind of distribution data meets. After such initial steps, some more program based test could be acommplished. One of the most simple and often used method for normality testing is Shapiro-Wilk test. This is a way to tell if a random sample comes from a normal distribution. The test gives a W value where small values indicate sample is not normally distributed. Also the value of p is given after test, for more clear decision. The null hypothesis of Shapiro's test is that the population is distributed normally. If the value of p is equal to or less than 0.05, then the hypothesis of normality will be rejected. On failing, the test can state that the data will not fit the distribution normally with 95% confidence. However, on passing, the test can state that there exists no significant departure from normality. The formula for the W value is:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where x_i are the ordered sample values, a_i are constants generated from the covariances, variances and means of the sample:

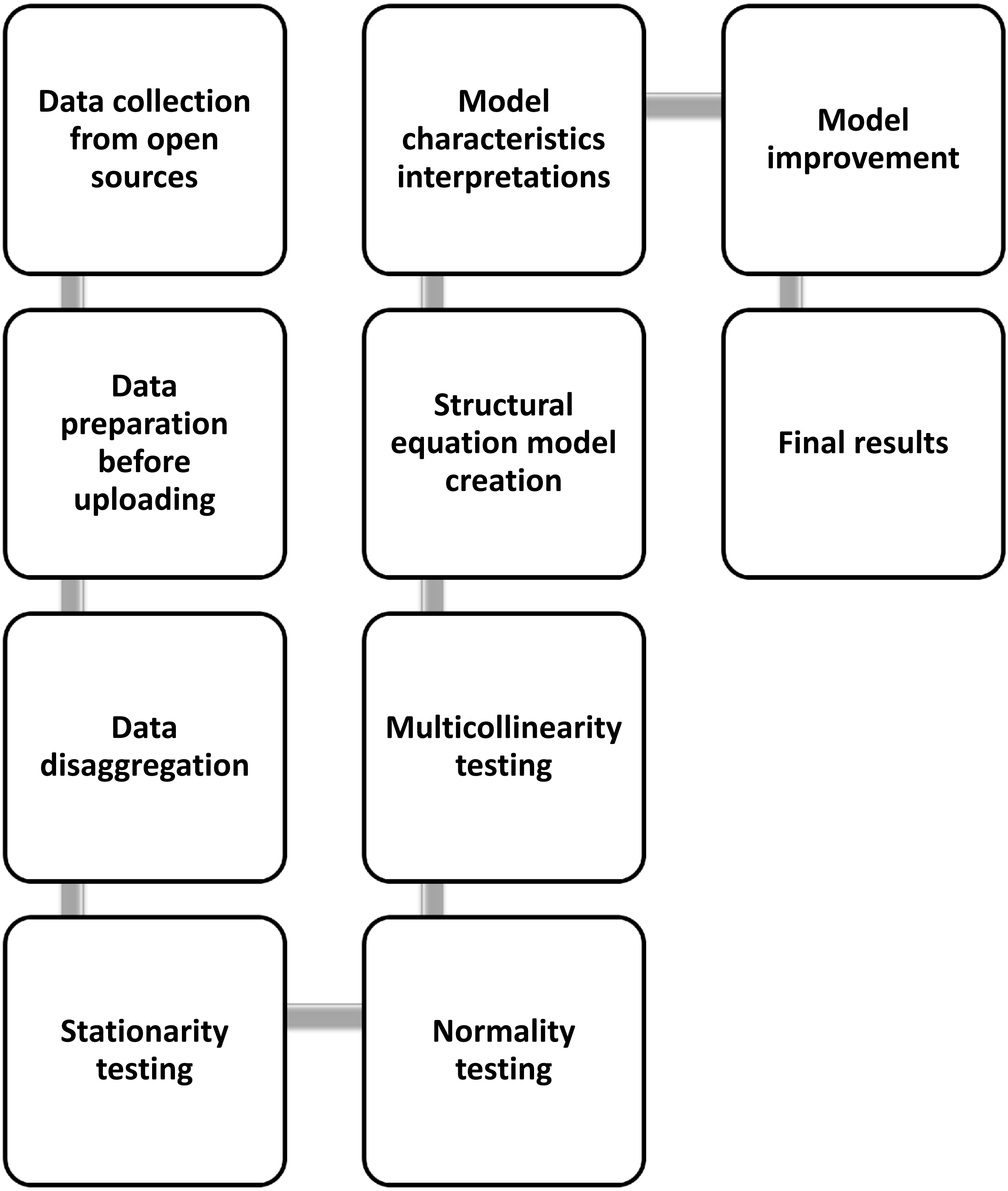
$$(a_1, \dots, a_n) = \frac{m^T V^{-1}}{C},$$

where C is a vector norm:

$$C = \|V^{-1}m\| = (m^T V^{-1} V^{-1} m)^{\frac{1}{2}},$$

and vector $m = (m_1, \dots, m_n)^T$ is made of the expected values. V is the covariance matrix. The test has limitations, most importantly that the test has a bias by sample size. The larger the sample, the more likely a statistically significant result will be.

2.2. Research scheme


Data collection from open sources
Data preparation before uploading
Data disaggregation
Stationarity testing
Normality testing
Multicollinearity testing
Structural equation model creation
Model characteristics interpretations
Model improvement
Final results

2 pic. Research scheme.

3. Research part

3.1. Data overview

To apply and to create a SEM for human capital, I choose Lithuanian data, it was agreed in the beginning, cause I am lithuanian and it will be interesting to work with my native country data. Also, not so many works about human capital or human development done with interfaces with Lithuania, so it will be not only interesting, but might be usefull as well. Real data was collected mostly from official statistics portal (https://osp.stat.gov.lt/), where data prepared and stored by "Environmental Protection Agency", "Finance Ministry", "Lithuanian Bank", "Lithuanian Department of Statistics", "State Labour Inspectorate", and a lot more official Lithuania departments, agency and organizations. HCI was taken from "United Nations Development Programme" reports about human development and capital. After collecting a bunch of raw data, of course preprocessing started. It is must be mentioned, that all data was collected on yearly basis, for that reason, from the very beginning I knew that data disaggregation will be applied. But to apply some kind of method, first of all I need a good and reliable base. Based on previous works and SEM specifics, it is known, that about 15-20 years of history for variables will be the perfect database, and then final number of records depends on how much variables are used. In my primal dataset I checked all the variables, and spotted, that almost all of them goes down to 2001, all other, which do not pass this criterion, were not included in further analyses. Of course by doing this, few variables were deleted, which might show some interesting facts, but they are just simply started to be accumulated too early — 3, 5, or up to 10 years ago, and this kind of period is too short to cause a significant impact, which then could be taken into consideration for HCI. After this step, my data having still lots of variables, and 19 records — starting from 2001 and ending in 2019. Making this, first SEM requirement for no missing values was fulfiled, and already second started — creating the correct number of records. Before moving forward, in the table below all used variables, which connect directly to human capital feature, explained a little bit (variables, which used for other than human capital latent attributes explanation, not included in this list):

Name	Full name	Units	Range	Explanation
hc	human capital index	index	0-1	indicating how countries mobilizing the economy and professional potential of their citizens
hhint	household having internet	percentage	0-100	showing how much of population have ability to use internet
hhcomp	household	percentage	0-100	showing how much of population have

	having computer			ability to use personal computer
pinsa	people involved in scientific activities	amount of people	up to 25000	number of how many people from country involved in some kind of scientific or experimental activities
gdp	gross domestic product	million, eur	up to 11500	the final market value of goods and services produced in a country over a period of time
tax	taxes	percentage	0-100	amount of fees per person from his income
edclvl	education level	index	0-10	showing mean value which represents how well educated people in 25-64 age group
hlvl	health level	index	0-100	showing mean value, which represents how good medical care and how healthy people in the society

1 table. Used variables.

Dataset was uploaded to R program, and all further actions, tests will be applied in that program. As already mentioned few lines above, second SEM requirement should be fulfiled, while first thoughts were to use about 10-15 variables in the final model, it is needed to have number of records in the period somewhere from 100 to 300. Yearly basis data have only 19 records, and temporal data disaggregation of time series must take part here. Denton-Cholette method used for that, it is explained in details in methodological part of this thesis. Were tested different steps — to change data from yearly to quarterly, monthly, bi-weekly, weekly and even some random number of intervals, but in the end decided to use monthly basis. Bigger period, quarterly data, did not work so well, I assume because of too small final number of records. And smaller periods, these are bi-weekly, weekly and other, seems to have no better effect to final results than monthly basis using, only to generate more records takes a bit more computational power and requires more time. For these reasons monthly data is the most optimal option when data quality does not suffer and disaggregation does not take too much time. After these actions I am having 228 records, second requirement for SEM is fulfilled and preprocesing is done, dataset is ready to take part in next model building steps.

3.2. Model creation

After succesfully performing initial data preparation steps, few more things need to be checked. That is stationarity, normality, and multicollinearity. Starting from stationarity, I used few most popular tests in R to understand whether my data is stationary or not. I wrote in more details about those tests in 2.1.2.2. section. Table below shows the results.

Variable	ADF	PP	KPSS
hc	0.4334679	0.9645479	0.01
hhint	0.9507070	0.9892446	0.01
hhcomp	0.8570821	0.9900000	0.01
pinsa	0.3977505	0.5554408	0.01
gdp	0.4221653	0.8447732	0.01
tax	0.9900000	0.9900000	0.01
edclvl	0.4541186	0.6879319	0.01
hlvl	0.3551152	0.6641993	0.01
comp	0.6253999	0.9296499	0.01
schl	0.0632912	0.9430167	0.01
educ	0.0383200	0.4850753	0.01
hexp	0.7473904	0.9402292	0.01
docperpop	0.4434795	0.7550208	0.01
patperday	0.9752943	0.9823695	0.01
mage	0.5412367	0.4516073	0.01

2 table. Stationarity tests.

Obvious that these variables did not pass the stationarity test, for ADF and PP almost all *p* values above 0.05, which shows, that the null hypothesis is not rejected, data cannot be considered as stationary. KPSS shows the same conclusion, all *p* values equals to 0.01 which shows that the null hypothesis is rejected, null hypothesis — data is stationary. Only one variable here, *educ* have ADF test *p* value = 0.03832 which shows weak, but still stationarity of the variable, despite that, PP and KPSS tests show that variable is not stationary, so final decision that this is also not stationary.

Normality was checked in quite similar way, for that I used one of the most popular way — Shapiro-Wilk test, it has been overviewed and explained in 2.1.2.2. section. Results in the table below.

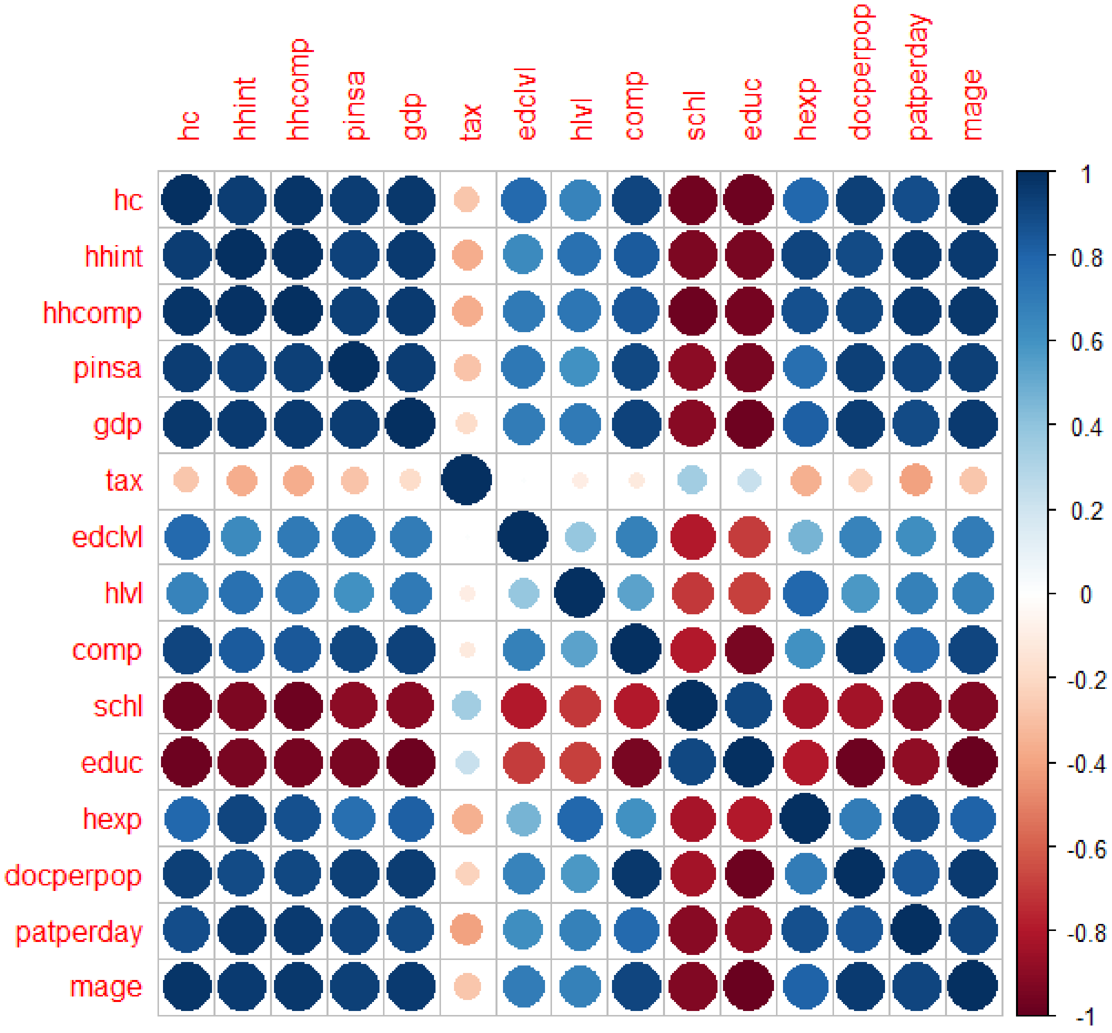
Variable	Shapiro-Wilk
hc	2.765994×10^{-8}

$hhint$	1.943707×10^{-13}
$hhcomp$	2.259039×10^{-11}
$pinsa$	1.334077×10^{-9}
gdp	9.947087×10^{-7}
tax	1.652823×10^{-16}
$edclvl$	2.665107×10^{-7}
$hlvl$	1.686309×10^{-5}
$comp$	9.947232×10^{-15}
$schl$	7.703224×10^{-14}
$educ$	7.156933×10^{-9}
$hexp$	4.378858×10^{-11}
$docperpop$	1.205056×10^{-10}
$patperday$	3.298197×10^{-13}
$mage$	6.694773×10^{-8}

3 table. Normality test.

While all attributes p values very low, <0.05, it means that data distribution differs statistically significant from normal distribution, and we cannot declare normality. Nevertheless, these kind of requirements not very necessary, but from theory known that it is better to have them, for improvement I tried linear data transformations, such as square root and logarithmic, but they did not improve the results significant, and both stationarity and normality remains rejected.

Last thing left — to check data multicollinearity. In picture below correlation for variables between each other shown.



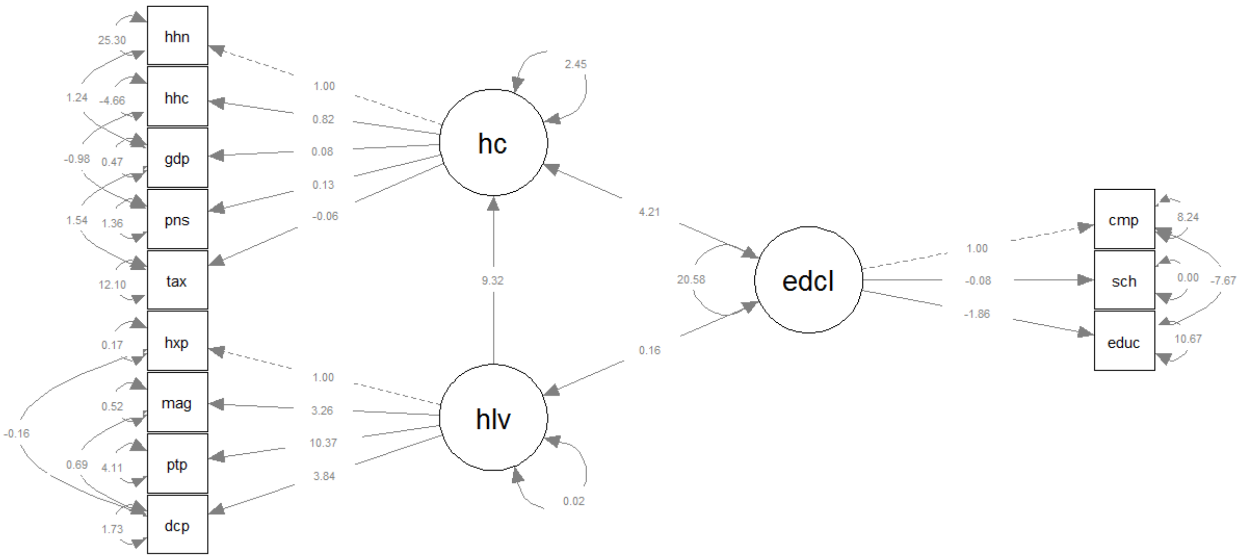
3 pic. Variables correlations.

It is understandable, that exist high correlation between some of the included variables, it is either near 1, either near -1. As known from the theory, it might complicate a bit further model creation, and even in worst case distort the final results.

These last three SEM requirements checking clearly showed, that data is nothing near the excellent, and work with it will be more difficult than it looked like from the beginning. I mentioned previously, that exist more than few works, where good SEM was created without those requirements, and it is completely possible to do that. Also, real world data not always must be perfect, but still possible to find some important and unique things in it.

3.3. Final results interpretations

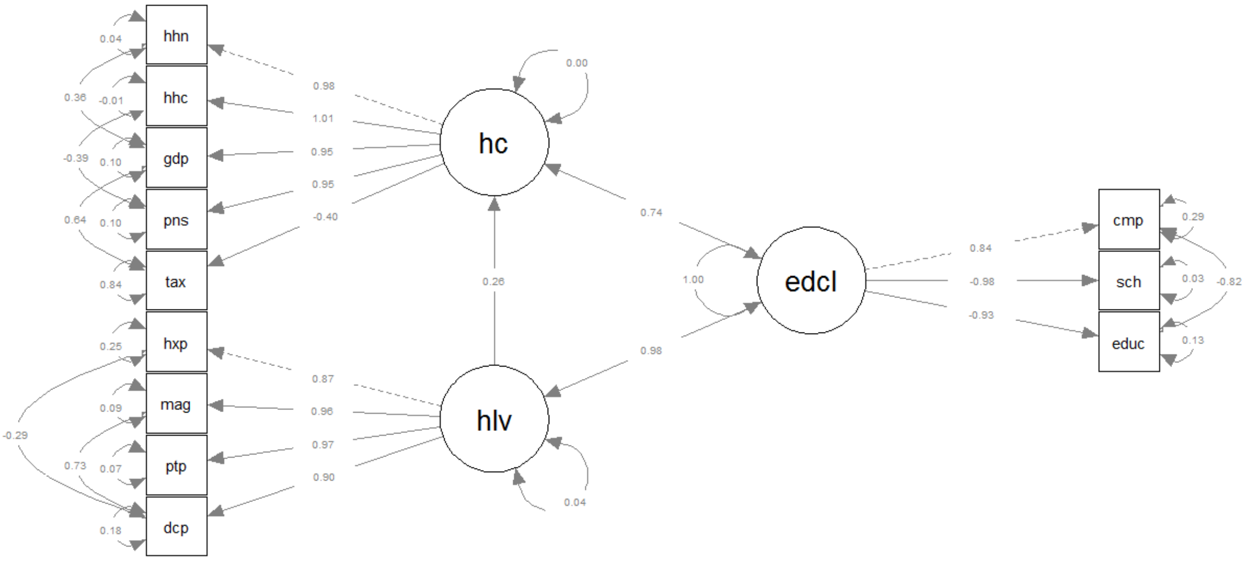
The model itself was created using R programm "lavaan" library, which lets to define model in a nice and understandable way for user. Scheme of how model looks like demonstrated below.



4 pic. Model scheme (with original parameter estimates).

While first variables were chosen relying on previous works and global researches, expecting that they must give acceptable results almost in no time for sure, but it seems that nothing is going to be easy here. Those variables were education level, health level, life expectancy and work costs. In the model correction phase tested quite a lot of different ways and interesting variables for seeking to improve primal its version and to receive better statistics.

Final model version consists three latent variables, those are human capital, education level and health level. Health level expressed using four measurable attributes — health care expenditure, in relation to gross domestic product, average age of the population, number of hospitalisation cases, number of doctors per patients. Biggest impact here comes from variable *patperday*, which is the number of hospitalized people per day, but other attributes do not differ very strong, estimates for all of them are 0.869, 0.956, 0.965, 0.903. Education level expressed using attributes such as how many computers used in schools and universities, number of educational institutions and number of teachers in these organizations. While all variables statistically significant, have standardized estimates respectively 0.845, -0.983 and -0.933. The main latent variable is of course human capital, it expressed using five latent attributes — household having internet, household having computer, gross domestic product (GDP), people involved in scientific activities, taxes. Standardized estimates for those 0.982, 1.000, 0.951, 0.947 and -0.396 respectively. Biggest impact comes from first two components and that was one of the most interesting thing. Some might discuss, that these two features could be assigned to relations to education, but tests show that they give more value when assigned straight to human capital variable, and model statistics increased tremendously by doing that. GDP is one of the most common variable which can be met in a lot of those kind researches, even distinct human capital models were built based only on this one feature, and it is of course giving plenty of explanation to my model, at the same time improving characteristics. Component named *pinsa*, which is showing how many people involved in some scientific activities must be correlating with education, but again, it was not giving so much needed impact to education level as it really giving that impact assigned to human capital latent variable. Since tested lots of, but *tax* is the only one feature which comes from work costs related topic and is statistically significant from data perspective. In model also exist two regressions, but interesting only the one related to human capital. Both *edclvl* and *hlvl* are statistically significant and have standardized estimates 0.741 for education level, 0.261 for health level. Full tables with all measurements added in Appendix B.



5 pic. Model visualization (with standardized parameter estimates).

Model visualization in R program with standardized parameter estimates could be seen in the 5th picture, and main characteristics is in the next table.

CFI	RMSEA	SRMR
0.794	0.399	0.069

4 table. Model evaluation.

From the statisctics it is clear, that model is not the best, could be improved and grown. SRMR is acceptable as it is at the moment, it is <0.08, and CFI is very close to acceptable — in best case it should be as close to 1 as possible. The major issue is RMSEA, this statistic is quite high, definitely too high for considering model as good one, and pointing out poor data fit to the model. I am not considering chi-square here, cause in all cases it was very low, meaning that it is statistically significant. In mathematical literature discussed that those problems with chi-square might occur, when number of records reaching over 200, also when non-normal distributed variables used in the analyses. Still in the Appendix B could be found all the characteristics and statistics.

So far, three main summaries could be made. Firstly, it is a bit harder to find publicly published Lithuania data than expected in the beggining of the whole work. And while not so much of data available, two main reasons oocurs — need to apply data disaggregation, this creates synthetic data, despite that data disaggregation methods is highly reliable these days, it is still hard to say that data reflects totally real world situation. Another reason — because it is not so easy to find all required data about some variables, or some data was collected for too short period, in the final model cannot be included at least few variables that could give some additional information and increase final statistics. Data quality definitely effects final results, and whether they are significant or not is hard to say, but in the end, I would assume that having better quality data it is possible to build a model with at least better RMSEA parameter. Second conclusion — this model characteristics do not mean, that model is poor, or not working. While I have human capital indicator already calculated, these final characteristics rather show the comparision between general model and my created model. Because my model was built while relying on all main areas of HC measurement, and still the statistics shows kinda poor data fit to the model, this could mean, that Lithuania data differs more than expected and while applying general model to it the results of HC might be not very close to reality. This requires specialized model creation — that is what I tried to do, whether my model describes better Lithuania HC — I really doubt about that (first conclusion also takes part in this), but it is definitely a very good start point for further investigations, which will be done in later stages (doctoral (PhD) level studies). Last one, while creating this model some interesting dependencies spotted. Information about how well citizens can access computers and internet, how much of them using smart gadgets gives significant increase in the final model evaluations, this means that nowadays fastly evolving inovations and things related to more smart lifestyle should be considered as valuable circumstance for human capital. Of course those things highly related to education quality, so they are not only increase HC so straight forward, but also increase education level, which in turn also increasing HC.

Conclusion

The main purpose of this thesis was to create a structural equation model based on Lithuania data for human capital, and to identify the main attributes for that, compare those with general approach variables, and make conclusion relying on what fortunate to achieve. As in most of the cases, data and its quality playing a major role in all further analyses, and Lithunia data is not the best, preprocessing, descriptive statistics clearly showed that. Taking into account final results from this paper, it is possible to state that work with Lithuania data could still give some valuable output. After a brief overview of human capital and its importance, relying on previous works, initial model was built. It was improved over time, adding and testing new components, trying to find new relations which improves evaluations. Despite that the final model is not meeting all the estimations, which described in theory as a good model measurements, it is possible to made few conclusions after the whole work. First, publicly published Lithuania data, which can be found, for some specfic needs sometimes not enough, and it will definitely affets the results of any kind of research. Part of the variables might be very biased due to low population, or just not recorded at all. Second, interesting relation was found between human capital and evolving innovations. Computers, internet usage, whole digital era really steping in fastly, and faster society adaptation to that will give significant increase in human capital. Third, this thesis, and this model without a doubt is a very good starting point for further investigations and analyses. More interesting activities could be made after delving into the subject more and more, and I believe those steps will be made in doctoral level studies later.

Bibliography

1. A. Karabayeva, I. Kuntuova, M. Webb. The role of education in realizing youths' human capital: social philosophical analysis, *Ensaio Avaliação e Políticas Públicas em Educação,* Rio de Janeiro, 2018, p. 969-971.
2. A. Maydeu-Olivares, D. Shi, Y. Rosseel. *Assessing fit in structural equation models: A Monte-Carlo evaluation of RMSEA vs. SRMR confidence intervals and tests of close fit*, 2018, p. 4-6.
3. A. Singer. *Investing in the Human Capital of Immigrants, Strengthening Regional Economies*, 2012, p. 1-2.
4. C. DiStefano, S. J. Finney. Non-normal and categorical data in structural equation modeling, *Structural equation modeling: a second course*, Greenwich, Connecticut: Information Age Publishing, 2006, p. 271-274.
5. C. Sax, P. Steiner. Temporal Disaggregation of Time Series, *The R Journal Vol. 5/2*, 2013, p. 80-83.
6. J. Austin, R. MacCallum. *Applications of Structural Equation Modeling in Psychological Research*, 2000, p. 218-219.
7. H. Mahbub. *Reflection on Human Development*, 1996.
8. I. Ercan, S. Cangur. Comparison of Model Fit Indices Used in Structural Equation Modeling Under Multivariate Normality, *Journal of Modern Applied Statistical Methods: Vol. 14*, 2015, p. 157-158.
9. K. Amadeo. *What is Human Capital?*, 2020.
10. S. Palachy. *Stationarity in time series analyses*, 2019.
11. The Task Force on Measuring Human Capital. *Guide on Measuring Human Capital*, 2016, p. 8-9.
12. World Bank Group, T*he Human Capital Index 2020 Update: Human Capital in the Time of Covid-19*, 2020, p. 13-14, 27-28.

Appendix A

Program code:

```
# Step 1. Initialization
rm(list = ls(all.names = TRUE))
gc()

library(readxl)
library(tibbletime)
library(dplyr)
library(tempdisagg)
library(stats)
library(tsbox)
library(lavaan)
library(semPlot)
library(forecast)
library(urca)
library(tseries)
library(ggplot2)
library(rockchalk)
library(corrplot)
library(PerformanceAnalytics)
library(mctest)

intervals = 12

# Step 2. Importing data
data_01 = read_excel('USER_PATH/data_final.xlsx', sheet = 'Sheet1')
# reading only first years, because in earlier records exist missing values
data_02 = data_01[6:24, ]

# Step 3. Data disaggregation
disagghc = td(data_02$hc~1, to = intervals, method = 'denton-cholette',
                 conversion = 'average')
dhc = predict(disagghc)
disaggyosch = td(data_02$yosch~1, to = intervals, method = 'denton-cholette',
                conversion = 'average')
dyosch = predict(disaggyosch)
disaggmage = td(data_02$mage~1, to = intervals, method = 'denton-cholette',
                conversion = 'average')
dmage = predict(disaggmage)
disaggexpage  =  td(data_02$expage~1,  to  =  intervals,  method  =  'denton-
cholette',
                conversion = 'average')
dexpage = predict(disaggexpage)
disaggpop = td(data_02$pop~1, to = intervals, method = 'denton-cholette',
                  conversion = 'average')
dpop = predict(disaggpop)
disaggmarr = td(data_02$marr~1, to = intervals, method = 'denton-cholette',
                  conversion = 'average')
dmarr = predict(disaggmarr)
disaggperomen = td(data_02$peromen~1, to = intervals, method = 'denton-
cholette',
                  conversion = 'average')
dperomen = predict(disaggperomen)
disaggurbn = td(data_02$urbn~1, to = intervals, method = 'denton-cholette',
                  conversion = 'average')
durbn = predict(disaggurbn)
disaggwork = td(data_02$work~1, to = intervals, method = 'denton-cholette',
                conversion = 'average')
dwork = predict(disaggwork)
disaggmeaninc = td(data_02$meaninc~1, to = intervals, method = 'denton-
cholette',
                conversion = 'average')
dmeaninc = predict(disaggmeaninc)
disaggcci = td(data_02$cci~1, to = intervals, method = 'denton-cholette',
                   conversion = 'average')
dcci = predict(disaggcci)
disaggpeominwag = td(data_02$peominwag~1, to = intervals, method = 'denton-
cholette',
                   conversion = 'average')
dpeominwag = predict(disaggpeominwag)
disaggareares = td(data_02$areares~1, to = intervals, method = 'denton-
cholette',
                   conversion = 'average')
dareares = predict(disaggareares)
disaggworkcost = td(data_02$workcost~1, to = intervals, method = 'denton-
cholette',
                   conversion = 'average')
dworkcost = predict(disaggworkcost)
disaggtax = td(data_02$tax~1, to = intervals, method = 'denton-cholette',
                    conversion = 'average')
dtax = predict(disaggtax)

disaggminwage = td(data_02$minwage~1, to = intervals, method = 'denton-
cholette',
               conversion = 'average')
dminwage = predict(disaggminwage)
disaggstdonstd = td(data_02$stdonstd~1, to = intervals, method = 'denton-cholette',
                   conversion = 'average')
dstdonstd = predict(disaggstdonstd)
disaggcomp = td(data_02$comp~1, to = intervals, method = 'denton-cholette',
                   conversion = 'average')
dcomp = predict(disaggcomp)
disaggschl = td(data_02$schl~1, to = intervals, method = 'denton-cholette',
                   conversion = 'average')
dschl = predict(disaggschl)
disaggstuds = td(data_02$studs~1, to = intervals, method = 'denton-cholette',
                   conversion = 'average')
dstuds = predict(disaggstuds)
disaggeduc = td(data_02$educ~1, to = intervals, method = 'denton-cholette',
                   conversion = 'average')
deduc = predict(disaggeduc)
disaggstdperpop = td(data_02$stdperpop~1, to = intervals, method = 'denton-cholette',
                   conversion = 'average')
dstdperpop = predict(disaggstdperpop)
disaggedclvl = td(data_02$edclvl~1, to = intervals, method = 'denton-cholette',
                   conversion = 'average')
dedclvl = predict(disaggedclvl)
disaggedcgdp = td(data_02$edcgdp~1, to = intervals, method = 'denton-cholette',
                   conversion = 'average')
dedcgdp = predict(disaggedcgdp)
disagghhcomp = td(data_02$hhcomp~1, to = intervals, method = 'denton-cholette',
                   conversion = 'average')
dhhcomp = predict(disagghhcomp)
disagghhint = td(data_02$hhint~1, to = intervals, method = 'denton-cholette',
                   conversion = 'average')
dhhint = predict(disagghhint)
disaggpinsa = td(data_02$pinsa~1, to = intervals, method = 'denton-cholette',
                 conversion = 'average')
dpinsa = predict(disaggpinsa)
disaggconper = td(data_02$conper~1, to = intervals, method = 'denton-cholette',
                 conversion = 'average')
```

```
dconper = predict(disaggconper)
disaggcrts = td(data_02$crts~1, to = intervals, method = 'denton-cholette',
                 conversion = 'average')
dcrts = predict(disaggcrts)
disaggpolperpop = td(data_02$polperpop~1, to = intervals, method = 'denton-cholette',
                 conversion = 'average')
dpolperpop = predict(disaggpolperpop)
disaggcrime = td(data_02$crime~1, to = intervals, method = 'denton-cholette',
                 conversion = 'average')
dcrime = predict(disaggcrime)
disagggdp = td(data_02$gdp~1, to = intervals, method = 'denton-cholette',
              conversion = 'average')
dgdp = predict(disagggdp)
disagghlvl = td(data_02$hlvl~1, to = intervals, method = 'denton-cholette',
              conversion = 'average')
dhlvl = predict(disagghlvl)
disagghexp = td(data_02$hexp~1, to = intervals, method = 'denton-cholette',
              conversion = 'average')
dhexp = predict(disagghexp)
disaggdocperpop = td(data_02$docperpop~1, to = intervals, method = 'denton-cholette',
              conversion = 'average')
ddocperpop = predict(disaggdocperpop)
disaggpharmperpop = td(data_02$pharmperpop~1, to = intervals, method = 'denton-cholette',
              conversion = 'average')
dpharmperpop = predict(disaggpharmperpop)
disaggpatperday = td(data_02$patperday~1, to = intervals, method = 'denton-cholette',
              conversion = 'average')
dpatperday = predict(disaggpatperday)

data_03 = data.frame(matrix(ncol = 1, nrow = (intervals * 19)))
data_03$hc = dhc
data_03$yosch = dyosch
data_03$mage = dmage
data_03$expage = dexpage
data_03$pop = dpop
data_03$marr = dmarr
data_03$peromen = dperomen
data_03$urbn = durbn
data_03$work = dwork
```

```
data_03$meaninc = dmeaninc
data_03$cci = dcci
data_03$peominwag = dpeominwag
data_03$areares = dareares
data_03$workcost = dworkcost
data_03$tax = dtax
data_03$minwage = dminwage
data_03$stdonstd = dstdonstd
data_03$comp = dcomp
data_03$schl = dschl
data_03$studs = dstuds
data_03$educ = deduc
data_03$stdperpop = dstdperpop
data_03$edclvl = dedclvl
data_03$edcgdp = dedcgdp
data_03$hhcomp = dhhcomp
data_03$hhint = dhhint
data_03$pinsa = dpinsa
data_03$conper = dconper
data_03$crts = dcrts
data_03$polperpop = dpolperpop
data_03$crime = dcrime
data_03$gdp = dgdp
data_03$hlvl = dhlvl
data_03$hexp = dhexp
data_03$docperpop = ddocperpop
data_03$pharmperpop = dpharmperpop
data_03$patperday = dpatperday
data_03 = data_03[ ,2:38]

# Step 4. Checking data for stationarity
stationarity_df = data.frame(matrix(ncol = 1, nrow = 4))
names(stationarity_df)[1] = 'Test type'
stationarity_df$'Test type' = c('ADF', 'PP', 'KPSS', 'KPSS2')
for(i in 1:(length(data_03)))
{
    data_for_check = data_03[[i]]
    col_name = names(data_03)[i]
    adf = adf.test(data_for_check)$p.value
    pp = pp.test(data_for_check)$p.value
    kpss = ur.kpss(data_for_check, type = 'tau')@teststat
    kpss2 = kpss.test(data_for_check)$p.value
    stationarity_df$col = c(adf, pp, kpss, kpss2)
    names(stationarity_df)[i + 1] = col_name
}

# Step 5. Data check for normality
normality_df = data.frame(matrix(ncol = 1, nrow = 1))
names(normality_df)[1] = 'Test type'
normality_df$'Test type' = c('Shapiro')
for(i in 1:(length(data_03)))
{
    data_for_check = data_03[[i]]
    col_name = names(data_03)[i]
    normality_df$col = c(shapiro.test(data_for_check)$p.value)
    names(normality_df)[i + 1] = col_name
}

# Step 6. Correlations matrix
corrplot(cor(data_03[ ,c(1, 26, 25, 27, 32, 15, 23, 33, 18, 19, 21, 34, 35,
37, 3)]))
chart.Correlation(data_03, histogram = TRUE, pch = 19)

# Step 7. Basic SEM
model_01 = '

    hc =~ hhint + hhcomp + gdp + pinsa + tax
    edclvl =~ comp + schl + educ
    hlvl =~ hexp + mage + patperday + docperpop

    hc ~ edclvl + hlvl
    hlvl ~ edclvl

    comp ~~ educ
    hexp ~~ docperpop
    tax ~~ gdp
    docperpop ~~ mage
    hhcomp ~~ pinsa
    hhint ~~ gdp

'

# Step 8. Model statistics and summary
data_04 = as_tibble(data_03)
fit1 = sem(model_01, data = data_04, estimator = 'mlr')
summary(fit1, standardized = TRUE)
```

```
fitMeasures(fit1, c('cfi', 'rmsea', 'srmr'))

# Step 9. Model figure
semPaths(fit1, style = 'mx', what = 'paths', whatLabels = 'est', layout =
'tree2',
         label.prop = 1.6, edge.label.cex = 1.2, sizeMan = 5, sizeLat = 9,
         rotation = 4, curve = TRUE, curvature = 3)

semPaths(fit1, style = 'mx', what = 'paths', whatLabels = 'std', layout =
'tree2',
         label.prop = 1.6, edge.label.cex = 1.2, sizeMan = 5, sizeLat = 9,
         rotation = 4, curve = TRUE, curvature = 3)
```

Appendix B

Final measurements:

```
lavaan 0.6-5 ended normally after 227 iterations
  Estimator: ML
  Optimization method: NLMINB
  Number of free parameters: 33
  Number of observations: 228

Model Test User Model:
```

	Standard	Robust
Test Statistic	1684.371	2185.115
Degrees of freedom	45	45
P-value (Chi-square)	0.000	0.000
Scaling correction factor, for the Yuan-Bentler correction (Mplus variant)		0.771

```
Parameter Estimates:
  Information: Observed
  Observed information based on: Hessian
  Standard errors: Robust.huber.white

Latent Variables:
```

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
hc =~							
	hhint	1.000				25.799	0.982
	hhcomp	0.823	0.011	74.131	0.000	21.223	1.000
	gdp	0.082	0.001	66.699	0.000	2.111	0.951
	pinsa	0.133	0.002	55.061	0.000	3.428	0.947
	tax	-0.058	0.009	-6.593	0.000	-1.502	-0.396
edclvl =~							
	comp	1.000				4.536	0.845
	schl	-0.079	0.004	-20.141	0.000	-0.361	-0.983
	educ	-1.860	0.029	-63.667	0.000	-8.439	-0.933
hlvl =~							
	hexp	1.000				0.722	0.869
	mage	3.261	0.086	37.778	0.000	2.356	0.956
	patperday	10.370	0.202	51.461	0.000	7.492	0.965
	docperpop	3.840	0.144	26.758	0.000	2.774	0.903

```
Regressions:
```

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all

hc ~							
	edclvl	4.213	0.716	5.881	0.000	0.741	0.741
	hlvl	9.316	4.296	2.168	0.030	0.261	0.261
hlvl ~							
	edclvl	0.156	0.006	24.207	0.000	0.981	0.981

Covariances:

		Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.comp ~~							
	.educ	-7.667	0.458	-16.757	0.000	-7.667	-0.818
.hexp ~~							
	.docperpop	-0.155	0.034	-4.597	0.000	-0.155	-0.287
.gdp ~~							
	.tax	1.541	0.352	4.381	0.000	1.541	0.644
.mage ~~							
	.docperpop	0.694	0.045	15.521	0.000	0.694	0.731
.hhcomp ~~							
	.pinsa	-0.977	0.142	-6.861	0.000	-0.977	-0.388
.hhint ~~							
	.gdp	1.237	0.238	5.202	0.000	1.237	0.358

Variances:

	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.hhint	25.295	2.702	9.362	0.000	25.295	0.037
.hhcomp	-4.662	0.817	-5.705	0.000	-4.662	-0.010
.gdp	0.473	0.045	10.625	0.000	0.473	0.096
.pinsa	1.359	0.133	10.246	0.000	1.359	0.104
.tax	12.098	2.903	4.168	0.000	12.098	0.843
.comp	8.244	0.619	13.322	0.000	8.244	0.286
.schl	0.005	0.001	5.231	0.000	0.005	0.034
.educ	10.666	0.667	15.987	0.000	10.666	0.130
.hexp	0.169	0.025	6.837	0.000	0.169	0.245
.mage	0.520	0.036	14.348	0.000	0.520	0.086
.patperday	4.112	0.449	9.157	0.000	4.112	0.068
.docperpop	1.733	0.100	17.418	0.000	1.733	0.184
.hc	2.453	2.444	1.004	0.316	0.004	0.004
edclvl	20.579	1.450	14.198	0.000	1.000	1.000
.hlvl	0.019	0.006	3.142	0.002	0.037	0.037

cfi	rmsea	srmr
0.794	0.399	0.069