Contents

1 Data Description

In this paper, I will be demonstrating how Structural Equation Modelling can be used to study relationships in various areas of social science. This process will be carried out on a dataset created through surveying young adults in Slovenia in 2013, which can be accessed through: `https://search.gesis.org/research_data/ZA5980`

This dataset contains responses, given by 907 individuals, on the following topics:

- Leisure and lifestyle
- Religious and social affiliations
- Family and friends
- Concerns and aspirations
- Education and employment
- Democracy and politics
- Governance and development
- National and world politics
- Demographic module

As a result, this dataset contains 386 variables spanning a broad range of topics, allowing for quite diverse sets of analysis, as well as highly complex analysis to be carried out. For the sake of this paper, I will initially consider 42 variables in order to answer numerous different questions. These variables will allow me to create latent variables relating to:

- Dietary health
- Lifestyle health
- Wealth
- Parents' level of education
- Religiosity
- Trust in social serving institutions

Additionally, I will utilize some standalone variables relating to:

- Desire to leave Slovenia - D2
- Health (or perceived level of health) - A17

2 Research Questions

As mentioned earlier, questions relating to several different topics will be explored:

1. Health:
   (a) Do people who are overall following healthy diets also tend to follow healthy lifestyles? (hypothesis: Yes)
   (b) Do people who follow healthy diets and lifestyles also tend to have good health? (hypothesis: Yes)
2. Does the education within a household have a substantial effect on the respondent's wealth? (hypothesis: Yes)
3. Religiosity and Trust in public serving institutions:
   (a) Is there a relationship between religiosity similar to people who trust public serving institutions? (hypothesis: No)
   (b) Does religiosity and trust in public services institutions affect the desire for an individual to leave Slovenia? (hypothesis: Religiosity has no effect, but trust in PS does)
   (c) Do the findings from parts a and b hold true if we were to consider males and females separately? (hypothesis: Yes)

3 Data Exploration and Pre-processing

Before beginning the analysis, I need to investigate and resolve any potential multicollinearity issues, as well as determining which variables don't satisfy the normality assumptions and finally dealing with missing values that may exist.

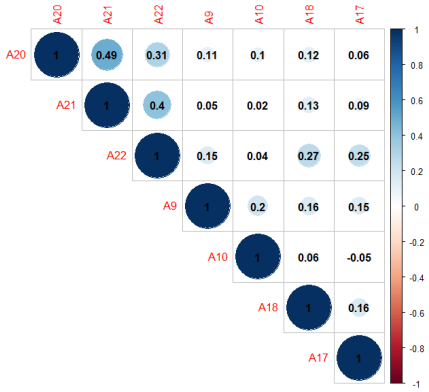

Figure 1: Q1 correlations

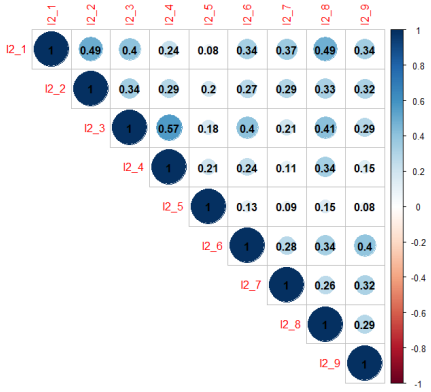

Figure 2: Q2 correlations

3.1 Multicollinearity

Investigating multicollinearity can be done through looking at correlation matrices. If two variables have a correlation coefficient greater than or equal to 0.7, then I should proceed to eliminate one of them. This is essential as such variables can increase the variance of coefficient estimates and make these estimates highly sensitive. With that being said, given that different variables will be looked at for each question, I will consider 3 separate correlation matrices.

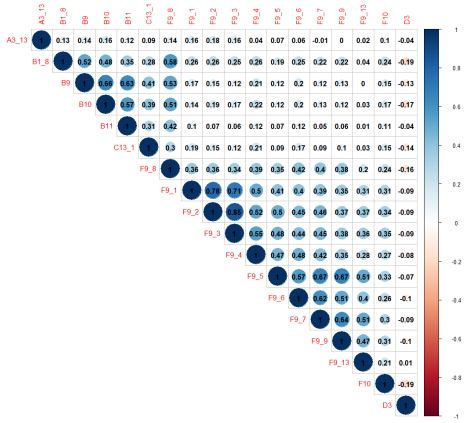

Figure 3: Q3 correlations

Looking at the correlation matrices above (figures 1 and 2), nothing very alarming pops up. The largest correlation (absolute value) in figure 1 is 0.49, and in figure 2 is 0.57, allowing us to conclude that we can keep all variables and should not face any issues. However, when we look at figure 3, we see that several potential problems pop up, where we are also being more strict as the number of variables interacting is quite large. Three main sets of variables need to be dealt with:

- F9_1, F9_2, F9_3: These variables represent the level of trust in political parties, parliament, and government respectively, sharing correlation coefficients of 0.76, 0.71, 0.85, which are extremely high. Therefore, eliminating two of them is a must, and I choose to keep F9_1
- F9_5, F9_7, F9_9: These variables represent the level of trust in the prosecutor's office, court of audit, and judiciary respectively, sharing correlation coefficients of 0.67, 0.67, 0.64. Similar to the other 2 cases, I eliminate two of the variables, keeping F9_5
- B9, B10, B11: These variables relate to the respondent's importance of God in his/her life, the regularity of attending religious service, and the regularity of praying outside of religious service. However, they share correlation coefficients of 0.66,0.63, and 0.57, prompting the decision to choose just one to stay on the safe side (in this case B9), as their are already several other variables relating to religiosity

3.2 Variable normality assumption

As we all know, Structural Equation Modelling makes several assumptions, one of which is multivariate normality, which if it were to be violated would distort statistical inference. Since we are dealing with categorical data, it automatically comes to reason that the variables will often not satisfy this normality assumption. To determine this we follow these 2 rules:

1. If a variable has less than 5 levels, then we will automatically assume it violates the normality assumption
2. If a variable has 5 or more levels, but less than 10 and is highly skewed (absolute(skew) is greater than 1) then it is assumed that the normality assumption is violated

In the case that a variable is determined to violate this assumption, I will treat the variable as an ordered categorical variable as can be seen in the Rcode. The significance of this is the way that the variable is dealt with when conducting SEM, where it will in-fact be represented by a latent variable that has its own threshold estimates w.r.t to the normal distribution.

With all that being said, after applying the above rules, 12 variables have been classified as categorical and so will be treated as such.

3.3 Dealing with missing values

When it comes to dealing with missing values it is of utmost importance to determine whether the values are: missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR).

Depending on the nature of the missingness of the variable, it would be dealt with differently. In the case of this dataset, given that this was a survey that was filled entirely by respondents, it is unlikely that they are MCAR or MAR, and so will be treated as MNAR. Therefore, the option of dealing with missing values using 'full information ML' is not available, and so I will resort to carrying out data imputations utilizing the missForest package (as the Amelia package can't deal with categorical variables).

Now that the missing data issue has been resolved with imputation, I can move on to the analysis of the data and explroing the research questions mentioned above.

4 Analysis

The approach that will be taken is the following:

1. Create a basic model with the base relationship that I want to test
2. Assess the model fit
3. Make adjustments according to MI if they are theoretically sound
4. Evaluate parameter estimates
5. Re-define model if necessary
6. Interpret results

It is worth noting that fit indices being used will be evaluated with the following rules of thumb:

- SRMR < 0.08 • RMSEA < 0.05 • CFI > 0.95 • TLI > 0.95

4.1 Topic 1 - Health

Part a - Model creation, assessment and parameter estimates assessment

In order to test whether people who follow healthy diets also tend to follow healthy lifestyles I must create latent variables representing each of these two concepts, and see how correlated they are. Therefore, the model initially defined is:

```
dietary_health =~ A20 + A21 + A22
lifestyle_health =~ A9 + A10 + A18
dietary_health ~~ lifestyle_health
```

where

- A20/A21 represent the number of fruit/vegetables servings consumed per day respectively
- A22 represents how healthy the respondent perceives his/her diet to be
- A9/10 represent how often the respondent smokes/consumes alcohol respectively
- A18 represents how often the respondent undergoes vigorous physical activity

This model has the following fit indices, indicating a lack of model fit:

SRMR	RMSEA	CFI	TLI
0.05471374	0.07432798	0.93485062	0.87784492

A quick look at modification indices showed that adding the relation $A18 \sim\sim A22$ would be beneficial to improving model fit. This makes intuitive sense when considering the output (not shown here, but can be seen in the Rcode) showing that the correlation between the two latent variables is as low as 0.2. Therefore, given that it is commonly seen that people who are active physically also try to follow healthy diets, it makes sense to make this addition. Moreover, the model is adjusted to incorporate that additional relationship giving us the final model with the following fit indices, indicating a good model fit:

SRMR	RMSEA	CFI	TLI
0.03664001	0.04748746	0.97673132	0.95013855

Part a - Results interpretation and conclusion

	Estimate	Std.Err	z-value	P($>\|z\|$)	Std.lv	Std.all
dietary_health =~						
A20	1				0.944	0.653
A21	1.137	0.109	10.383	0	1.072	0.739
A22	0.497	0.044	11.34	0	0.469	0.522
lifestyle_heal =~						
A9	1				0.626	0.626
A10	0.456	0.12	3.814	0	0.286	0.308
A18	1.187	0.303	3.911	0	0.743	0.348
dietary_health ~~						
lifestyle_hl	0.199	0.043	4.664	0	0.338	0.338
.A22 ~~						
.A18	0.395	0.057	6.936	0	0.395	0.257

A look at the standardized parameters shows that A18 and A10 have standardized loadings indicating they are not appropriate. Therefore, I will adjust the model, removing A10, but retaining A18, as I believe it is an essential variable for overall health, and seeing if this results in better estimates.

Part a - Model re-creation, assessment, and interpretation

The updated model is:

```
dietary_health =~ A20 + A21 + A22
lifestyle_health =~ A9 + A18
dietary_health ~~ lifestyle_health
```

and its path diagram can be seen in the which can be seen in the appendix[4] and it has the following fit indices:

SRMR	RMSEA	CFI	TLI
0.03101428	0.04781467	0.98897088	0.96323627

and the following parameter estimates:

	Estimate	Std.Err	z-value	P(> |z|)	Std.lv	Std.all
dietary_health =∼						
A20	1				0.92	0.637
A21	1.191	0.113	10.51	0	1.096	0.755
A22	0.512	0.044	11.684	0	0.471	0.523
lifestyle_heal =∼						
A9	1				0.46	0.46
A18	2.035	0.654	3.114	0.002	0.936	0.438
dietary_health ∼∼						
lifestyle_hl	0.169	0.045	3.744	0	0.4	0.4
.A22 ∼∼						
.A18	0.337	0.056	5.981	0	0.337	0.229

Before coming to any conclusions, we see that all the indicator variables are appropriate (surpassing the minimum 0.4 standardized loading).

Moving on, we see that A21 (the servings of vegetables loads the most strongly on dietary health), which makes sense as it is one of the areas where there is no controversy regarding its upside and close to complete lack of downside. Furthermore, looking at the lifestyle health variable, neither variable loads very highly, telling us that this is not an ideal representative for the concept 'lifestyle health', but one we can still work with.

With regards to the structural relationship, we see a moderately low positive correlation present between dietary health and lifestyle health, telling us that while most people don't seem to do both, **it is more likely that if one follows a healthy diet then they also follow a healthy lifestyle and vice versa...**

Part b - Model construction

Building on part a, we add to our model the relationship $A17\ dietary\_health + lifestyle\_health$, and obtain the following fit indices:

SRMR	RMSEA	CFI	TLI
0.04408654	0.07124679	0.95861194	0.89652985

However, when looking at the modification indices, no theoretically sound additions are present. Therefore, we can't proceed to rely on this model nor on updating it to be more reliable, and so no conclusion can be reached.

4.2 Topic 2 - Education and Wealth

Model creation, assessment and parameter estimates assessment

The goal here is to test the direct relationship that the level of education within a household (mainly parents as the respondent is too young to have his/her education effect his/her wealth) affects the wealth of the household. To do this the following model is specified:

```
wealth =~ I2_1 + I2_2 + I2_3 + I2_4 + I2_5 + I2_6 + I2_7 + I2_8 + I2_9
parents_education =~ I1_2 + I1_3
wealth ~ parents_education
```

where:

- I2_1,... I2_9 represent the number of each item owned by the respondent
- I2_2 represents the mother's level of education and I2_3 represents the father's level of education

I initially obtain a poor model fit as can be seen by these fit indices:

SRMR	RMSEA	CFI	TLI
0.06630370	0.07146349	0.95001929	0.93607119

A quick look at modification indices shows that adding the relation $I2\_3 \sim\sim I2\_4$ and $I2\_6 \sim\sim I2\_9$ would be beneficial. These two additions make sense since:

- I2_3 represents the number of TV's owned, and I2_4 represents the number of LCD TV's owned (which are obviously correlated, due to one being a subset of the other)
- I2_6 represents the number of refrigerators and I2_9 represents the number of rooms in the apartment, which again makes sense as the larger an apartment/house gets the greater the number of refrigerators would be

As a result, I end up with the final model which can be seen in the appendix[5] with the following fit indices, indicating an acceptable model fit:

SRMR	RMSEA	CFI	TLI
0.05355250	0.05713077	0.96954284	0.95914284

Results Interpretation and Conclusion

	Estimate	Std.Err	z-value	P(> |z|)	Std.lv	Std.all
wealth =~						
I2_1	1				0.855	0.709
I2_2	0.866	0.049	17.664	0	0.74	0.628
I2_3	0.694	0.046	14.995	0	0.593	0.581
I2_4	0.435	0.04	10.956	0	0.372	0.408
I2_5	0.357	0.052	6.829	0	0.305	0.305
I2_6	0.411	0.029	14.358	0	0.351	0.536
I2_7	1.146	0.073	15.731	0	0.979	0.464
I2_8	0.729	0.042	17.279	0	0.623	0.627
I2_9	1.875	0.142	13.232	0	1.603	0.506
parents_education =~						
I1_2	1				0.584	0.648
I1_3	1.511	0.422	3.585	0	0.883	0.883
wealth ~						
parents_eductn	0.257	0.07	3.694	0	0.176	0.176

We first look at the measurement part of the model, to see if our indicator variables are appropriate for our latent concepts.

Starting off with wealth, it is clear that out of the 9 indicators, 8 are satisfactory (having std.all values greater than 0.4). Furthermore, it can be seen that the items I2_1, I2_2, and I2_8 items have the most substantial impact on wealth, i.e. the number of phones, computers and cars, which theoretically makes

a lot of sense. Furthermore, the one variable that is not a good representative is the I2_5 one, i.e. the number of ACs in the household. This could be because Slovenia is not a country that typically gets very hot, and so not one that needs ACs.

With regards to parent's education, both indicators are very valuable, as one would expect, and even more so the loading of the father's education is close to 0.9, being extremely high, as one would expect in past generations that the male is more likely to be educated.

Finally, looking at the effect parents' education has on wealth we see a standardized parameter of 0.176, wherein an increase of 1 standard deviation in parent's education would result in an increase of 0.176 standard deviations in wealth. This tells us that clearly **there is quite a substantial impact of the level of education of the parents with the wealth of the household as one would expect**.

4.3 Topic 3 - Religiosity and Trust in public serving institutions

Part a - Model creation and assessment

The goal here is to study the relationship between religiosity and trust in public services institutions. Initially the following simple model is specified:

```
trust_ps =~ F9_1 + F9_4 + F9_5 + F9_6 + F9_13
religosity =~ A3_13 + B1_8 + B9 + C13_1 + F9_8
religosity ~~ trust_ps
F9_5 ~~ F9_13
B1_8 ~~ B9
```

where:

- F9_1,...F9_13 are variables indicating the level of trust an individual has for different types of public services institutions
- A3_13 is the frequency that the respondent watches religious shows
- B1_8 is how much the respondent trust religious leaders
- B9 is how important God is in the respondent's life
- C13_1 is how important religious affiliation is for the respondent when seeking a partner/spouse
- F9_8 is the level of trust the individual has towards religious leaders (similar to B1_8 but somehow doesn't cause multicollinearity issues)

Initially, when fitting this model a non-ideal model fit is obtained, as can be seen with the following global fit indices:

SRMR	RMSEA	CFI	TLI
0.05839023	0.07198360	0.97778207	0.97059392

However, an evaluation of modification indices suggests adding several relationships, of which the following two are theoretically consistent: $F9\_5 \sim\sim F9\_13$ and $B1_8 \sim\sim B9$. Therefore, the model is updated accordingly and has the following fit indices, indicating good model fit:

SRMR	RMSEA	CFI	TLI
0.04413120	0.05022556	0.98981978	0.98568407

Part a - Results interpretation and conclusion

	Estimate	Std.Err	z-value	P(> |z|)	Std.lv	Std.all
religosity =∼						
A3_13	1				0.13	0.198
B1_8	4.539	0.822	5.521	0	0.59	0.59
B9	3.351	0.625	5.364	0	0.435	0.435
C13_1	3.312	0.633	5.23	0	0.43	0.375
F9_8	7.695	1.332	5.778	0	1	1
trust_ps =∼						
F9_1	1				0.77	0.77
F9_4	0.975	0.041	23.896	0	0.75	0.75
F9_5	0.982	0.042	23.543	0	0.756	0.756
F9_6	1.046	0.04	25.848	0	0.805	0.805
F9_13	0.644	0.045	14.358	0	0.496	0.496
religosity ∼∼						
trust_ps	0.061	0.011	5.475	0	0.607	0.607
.F9_5 ∼∼						
.F9_13	0.228	0.028	8.202	0	0.228	0.4
.B1_8 ∼∼						
.B9	0.192	0.03	6.418	0	0.192	0.263

Starting off with the measurement models, looking at the standardized loadings, all indicators for trusting ps are informative, most of which load very highly. However, with regards to religiosity, 2 variables (C13_1 and A3_13) don't meet the 0.4 minimum that we seek, therefore are not appropriate indicators for religiosity. It is not clear why that would be the case and it warrants further investigation. Regardless I will recreate the model excluding these two variables.

Part a - Model re-creation, assessment, and interpretation

The model has been updated to the following:

```
religosity =~  B1_8 + B9  + F9_8
trust_ps =~ F9_1 + F9_4 + F9_5 + F9_6 + F9_13
religosity ~~ trust_ps
F9_5 ~~ F9_13
B1_8 ~~ B9
```

and its path model can be seen in the appendix[6]. The models fit indices indicate a good model fit, and can be seen in the table below:

SRMR	RMSEA	CFI	TLI
0.03783239	0.04858185	0.99462646	0.9911494

The table below presents the updated parameter estimates:

	Estimate	Std.Err	z-value	P(> |z|)	Std.lv	Std.all
religosity =∼						
B1_8	1				0.572	0.572
B9	0.726	0.06	12.033	0	0.415	0.415
F9_8	1.817	0.162	11.231	0	1.04	1.04
trust_ps =∼						
F9_1	1				0.759	0.759
F9_4	0.984	0.041	23.837	0	0.747	0.747
F9_5	1.008	0.042	23.88	0	0.764	0.764
F9_6	1.06	0.041	25.631	0	0.804	0.804
F9_13	0.669	0.045	14.823	0	0.508	0.508
religosity ∼∼						
trust_ps	0.258	0.027	9.596	0	0.595	0.595
.F9_5 ∼∼						
.F9_13	0.214	0.027	7.859	0	0.214	0.385
.B1_8 ∼∼						
.B9	0.21	0.031	6.695	0	0.21	0.282

Now, it is quite clear that all the indicators are appropriate for the latent variables. What is left is to look at the relation between religiosity and trusting public serving institutions. We see a correlation of 0.26, indicating a weak but existing positive correlation. The reason for this could be that religious people tend to be less skeptical than non-religious people.

Part b - Model creation

Building on part a, I will add a variable the variable D3, which represents the degree to which the respondent desires to leave Slovenia, as I would like to assess the effect that the two previously seen latent variables have on this variable. Therefore, the model constructed is:

```
D3 ~ religosity + trust_ps
religosity =~  B1_8 + B9  + F9_8
trust_ps =~ F9_1 + F9_4 + F9_5 + F9_6 + F9_13
religosity ~~ trust_ps
F9_5 ~~ F9_13
B1_8 ~~ B9
```

and its path model can be seen in the appendix[7]. Additionally, it has the following fit indices are obtained, indicating a good model fit:

SRMR	RMSEA	CFI	TLI
0.03925010	0.04492384	0.99386890	0.99040350

Part b - Results interpretation and conclusion

Given that the latent variables have the same indicators as before, I will not present the results for the measurement part of the model, but simply present the new part (the relation of the two constructs with the desire to leave Slovenia):

	Estimate	Std.Err	z-value	P(> |z|)	Std.lv	Std.all
D3 ∼						
religosity	-0.538	0.107	-5.024	0.000	-0.314	-0.314
trust_ps	0.103	0.077	1.342	0.179	0.078	0.078

The output in the table above shows there is no statistically significant relationship between the desire to leave Slovenia and the level of trust one has towards public servicing institutions. However, religiosity

not only has a statistically significant effect, but quite a strong one, where an increase of 1 standard deviation in religiosity results in a decrease in 0.314 standard deviations in desire to leave Slovenia. This could be due to religious people being more spiritual, and so less being less concerned with the material/physical world, and/or more concerned with practicing their faith, community, and family.

Part c - Do the previous findings hold for each gender separately?

In order to test whether these findings hold if we were to model separately on the basis of gender, I will test for measurement invariance/equivalence. The different levels, from weakest to strongest, of invariance are:

- Configural
- Metric
- Scalar
- Strict
- Structural

To test for each level of equivalence, we create models imposing restrictions according to each level of equivalence, and then evaluate the goodness of fit of each model. Additionally, we will consider the Chisq change from one level to another.

For part a (correlation between religiosity and trusting ps) we obtain the following results:

	df	cfi	tli	rmsea	rmsea.pvalue	srmr	Chisq	Chisq diff	Df diff	Pr(>Chisq)
Configural	34	1	0.99	0.03	0.92	0.04	52.03			
Metric	40	1	0.99	0.04	0.92	0.05	63.175	6.8824	6	0.33186
Scalar	65	1	1	0.01	1	0.04	70.796	17.082	25	0.8788
Strict	66	1	1	0.01	1	0.04	70.802	0.0479	1	0.82672
Structural	69	0.99	0.99	0.03	0.98	0.05	106.202	10.7475	3	0.01317

All fit measures don't have any significant change until we reach structural invariance, where only the Chisq difference test indicates invariance. Furthermore, we note that even achieving partial structural invariance is impossible, as the score test we carried out failed to be rejected, indicating that even if we release all constraints the model fit would not be improved enough.

With regards to part b (religosity and trust in ps effect on desire to leave Slovenia) we obtain the following results:

	df	cfi	tli	rmsea	rmsea.pvalue	srmr	Chisq	Chisq diff	Df diff	Pr(>Chisq)
Configural	46	1	0.99	0.03	0.01	0.05	0.96	0.04		
Metric	52	1	0.99	0.03	0.02	0.05	0.96	0.05	6	0.31182
Scalar	79	1	1	0.02	0	0.03	1	0.05	27	0.93982
Strict	80	1	1	0.01	0	0.03	1	0.05	1	0.8243
Structural	83	0.99	0.99	0.03	0.02	0.04	0.99	0.05	3	0.01334

Once again, all fit measures don't have any significant change until we reach structural invariance, where only the Chisq difference test indicates invariance. Furthermore, we once again note that even achieving partial structural invariance is impossible for the same reasons as before.

However, we can conclude that both findings would hold for both genders, as the only thing that varies between groups is latent variables variance/covariance.

5 Conclusion and Next Steps

This paper demonstrates how structural equation modelling can be used to test hypothesis a person might have in various different areas of social science. As we were able to create models that allowed us to answer questions regarding health, wealth, education, religiosity and trust.

Furthermore, much more complex analyses could be done, I have only managed to scratch the surface here. Ideally, one would continue to utilize datasets, such as this one, to test even more complex hypothesises, testing relationships between all the topics we saw and more.

6 Appendix

6.1 Model 1a

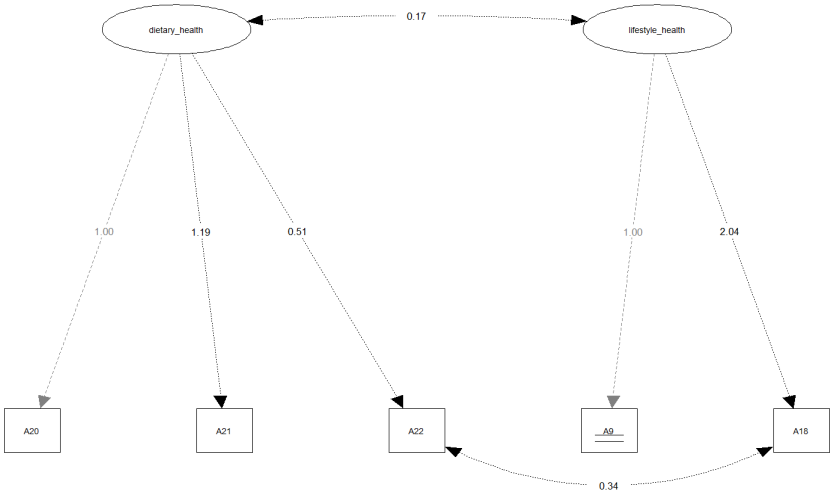

Figure 4: Dietary/lifestyle health relationship model

6.2 Model 2

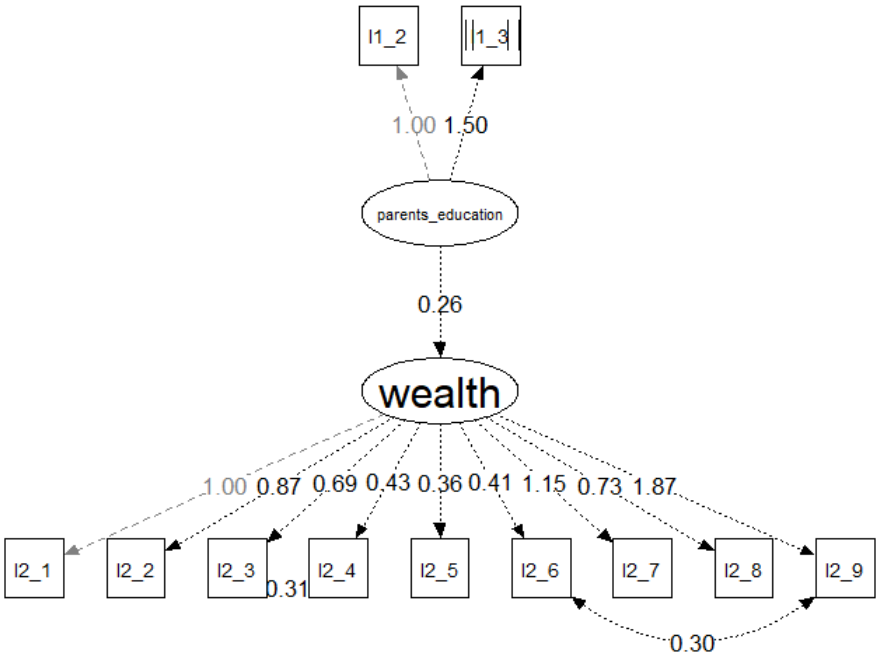

Figure 5: Household education and wealth model

6.3 Model 3a

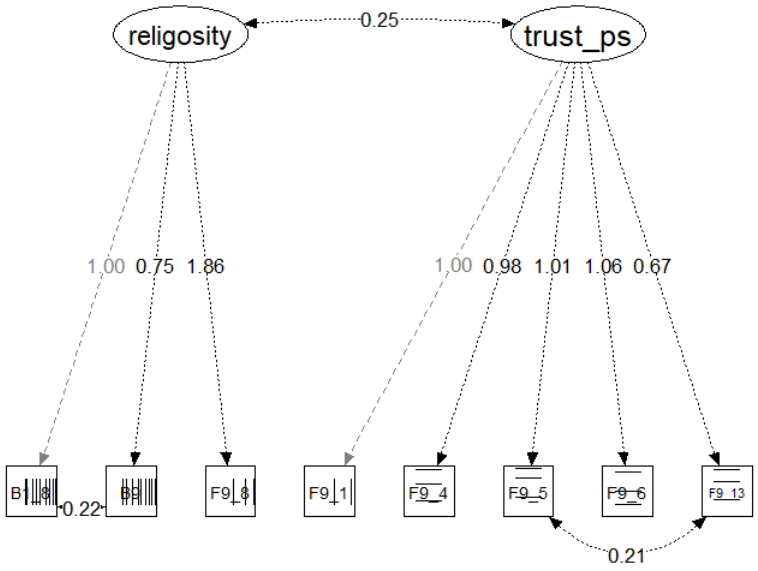

Figure 6: Religiosity and Trust in PS institutions

6.4 Model 3b

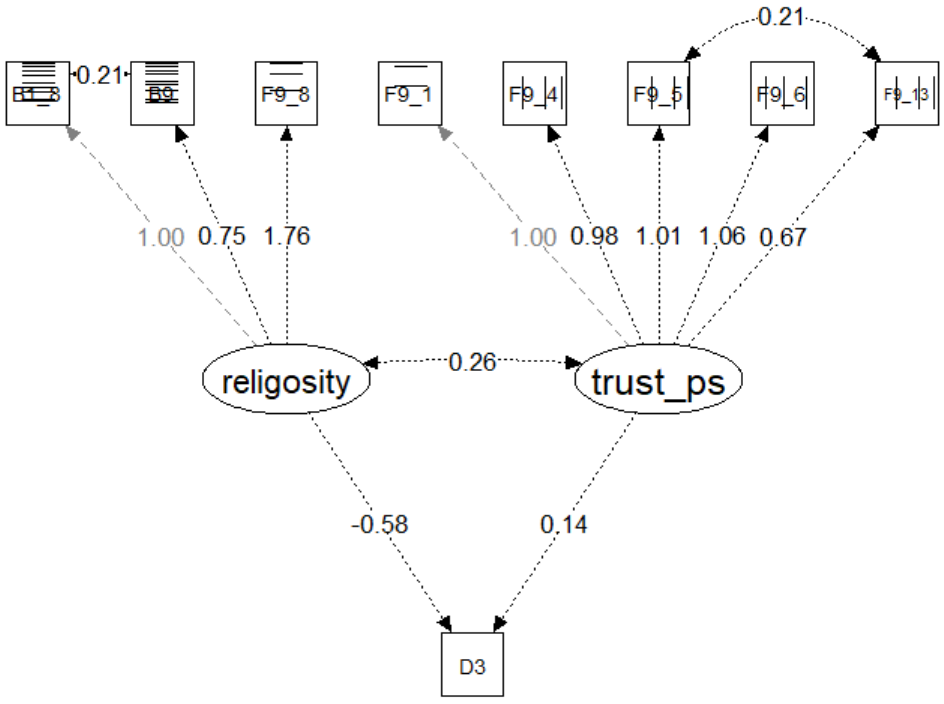

Figure 7: Religiosity and Trust in PS institutions effect on desire to leave Slovenia