Principles of Data Science Coursework
Emanuele AM Cappella, student 160013431
Word count (without references): 2703 words

"Cross-cultural environmental concerns"

1) Analysis Domain, Questions, Plan

1.1 Data

Data was downloaded from the World Values Survey (WVS) website (www.worldvaluessurvey.org). The WVS is a project that has been investigating human belief and values - from environmental concern to religion and political beliefs - for the last 30 years. The dataset I will analyse is a cross-sectional selection derived from the much larger WVS complete database, which spans from 1981 to 2014. My data belong to the 2005 wave and can boast 83975 respondents. Each WVS variable has been measured by multiple survey questions. Socio-demographic variables were also collected: marital status, gender, age, education, employment status, socioeconomic status.

The domain I would like to focus on is Environmental Concerns (EC). EC has been measured in the interview by a set of 9 items on a four-step Likert Scale. These items are thought to load on three different facets of EC: concerns about one's own community - investigating satisfaction about water quality, air quality and sanitation; concerns about the world at large, exploring fears about global warming, loss of biodiversity and ocean pollution. The last dimension is Willingness to Pay (WTP), with the following items: "I would give part of my income if I were certain that the money would be used to prevent environmental pollution", "I would agree to an increase in taxes if the extra money were used to prevent environmental pollution", and "The Government should reduce environmental pollution, but it should not cost me any money".

1.2 Questions & Analysis Plan

My project will focus on two major research questions:

1.2.1 What is the factorial structure of Environmental Concerns?

A. Is the three-factor model implicitly proposed by the structure of items or we can assume a one-factor structure?

To answer this question, I will investigate the factorial structure of EC with Exploratory Factor Analysis which I will then test with Confirmatory Factor Analysis.

I will use Factor Analysis instead of Principal Component Analysis (PCA) because in this case I am more interested in exploring the latent structure of EC rather than obtaining a dimensionality reduction. I will also test the reliability (Cronbach alpha) and validity of the questionnaire.

B. Does the EC structure differ for different countries?

I am interested in investigating the structure of environmental concern in different countries. In fact, it seems that these questions have been generated mostly with western countries in mind. I think that the environmental awareness may differ markedly across countries. For this reason, I will select a subset of three countries - one from Europe (Norway), one from Africa (Ethiopia), and one from Asia (India) – and test the changes in the factorial structure found for the whole dataset.

1.2.2 How much do income and age predict WTP for environmental concerns?

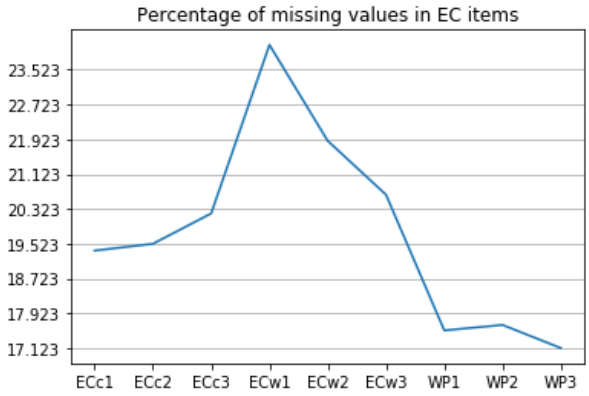
A. Regression analysis
For this analysis I will consider the sub-domain of Willingness to Pay (WTP). WTP describes how much a person is prepared to sacrifice in terms of money to act upon environmental concerns. In my view, WTP is a potentially interesting concept for both governments and environmental no-profit organizations.
I will test the role of two socio-demographic variables in predicting WTP: age and income. I will inspect regression assumptions, compare the fit of different models, and interpret the model parameters of the best-fitting model, whose results will be plotted.

B. Missing data analysis
I would also like to compare the results obtainable in the regression analyses with a simple listwise deletion against a multiple imputation procedure. Hopefully, this step will improve the explicatory power of my model.

2) Analytical Process

2.1 Data pre-processing

I focused on two pre-processing issues:
1. I recoded the EC reversed items for the reliability estimate, and computed total scores for EC variables.
2. Missing values imputation, as the percentage of missing values in this database is quite high for many key values. For instance, the missing proportion for EC variables varies between 17% and a peak of 24%. I will detail this step more thoroughly in the section of regression analysis.



2.2 Tools

I mostly used R in this assignment, because of the richness of its libraries. I used the following packages:

- *mice* and *VIM* dealt with the imputation of missing values
- *ggplot2* for visualizing the regression interaction
- *lavaan* estimated the Confirmatory Factor Analysis
- *nFactor* for drawing the Scree Plot

2.3. Data analysis

2.3.1 What is the factorial structure of Environmental Concerns?

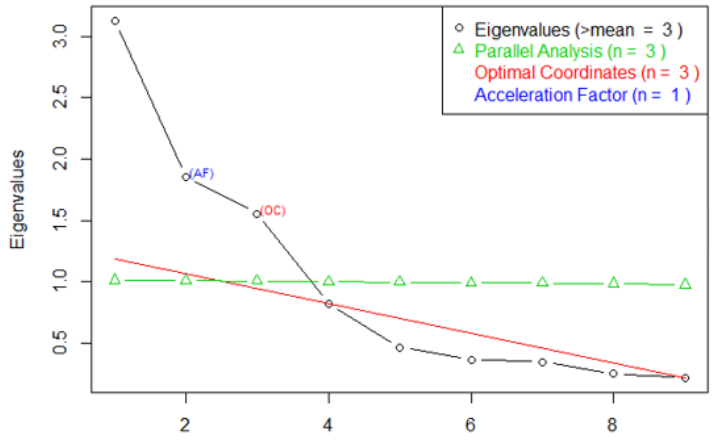
A. Exploratory Factor Analysis (EFA)

First, I computed a correlation matrix for the nine items of EC. Correlations seem generally low, especially between items of different subscales. The strongest relationship exists between the envcom items, while the item wtp3 do not seem to share a strong relationship with any of the others.

	envcom1	envcom2	envcom3	envwrld1	envwrld2	envwrld3	wtp1	wtp2	wtp3
envcom1	1.00000000	0.76213105	0.772213101	0.17679930	0.17793892	0.20903984	0.09482112	0.06575280	-0.12318480
envcom2	0.76213105	1.00000000	0.741881773	0.22988964	0.22697895	0.26345870	0.09853998	0.06765034	-0.11713305
envcom3	0.77221310	0.74188177	1.000000000	0.17677145	0.20022210	0.24028690	0.09131556	0.06262834	-0.13972561
envwrld1	0.17679930	0.22988964	0.176771450	1.00000000	0.58211306	0.53847374	0.11681536	0.08626154	-0.01201744
envwrld2	0.17793892	0.22697895	0.200222097	0.58211306	1.00000000	0.62459986	0.13997907	0.12664323	-0.01578177
envwrld3	0.20903984	0.26345870	0.240286898	0.53847374	0.62459986	1.00000000	0.09637610	0.07713695	-0.02765176
wtp1	0.09482112	0.09853998	0.091315557	0.11681536	0.13997907	0.09637610	1.00000000	0.63713726	0.23987031
wtp2	0.06575280	0.06765034	0.062628337	0.08626154	0.12664323	0.07713695	0.63713726	1.00000000	0.27021132
wtp3	-0.12318480	-0.11713305	-0.139725608	-0.01201744	-0.01578177	-0.02765176	0.23987031	0.27021132	1.00000000

Next, I calculated a series of exploratory models. My hypothesis is that the structure of the data should reflect three distinct factors which correlate, even if lowly, as shown by both the scree plot and the parallel analysis in Figure 1.

Figure 1. Scree Plot & Parallel Analysis



I thus computed an EFA with three factor and oblique rotation ('promax' method), and compared in terms of model fit against alternative models.

Table 1. EFA with three factors

	Factor 1	Factor 2	Factor 3
envcom1	0.908		
envcom2	0.844		
envcom3	0.868		
envwrld1		0.713	
envwrld2		0.835	
envwrld3		0.756	
wtp1			0.757
wtp2			0.850
wtp3	-0.159		0.339

Again, wtp3 was the only problematic item, because it loaded on two different factors showing low discriminant validity. Moreover, the model fit is unsatisfactory: $\chi^2(df) = 368.74$ (12), $p < .001$. To obtain a better fit to the data, I tried other solutions, varying both the number of factors (between 1 and 4) and the number of items (deleting wtp3), but was unable to find a good fit. I chose to accept the model with three factors, without deleting wtp3, because it is the solution more in line with the structure of the questionnaire.
Finally, I calculated Cronbach alpha, whose 0.72 value attested an acceptable reliability rate for the scale.

B. Confirmatory Factor Analysis (CFA)
I tested the structure of the three-factor model with CFA, and compared it against the two-factor and one-factor models. The models were evaluated in terms of four statistics, whose indicative cut-off has been reported inside brackets in Table 2: Root Mean Square Error of Approximation, Standardized Root Mean Square Residual, Comparative Fit Index (CFI), and Tucker-Lewis Index (TLI).

Table 2. CFA models

	RMSEA (<.08)	SMRM (<.05)	CFI (>=.95)	TLI (>=.95)
3 Factors	**0.05**	**0.046**	**0.984**	**0.975**
2 Factors	0.191	0.141	0.746	0.648
1 Factor	0.238	0.174	0.585	0.447

The three-factor model is confirmed to be the best one, with a good-enough fit testified by all the parameters reported in Table 3.
Afterwards, I tested the best fitting model separately for three different countries, one from Europe (Norway), one from Asia (India) and one from Africa (Ethiopia).

Table 3. CFA – comparison of different countries

	RMSEA (<.08)	SMRM (<.05)	CFI (>=.95)	TLI (>=.95)
Norway	0.037	0.032	0.992	0.988
Ethiopia	0.121	0.077	0.933	0.899
India	0.066	0.043	0.965	0.948

Table 3 shows clearly that the model varies significantly between different countries. For instance, it is perfect for Norway, whose item wtp3 incidentally loads perfectly on factor 3 (item load = 0.67). It is adequate for India, but in its case item wtp3 is completely uncorrelated with factor 3 (item load = 0.003). On the other hand, it is very unstable in the case of Ethiopia, because all the items do not load very well on their factors.

2.3.2 Do income and age predict willingness to pay (WTP)?

My hypothesis is that WTP is influenced mainly by age and income: on the one hand, I think that the younger you are, the more willingly you would pay for environment protection; on the other, I expect a linear relationship between income and WTP, that is the richer you are, the more easily you may pay. I will also test the eventual interaction between the two predictors.

A. Data exploration

First, concerning the nature of the data:

- age was measured in years;
- income was measured with a Likert item, ranging from 0 to 10, so it is clearly an ordinal variable. The item asked the following question: 'On this card is a scale of incomes on which 1 indicates the "lowest income decile" and 10 the "highest income decile" in your country. We would like to know in what group your household is';
- willingness to pay is a total score of three Likert items.

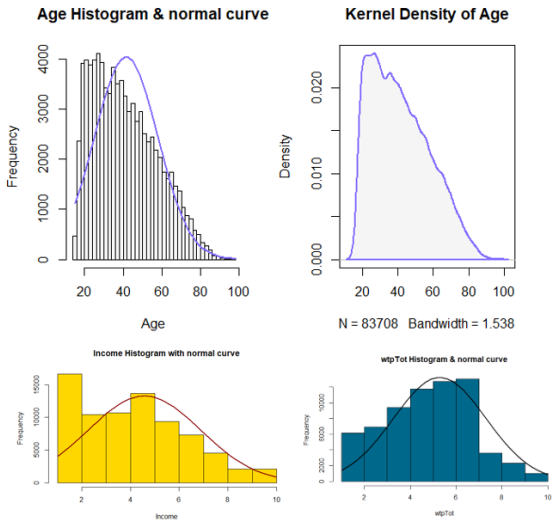
Age spans from 15 to 98 years, while the two ordinal variables have a much less variability, as can be seen from their range and skewed distribution. The distribution of age approaches normality, even if it is skewed to the left.

WTP is positively associated with income (r = 0.12), and negatively correlated with age (r = -0.06). Though these coefficients are very low, all the relationship are significant at the p< .001 - but this is mainly attributable to the huge sample.

Table 4. Descriptive Statistics

	Mean (SD)	Median	Min-Max	Skewness	Kurtosis
Age	41.5 (16.5)	39	15-98	0.51	2.42
Income	4.57 (2.31)	5	1-10	0.3	2.42
WTP	5.28 (2.02)	5	1-9	-0.19	2.7

Figure 2. Distribution of regression variables

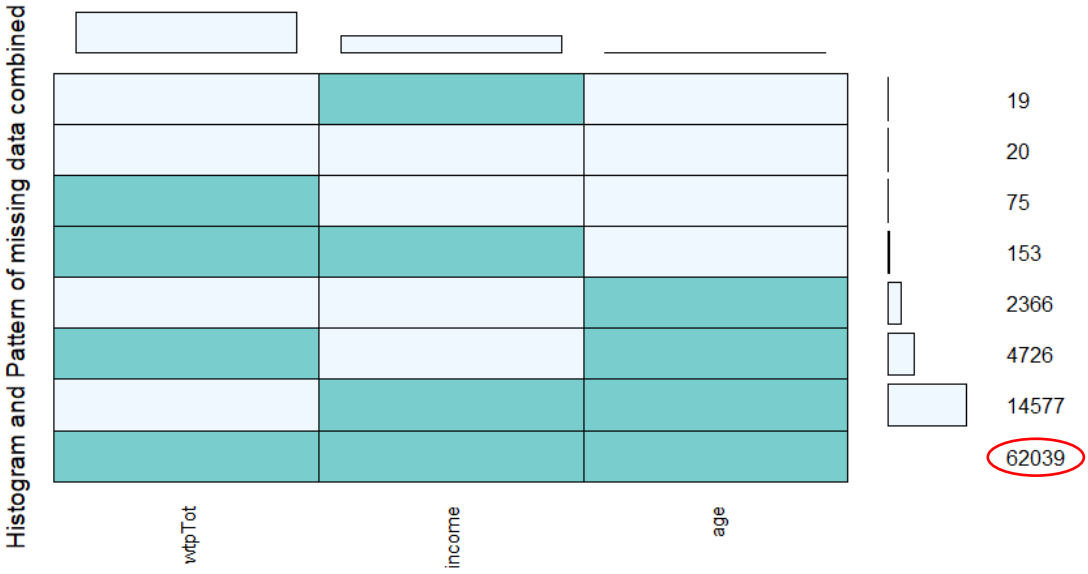


B. Missing values visualization

Figure 3 shows that most missing values come from WTP. A remarkable 25.9% of dataset rows have at least one value missing in one of the five features, that is only 74.1% (62039/83975) of my data is complete. Should I use listwise deletion, I would give up this proportion.

To understand better the issue, in the following section I will compare two multiple regression models, the first one with listwise deletion, the second one with multiple imputation.

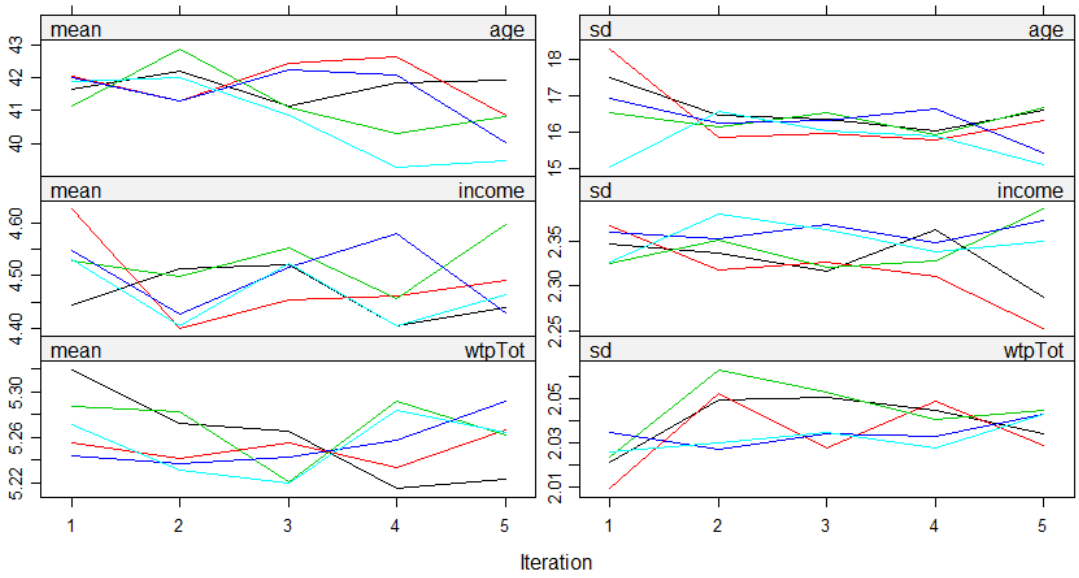
Figure 3. Missing data visualization



C. Data imputation

To study how convergence was reached, I visualized the five iterations the algorithm produced. The desired output of the following Trace Line plots consists in five overlapping lines without any evident trend on convergence (Buuren, S., & Groothuis-Oudshoorn, K. 2011). In this case, the results seem fine.

Figure 5. mice imputation



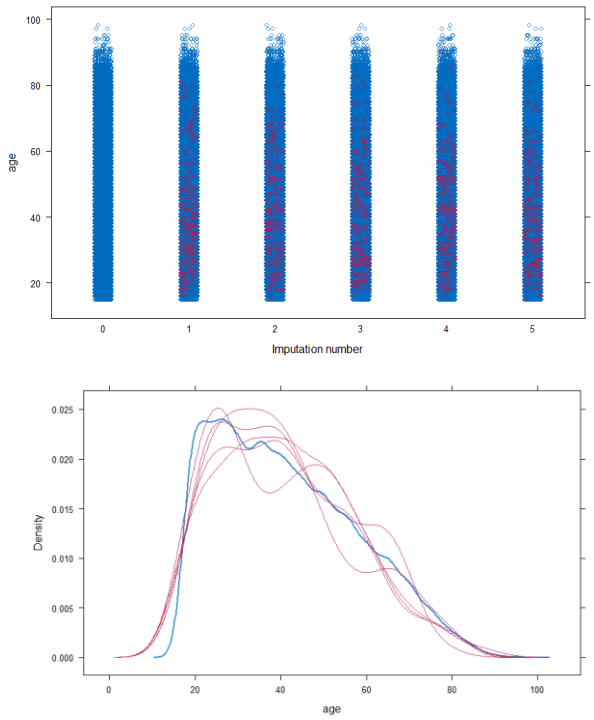


Figure 6. Age imputed distribution

Imputed data should be compatible with what has been observed. Red dots/lines represent the imputations, the blue ones the observed values. It seems that most imputation take place for lower age values.

D. Multiple Regression Models

In the following table I show the pooled results of the five estimated models.

The model is highly significant, though this should not be a surprise, considering the extensive sample. However, the percentage of willingness to pay that age and income explain together is not particularly high (~19%).

Table 6. Multiple Imputation Regression model

	estimate	se	t (df)	p	R^2
Intercept	5.45	0.06	94.75 (17.24)		**0.019**
Age	-0.01	0.001	-12.89 (21.73)	< .001	
Income	0.015	0.01	1.18 (12.15)	0.26	
Age*income	0.002	0.0002	7.92 (14.25)	< .001	

I then compared the pooled model against the model obtained thorough listwise deletion.

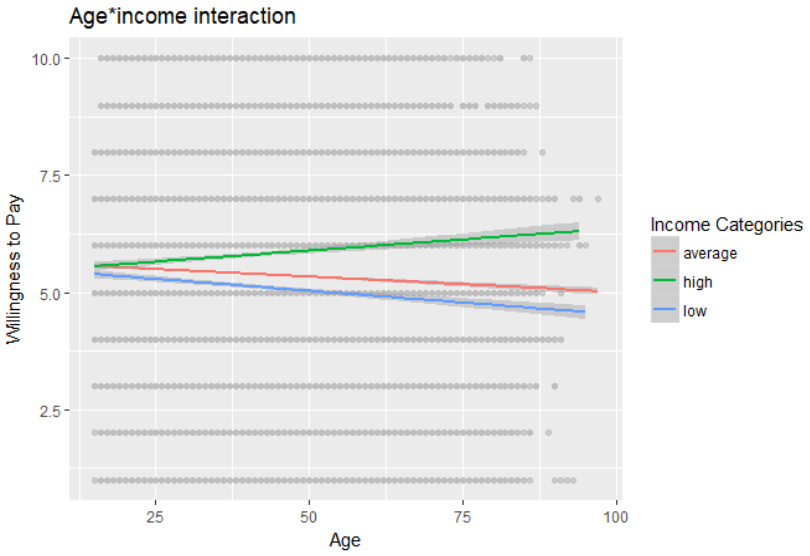
Table 7. Listwise Deletion Regression Model

	estimate	se	t	p	R^2
Intercept	5.56	0.06	93.2	< .001	**0.018**
Age	-0.016	0.001	-12.15	< .001	
Income	0.005	0.011	0.45	0.65	
Age*income	0.02	0.0003	9.31	< .001	

It seems that the listwise model is performing as good as the one with multiple imputation: both models find an interesting interaction between age and income, and have similar predictive power (R^2 = 0.019 vs 0.018).

We can see from the estimates of the regression model, that there is a clear effect of age on WTP, whereas income do not seem to influence the dependent variable. However, the main result concern the interaction between the two independent predictors. This interaction indicates that the effect of age on WTP is moderated by income. Without the interaction term the fit of the regression model significantly worsens (F= 338.82, p<. 001).

To understand better this relationship, I created a graph with ggplo2. As income would be a nominal moderator, its effect on age would be difficult to visualize. Therefore, I decided to split income into three categories: 'high' (> 1 sd over the mean, N= 5900), 'average' (±1 sd from the mean, N= 33009) and 'low' (<1 sd, N= 7791).



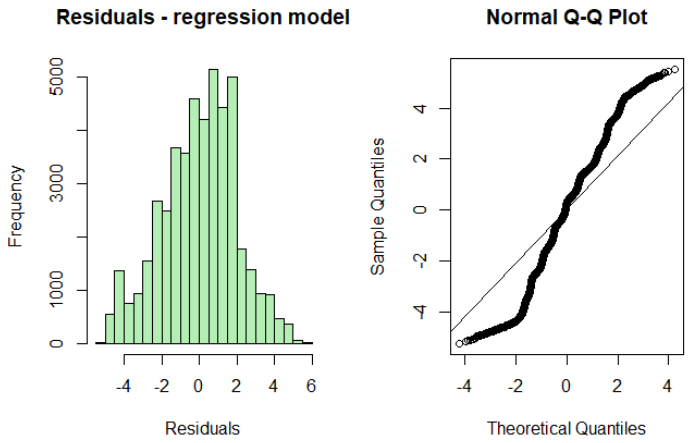
The graph shows that WTP tends to decrease as age increases. However, this insight is not correct for people with high incomes. In this case, as age increases, WTP shows a marked surge. Thus, income is a moderator of the age effect on willingness to pay, and we can visualize this interaction from the different slopes of the three regression lines, corresponding to different levels of income.

E. Checking Regression Assumption

I checked the underlying assumptions of the regression model to test its generalizability (Field 2012).

- *Linear distribution of residuals*

The graphs display a problem with the skewness of residuals distribution. This may be due to the quite unbalanced distribution of the three variables, especially the two ordinal predictors.



- *Independence of errors*

The value of Durbin–Watson test is close to 2 (1.43), confirming the assumption.

- *Multicollinearity*

As previously reported, the correlation coefficient between the two predictors is significant but quite low (r = -0.09, p<.001), so multicollinearity should not be a problem. In fact, the Variance Inflation Factor (VIF) values are below the 10 cut-off - except for the interaction term, which however is unaffected by problems of multicollinearity (Allison, 2012).

3) Findings and Reflections

3.1 EC Questionnaire

First, I was able to confirm the three-factor structure of the EC questionnaire. From my analyses a problem concerning the only negative-worded item (wtp3) emerges, as it does not load well on any factor.
A second interesting aspect to consider regarding wtp3 is that it is the only item that specifically refers to the 'Government'. The use of this term may trigger differences in the responses depending on the perceived effectiveness of the government. I think this is a second important reason for its inefficacy, which may partially explain why for some countries (like Norway) the item did not cause any problems, while in others (like India) it does not fit well at all.
Considering these findings, my recommendation would be:
- avoid any reverse items, because they complicate the factorial structure of the questionnaire and may increase the risk of misinterpretation;
- the reference to 'Government' may raise issues that are in my view partially separate from handling environmental problems. I would change the wording of wtp3 or create a separate scale around it.

*3.2 Age*income*

The main finding of the regression analyses is the interaction between age and income. This insight may suggest appropriate marketing strategies for different groups based on two simple socio-demographic variables. In fact, the two groups more responsive to environmental themes are tendentially younger people regardless of income and older people with higher incomes.
Further analyses should concentrate in finding a more accurate age threshold to understand after what age people are less willing to pay for environmental protection.
Moreover, it is interesting to note that the income variable does not reflect real income, but rather the perception of income related to one's own country levels. I would suggest to test whether this relationship holds even for real income estimates and to create a more robust scale of income perception, adding more items.

3.3 Listwise deletion

Apart for a problem of skewness reflected in both the actual data and their residuals, the regression model computed after listwise deletion seemed solid in terms of results and assumptions, and had more or less the same predictive power of the imputed one. Another important issue to consider when choosing how to handle missing data is the computationally expensive and time-consuming nature of the imputation procedure, especially with large databases like the current one.
In conclusion, as Paul Allison (2014) wrote: "*If listwise deletion still leaves you with a large sample, you might reasonably prefer it over maximum likelihood or multiple imputation*".

3.4 Ordinal scales

The World Value Survey format includes many different Likert scales investigating a wide range of issues. However, all such scales are very short. The result is a series of ordinal variables with too few levels. I would suggest either to increase the number of items per scale, so that the total score may border on being a continuous variable like age; or to use Visual Analogue Scales for key topics, which treats items as continuous dimensions.

References

- Allison, P. (2012). When Can You Safely Ignore Multicollinearity? https://statisticalhorizons.com/multicollinearity
- Allison, P. (2014). Listwise deletion: it's NOT evil. https://statisticalhorizons.com/listwise-deletion-its-not-evil
- Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, *45*(3).
- Field, A., Miles, J., & Field, Z. (2012). Discovering Statistics Using R. Sage Publications
- Oberski, D. L. (2014). lavaan. survey: An R package for complex survey analysis of structural equation models. Journal of Statistical Software, 57(1), 1-27. https://cran.r-project.org/web/packages/lavaan/lavaan.pdf
- Prantner, B. (2011). Visualization of imputed values using the R-package VIM.
- R blogger post, Imputing missing data with R 'mice' package https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/
- Sauer, S. (2017) Visualizing Interaction Effects with ggplot2 https://sebastiansauer.github.io/vis_interaction_effects/
- Vink, G. & van Buuren S. (2017) - Ad hoc methods and mice (online Vignettes) https://gerkovink.github.io/miceVignettes/Ad_hoc_and_mice/Ad_hoc_methods.html https://gerkovink.github.io/miceVignettes/Convergence_pooling/Convergence_and_pooling.html