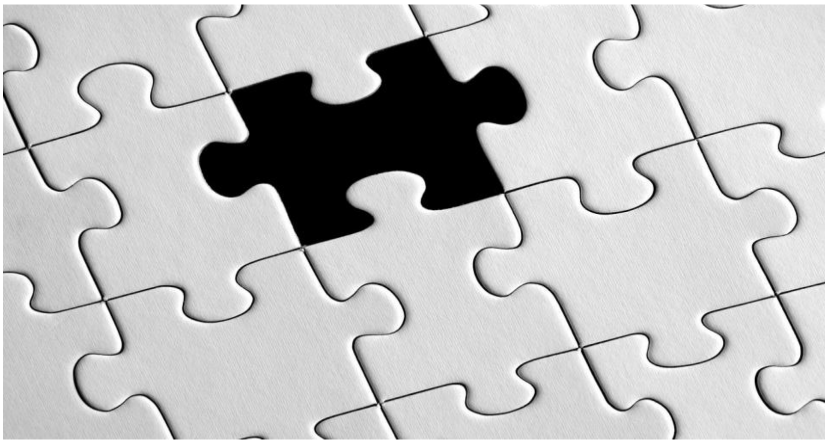
Methodological Review SGI

Treatment of Missing Values

Dr. Margit Kraus, Julia Schmidt



Missing values – why do we care about them?

“Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest” (Kang 2013). Ignoring missing data might reduce the representativeness of the sample and generate biased estimates (Rubin 1987, Schafer 1997). Furthermore, ignoring missing data leads to a loss of information which inturn might decrease statistical power and increase standard errors. In the case of the SGI, the final index cannot be calculated at all if missing data are not imputed – at least not without changing the aggregation procedure for the respective countries. Applying the latter case would introduce a different kind of bias, as it means having different weights for countries with missing data. Albeit a rare phenomenon in databases of OECD and EU countries, the SGI identifies missing data in particular in three categories of the Political Performance Index (P6, P9, P13) as well as in the Governance Index in the category of media accountability (G11). Before a data set with missing values can be analyzed by statistical procedures, it needs to be edited in some way into a “complete” data set.

After testing several methods and implementing cross-validation checks, we decide to impute missing values by full information maximum likelihood estimations (FIML) as recommended by the EU Commission (OECD/EU/JRC, 2008). We chose a maximum likelihood estimation because it is a comprehensive, well-designed imputation method. The FIML approach was first introduced by Hartley and Hocking (1971). Unlike multiple imputation, FIML does not directly impute any missing data. It estimates parameters directly using all the information that is already contained in the incomplete data set. FIML is easy to reproduce since it requires fewer decisions on the calculation process and – contrary to multiple imputation – it produces deterministic results everytime you run the model.[1] For our data, it produced the best fitting values when we compared it to the original data in our cross-validation checks with known values.

[1] Further readings on comparing maximum likelihood estimations with multiple imputation read Allison (2012), Mazza et al (2015).

Steps to identify a suitable methodology

There exist various methods in the literature to address the problem of missing data in a composite indicator (OECD/EU/JRC 2008). In order to choose the best possible option for the SGI, we tested and compared hotdeck methods, single imputation via OLS as well as a Full Information Maximum Likelihood Estimation (FIML). We did not consider a listwise or pairwise deletion of missing values since we need all countries in the process of evaluation of our index.

Pre-imputation steps

Before starting with imputation, it is recommended to identify proportions and patters for missing data (Rubin 1976). According to Schafer (1997), a missing rate of 5% or less is inconsequential. Bennett (2001) argues that statistical analysis is likely to be biased when more than 10% of data are missing. In the case of composite indicators, the OECD/EU/JRC (2008) recommend to have at least 65% of the countries with valid data at the indicator level. At the country level, at least 65% of the indicators should have valid data. In our case, missing data is a rare phenomenon since we work with OECD and EU countries (see Table 1).

Table 1: Missing data per indicator in the SGI database

Indicator	Country with missing data
P6.2 Tier 1 Capital Ratio	New Zealand
P9.2 Spending on Health Programs	Chile, Malta
P9.5 Perceived Health Status	Mexico
P13.3 Personal security	Malta
G11.2 Newspaper Circulation	Cyprus, Malta

Moreover, we assessed the distribution of missing data and identified the type of "missingness" (Rubin 1976). In the case of the SGI, the data is either missing completely at random (MCAR), indicating that "missingness" is not related to any other variable, or missing at random (MAR), indicating that it is possible to control for the factors of "missingness" (OECD/EU/JRC 2008).

Hotdeck methods

Hotdeck methods substitute missing values with the value(s) of similar countries. In order to do so we calculate a dissimilarity matrix for each country. This matrix specifies input variables and distance (dissimilarity) measures for each country. We apply different dissimilarity measures to our database (Euclidian Distance, City Block/Manhattan Distance) by using the Stata[R] command *matrix dissimilarity matname = [varlist] [if] [in] [,options].* On the basis of the dissimilarity matrix, we find the nearest neighbour for each country by identifying the smallest value of distance measure among all other countries. In a last step, we assign the imputed value of our variables to each country as the value belonging to the country's nearest neighbour. For our test variable P6.2 Tier I capital ratio (FSI2), the results vary strongly compared to the actual values in our database (see table 4 in the appendices). We therefore decide not to use imputation by hotdeck methods.

Single imputation via OLS

In the category of single imputation methods, we test OLS regression models. In order to illustrate the way we construct the different models, we use data for the category P6 Global Financial Systems. The indicator P6.2. Tier I capital ratio (FSI2) has missing data for all time series for the country New Zealand. We impute the missing values on the basis of the SGI 2019 database (see table 2, p. 4). Here, we test four models. Each of the models include a different combination of the chosen explanatory variables (FSI1, BCAR, ZSCORE, NPL).[2] In order to test the quality of the chosen models, we conduct a series of tests.

First, we apply a Tukey/Perigbon linktest in order to test the specification of the model. It is based on the idea that if a regression is properly specified, there should be no additional independent variables that are significant except by chance. If our models are specified correctly, then the prediction squared would have no explanatory power. For the SGI 2019 data set, the linktest proved a correct specification of all four models. In addition, we apply the Ramsey Reset test as a functional form test. It tests for the null hypothesis that the model is properly specified. A significant F-statistic would suggest some kind of functional form problem. In our case, none of the models in the datset of 2019 have significant F statistics. This changed naturally when we included all years in the analysis, extending the database to a larger time period, due to the presence of autocorrelation and atypical influences in years of financial crises. Variance inflation factors did not indicate any multicollinearity in the data. When testing for the assumption of a normal distribution, we could identify countries such as Estonia, Iceland, Cyprus, Greece and Luxembourg as outliers in the sample. Including these five countries in the sample provides problematic results for normality as well as heteroscedasticity tests. We therefore decide not to include these five countries when estimating the missing value for New Zealand. When plotting a kernel density function as well as the standardized normal probability and the quintile normal probability, we did not detect significant outliers. This was also confirmed for outliers in the residuals analysing the inner and outer fences of the interquartile ranges. The Shapiro-Wilk test for normal data attested a normal distribution. Finally, heteroscedasticity, skewness and kurtosis were tested using Cameron &Trivedi's decomposition of IM-test as well as the Breusch-Pagan/Cook-Weisberg test for heteroscedasticity. In the case of FSI2, model 4 predicts the best results for the SGI data 2019. The quality of the model is confirmed by all tests conducted with the SGI data set 2019.

[2] For an overview of the regression results of all models and their specification parameters see tables 7 and 8 in the appendix.

Table 2: Regression diagnostics OLS for 2019 data (no outliers)

Regression Diagnostics					
		Model 1	**Model 2**	**Model 3**	**Model 4**
Specification, Omitted Variables, Multicollinearity					
Tukey/Pregibon Linktest, Signifikanz hat²	Prob > |t|	0,1130	0,0940	0,2210	0,1380
Ramsey Omitted Variables Test	Prob > F	0,1882	0,2243	0,4485	0,1621
Variance Inflation Factors > 5	y/n	no	no	no	no
Outliers, Influential Observations					
Leverage > 3k/n	n_obs	1	5	3	3
Cooks Distanzmaß > 4/n	n_obs	1	1	3	2
Distance between fitted values (DFITS) > 2√(k/n)	n_obs	1	0	2	1
Residuals - Normality, Heteroscedasticity					
IQR, Severe Outliers in Residuals	n_obs	0	0	0	0
Shapiro-Wilk Normality test	Prob > z	0,1603	0,0282	0,1104	0,4279
Cameron & Trivedi's decomposition of IM-test	Prob > Chi²	0,5471	0,4167	0,2517	0,6674
Breusch-Pagan / Cook-Weisberg heterosk. test	Prob > Chi²	0,4640	0,5535	0,5434	0,4316

Note: FSI1 = Bank regulatory capital to risk-weighted assets ratio in % (IMF), BCAR = Bank capital and reserves to total assets ratio in % (World Bank), ZSCORE = Bank Z Score, probability of default of a country's banking system (World Bank), NPL = Non-performing loans (IMF)

For reasons of consistency, we did impute the missing values for the full database 2008-2018 (see table 8 in the appendices). However, due to trends in the data, tests do attest a biased distribution, heteroscedasticity and multicollinearity. We therefore decided, to calculate the missing data separately for each year missing on the basis of the above outlined values.

Full Information Maximum Likelihood Estimation

Full information maximum likelihood function adjusts the likelihood function so that each case contributes information on the variables that are observed. It is mostly used in the case of linear structural equation models. This method follows the assumptions that there is multivariate normality in the data and that the data is either missing completely at random or missing at random. We calculated the FIML using the *sem* command in Stata^{R} for structural equation modelling for all four models we identified in the OLS regression section (Medeiros 2016). In a second step we conducted the Pearson X^2 test for goodness of fit. It tests the observed against the expected number of responses using cells defined by the covariate patterns.

In order to estimate the values for the indicators that have missing values for some countries, we use the explanatory variables as outlined in table 3. Similar to our example FSI2, several models have been tested in order to find out the best quality of the models.

Table 3: FIML imputation with explanatory variables for missing value

Indicators with missing values	**Explanatory Variables**
P6.3 Tier I Capital Ratio	Bank regulatory capital to Risk-weighted assets ratio in % (IMF) Bank capital and reserves to total assets ratio in % (World Bank) Bank Z Score (World Bank)

P9.5 Perceived Health Status Quintile Ratio	Healthy Live Years (WHO)) Live Expectancy (Eurostat, OECD) Life Satisfcation (World Happiness Report) Health Expenditures (Eurostat, OECD) Income Quintile Share Ratio, Poverty Rate (Eurostat, OECD)
P13.3 Personal Security	Reliability of Police Services (WEF) Property Rights (WEF) Homicide Rate (UNODC) Further indicators on crime (UNOCD)
G11.2 Quality of Newspapers	Total paid-for daillies per 1,000 inhabitants (World Press Trends) Population with at least upper secondary attainment, age 25-64 (Eurostat) Individuals using the Internet (% of population) (World Bank)

The cross-validation check for FIML provided the exact same results as the best OLS model (see table 5 in the appendices). Under the assumptions outlined above, the OLS estimator with the correct vairables and the correct specified model is BLUE and thus equals the FIML. For consistency we calculated OLS models as well as FIML estimates for all indicators with missing values. We then used the FIML estimates to impute the missing values (see table 9 in the appendices).

Limitations

Our approach using FIML is still based on certain assumptions that come along with full information maximum likelihood models. The linear structural equation models are based on available data for explanatory variables. Although we have tested extensive subsets of possible explanatory variables and the models have proved to be correctly specified, no imputation method is able to reproduce precise values without imposing certain structures on the data. Facing the dilemma of on the one side aiming at a scientific sound and comprehensive imputation of missing data, and on the other side a method that can be communicated to a wider public we decided to follow the above outlined process.

Literature

Allison, P. (2012). Handling Missing Data by Maximum Likelihood. SAS Global Forum 2012, Paper 312-2012 https://statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf, accessed 14.12.2018

Bennett, D. A. (2001). How can I deal with missing data in my study? Australian and New Zealand Journal of Public Health, 25(5), 464-469. doi:10.1111/j.1467-842X.2001.tb00294.x

Hartley, H., & Hocking, R. (1971). The Analysis of Incomplete Data. Biometrics, 27(4), 783-823. doi:10.2307/2528820

Kang, H., (2013). The prevention and handling of the missing data. The Korean Journal of Anesthesiology. 64(5): 402–406.

Mazza, G. L., Enders, C. K., & Ruehlman, L. S. (2015). Addressing item-level missing data: A comparison of proration and full information maximum likelihood estimation. Multivariate Behavioral Research, 50(5), 504-519. doi:10.1080/00273171.2015.1068157

Medeiros, R. (2016). Handling Missing Data in STATA: Imputation and likelihood-based approaches, https://www.stata.com/meeting/switzerland16/slides/medeiros-switzerland16.pdf , accessed 14.12.2018

OECD/EU/JRC (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*, Éditions OCDE, Paris, https://doi.org/10.1787/9789264043466-en.

Rubin, D. (1976). Inference and Missing Data. Biometrika, 63(3), 581-592. doi:10.2307/2335739

Schafer, Joseph L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Woolridge, Jeffrey M. (2012). Introduction to Econometrics: Missing Data and Nonrandom Samples, p. 324 – 332

Contacts

Dr. Margit Kraus

Calculus Consult

Email: kraus@calculus-consult.com

Julia Schmidt

Program Shaping Sustainable Economies, Bertelsmann Stiftung

Email: julia.schmidt@bertelsmann-stiftung.de

Appendices

Table 4: Cross-validation Check Hotdeck Method

		L1 (Euclidian Distance) Imputed Value			L2 (Manhattan Distance) Imputed Value		
	FSI2	FSI1 BCAR	FSI1 ZSCORE	FSI1 BCAR ZSCORE	FSI1 BCAR	FSI1 ZSCORE	FSI1 BCAR ZSCORE
AUS	12,33	14,47	14,47	14,47	10,86	14,47	14,47
AUT	15,31	16,77	16,77	16,77	16,38	16,77	16,77
BEL	16,38	15,72	16,55	15,72	15,31	16,55	15,72
BGR	19,75	20,09	20,09	20,09	18,79	20,09	20,09
CAN	13,21	12,33	14,67	12,33	14,35	14,67	12,33
CHL	10,42	13,11	15,78	13,11	10,86	15,78	13,11
HRV	21,66	20,09	20,09	20,09	22,62	20,09	19,75
CYP	14,67	13,11	13,21	13,11	14,19	13,21	13,11
CZE	17,12	14,91	14,91	14,91	15,31	14,91	14,91
DNK	18,85	16,77	15,20	16,77	17,10	15,20	16,77
EST	28,24	20,09	20,09	20,09	22,62	21,66	20,09
FIN	19,62	14,91	17,10	14,91	18,32	17,10	17,10
FRA	15,20	16,77	15,20	16,77	16,77	15,20	16,77
DEU	16,77	16,77	15,50	16,77	15,20	15,20	16,77
GRC	15,78	13,51	13,51	13,51	13,51	13,51	13,51
HUN	15,30	16,23	13,51	16,23	16,23	13,51	16,23
ISL	22,26	21,66	21,66	21,66	21,66	21,66	21,66
IRL	22,62	21,66	23,00	21,66	21,66	23,00	19,75
ISR	10,86	13,63	13,63	13,63	12,33	13,63	13,63
ITA	14,35	13,11	13,11	13,11	13,21	13,11	13,11
JPN	14,91	17,12	17,12	17,12	14,35	17,12	15,72
KOR	13,11	14,35	14,35	14,35	14,47	14,35	14,35
LVA	20,09	17,55	21,66	17,55	19,75	19,75	19,75
LTU	18,79	17,55	17,55	17,55	16,55	17,55	17,55
LUX	24,56	15,31	15,31	15,31	23,00	13,63	13,63
MLT	15,25	16,77	15,20	16,77	17,12	15,20	15,20
MEX	14,19	16,55	15,72	16,55	13,51	15,25	16,55
NLD	18,32	19,06	19,06	19,06	19,62	19,06	19,06
NZL	.	16,77	15,20	16,77	16,38	15,20	16,77
NOR	19,06	18,32	18,32	18,32	18,85	18,32	18,32
POL	16,23	15,30	15,30	15,30	15,30	15,30	15,30
PRT	14,47	12,33	12,33	12,33	13,11	12,33	12,33
ROU	17,55	18,79	18,79	18,79	18,61	18,79	18,79
SVK	16,55	16,38	16,38	16,38	18,79	16,38	14,19
SVN	18,61	17,55	18,79	17,55	16,38	18,79	17,55
ESP	13,44	15,25	15,25	15,25	13,11	15,25	15,25
SWE	23,00	19,62	22,62	19,62	24,56	22,62	19,62
CHE	15,73	16,38	14,47	16,38	15,25	14,47	14,91
TUR	13,51	15,78	15,30	15,78	15,78	15,30	15,78
GBR	17,10	16,23	18,32	19,06	18,85	18,32	18,32
USA	13,63	10,86	10,86	10,86	14,67	10,86	10,86

Table 5: Cross-validation check OLS regression data set 2019 (no outliers)

	Fitted Values				
	FSI2	**Model 1**	**Model 2**	**Model 3**	**Model 4**
		FSI1, BCAR, ZSCORE, NPL	FSI1, BCAR, NPL	FSI1, ZSCORE, NPL	FSI1, BCAR, ZSCORE
AUS	12,33	12,61	12,53	12,71	12,67
AUT	15,31	15,74	16,00	15,73	15,74
BEL	16,38	16,10	16,15	16,19	16,11
BGR	19,75	19,31	19,27	19,12	19,20
CAN	13,21	13,13	13,00	13,38	13,19
CHL	10,42	11,81	11,54	11,76	11,89
HRV	21,66	21,27	21,22	20,89	21,16
CZE	17,12	15,50	15,45	15,56	15,55
DNK	18,85	18,68	18,90	18,75	18,62
FIN	19,62	18,32	18,32	18,56	18,36
FRA	15,20	16,20	16,34	16,42	16,14
DEU	16,77	16,41	16,63	16,55	16,42
HUN	15,30	15,57	15,40	15,49	15,64
IRL	22,62	23,30	23,44	22,92	23,18
ISR	10,86	12,01	12,30	11,97	12,02
ITA	14,35	14,54	14,46	15,14	14,06
JPN	14,91	14,80	14,75	15,02	14,83
KOR	13,11	13,20	13,03	13,14	13,33
LVA	20,09	20,41	20,31	20,32	20,41
LTU	18,79	17,40	17,22	17,18	17,50
MLT	15,25	14,94	15,09	14,99	14,88
MEX	14,19	14,28	14,39	13,94	14,37
NLD	18,32	19,38	19,30	19,64	19,42
NZL	.	16,46	16,64	16,43	16,54
NOR	19,06	19,50	19,41	19,66	19,55
POL	16,23	16,40	16,26	16,30	16,44
PRT	14,47	13,95	14,03	14,16	13,55
ROU	17,55	18,23	18,07	18,26	18,21
SVK	16,55	16,69	16,79	16,35	16,75
SVN	18,61	17,08	16,81	17,11	17,17
ESP	13,44	13,42	13,60	13,43	13,34
SWE	23,00	23,17	23,21	23,41	23,26
CHE	15,73	14,00	14,00	14,06	14,07
TUR	13,51	14,92	14,75	14,64	15,02
GBR	17,10	18,05	17,98	18,12	18,16
USA	13,63	12,96	13,32	12,42	13,08

Table 6: Cross-validation check FIML SGI data set 2019 (no outliers)

	Fitted Values				
	FSI2	Model 1	Model 2	Model 3	Model 4
		FSI1, BCAR, ZSCORE, NPL	FSI1, BCAR, NPL	FSI1, ZSCORE, NPL	FSI1, BCAR, ZSCORE
AUS	12,33	12,61	12,53	12,71	12,67
AUT	15,31	15,74	16,00	15,73	15,74
BEL	16,38	16,10	16,15	16,19	16,11
BGR	19,75	19,31	19,27	19,12	19,20
CAN	13,21	13,13	13,00	13,38	13,19
CHL	10,42	11,81	11,54	11,76	11,89
HRV	21,66	21,27	21,22	20,89	21,16
CZE	17,12	15,50	15,45	15,56	15,55
DNK	18,85	18,68	18,90	18,75	18,62
FIN	19,62	18,32	18,32	18,56	18,36
FRA	15,20	16,20	16,34	16,42	16,14
DEU	16,77	16,41	16,63	16,55	16,42
HUN	15,30	15,57	15,40	15,49	15,64
IRL	22,62	23,30	23,44	22,92	23,18
ISR	10,86	12,01	12,30	11,97	12,02
ITA	14,35	14,54	14,46	15,14	14,06
JPN	14,91	14,80	14,75	15,02	14,83
KOR	13,11	13,20	13,03	13,14	13,33
LVA	20,09	20,41	20,31	20,32	20,41
LTU	18,79	17,40	17,22	17,18	17,50
MLT	15,25	14,94	15,09	14,99	14,88
MEX	14,19	14,28	14,39	13,94	14,37
NLD	18,32	19,38	19,30	19,64	19,42
NZL	.	16,46	16,64	16,43	16,54
NOR	19,06	19,50	19,41	19,66	19,55
POL	16,23	16,40	16,26	16,30	16,44
PRT	14,47	13,95	14,03	14,16	13,55
ROU	17,55	18,23	18,07	18,26	18,21
SVK	16,55	16,69	16,79	16,35	16,75
SVN	18,61	17,08	16,81	17,11	17,17
ESP	13,44	13,42	13,60	13,43	13,34
SWE	23,00	23,17	23,21	23,41	23,26
CHE	15,73	14,00	14,00	14,06	14,07
TUR	13,51	14,92	14,75	14,64	15,02
GBR	17,10	18,05	17,98	18,12	18,16
USA	13,63	12,96	13,32	12,42	13,08

Table 7: OLS regression results and diagnostics for 2019 data set (no outliers)

	Regression Results							
	Model 1		**Model 2**		**Model 3**		**Model 4**	
	FSI1, BCAR, ZSCORE, NPL		FSI1, BCAR, NPL		FSI1, ZSCORE, NPL		FSI1, BCAR, ZSCORE	
	Val. / Coeff.	Prob>F / |t|	Val. / Coeff.	Prob>F / |t|	Val. / Coeff.	Prob>F / |t|	Val. / Coeff.	Prob>F / |t|
adj. R^2	0,9199	0,0000	0,9195	0,0000	0,9162	0,0000	0,9206	0,0000
FSI1	0,8828	0,0000	0,8942	0,0000	0,8862	0,0000	0,8876	0,0000
BCAR	0,1114	0,1310	0,1228	0,0940	--	--	0,1339	0,0530
ZSCORE	-0,0259	0,2900	--	--	-0,0314	0,2060	-0,0295	0,2200
NPL	0,0401	0,3940	0,0488	0,2950	0,0662	0,1440	--	--
_CONS	-0,5391	0,6510	-1,2473	0,2130	0,2958	0,7850	-0,6095	0,6070

Regression Diagnostics					
		Model 1	**Model 2**	**Model 3**	**Model 4**
Specification, Omitted Variables, Multicollinearity					
Tukey/Pregibon Linktest, Signifikanz hat²	Prob > |t|	0,1130	0,0940	0,2210	0,1380
Ramsey Omitted Variables Test	Prob > F	0,1882	0,2243	0,4485	0,1621
Variance Inflation Factors > 5	y/n	no	no	no	no
Outliers, Influential Observations					
Leverage > 3k/n	n_obs	1	5	3	3
Cooks Distanzmaß > 4/n	n_obs	1	1	3	2
Distance between fitted values (DFITS) > 2√(k/n)	n_obs	1	0	2	1
Residuals - Normality, Heteroscedasticity					
IQR, Severe Outliers in Residuals	n_obs	0	0	0	0
Shapiro-Wilk Normality test	Prob > z	0,1603	**0,0282**	0,1104	0,4279
Cameron & Trivedi's decomposition of IM-test	Prob > Chi²	0,5471	0,4167	0,2517	0,6674
Breusch-Pagan / Cook-Weisberg heterosk. test	Prob > Chi²	0,4640	0,5535	0,5434	0,4316

Table 8: OLS regression results and diagnostics for extended data set (2008-2018)

Regression Results								
	Model 1		Model 2		Model 3		Model 4	
	FSI1, BCAR, ZSCORE, NPL		FSI1, BCAR, NPL		FSI1, ZSCORE, NPL		FSI1, BCAR, ZSCORE	
	Val. / Coeff.	Prob>F / |t|	Val. / Coeff.	Prob>F / |t|	Val. / Coeff.	Prob>F / |t|	Val. / Coeff.	Prob>F / |t|
adj. R^2	0,9301	0,0000	0,9296	0,0000	0,924	0,0000	0,9247	0,0000
FSI1	0,9884	0,0000	0,9907	0,0000	1,0253	0,0000	0,9941	0,0000
BCAR	0,1306	0,0000	0,1307	0,0000	--	--	0,1259	0,0000
ZSCORE	-0,0126	0,0990	--	--	-0,0140	0,0690	-0,0216	0,0020
NPL	0,0219	0,0030	0,0268	0,0000	0,0266	0,0000	--	--
_CONS	-2,9491	0,0000	-3,1689	0,0000	-2,5792	0,0000	-2,7334	0,0000

Regression Diagnostics					
		Model 1	Model 2	Model 3	Model 4
Specification, Omitted Variables, Multicollinearity					
Tukey/Pregibon Linktest, Signifikanz hat²	Prob > |t|	**0,0020**	**0,0010**	**0,0030**	**0,0180**
Ramsey Omitted Variables Test	Prob > F	**0,0113**	**0,0052**	**0,0246**	**0,0853**
Variance Inflation Factors > 5	y/n	no	no	no	no
Outliers, Influential Observations					
Leverage > 3k/n	n_obs	25	25	23	21
Cooks Distanzmaß > 4/n	n_obs	25	21	27	23
Distance between fitted values (DFITS) > 2√(k/n)	n_obs	16	18	18	14
Residuals - Normality, Heteroscedasticity					
IQR, Severe Outliers in Residuals	n_obs	**1**	**1**	0	**1**
Shapiro-Wilk Normality test	Prob > z	**0,0001**	**0,0001**	**0,0004**	**0,0000**
Cameron & Trivedi's decomposition of IM-test	Prob > Chi²	**0,0387**	0,0639	0,0562	0,0508
Breusch-Pagan / Cook-Weisberg heterosk. test	Prob > Chi²	**0,0325**	0,0654	**0,0013**	0,0911

Table 9: FIML results for SGI data set 2019 (no outliers)

	Regression Results							
	Model 1		Model 2		Model 3		Model 4	
	FSI1, BCAR, ZSCORE, NPL		FSI1, BCAR, NPL		FSI1, ZSCORE, NPL		FSI1, BCAR, ZSCORE	
	Val. / Coeff.	Prob>Chi² / |t|	Val. / Coeff.	Prob>Chi² / |t|	Val. / Coeff.	Prob>Chi² / |t|	Val. / Coeff.	Prob>Chi² / |t|
LR-Test		0,0000		0,0000		0,0000		0,0000
FSI1	0,8828	0,0000	0,8942	0,0000	0,8862	0,0000	0,8876	0,0000
BCAR	0,1114	0,0930	0,1228	0,0660	--	--	0,1339	0,0320
ZSCORE	-0,0259	0,2450	--	--	-0,0314	0,1700	-0,0295	0,1840
NPL	0,0401	0,3500	0,0488	0,2580	0,0662	0,1110	--	--
_CONS	-0,5391	0,6220	-1,2473	0,1770	0,2958	0,7700	-0,6095	0,5810

BertelsmannStiftung