Paper : 2027-J2SP-2017

SAS® GLOBAL FORUM STUDENT SYMPOSIUM 2017

An Investigation Into Social Factors That May Influence National GDP

Team : John Eacott
Sid Grover
Jayant Sharma
Par Aravazhi

Oklahoma State University

Introduction

This report details analysis that was conducted on the "Worldwide Development Indicators'" dataset, obtained from the World Bank Information Repository. The intent was to identify specific social initiatives that may influence a country's GDP (Gross Domestic Product). GDP is defined as the total of goods and services which are produced within the borders of a country. "GDP is commonly used as an indicator of the economic health of a country, as well as a gauge of a country's standard of living. Since the mode of measuring GDP is uniform from country to country, GDP can be used to compare the productivity of various countries with a high degree of accuracy"[1]. GDP therefore has an important and direct influence on the quality of life within a country.

Problem Definition

The aim of this analysis, was to provide useful insight into how countries, particularly developing countries such as those in South America and Asia, can use social investment programs to grow their GDP.

Strategies for growing GDP can include: fiscal policy, government borrowing and forging trade agreements. However, many developing countries may not have the expertise or infrastructure, to grow their GDP through investment in these areas. Identifying social factors that could influence GDP growth provides countries with alternative information on which to base their investment decisions. Investment in these areas may also be more realistic and attainable for developing countries.

Data

This analysis is based on the Worldwide Development Indicators dataset (WDI) [2]. This was obtained from the World Bank Information Repository, and is publicly available data.

Data was available from time period 1960 to 2016, and includes a large number of different variables, although, many of these could be classified as belonging to the same 'family' of variables. Such as: Education, HealthCare, Sanitation, Energy Consumption, Exports, Imports & Financial metrics. Examples of the types of variables in the dataset are included as Appendix 1.

Data Review (Cleaning/Validation)

The raw data was structured with separate rows for each variable, repeated for each country, covering the time period 1960 – 2016. An example of the raw data is included as Appendix 2. This structure resulted in the dataset containing 380,160 individual measurements, representing each individual country replicated by the number of variables.

Data in the original structure posed two major challenges: 1) the format did not fit the required format for a lot of modeling techniques; 2) many variables had a high percentage of missing data.

To overcome the issues presented by the original data structure, data was transposed centered around the country, with the transpose occurring on the potential predictors and the individual years. This created rows which contained the observed value for all the potential predictor variables for a given country and a given year. The data now had a structure that was appropriate for a variety of modeling techniques. An example of the amended data structure is included as Appendix 3.

The Target variable for the models that were built was '*GDP per Capita in US$*'. This was considered to be the most appropriate outcome variable to use. As it is calculated on a 'per capita' basis, it provides an accurate and meaningful measure of comparison across different Countries.

The new structure had 1,439 columns/variables, in which many columns had large percentages of missing data. Variables that were not considered to be of interest to this particular project, as they were not 'social' type variables, were eliminated from the dataset. This reduced the available variables to 456. Missing values were still an issue, as many modeling techniques will automatically exclude measurements that have missing values for any of the input variables.

Data for years 2000 onwards had lower percentages of missing values. Hence, only data from the year 2000 onwards was used. To overcome issues presented by the remaining measurements, that contained missing data values, it was decided to create a binary outcome variable, based on whether a record was above or below the average value for the Target variable (*GDP per Capita in US$)*. Measurements below the average were coded 0, and those equal to or above the average were coded 1. This enabled each variable to be screened for potential usefulness, using Weight of Evidence (WOE) and Information Value (IV), using this binary variable as a pseudo good/bad outcome. WOE and IV are techniques that can be used to provide preliminary assessment of how effective a variable is in predicting a binary target variable. These techniques can be used even when the candidate variables for prediction contain missing values.

Additionally, this would provide a basis for missing value imputation as missing data values would be classified separately. The first attempt at missing value imputation, was therefore performed using Data Ranges that had an equivalent 'Bad Rate' to the Missing group. Imputation of up to 30% of Data values was considered acceptable.

Variables with low Information Values were reviewed, and Shape transformations applied as necessary. Information Values were then recalculated for these transformed variables. Appendices 4 to 6 provide a visual illustration of the types of data issues encountered.

Analysis

The objective of the analysis performed was to identify the most important social factors for countries to allocate resource and investment towards, in an attempt to influence their GDP. This was considered to be of particular importance for developing countries where effective use of available resources, is most important and will provide the most benefit.

Initial review and modeling was performed on all variables with an Information Value of 1.5, or greater, and with no more than 30% missing values.

Additionally, a random subset of variables with Information Value below 1.5 was selected. The purpose of this was to confirm that the basis for initial selection (based on the Pseudo Good/Bad Outcome) was appropriate. None of these variables were found to be having good predictive power.

All variables were reviewed for Distribution and Shape. Transformations were performed where required. Target variable for our analysis is continuous and Linear Regression was therefore selected for model building.

The data was split into Training and Validation datasets on a ratio of 60% to the Training dataset and 40% to the Validation dataset. This initial stage of model building arrived at what appeared to be a reasonable model. However, a review of Diagnostic Plots such as Residual Plot and Cook's D statistics, revealed patterns in the data that indicated underlying issues that needed to be resolved before proceeding. Further investigation revealed that the imputation process was not sufficiently accurate. This resulted in imputed values that were considerably different and out-of-pattern, compared to values that were populated for the same variable and country. Diagnostic plots using this imputation method are included as Appendix 7.

In an attempt to resolve the issues with imputed values, alternative imputation methods were explored, including Mean value imputation, Median value imputation and imputation by Interpolation.

Separate Linear Regression models were built, using data from all 3 methods of imputation. Different model selection techniques were also used (Forwards, Backwards and Stepwise), Stepwise was used for final model selection. Results were compared to assess similarities in variable selection, and to assess the effect of the three missing value imputation methods used.

Results were similar for the two models built using Mean and Median imputation, but the imputation by Interpolation process produced a different model. All three models were reviewed for Correlations, Interactions and individual data point influence. Review of the Residual Plots and model fit statistics was also performed to identify potential areas of concern. It was concluded that the Mean and Median imputation approaches were resulting in models with less predictive power. Best imputation approach was identified as Interpolation, this produced a model with better predictive ability on both the Training and Validation samples. Diagnostic plots for the model build using Interpolation are included as Appendix 8.

As an alternative basis for comparison, a Decision Tree model was also built to see if a different modeling approach would produce different results. This was built with no imputation performed. The Decision Tree model although slightly different, still had some core similarities with the Regression model built using data generated using imputation by Interpolation.

Variables selected in the regression model using interpolation were:

1) Health Expenditure Per Capita
2) Gross Enrollment Ratio in Secondary Education (Both Sexes)
3) % of the population with access to Improved Sanitation Facilities.

Variables selected in the Decision Tree model were:

1) Health Expenditure Per Capita
2) Adjusted Savings – Education Expenditure (% of GNI).
3) Mortality Rate – Neonatal per 1,000 live births
4) School Enrollment, Tertiary (gross), Gender Parity Index (GPI).

The observed differences between the regression models are believed to be due to the imputation methods used. The differences illustrate the challenges faced in dealing with this dataset, and the impact of differing imputation methods.

Considering the similarity in the regression model using Interpolation for dealing with missing values, and the Decision Tree model, it was decided that these two models would form the basis of our conclusions.

Rather than selecting one model over the other, we have instead used the 'family' of the variables selected to identify the overall direction for providing our recommendations. Both models select variables that relate to Education and variables that can be considered as areas of Social Infrastructure. Both models select Healthcare expenditure. Model fit statistics for the regression model using Interpolation is included as Appendix 9. Model fit statistics for the Decision Tree model is included as Appendix 10.

Conclusions

Useful models were obtained to predict Per Capita GDP based on social factors.

It appears that three main social data elements can be used to predict Per Capita GDP. These are:

1) Health Care expenditure per Capita – Higher investment results in higher Per Capita GDP.
2) Access to Education – Higher participation results in higher Per Capita GDP.
3) Social Infrastructure – Higher investment in this area leads to higher Per Capita GDP.

These areas are recommended as the social investment areas that developing countries should explore, to make effective use of any funds when investing for economic growth.

It is acknowledged that some countries may already incorporate some, perhaps all, of these elements into their economic policies. However, this analysis may have helped to confirm the importance of using the 3 distinct areas mentioned, in conjunction with each other.

Suggestions For Future Studies

At this point, it has been decided not to publish an exact formula for investment level and expected Per Capita GDP growth associated with that. This is mainly due to the data issues encountered, it is believed that additional work would be beneficial to research and establish higher instances of populated data as opposed to relying on missing value imputation to achieve sufficient case volume for modeling purposes.

Additionally, specific Countries could be selected for individual Case Studies. The purpose of this would be to explore and validate on a comprehensive basis the effect of the Predictor variables identified through this analysis. This would help to answer questions such as, whether the variables identified are truly predictive, or whether they correlate to other data which actually predict more effectively, and were not available to in the Data sample used for this Report.

Further areas for future study/analysis include: 1) clustering to identify similarity between countries and how this relates to GDP 2) analysis of how variable importance changes over time.

References

[1] http://www.investopedia.com/terms/g/gdp.asp

[2] http://data.worldbank.org/data-catalog/world-development-indicators

Appendix 1

(Example of available variables)

231	SE_ADT_1524_LT_FM_ZS	Num	8	BEST12.		Literacy rate, youth (ages 15-24), gender parity index (GPI)
233	SE_ADT_1524_LT_MA_ZS	Num	8	BEST12.		Literacy rate, youth male (% of males ages 15-24)
463	SE_ADT_1524_LT_ZS	Num	8	BEST12.		Youth literacy rate, population 15-24 years, both sexes (%)
229	SE_ADT_LITR_FE_ZS	Num	8	BEST12.		Literacy rate, adult female (% of females ages 15 and above)
230	SE_ADT_LITR_MA_ZS	Num	8	BEST12.		Literacy rate, adult male (% of males ages 15 and above)
14	SE_ADT_LITR_ZS	Num	8	BEST12.		Adult literacy rate, population 15+ years, both sexes (%)
87	SE_COM_DURS	Num	8	BEST12.		Duration of compulsory education (years)
373	SE_ENR_PRIM_FM_ZS	Num	8	BEST12.		School enrollment, primary (gross), gender parity index (GPI)
374	SE_ENR_PRSC_FM_ZS	Num	8	BEST12.		School enrollment, primary and secondary (gross), gender parity index (GPI)
379	SE_ENR_SECO_FM_ZS	Num	8	BEST12.		School enrollment, secondary (gross), gender parity index (GPI)
384	SE_ENR_TERT_FM_ZS	Num	8	BEST12.		School enrollment, tertiary (gross), gender parity index (GPI)
328	SE_PRE_DURS	Num	8	BEST12.		Preprimary education, duration (years)
362	SE_PRE_ENRL_TC_ZS	Num	8	BEST12.		Pupil-teacher ratio in pre-primary education (headcount basis)
167	SE_PRE_ENRR	Num	8	BEST12.		Gross enrolment ratio, pre-primary, both sexes (%)
168	SE_PRE_ENRR_FE	Num	8	BEST12.		Gross enrolment ratio, pre-primary, female (%)
169	SE_PRE_ENRR_MA	Num	8	BEST12.		Gross enrolment ratio, pre-primary, male (%)
427	SE_PRE_TCAQ_FE_ZS	Num	8	BEST12.		Trained teachers in preprimary education, female (% of female teachers)
428	SE_PRE_TCAQ_MA_ZS	Num	8	BEST12.		Trained teachers in preprimary education, male (% of male teachers)
426	SE_PRE_TCAQ_ZS	Num	8	BEST12.		Trained teachers in preprimary education (% of total teachers)
294	SE_PRM_AGES	Num	8	BEST12.		Official entrance age to primary education (years)
351	SE_PRM_CMPT_FE_ZS	Num	8	BEST12.		Primary completion rate, female (% of relevant age group)
352	SE_PRM_CMPT_MA_ZS	Num	8	BEST12.		Primary completion rate, male (% of relevant age group)
353	SE_PRM_CMPT_ZS	Num	8	BEST12.		Primary completion rate, total (% of relevant age group)

Appendix 2

(Initial Data Structure)

	Country_Name	Country_Code	Indicator_Name	Indicator_Code	_1960	_1961	_1962	_1963	_1964	_1965	_1966
64649	China	CHN	Surface area (sq. km)	AG.SRF.TOTL.K2	.	9562950	9562950	9562950	9562950	9562950	9562950
64650	China	CHN	Survey mean consumption or income per capita, bottom 40% of population (2011 PPP $ per day)	SI.SPR.PC40	.	.	.	.	.	.	.
64651	China	CHN	Survey mean consumption or income per capita, total population (2011 PPP $ per day)	SI.SPR.PCAP	.	.	.	.	.	.	.
64652	China	CHN	Survival rate to Grade 5 of primary education, both sexes (%)	SE.PRM.PRS5.ZS	.	.	.	.	.	.	.
64653	China	CHN	Survival rate to the last grade of primary education, both sexes (%)	SE.PRM.PRSL.ZS	.	.	.	.	.	.	.
64654	China	CHN	Survival rate to the last grade of primary education, female (%)	SE.PRM.PRSL.FE.ZS	.	.	.	.	.	.	.
64655	China	CHN	Survival rate to the last grade of primary education, male (%)	SE.PRM.PRSL.MA.ZS	.	.	.	.	.	.	.
64656	China	CHN	Survival to age 65, female (% of cohort)	SP.DYN.TO65.FE.ZS	34.79281	34.84841	34.90402	38.4616	42.01919	45.57677	49.13435
64657	China	CHN	Survival to age 65, male (% of cohort)	SP.DYN.TO65.MA.ZS	25.47925	25.62367	25.76809	29.62284	33.4776	37.33235	41.18711
64658	China	CHN	Tariff rate, applied, simple mean, all products (%)	TM.TAX.MRCH.SM.AR.ZS	.	.	.	.	.	.	.
64659	China	CHN	Tariff rate, applied, simple mean, manufactured products (%)	TM.TAX.MANF.SM.AR.ZS	.	.	.	.	.	.	.

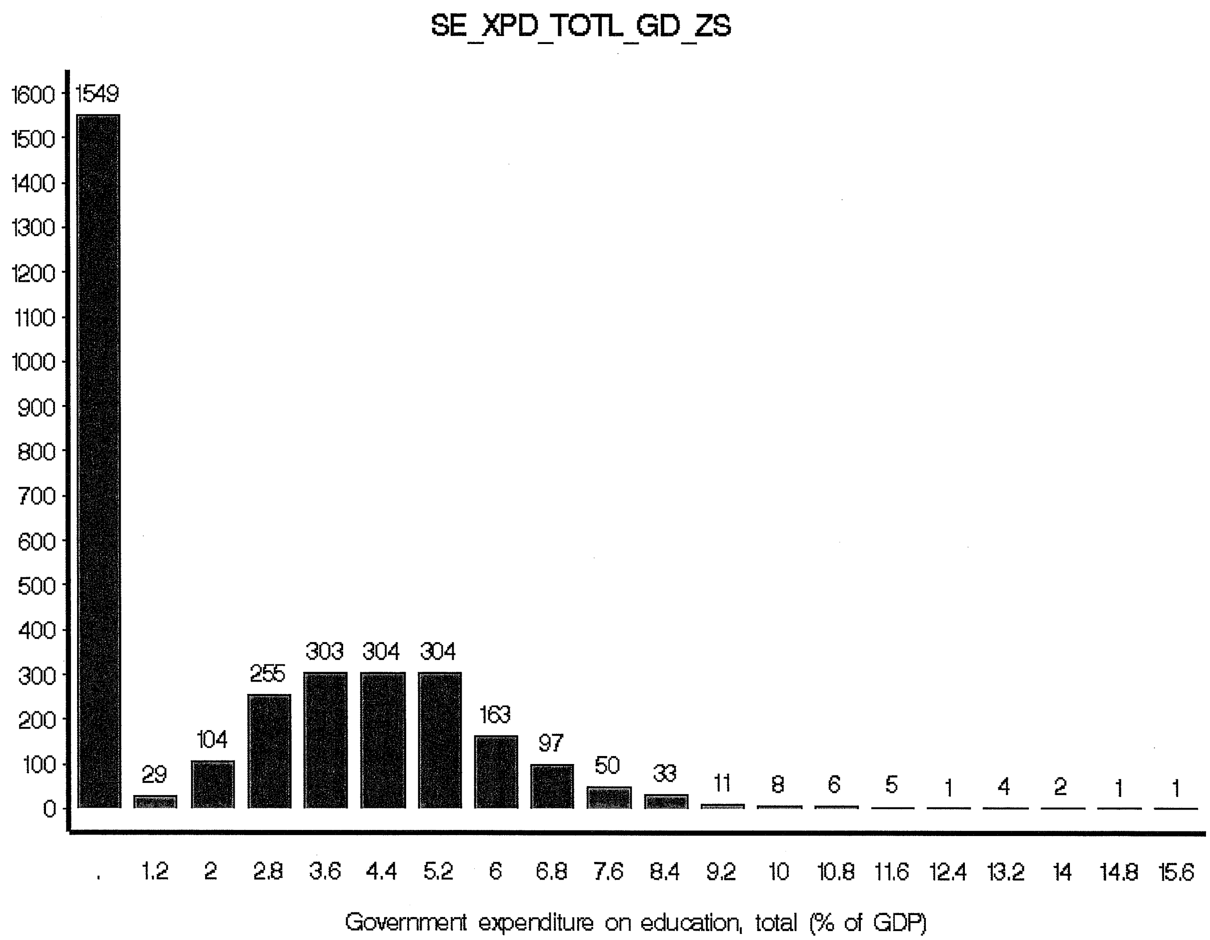
Appendix 3

(Amended Data Structure)

	Country_Name	NAME OF FORMER VARIABLE	2005 PPP conversion factor, GDP (LCU per international $)	2005 PPP conversion factor, private consumption (LCU per international $)	ARI treatment (% of children under 5 taken to a health provider)	Access to electricity (% of population)	Access to electricity, rural (% of rural population)	Access to electricity, urban (% of urban population)	Access to non-solid fuel (% of population)	Access to non-solid fuel, rural (% of rural population)	Access to non-solid fuel, urban (% of urban population)	Account at a financial institution (% age 15+) [ts]	Account at a financial institution, female (% ag 15+) [ts]
82	Germany	_2000	.	.	.	100	100	100	100	.	.	.	
83	Ghana	_2000	.	.	.	45	20.9	75.730049597	8.4789097309	.	.	.	
84	Gibraltar	_2000	.	.	.	.	.	.	.	.	.	.	
85	Greece	_2000	.	.	.	100	100	100	100	.	.	.	
86	Greenland	_2000	.	.	.	100	100	100	.	.	.	.	

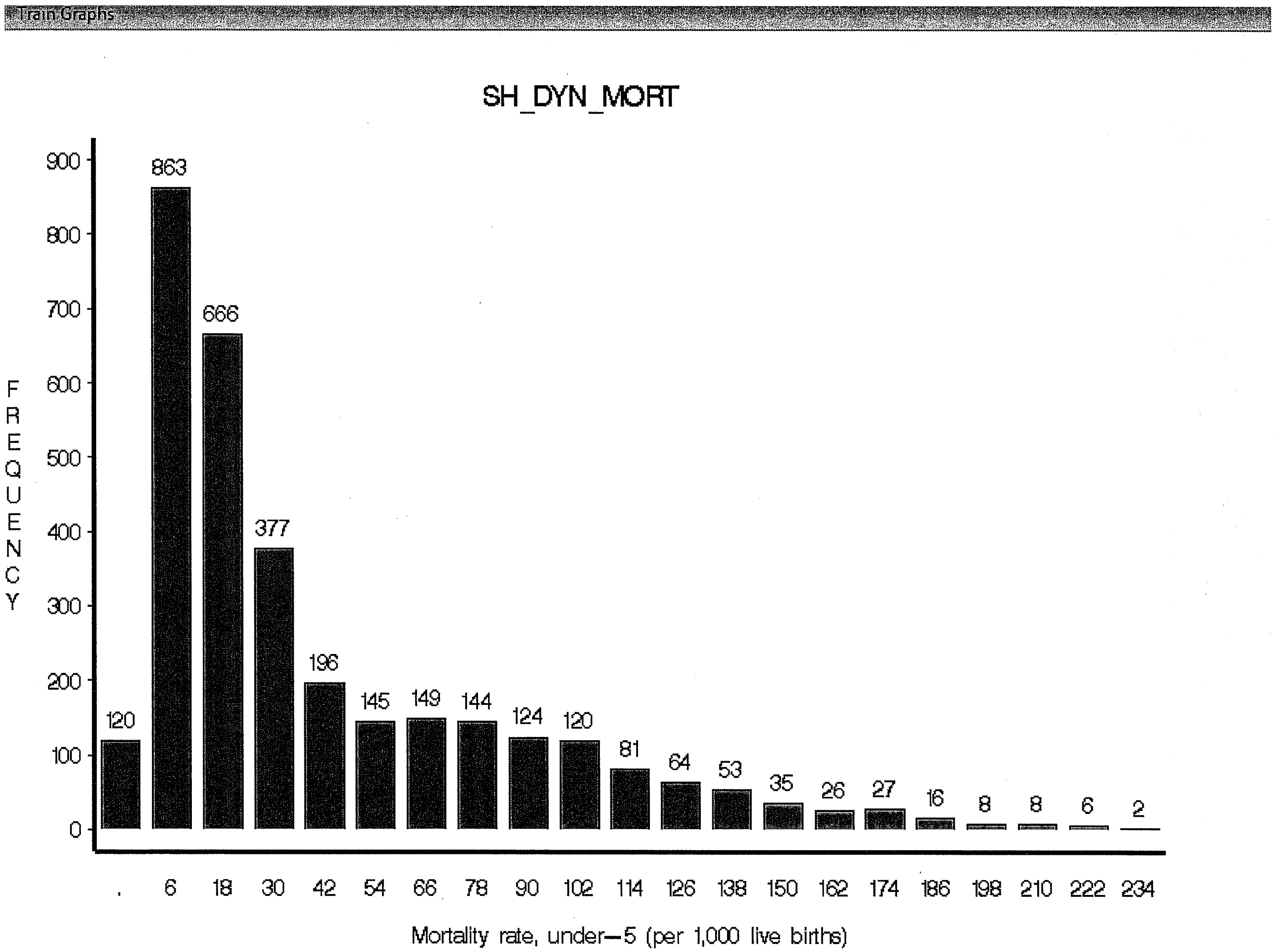
Appendix 4

(Example of variable showing high instances of missing values, first column)



Appendix 5

(Example of variable possibly requiring shape transformation)



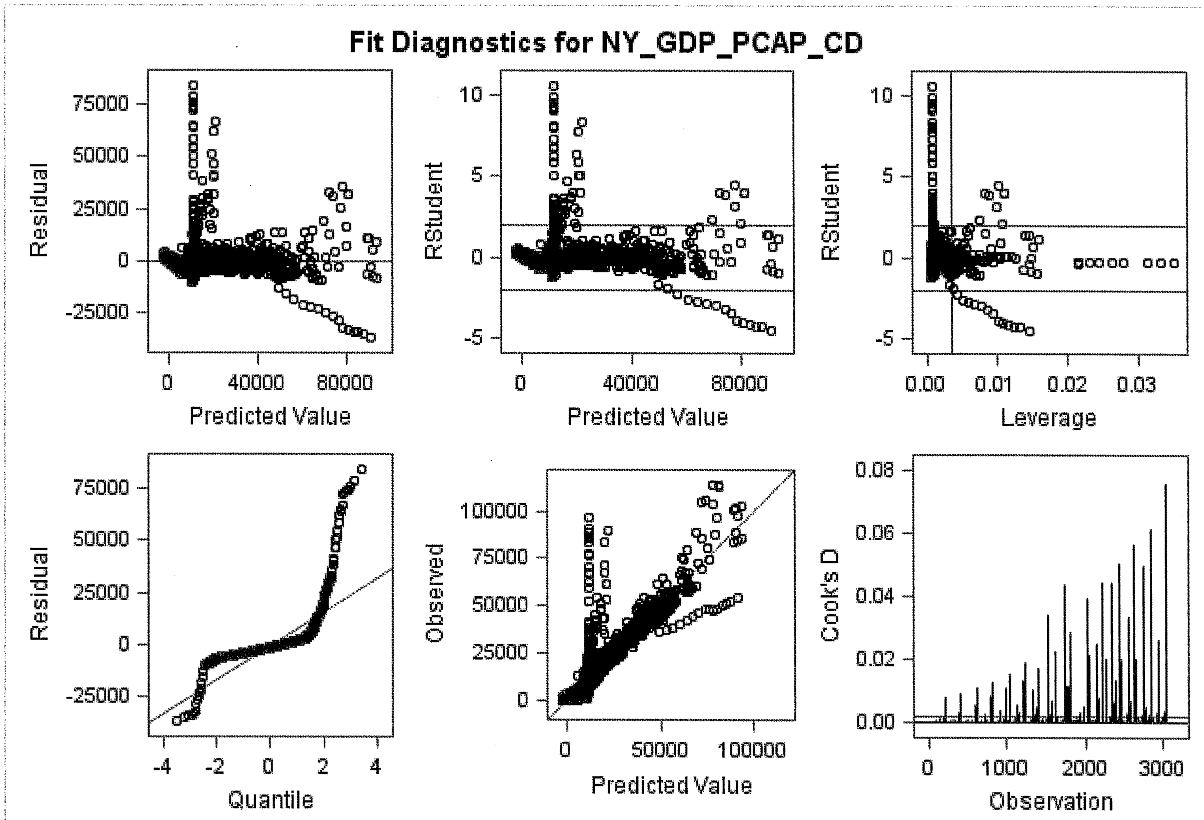
Appendix 6

(Missing Data Illustration - 'N Miss' indicates instances of measurements with missing data)

Variable	Label	N	N Miss
SE_ADT_LITR_ZS	Adult literacy rate, population 15+ years, both sexes (%)	632	2598
EA_PRD_AGRI_KD	Agriculture value added per worker (constant 2010 US$)	2745	485
NY_GDP_PCAP_CD	GDP per capita (current US$)	3230	0
SE_XPD_TOTL_GD_ZS	Government expenditure on education, total (% of GDP)	1681	1549
SE_XPD_SECO_PC_ZS	Government expenditure per student, secondary (% of GDP per capita)	1087	2143
BX_GRT_EXTA_CD_WD	Grants, excluding technical cooperation (BoP, current US$)	2141	1089
SE_SEC_ENRR	Gross enrolment ratio, secondary, both sexes (%)	2235	995
SH_XPD_PCAP	Health expenditure per capita (current US$)	2909	321
NE_CON_PETC_ZS	Household final consumption expenditure, etc. (% of GDP)	2884	346
SH_IMM_IDPT	Immunization, DPT (% of children ages 12-23 months)	3083	147
SH_IMM_MEAS	Immunization, measles (% of children ages 12-23 months)	3083	147
SH_STA_ACSN	Improved sanitation facilities (% of population with access)	3076	154
SH_STA_ACSN_RU	Improved sanitation facilities, rural (% of rural population with access)	3059	171
SH_STA_ACSN_UR	Improved sanitation facilities, urban (% of urban population with access)	3077	153
SH_H2O_SAFE_ZS	Improved water source (% of population with access)	3077	153
SH_H2O_SAFE_RU_ZS	Improved water source, rural (% of rural population with access)	3060	170
SH_H2O_SAFE_UR_ZS	Improved water source, urban (% of urban population with access)	3095	135
SH_TBS_INCD	Incidence of tuberculosis (per 100,000 people)	2899	331
SI_DST_05TH_20	Income share held by highest 20%	901	2329
SH_MMR_RISK_ZS	Lifetime risk of maternal death (%)	3000	230
SH_STA_MMRT	Maternal mortality ratio (modeled estimate, per 100,000 live births)	3000	230
SP_DYN_AMRT_FE	Mortality rate, adult, female (per 1,000 female adults)	2819	411
SP_DYN_AMRT_MA	Mortality rate, adult, male (per 1,000 male adults)	2819	411
SH_DYN_NMRT	Mortality rate, neonatal (per 1,000 live births)	3110	120
SH_DYN_MORT	Mortality rate, under-5 (per 1,000 live births)	3110	120
SH_STA_ODFC_ZS	People practicing open defecation (% of population)	2943	287
SH_STA_ODFC_RU_ZS	People practicing open defecation, rural (% of rural population)	2927	303
SH_STA_ODFC_UR_ZS	People practicing open defecation, urban (% of urban population)	2967	263
GC_REV_SOCL_ZS	Social contributions (% of revenue)	1133	2097

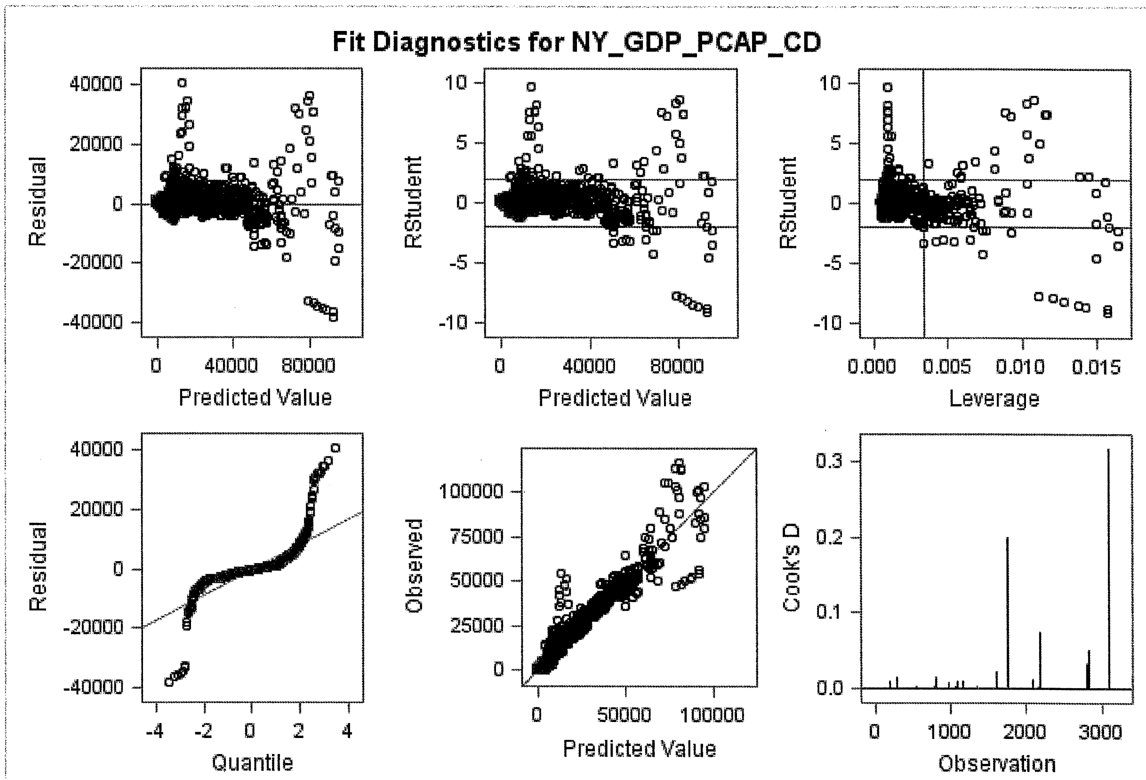
Appendix 7

(Fit Diagnostics using initial imputation method for missing values)



Appendix 8

(Fit Diagnostics using Interpolation imputation method for missing values)



Appendix 9

(Regression using Interpolation - Model Fit Statistics)

The selected model, based on Validation ASE, is the model at Step 3.

Effects:	Intercept SH_XPD_PCAP SH_STA_ACSN SE_SEC_ENRR

Analysis of Variance				
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**
Model	3	3.566676E11	1.188892E11	6678.81
Error	1385	24654313737	17800949	
Corrected Total	1388	3.813219E11		

Root MSE	4219.11703
Dependent Mean	10474
R-Square	0.9353
Adj R-Sq	0.9352
AIC	24584
AICC	24584
SBC	23214
ASE (Train)	17749686
ASE (Validate)	19154746

Parameter Estimates				
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**
Intercept	1	-1551.404383	331.797559	-4.68
SH_XPD_PCAP	1	9.518778	0.083047	114.62
SH_STA_ACSN	1	41.858332	7.691206	5.44
SE_SEC_ENRR	1	12.675646	7.841633	1.62

Appendix 10

(Decision Tree - Model Fit Statistics)

Fit Statistics

Target=NY_GDP_PCAP_CD Target Label=' '

Fit Statistics	Statistics Label	Train	Validation
NOBS	Sum of Frequencies	2261.00	969.00
MAX	Maximum Absolute Error	46481.22	56347.25
SSE	Sum of Squared Errors	46055779740.72	32985495025.17
ASE	Average Squared Error	20369650.48	34040758.54
RASE	Root Average Squared Error	4513.27	5834.45
DIV	Divisor for ASE	2261.00	969.00
DFT	Total Degrees of Freedom	2261.00	.

Assessment Score Rankings

Data Role=TRAIN Target Variable=NY_GDP_PCAP_CD Target Label=' '

Depth	Number of Observations	Mean Target	Mean Predicted
5	118	61202.41	61202.41
10	300	31396.03	31396.03
20	50	21883.87	21883.87
25	98	13463.57	13463.57
30	135	10829.81	10829.81
35	121	7490.21	7490.21
40	88	6539.50	6539.50
45	235	4720.70	4720.70
55	129	3504.66	3504.66
60	115	2564.22	2564.22
65	90	2312.55	2312.55
70	137	1734.18	1734.18
75	209	1130.58	1130.58
85	202	716.89	716.89
90	234	370.88	370.88

Data Role=VALIDATE Target Variable=NY_GDP_PCAP_CD Target Label=' '

Depth	Number of Observations	Mean Target	Mean Predicted
5	63	60180.88	61199.63
10	127	33772.84	31966.44
20	10	24348.21	29231.20
25	51	16995.47	16195.49
30	67	11106.86	10846.62
35	43	6944.92	7490.21
40	41	6298.65	6517.18
45	45	4993.03	5070.73
50	41	4668.79	4447.58
55	64	3382.32	3464.68
60	47	2669.44	2557.57
65	45	2225.30	2312.55
70	43	1847.72	1744.35
75	73	1093.30	1130.58
80	107	716.24	716.89
90	102	372.77	370.88