Final Project Report

Aiman Haider

November 23, 2020

SUMMARY

The report attempts to understand whether and how factors in the social environment are associated with the odds of being a happier nation, in general and, regionally. To do so, the report develops a region-level proportional odds model on happinesslevel using the World Happiness Report based on the AIC stepwise single-level model of one of its imputed datasets, founded on plausibility and significance. The model so formed finds that Log GDP per capita, Life Expectancy, Generosity, Freedom to make choices, Social Support, Life Expectancy combined with Log GDP per capita, Social Support combined with Log GDP per capita and Generosity given Social Support are important associated factors; besides region, and Log GDP per capita and Freedom to make choices given the region. It particularly finds that wealth, health and social support are crucial factors, with wealth being the most influential. Further, it also discovers that the associated odds are higher for those in the Americas, followed by those in Europe, Middle East and Africa and South and East Asia. It also finds that higher income seems to be more favorable in the Americas and Middle East and Africa, and that greater freedom to make choices in South and East Asia shows a surprising non-favorable additional effect.

INTRODUCTION

Happiness is considered both a cherished goal and an outcome of development. In this light, the Gallup World Poll supported by the Sustainable Development Solutions Network publishes the popular World Happiness Report (WHR) that calculates happiness scores for most countries by averaging out the response scores, on the Cantril ladder (Refer Appendix), of roughly (at least) around 1,000 nationally sampled registered citizens each year. It also concurrently publishes statistics for a number of social environment variables such as Log GDP per capita, Life Expectancy, Perceptions on Corruption, Social Support, Freedom to make choices and Generosity; and provides an opportunity to understand Happiness.

Using the WHR and especially, the Happiness Scores, this report attempts to find the "secret" of happy nations in their social environments. In order to do so, it attempts to understand which of the above-mentioned environmental variables are important i.e. significantly associated, and how do they relate to the odds of being happier. It also strives to understand the more influential ones. Further, it aspires to unravel the regional variations. It tries to understand whether belonging to one region or the other makes a difference. Also, it attempts to understand which factors matter differently across regions and how does the belonging to one region or the other change their associations to the odds of being happier.

The report uses a region-level proportional odds model for the analysis. It begins with data preprocessing followed by EDA, and tries developing the model aided by an imputed data AIC Stepwise-built single-level model, guided by statistical significance and plausibility.

DATA

The data is obtained using two files-one containing the data for the years 2005-2018 from WHR 2019, and the other containing data for the year 2019 from WHR 2020. The data consists of a number of pertinent and infrequent non-pertinent columns. Of these, only the six key (published) variables (Codebook in Appendix) - Log GDP per capita (Log.GDP), Life Expectancy, Perceptions on Corruption (Perceptions.Corruption), Social Support, Freedom to make choices (Freedom.Choices) and Generosity, and Country and Year are used. The files are then merged after streamlining the names and order of columns to give a total of 1831 rows and the above 8 columns. From the 2020 report, a "region" table is created to map each country with its respective region using the country and region columns. The merged file is then left joined with this to give the final table with the columns Log.GDP, Life.Expectancy, Perceptions.Corruption, Social.Support, Freedom.Choices, Generosity and Region[1].

However, it is found that this final table has some missing data - 5.24% for Corruption, 4.44% on Generosity, around 1.5% on Log.GDP, Freedom.Choices and Life.Expectancy and less than 1% for Social.support. Further, it is observed that a few countries consistently do not report certain metrics while most only have a few missing values. On further observation, it is surmised that the missing data could be a combination of missing at random and missing not at random. However, as most seem to be a case of MAR, MAR is assumed and five imputations using the pmm mice method are implemented. The missing values are then imputed using all the variables[2] and are found to fare well (Plot in Appendix).

After imputing, the data is then preprocessed for modeling. The variable Happiness Score (Ref. Appendix) is binned to get its base whole number (e.g. 3.99 as 3) and is found to have a normal distribution lying between 2-8. However, as the regional distributions show some gaps (no observations for certain scores), this variable is further binned to create a more comprehensive ordinal variable Happiness Level, with scores below 5 categorized as “Low”, 5 as “Mid” and above 5 as “High”. This is then used as the response variable. The regions are also combined to belong to “Americas”, “Europe”, “Middle East and Africa” and “Southern and Eastern Asia”. Besides, for the analysis, variables Social.support, Freedom.Choices, Perceptions.Corruption and Generosity are multiplied by 100 and mean centered, and Log.GDP and Life.Expectancy mean centered, to avoid colinearity and ease model building.[3] The data is then explored for any interesting patterns using one of the imputed datasets and checked for consistency across others. It is found that the number of observations with “Low” happinesslevel appears to be the highest followed by “Mid” and “High”. The distribution of happinesslevel across regions, however, is particularly noteworthy - Different regions have different distributions, indicating the possibility of a region-level model. It can also be seen that higher Log.GDP, Life.Expectancy, Freedom.Choices and Social.support seem to show higher happinesslevels, with the latter two showing some saturating trends at higher levels. Intriguingly, the other two do not show increasing trends. Perceptions.Corruption seems to have nearly the same range for “Low” and “Mid” levels. Generosity, too, doesn’t follow the expected trend with a slightly lower median for “Mid” level followed by “Low” and “High”.

[1] Country and Year are dropped as the former is not required in the analysis and acts as "observations", and the latter causes issues of non-convergence in imputations and modeling.

[2] Imputations seem to worsen with "norm" method and/or different predictor matrices.

[3] The scale of the first four variables are on a scale from 0-1.

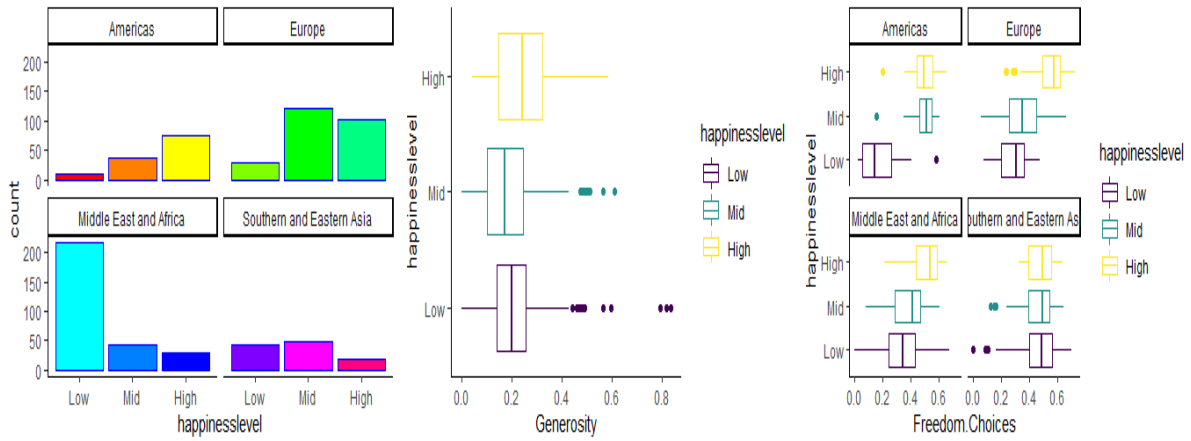

Figure 1: EDA Plots using Imputed Dataset 1

The report then looks for possible interactions, particularly with region, and observes that Log.GDP, Life.Expectancy, Social.Support (slight difference in trends), Freedom.Choices, Perceptions.Corruption and Generosity show differences in trends across regions. It is striking how all happinesslevels in South and East Asia show nearly the same freedom to make choices. The trends for Generosity and Perceptions on Corruption, too, show noticeable difference in the trends for different regions, and Americas and Europe respectively. Log GDP per capita also shows discernible difference in trends and so does Life Expectancy.

MODEL

The model is developed via a two-step iterative process. First, a single-level prop. odds model is developed using the same imputed dataset as in EDA, taking region as a categorical variable and using the AIC-stepwise criteria. The Null Model used is happinesslevel ~ Region (requisite) and the Full Model used is: *happinesslevel ~ (Life.Expectancy+ Freedom.Choices+ Perceptions.Corruption+ Social.support+ Generosity+Log. GDP)*Region+ Log. GDP:Life.Expectancy+ Log.GDP: Perceptions. Corruption+ Log.GDP:Social.support +Life. Expectancy:Social.support+Log.GDP:Freedom.Choices +Generosity: Social.support +Generosity: Life.Expectancy*,which includes variables that show interesting patterns in the EDA and those that made intuitive sense[4] . Feeding these models, the AIC Stepwise Model found is: *happinesslevel ~ Log.GDP + Region + Freedom.Choices + Generosity + Social.support + Life.Expectancy + Perceptions.Corruption + Region:Freedom.Choices + Log.GDP:Social.support + Log.GDP:Region + Social.support:Life.Expectancy + Generosity: Social.support + Log.GDP: Freedom.Choices + Region:Generosity + Region:Life.Expectancy + Log.GDP:Life.Expectancy + Region: Perceptions.Corruption.* The variables returned are then vetted using the single-level (polr) and a parallel mixed model(clmm; for direction)[5], where region is converted into the random intercept and the variables having interaction with region into the random slopes, to ensure a consistent working mixed model, low multicolinearity, low random slope correlations and statistical significance (with the help of ANOVA and AIC; Ref. Appendix). The model thus found is highly statistically significant (p<2.2e-16, Null Model) with fair AIC (1932, a little higher than the AIC Model, but balanced). And is: happinesslevel ~ Log.GDP + Generosity

[4]Log.GDP:Life.Exp denoting a healthy-wealthy life, Log.GDP:Per.Corruption unfair wealth appropriation, Log.GDP:Social.support a well-off society, Life.Exp:Social.support supportive society in old age, Log.GDP:Freedom.Choices independence with wealth,Generosity:Social.support a caring society, and Generosity:Life.Exp a society caring for the old. Explored here as continuous.

[5]On a number of occasions, variables found fine with one did not work with the other and vice versa. Thus, making it an iterative process. Parallel checks ensured a consistent model.

+ Social.support + Life.Expectancy+Freedom.Choices+ Log.GDP:Life.Expectancy+Log.GDP: Social.support+Generosity:Social.support+(Log.GDP+ Freedom.Choices|Region). It is then assessed using its polr version[6]. The binned residuals seem to be well within the 95% limits and the ROC curves too perform well (Ref. Appendix). The AUC for countries with "Low" happinesslevel is 94.05%, "Mid" is 83.53% and "High" is 96.5%. The respective sensitivities and specificities are 0.8161 and 0.8793, 0.6189 and 0.8329, and 0.8396 and 0.9279. All categories seem to fare well, generally speaking, except for a comparatively low sensitivity for "Mid" Happiness Scores, which could possibly be due to anomalous trends in Generosity and Perceptions.Corruption and other external factors. VIFs for all the variables in the clmm version are lesser than 8, while for the polr version only categorical ones are high (and Life.Expectancy at the margin with 10.39). Thus, it appears that the model seems to be faring well. Further, the model assessment seems to be consistent across various imputed datasets and the EDA, too. The model also seems to be plausible as the main effects, interactions and random effects are intuitive. While there might be terms which could have been added, the predictors present in a sense explain them due to their high correlation e.g. Region:Log.GDP and Region:Life.Expectancy. Hence, the model seems to be a succinct, significant and informative one that balances the various constraints. Thus, the final model is:

$$\log(\frac{Pr(happinesslevel_i <= j|x_i)}{Pr(happinesslevel_i > j|x_i)}) = (\beta_{0j}+\gamma_{0jregion}) - (\beta_1+\gamma_{1jregion})Log.GDP_i - \beta_2 Generosity_i - \beta_3 Social.support_i - \beta_4 Life.Expectancy_i + (\beta_5+\gamma_{2jregion})Freedom.Choices_i - \beta_6 Log.GDP_i : Life.Expectancy_i - \beta_7 Log.GDP_i : Social.support_i - \beta_8 Generosity_i : Social.suport_i$$

This model is then pooled using Rubin's method to obtain the estimate of the fixed effects. The random effects are found by averaging out the random effects of all the imputed datasets with the final model[7] applied. The results are as in the table below. From the table,

Output of Final Model on Imputed Data					
Term	estimate	std.error	statistic	df	p.value
Low|Mid	-1.3943351	0.5408498352	-2.5780	1818.813	1.001415e-02
Mid|High	2.456306939	0.5462844333	4.496388	1818.813	7.347955e-06
Log.GDP	1.666103667	0.2782021217	5.988824	1818.813	2.541065e-09
Generosity	0.022749379	0.0045574605	4.991679	1818.813	6.557659e-07
Social.support	0.071884262	0.0088321217	8.138957	1818.813	8.881784e-16
Life.Expectancy	0.090350083	0.0183266813	4.929975	1818.813	8.971664e-07
Freedom.Choices	0.029236973	0.0152876897	1.912452	1818.813	5.597516e-02
Log.GDP:Life.Exp	0.039947391	0.0146825898	2.720732	1818.813	6.575802e-03
Log.GDP:Social.sup	0.048895892	0.0076778084	6.368470	1818.813	2.411871e-10
Generosity:Social.s	0.001261336	0.0004251866	2.966546	1818.813	3.050924e-03

Random Effects (Normal scaled)			
Term	Intercept	Log.GDP	Freedom.Choices
Americas	5.81314	1.71285	1.02069
Europe	0.81643	0.66635	1.01988
Middle East and Africa	0.49932	1.43674	1.00506
Southern and Eastern Asia	0.42467	0.58806	0.95455

it can be seen that all the above-mentioned variables show strong association with the log-odds of having higher happinesslevels. It can be understood that at the significance level of

[6] clmm contains random slopes that hinders abstraction of residuals for assessment

[7] The estimates and standard errors are consistent across imputations.

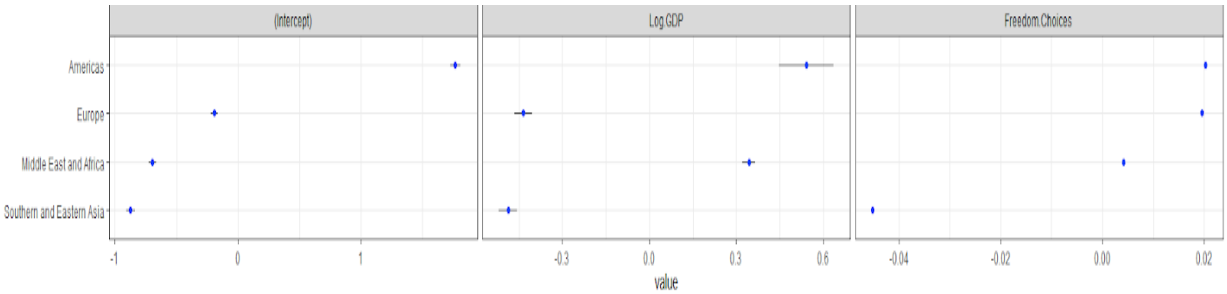

Figure 2: Dotplot of Random Effects using Imputed Dataset 1

0.01, the odds of a country to belong to a higher happiness level increases by 5.29 times for every $2.7 increase in Log.GDP over $9976, by 9.45% for a year increase in Life.Expectancy over 63 yrs and by 2.3%, 7.45% and 2.97% for every unit increase (Ref. Appendix) in Generosity, Social.support and Freedom.Choices, keeping other variables constant. Further, there is an additional increase in the odds by 4.08% for every unit increase (Ref.Appendix) in combined Log.GDP and Life.Expectancy, 5.01% for Log.GDP and Social.support and minuscule for Generosity given Social.support. Thus, it can be seen that wealth, health, generosity, independence, social support, and combinations of a healthy-wealthy life, wealthy supportive society and a caring society are important factors associated with happiness, generally. Of these, wealth, health and social support have particularly strong associations (t-values), with wealth (highest coeff.) being the most influential factor. From the model, it can also be understood that belonging to one region or the other does impact the odds of being happier. Based on the output of dataset 1 (similar in others), the intercepts seem to vary by 1.04 (Std dev on log scale) across regions. From the pooled random effects, it can be seen that the odds additionally increase by 5.8 times for those in the Americas and decreases by 18.37% in Europe, 50.06% in Middle East and Africa and 57.54% in South and East Asia over and above the resp. level (low or mid) threshold intercepts. Further, it is seen that Log.GDP (std dev:0.5, log scale, imp data 1) and Freedom.Choices (std dev: 0.07, log scale, imp data 1) matter differently across regions as can also seen from Fig 2. A unit increase in Log.GDP increases the odds by an additional 70% in Americas and 43.67% in Mid East and Africa, and decreases by an additional 33% in Europe and 41.2% in South And East Asia over and above the main effect. There is not much difference in terms of Freedom.Choices, though, which has an additional increase of 2.07% in the Americas,1.99% Europe and 0.51% in Mid East and Africa over and above its main effect; but, a surprising decrease of 4.6% for those in South And East Asia. In other words, the associated odds of being happier are highest for those in Americas, followed by Europe, Mid East and Africa, and South and East Asia. Besides, higher income seems to have the highest odds for those in the Americas, followed by Mid East and Africa, Europe and South and East Asia. Freedom to make choices in South and East Asia, however, seems to show a surprising but slight negative additional effect. Clearly, regions matter.

CONCLUSION

The findings suggest that social environment does have a strong association with happiness. In particular, wealth and region seem to be unparalleled factors related to the odds of being a happier nation. Inclusive development, thus, seems to be important for a happy society.

The report, though, is limited by the imputations used (esp. the present run and assumption) which could adversely affect the accuracy and reliability, and calls for a more advanced technique. The CLMM method, too, posed another technical limitation. Besides, the quality of data collection and the subjectivity of the topic, too, have been other major constraints. Nonetheless, the data does help one get an initial insight into the topic.

APPENDIX

MODEL INTERPRETATION

Every unit increase per variable, is defined as follows:

- Log.GDP per capita: A unit increase refers to an increase of $2.7 above $9976.62.
- Life Expectancy : A unit increase refers to the Life Expectancy of the nation improving by an additional year over 63.09 years.
- Social.Support: A unit increase refers to 1 more person being helped, over and above 81 person already being supported. Assuming 1000 respondents per country.
- Freedom.Choices: A unit increase refers to 1 more person having the freedom to make life choices, over and above 74 people already asserting that.Assuming 1000 respondents per country.
- Generosity : A unit increase refers to 1 more person donating to charity. Assuming 1000 respondents per country.
- Log.GDP per capita and Life. Expectancy: A unit increase refers to a combined increase of $2.7 above $9976.62 in GDP per capita and the Life Expectancy of the nation improving by an additional year over 63.09 years.
- Log.GDP per capita and Social.Support: A unit increase refers to a combined increase of $2.7 above $9976.62 in GDP per capita and 1 more people being helped, over and above 81 people already supported. Assuming 1000 respondents per country.
- Generosity and Social Support: A unit increases refers to 1 more person donating to charity and 1 more person being helped, over and above 81 people already. Assuming 1000 respondents per country.

Additional Interpretation for the model:

- Low Threshold: Given a country with GDP per capita $9976.62, Life expectancy of 63.09 years, 81 people supported, 74 people free to make choices and none spending on charity, the odds of belonging to a low happinesslevel country is 0.25.
- Mid Threshold: Given a country with GDP per capita $9976.62, Life expectancy of 63.09 years, 81 people supported, 74 people free to make choices and none spending on charity, the odds of belonging to a mid to low happinesslevel country is 12.11.
- Region: The estimated standard deviation (1.04, added on log scale) describes the across region variation attributed to the random intercept. It adds over the low or mid threshold baselines. This number is from the results of imputed dataset 1. The results are however, consistent across imputations.
- Log.GDP: Region : The standard deviation (0.5, on log scale) describes the variation in Log.GDP across region that adds over the main effect of Log.GDP. It shows that GDP does vary across regions.
- Log.GDP: Freedom.Choices: Similarly but with standard deviation of 0.07.

CODE BOOK

Variable	Description	Type
Happiness Score	Averaged out response scores to the question, "Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?"	Continuous (Converted to the ordinal variable *happinesslevel* by binning Scores)
Log.GDP	Natural Log of GDP Per Capita.GDP per capita is in terms of Purchasing power parity (PPP) adjusted to constant 2011 international dollars, taken from the World Development Indicators (WDI) released by the World Bank on November 28, 2019. GDP data for 2019 were extended from the GDP time series from 2018 to 2019 using country-specific forecasts of real GDP growth from the OECD Economic Outlook No. 106 (edition November 2019) and the world bank's Global Economy Prospects, after adjustment for population growth.	Continuous
Life.Expectancy	Life Expectancy, constructed based on the data from the World Health Organization (WHO) Global Health Observatory data repository.	Continuous
Social.Support	Social support is the national average of the binary responses (0=no, 1=yes) to the Gallup World Poll (GWP) question, "if you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"	Continuous (Scale:0-1)
Freedom.Choices	Freedom to make life choices is the national average of binary responses to the GWP question, "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?"	Continuous (Scale:0-1)
Generosity	Generosity is the residual of regressing the national average of GWP responses to the question, "Have you donated money to a charity in the past month?" on GDP per capita.	Continuous (Scale:-1 to 1)
Perceptions.Corruption	Perceptions of corruption are the average of binary answer to two GWP questions: "Is Corruption widespread throughout the government or not?" and "is corruption widespread within business or not?" Where data for government corruption are missing, the perception of business corruption is used as the overall corruption-perception measure.	Continuous (Scale:0-1)

Reference:https://worldhappiness.report/ed/2020/social-environments-for-world-happiness/

TABLES

THE POLR MODEL OUTPUT FOR IMPUTED DATASET 1

Predictors	*Odds Ratios*	*CI*	*p*
Low|Mid	0.04	0.02 – 0.07	**<0.001**
Mid|High	1.94	1.92 – 1.95	**<0.001**
Log.GDP	10.10	9.92 – 10.27	**<0.001**
Generosity	1.02	0.99 – 1.06	**<0.001**
Social.support	1.08	1.05 – 1.11	**<0.001**
Life.Expectancy	1.09	0.72 – 1.65	**<0.001**
Freedom.Choices	1.05	0.65 – 1.70	**0.001**
Region [Europe]	0.14	0.09 – 0.22	**<0.001**
Region [Middle East and Africa]	0.08	0.08 – 0.08	**<0.001**
Region [Southern and Eastern Asia]	0.07	0.07 – 0.07	**<0.001**
Log.GDP * Life.Expectancy	1.05	1.04 – 1.05	**0.002**
Log.GDP * Social.support	1.05	0.57 – 1.93	**<0.001**
Generosity * Social.support	1.00	0.55 – 1.81	**0.016**
Log.GDP * Region [Europe]	0.32	0.17 – 0.59	**<0.001**
Log.GDP * Region [Middle East and Africa]	0.81	0.78 – 0.83	0.475
Log.GDP * Region [Southern and Eastern Asia]	0.32	0.31 – 0.34	**<0.001**
Freedom.Choices * Region [Europe]	1.01	0.97 – 1.04	0.718
Freedom.Choices * Region [Middle East and Africa]	0.99	0.66 – 1.47	0.507
Freedom.Choices * Region [Southern and Eastern Asia]	0.93	0.67 – 1.30	**<0.001**
Observations	1831		
R^2 Nagelkerke	0.777		

THE CLMM MODEL OUTPUT ON AN IMPUTED DATASET

	happinesslevel		
Predictors	*Odds Ratios*	*CI*	*p*
Low|Mid	0.25	0.09 – 0.72	**0.010**
Mid|High	12.11	4.21 – 34.86	**<0.001**
Log.GDP	5.28	3.03 – 9.19	**<0.001**
Generosity	1.02	1.01 – 1.03	**<0.001**
Social.support	1.08	1.06 – 1.10	**<0.001**
Life.Expectancy	1.09	1.06 – 1.13	**<0.001**
Freedom.Choices	1.03	1.00 – 1.06	**0.036**
Log.GDP * Life.Expectancy	1.05	1.02 – 1.08	**<0.001**
Log.GDP * Social.support	1.05	1.03 – 1.06	**<0.001**
Generosity * Social.support	1.00	1.00 – 1.00	**0.003**
Random Effects			
σ^2	3.29		
$\tau_{00\ Region}$	1.09		
$\tau_{11\ Region.Log.GDP}$	0.25		
$\tau_{11\ Region.Freedom.Choices}$	0.00		
ρ_{01}	0.57		
	0.61		
ICC	0.34		
N_{Region}	4		
Observations	1831		
Marginal R^2 / Conditional R^2	0.675 / 0.785		

POOLED RANDOM EFFECTS

id	*variable*	*value*	*se*
Americas	(Intercept)	1.75	0.02
Europe	(Intercept)	-0.20	0.01
Middle East and Africa	(Intercept)	-0.68	0.02
Southern and Eastern Asia	(Intercept)	-0.87	0.02
Americas	Log.GDP	0.54	0.05
Europe	Log.GDP	-0.42	0.02
Middle East and Africa	Log.GDP	0.35	0.01
Southern and Eastern Asia	Log.GDP	-0.51	0.02
Americas	Freedom.Choices	0.02	0.00
Europe	Freedom.Choices	0.02	0.00
Middle East and Africa	Freedom.Choices	0.00	0.00
Southern and Eastern Asia	Freedom.Choices	-0.05	0.00

Note: The Value for a random effect is calculated by finding the average of the resp. random effect of the CLMM model using all Imputed Datasets. The formula used for standard rror is $S_{pooled} = \sqrt{\left(\frac{(S_1)^2+(S_2)^2+(S_3)^2+(S_4)^2+(S_5)^2}{5}\right)}$ where S is the standard error of the resp. imputed dataset. The standard errors here are minuscule and hence, the respective groups don't show much deviation by themselves. These values are calculated using a different set of imputations as used above but are remarkably similar.

Model Selection

AIC Stepwise Model on an Imputed Dataset

```
Call:
polr(formula = happinesslevel ~ Log.GDP + Region + Freedom.Choices +
    Generosity + Social.support + Life.Expectancy + Perceptions.Corruption +
    Region:Freedom.Choices + Log.GDP:Social.support + Log.GDP:Region +
    Social.support:Life.Expectancy + Generosity:Social.support +
    Log.GDP:Freedom.Choices + Region:Generosity + Region:Life.Expectancy +
    Log.GDP:Life.Expectancy + Region:Perceptions.Corruption,
    data = d[[i]])

Coefficients:
                                                      Value Std. Error  t value
Log.GDP                                            2.186636  0.3616698  6.04595
RegionEurope                                      -2.125939  0.2602903 -8.16757
RegionMiddle East and Africa                      -2.645701  0.3013416 -8.77974
RegionSouthern and Eastern Asia                   -2.749024  0.2893031 -9.50223
Freedom.Choices                                    0.034479  0.0156051  2.20947
Generosity                                         0.022505  0.0159483  1.41115
Social.support                                     0.089761  0.0102696  8.74050
Life.Expectancy                                    0.127821  0.0542654  2.35547
Perceptions.Corruption                            -0.050527  0.0197557 -2.55760
RegionEurope:Freedom.Choices                       0.015680  0.0182303  0.86008
RegionMiddle East and Africa:Freedom.Choices      -0.005232  0.0192084 -0.27236
RegionSouthern and Eastern Asia:Freedom.Choices   -0.048243  0.0195465 -2.46812
Log.GDP:Social.support                             0.023590  0.0125476  1.88002
Log.GDP:RegionEurope                              -0.907797  0.4102371 -2.21286
Log.GDP:RegionMiddle East and Africa              -0.006512  0.4091157 -0.01592
Log.GDP:RegionSouthern and Eastern Asia           -0.315696  0.4683384 -0.67408
Social.support:Life.Expectancy                     0.006063  0.0018610  3.25798
Generosity:Social.support                          0.001702  0.0005103  3.33505
Log.GDP:Freedom.Choices                           -0.018162  0.0062909 -2.88706
RegionEurope:Generosity                            0.002018  0.0180598  0.11171
RegionMiddle East and Africa:Generosity            0.021621  0.0199055  1.08620
RegionSouthern and Eastern Asia:Generosity        -0.020856  0.0176840 -1.17938
RegionEurope:Life.Expectancy                      -0.058125  0.0727688 -0.79877
RegionMiddle East and Africa:Life.Expectancy      -0.030392  0.0581869 -0.52231
RegionSouthern and Eastern Asia:Life.Expectancy   -0.172068  0.0740937 -2.32230
Log.GDP:Life.Expectancy                            0.036393  0.0187917  1.93664
RegionEurope:Perceptions.Corruption                0.039978  0.0213865  1.86929
RegionMiddle East and Africa:Perceptions.Corruption  0.040178  0.0213266  1.88392
RegionSouthern and Eastern Asia:Perceptions.Corruption  0.057293  0.0217417  2.63516

Intercepts:
         Value    Std. Error t value
Low|Mid   -3.4802   0.2437   -14.2820
Mid|High   0.4871   0.2105     2.3138

Residual Deviance: 1825.848
AIC: 1887.848
```

Low Correlation

	Parameter	VIF	Increased SE
	Generosity	2.09	1.45
Social.support:Life.Expectancy		1.66	1.29
Generosity:Social.support		1.48	1.22

Moderate Correlation

	Parameter	VIF	Increased SE
	Log.GDP	6.12	2.47
Perceptions.Corruption		6.70	2.59
Log.GDP:Social.support		8.38	2.89
Log.GDP:Freedom.Choices		7.62	2.76

High Correlation

	Parameter	VIF	Increased SE
	Region	178.65	13.37
Freedom.Choices		17.02	4.13
Social.support		15.31	3.91
Life.Expectancy		26.06	5.11
Region:Freedom.Choices		63.02	7.94
Log.GDP:Region		238.12	15.43
Region:Generosity		133.38	11.55
Region:Life.Expectancy		408.05	20.20
Log.GDP:Life.Expectancy		11.64	3.41
Region:Perceptions.Corruption		155.66	12.48

VARIABLE SELECTION FOR THE REGION-LEVEL MODEL

Variable	Selection/Removal
Log.GDP	Retained; Has high t-value; Highly significant as per ANOVA test.
Social.Support	Retained; Has high t-value; Highly significant as per ANOVA test.
Freedom.Choices	Retained; Highly significant as per ANOVA test.
Perceptions.Corruption	Dropped; Inconsistent across Imputations; Entailing info better explained by more consistent, less correlated and informative combination of variables which have higher t-values; Also, vetted through AIC. Besides, has very high correlation with Log.GDP (which has much higher t-value all through) in a number of datasets and is not significant as per ANOVA in the final model here(p=0.38; final model).
Generosity	Retained; Highly significant as per ANOVA test.
Life.Expectancy	Retained; Highly significant as per ANOVA test.
Region:Log.GDP	Retained; Highly significant as per ANOVA test for polr and improves the AIC of the clmm model by over 1000 units.
Region:Freedom.Corruption	Retained; Highly significant as per ANOVA test for polr and improves the AIC of the clmm model of around 380 units.
Region:Perceptions.Corruption	Dropped; Creates problem of high correlation and multicolinearity. Also, has less improvement in AIC (270 units).
Region: Generosity	Dropped; Has low t-value and does not improve the model on addition. Also, the model can't accommodate for many random slopes. So, only the most important random slopes are kept.
Region: Life.Expectancy	Dropped; Has low t-value and creates a problem of high corr.
Log.GDP:Perceptions.Corruption	Dropped; Prevents a well-defined var-covar matrix.
Generosity:Social.Support	Retained; Significant as per the ANOVA test(p=0.00425).
Life.Expectancy:Social.Support	Dropped; Blows up the VIFs when added to the combination of Log.GDP:Social.Support and Log.GDP:Life.Expectancy. Also, has worse multicolinearity, correlation among the random slopes and is less informative in comparison.
Log.GDP:Freedom.Choices	Dropped; Not significant as per the ANOVA test (p=0.7185) and creates the problem of multicolinearity and high correlation among random slopes.
Log.GDP: Social.Support	Retained; Highly Significant and does not create problems with either multicolinearity or random slopes.
Log.GDP: Life.Expectancy	Retained; Highly Significant and does not create problems with either multicolinearity or random slopes.
Freedom.Choices:Social.Support	Dropped; Creates problem of multicolinearity and high correlation among random slopes.

Note: The variables above are from the AIC Stepwise single-level model using imputed dataset 1. The results of the final selection are, however, consistent across different imputed datasets. Highly significant variables have p values in the range $< 10^{-7}$ (ANOVA). P-values cited are from CLMM version; P-values from polr version are in consonance. The low t-valued region-interactions are removed first to ensure a running clmm model as the model can't include many random slopes. Then, the low t-valued variables creating issues of non-convergence or non-definition of var-covar matrix are removed. Then the remaining variables and their combinations tested.

Some Additional Plots

Missing Data

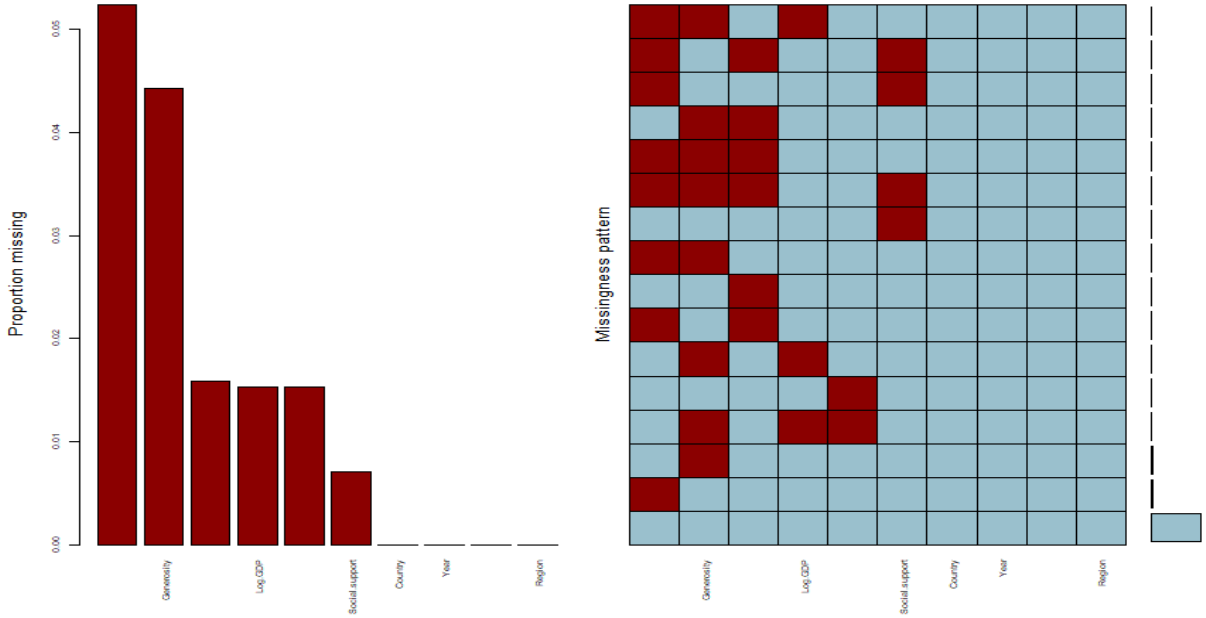

Figure 3: Missing Data Patterns

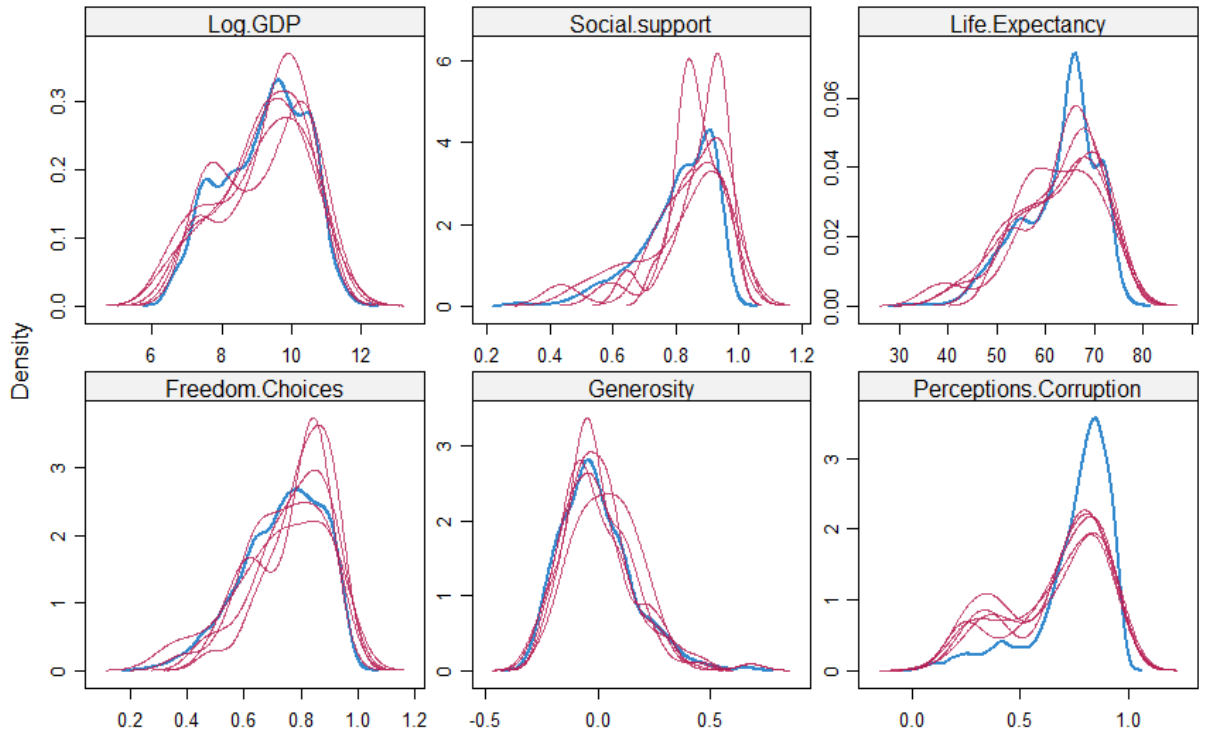

Figure 4: Imputed Datasets

As can be seen, the variables don't seem to show a normal distribution. Hence, the pmm method is used to ensure closeness to the original trends as far as possible.

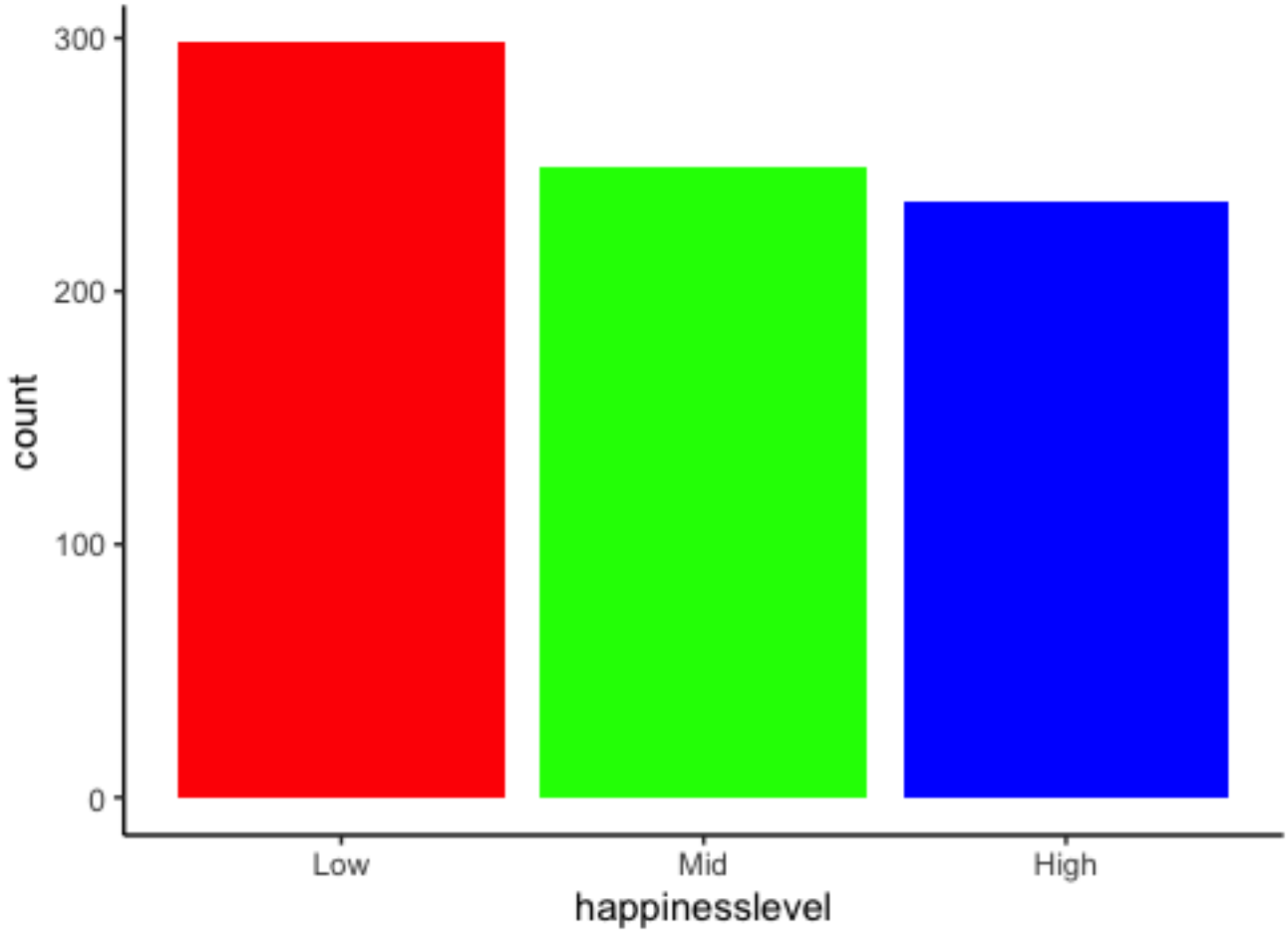

Figure 5: Distribution of Happinesslevel

The distribution of the response variable here provides a cue for findings. It is, however, not a requirement for fitting in a proportional odds model.

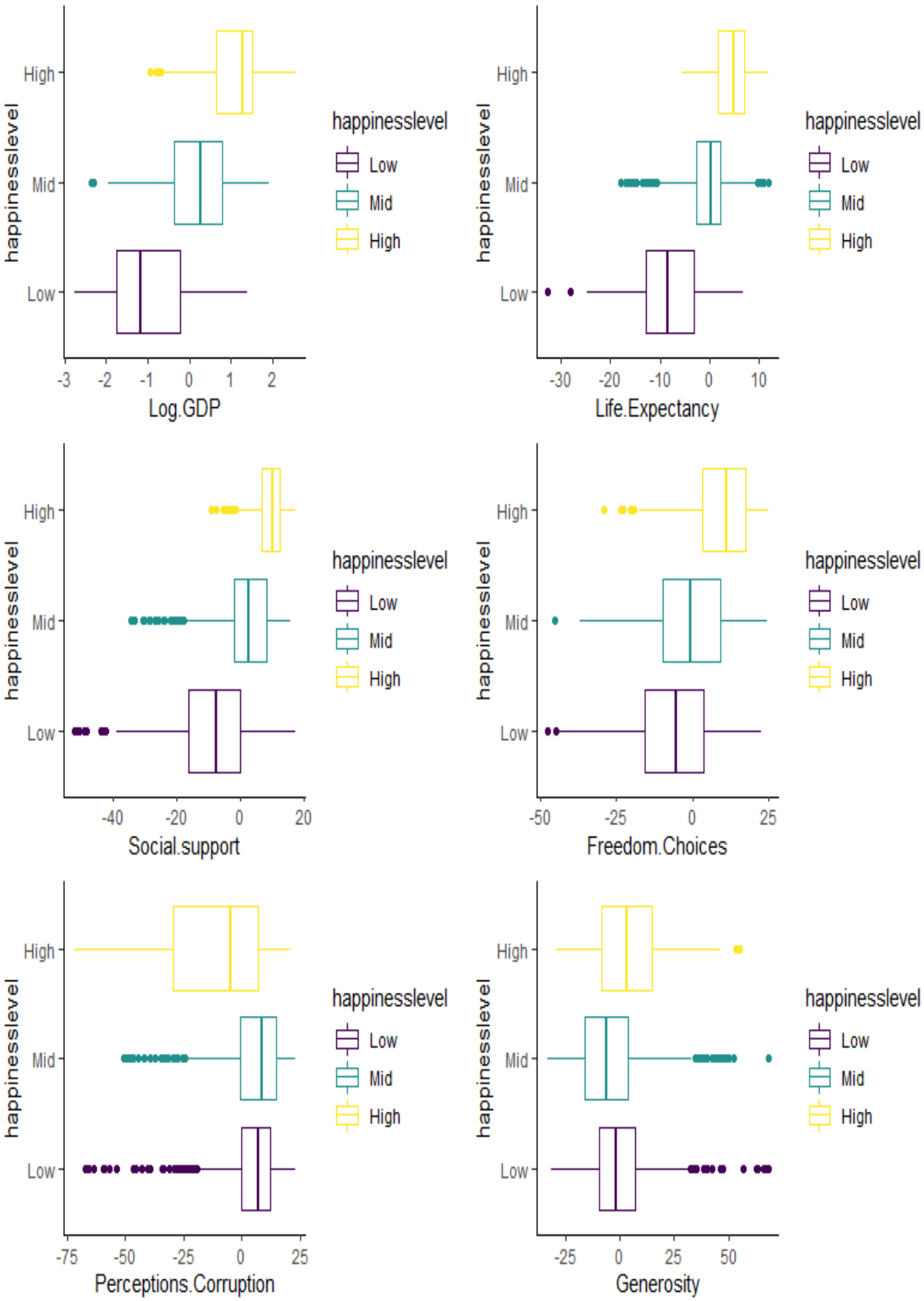

Figure 6: EDA for Main Effects using Imputed Datset 1

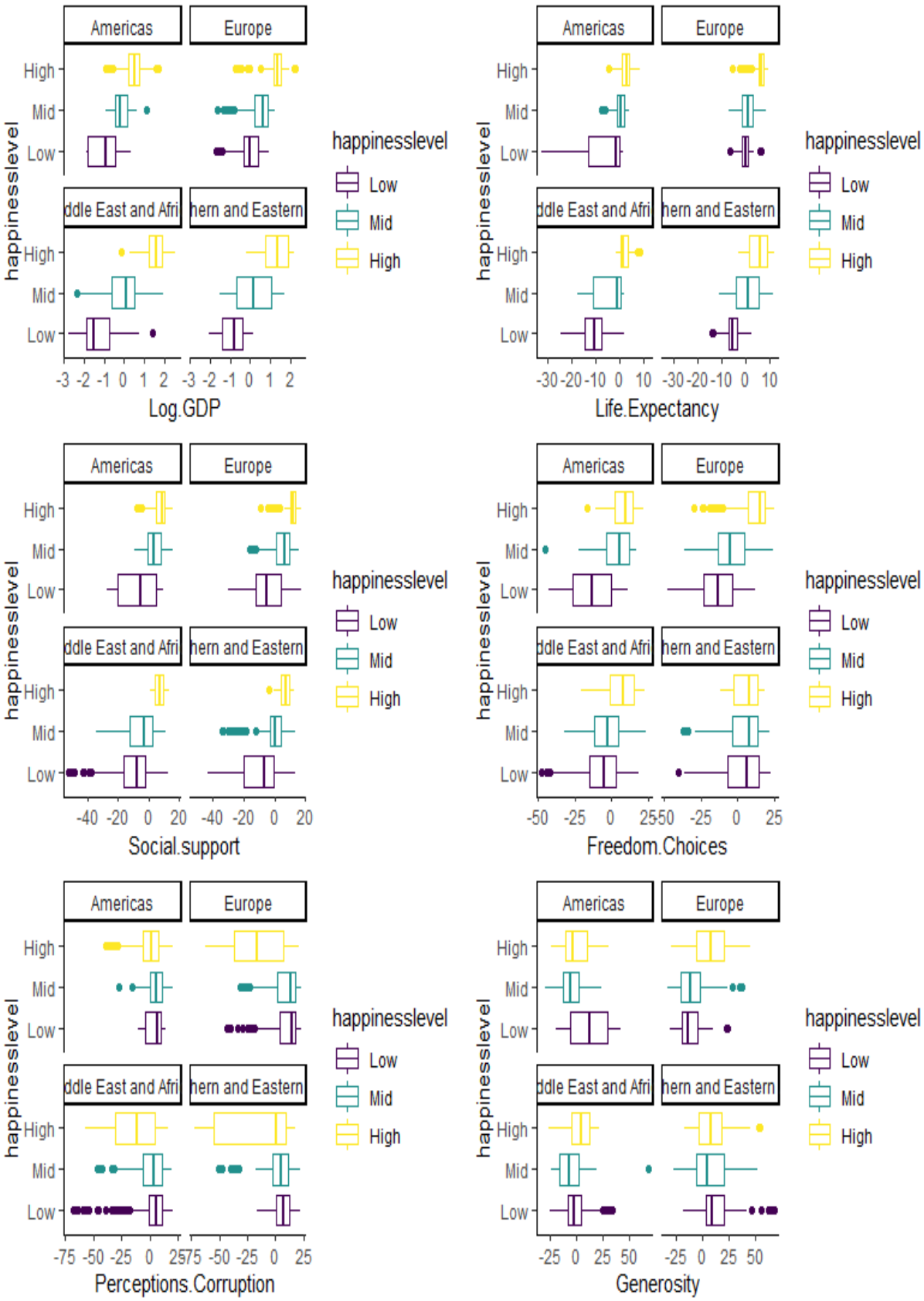

Figure 7: EDA for Interactions using Imputed Datset 1

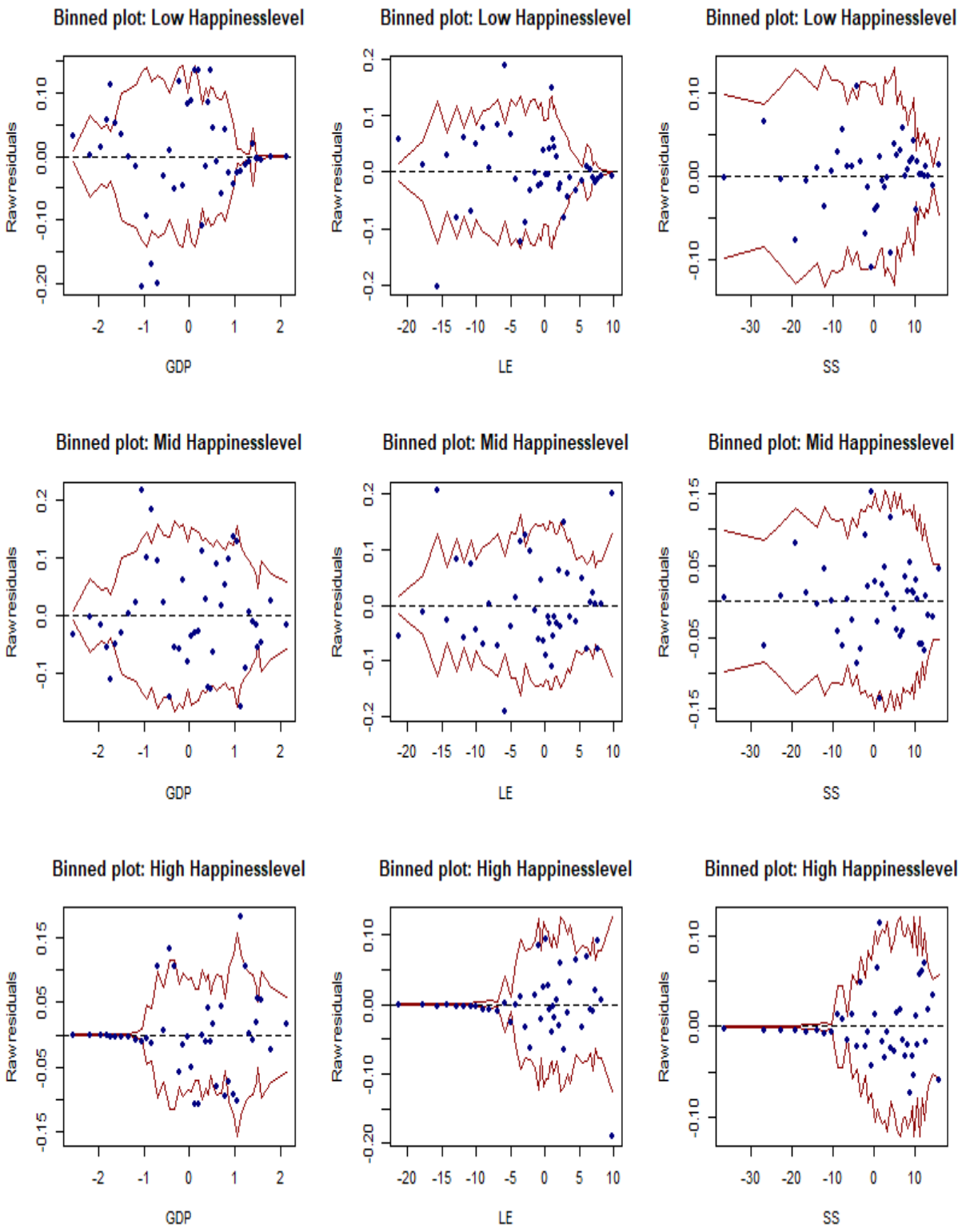

Figure 8: Model Assessment using Imputed Datset 1

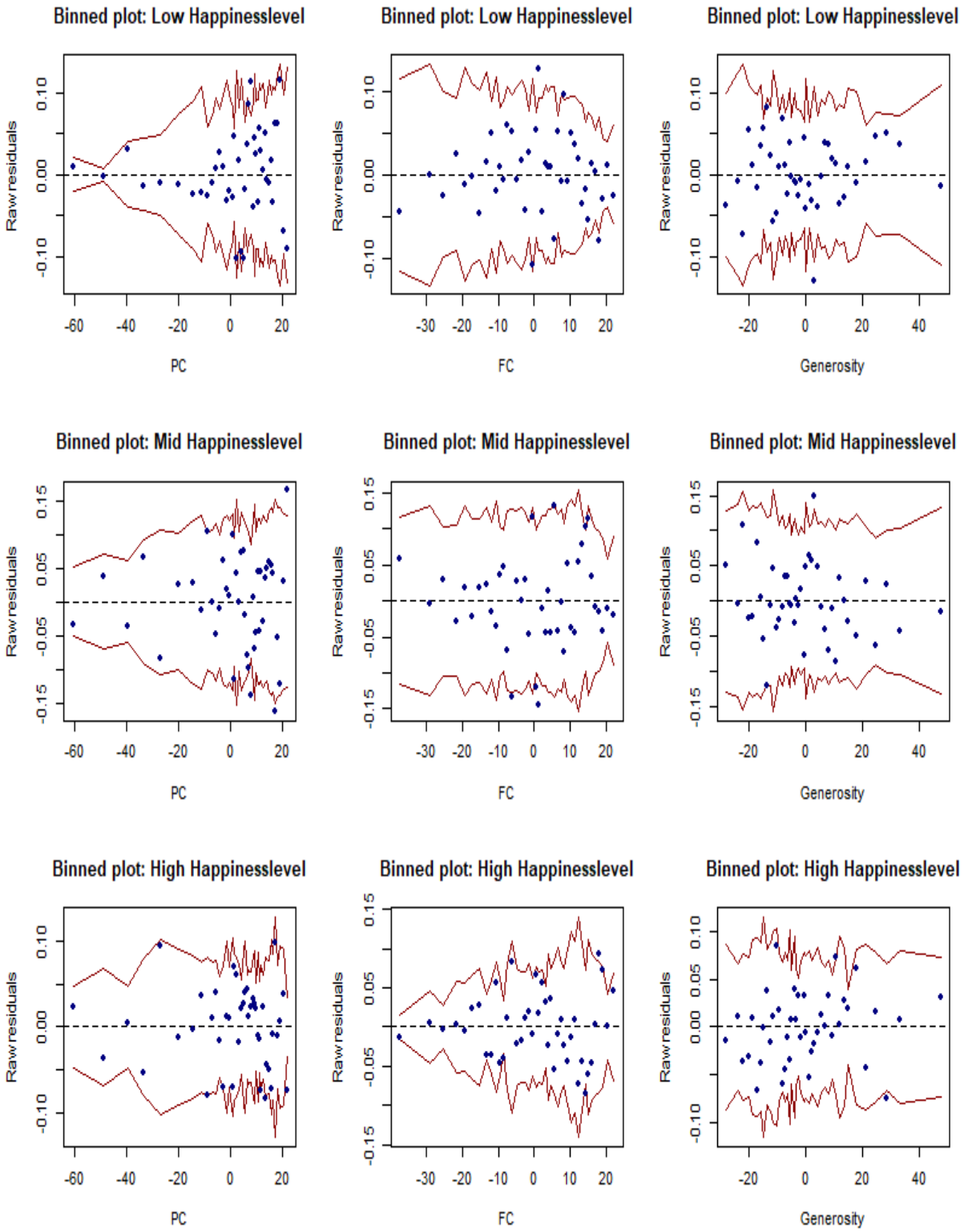

Figure 9: Model Assessment using Imputed Datset 1

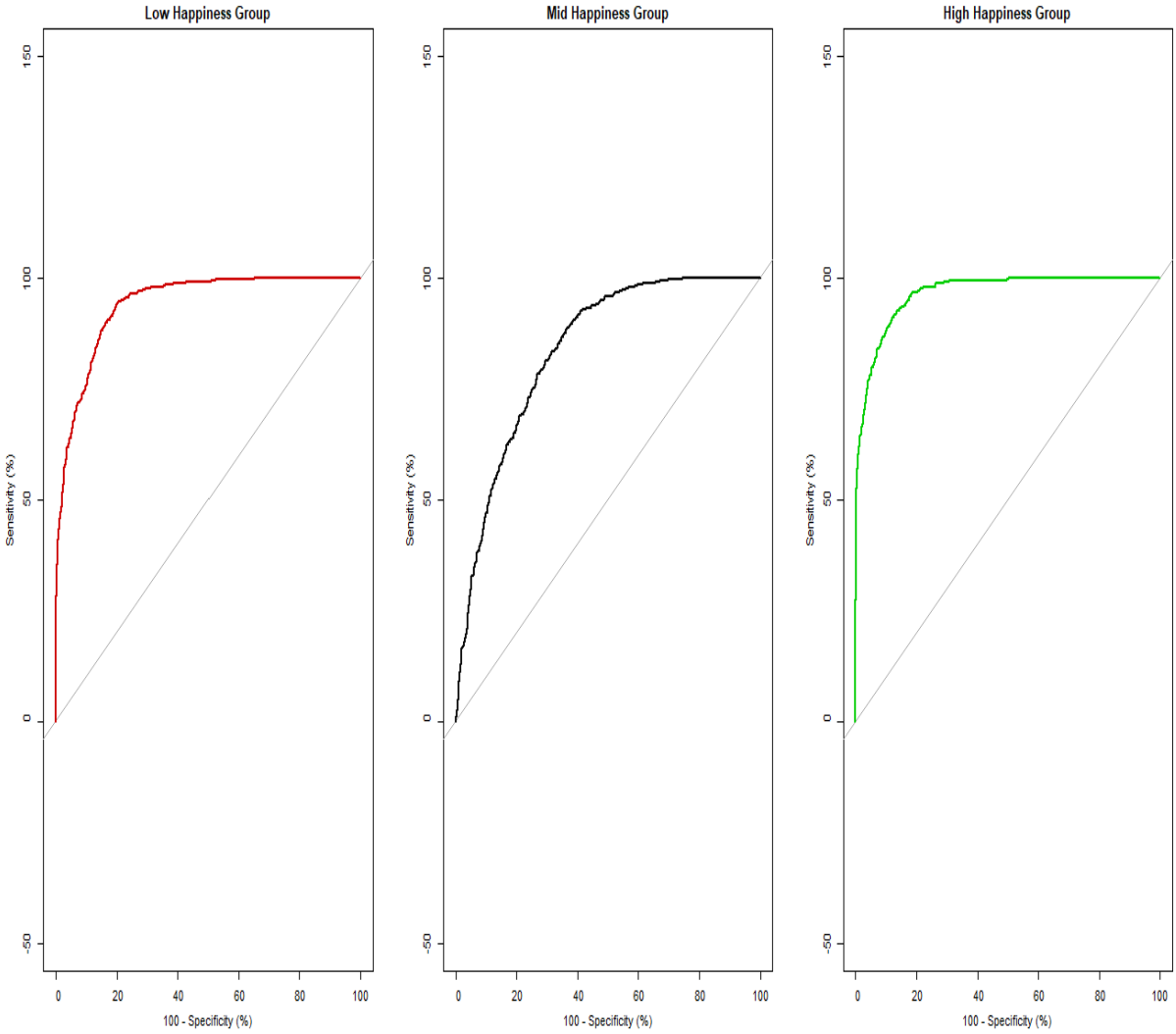

Figure 10: ROC Plot using Imputed Datset 1