Final Project Part I - (Generalized) Linear models

Alpha

11/15/2020

1 Introduction

Happiness is the essence of human pursuit across time and regions. Although the components that determines different countries' happiness index varies, we believe that there are some common underlying factors that affect people's happiness index across the world. With the data provided by United Nations, this project aims to provide insight into the factors that affect people's happiness index in different regions (Part I) as well as predicting happiness index given relevant information (Part II). Combining with each country/region's factual information, such as `GDP`, `population`, and `Health`, the `happiness dataset` consists variables generated based on questionnaires that do not have assigned units and reflect their relative magnitude only.

Since Part I of the project focuses on interpreting the factors' impacts on the happiness index, the crux of the model in Part I is about finding influential factors on happiness index rather than striving for the best predictive model. Hence, we will seek a generalized linear model on the `happiness dataset` collected from 2011 to 2018 for better variable interpretation.

First, we will explore the data in section 2 to get a general idea of the overall relationship among the variables. In section 3, we will conduct a series of modeling-building and feature manipulation/selection to get a final linear model. Without specified or fixed units, the generated variables are scores indicating their values relative to other countries'. This gives us more freedom on data transformation so that we can transform some variables for multinormality while making a flexible balance to their interpretability. We will present the final model to UN in section 4 along with some special findings according to our model. Finally, to evaluate our model, we will apply it on the test dataset to see how well the model fits the data via criteria listed in section 5, which includes bias, maximum deviation, Root Mean Square Error (RMSE), and coverage.

2 Exploratory data analysis

2.1 Data Summaries

The dataset includes happiness data from 2011 to 2018 for worldwide countries. The raw data contains 14 variables and 1110 observations. As summarized in the table 1, all variables except `Country.Territory` are quantitative. Among the quantitative variables, all but `Year` can be viewed as continuous. Besides, clearly the predictors are on different scales; `Population` and `GDP` have much larger orders of magnitude compared to other predictors. It implies that potential transformation may be needed in the following modeling.

Besides, not all countries have full 8-year records. There are a total of 157 countries in the raw data, among which 42 don't have full records over the 8 years from 2011 to 2018 (as shown in table 2). For simplicity, we only keep the countries that have full 8-year records. After the filtering, there remains 115 countries and a total of 920 records with no missing values.

In addition, the countries in the dataset belong to 10 regions, table 3 summarizes the number of records within each region.

Table 1: Descriptive summaries of the raw data

	Variable	Type	Min	1stQuantile	Mean	3rdQuantile	Max
1	Country.Territory	factor	NA	NA	NA	NA	NA
2	Year	integer	NA	NA	NA	NA	NA
3	Population	numeric	320716	5270167	50194829	36270746	1392730000
4	GDP	numeric	1.016	17.199	545.532	312.808	20529.049
5	Support	numeric	0.290	0.742	0.807	0.905	0.987
6	Health	numeric	36.860	58.100	63.526	68.400	76.800
7	Freedom	numeric	0.304	0.652	0.749	0.861	0.985
8	Generosity	numeric	-0.332	-0.120	0.000	0.094	0.680
9	Corruption	numeric	0.047	0.691	0.745	0.866	0.977
10	Positive	numeric	0.369	0.620	0.707	0.798	0.944
11	Negative	numeric	0.095	0.209	0.275	0.327	0.705
12	Government	numeric	0.080	0.340	0.500	0.640	0.994
13	Gini.Index	numeric	0.217	0.368	0.451	0.521	0.961
14	Happiness	numeric	2.662	4.583	5.430	6.247	7.858

Table 2: Frequency table for the recorded years of each country

	1	2	3	4	5	6	7	8
Count of Recorded Years	4	2	9	6	5	6	10	115

Table 3: Number of records within each region

Region	n
Central and Eastern Europe	136
Commonwealth of Independent States	96
East Asia	32
Latin America and Caribbean	144
Middle East and North Africa	88
North America and ANZ	32
South Asia	40
Southeast Asia	40
Sub-Saharan Africa	176
Western Europe	136

2.2 Visualizations

2.2.1 Empirical Distributions

First of all, we would like to get an idea of how the variables are distributed, as shown in figure 1. There are several noticeable patterns:

- Roughly speaking, the continuous response `Happiness` is symmetric and nearly normal, implying that the normal assumption may be basically satisfied (after potential transformation) if we are going to build a Gaussian linear model on it.
- The predictors `GDP` and `Population` have extremely wide ranges and heavy tails on the right, meaning that while most of the countries have a moderate level of population and GDP, there are few countries having extremely large population and GDP. This result is a consistent to the summary statistics in table 2, where the maximum values of `GDP` and `Population` are much larger than their correponding 3nd quantiles. Two possible treatments may be desired before these two predictors enter the model:

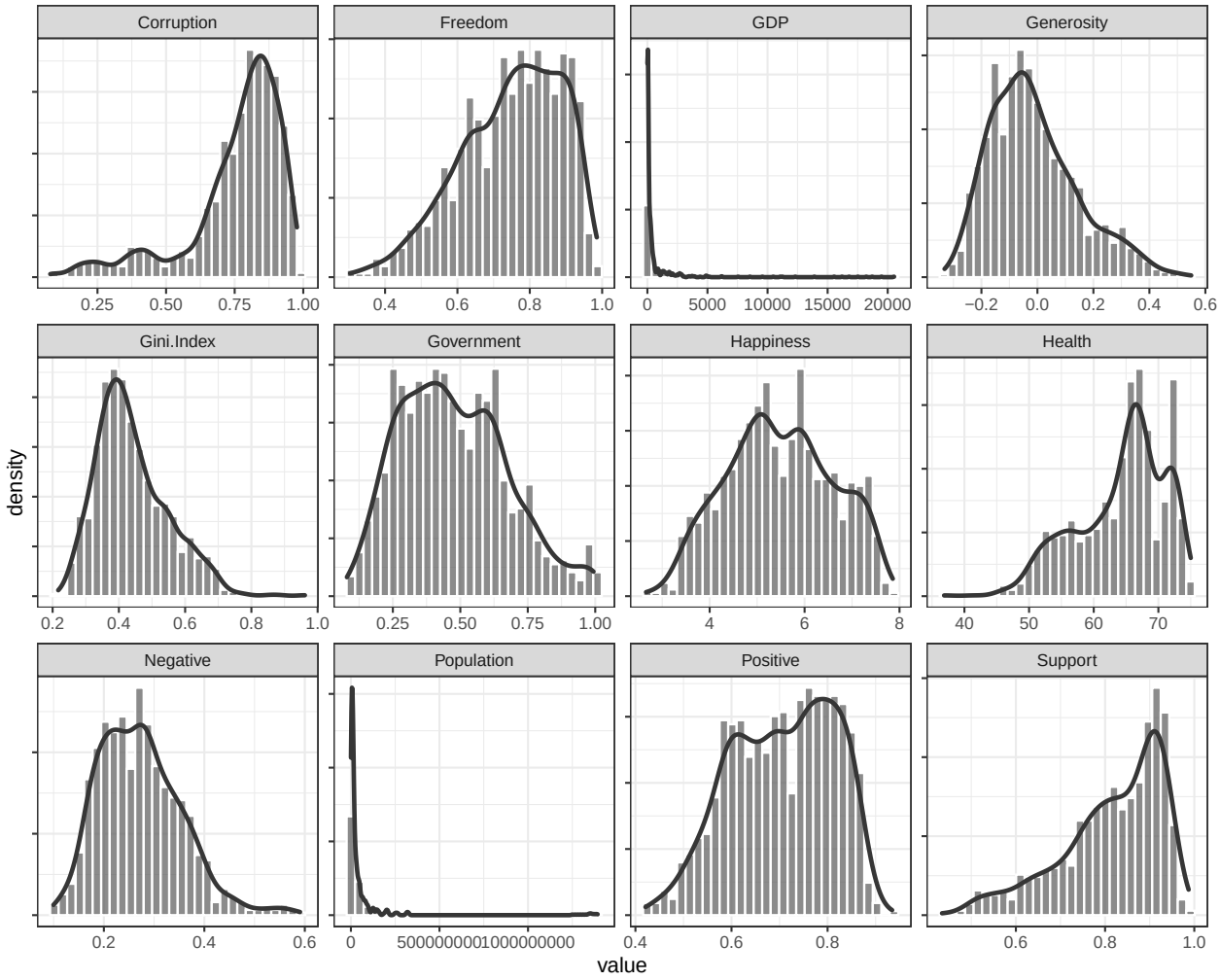

Figure 1: Histogram of variables

1) log-transformation, which is helpful for scaling down strictly positive variables with large order of magnitude, and 2) computing the GDP per capita, which is a commonly used approach to handle this kind of economics variables and also has very good interpretability.

- The empirical distributions of other variables are not too bad; they are all roughly unimodel, lying within a moderate range without any clear outlying values.

Also, we look into whether `Happiness` is distributed differently across different countries, regions and years, as shown in figure 2 and 3. As a result, different countries have different "baseline levels" of `Happiness` and each country's `Happiness` is basically varying around its mean. Besides, `Happiness` also shows clear difference across the 10 different regions. But for years, there is no clear pattern of the response `Happiness` across different years in terms of the mean, although the median seems to have a weakly increasing trend.

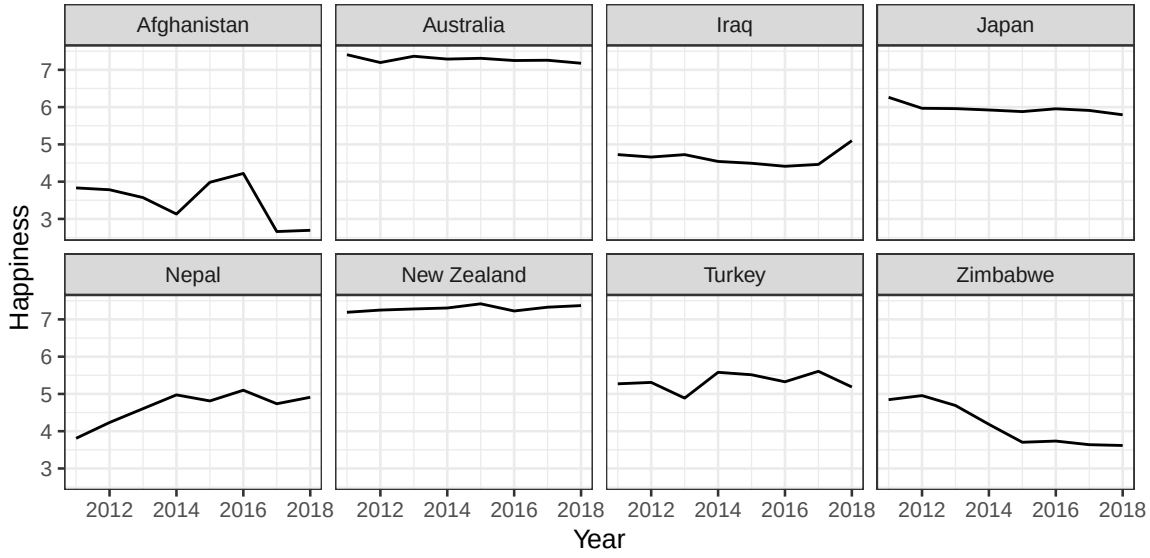

Figure 2: Happiness on different years for different countries

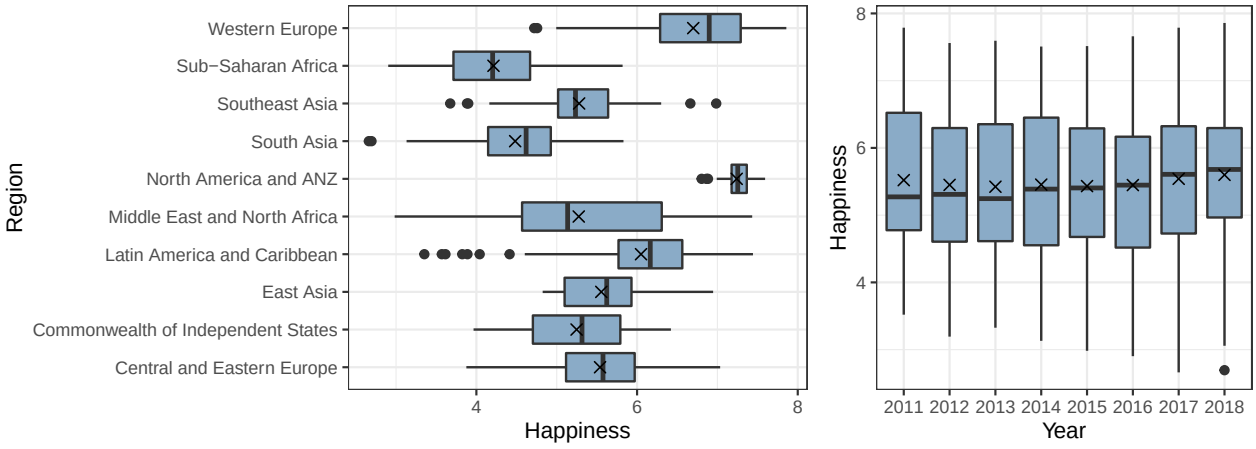

Figure 3: Happiness for different regions and years (crosses denote the means)

2.2.2 Correlation Structure Analysis

In addition to the variables empirical distributions, the correlation structure among the dataset is also of particular interests. Since we have 13 numerical variables here, instead of making all possible pairwise scatterplots, which will be extremely tedious, we plot a correlation map, as shown in figure 4.

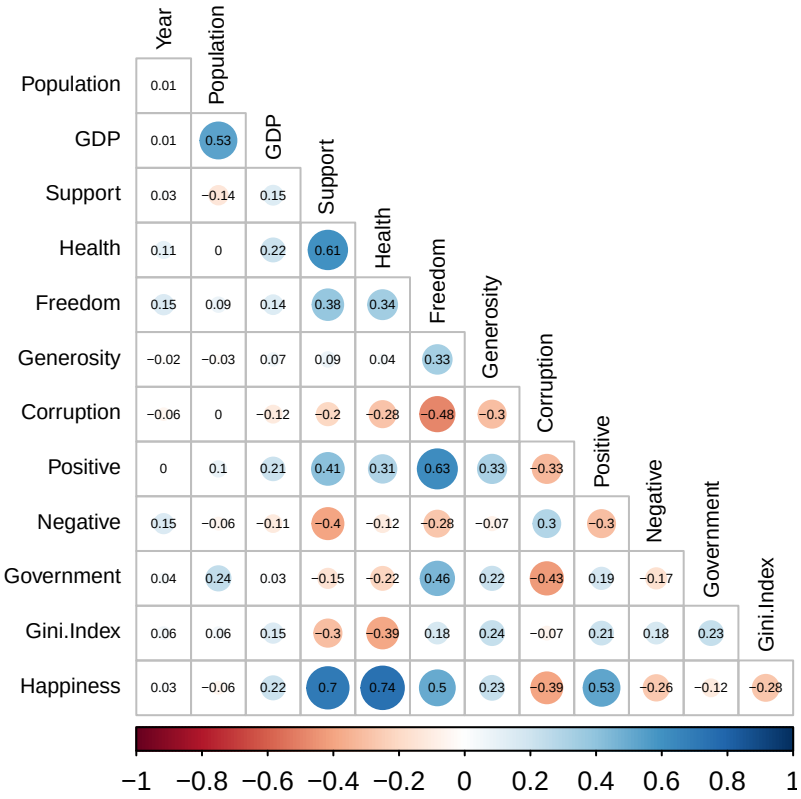

Figure 4: Correlation among variables

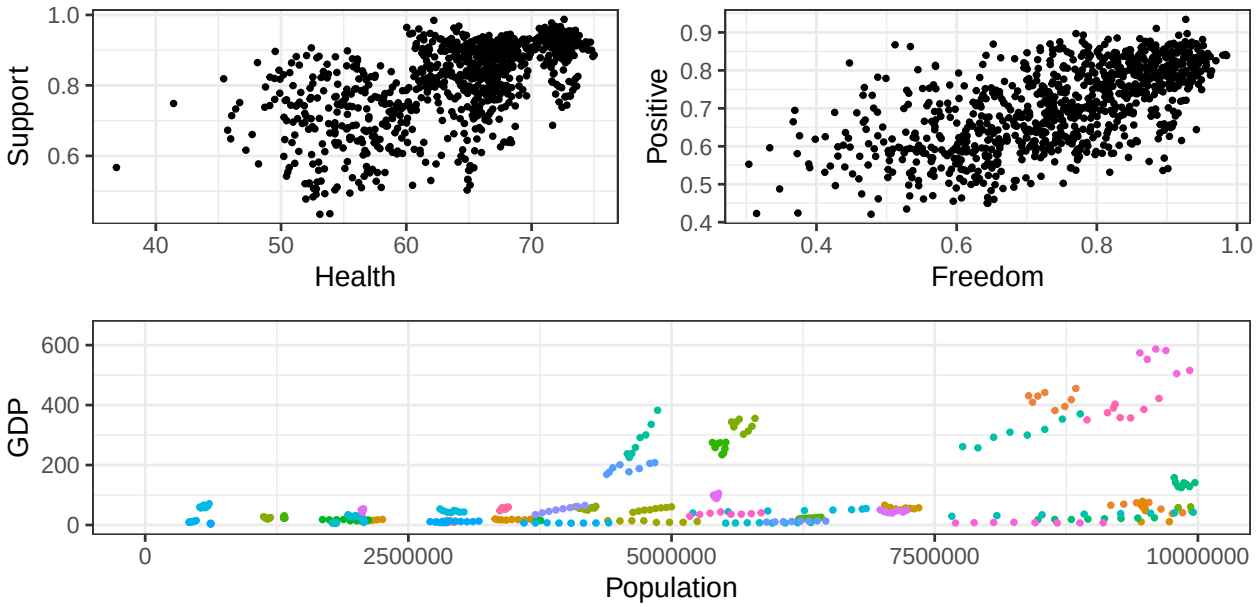

Figure 5: Scatterplots of three groups of predictors

As for the relationships between the response `Happiness` and its predictors, `Support`, `Health` and `Positive` demonstrate some associations with `Happiness` in terms of the linear correlation coefficient. Among the predictors, there are three pairs of predictors having clearly higher correlations than the other predictors; that is, `Population` vs `GDP`, `Health` vs `Support`, and `Positive` vs `Freedom`.

We further draw the pairwise scatterplots for these three pairs, as shown in figure 5. For the scatterplot of `Population` vs `GDP`, we color each point with its belonging country (i.e., `Country.Territory`) and limit the ranges of the axis to make the plot clear.

It turns out that: 1) higher population tends to be associated with higher GDP, and 2) different countries are "well separated" according its population and GDP. In other words, different countries have different scale of population and GDP, which can be totally non-comparable. An advisable choice is to replace `Population` and `GDP` with GDP per capita, which scales down the range as well as provides better interpretation.

As for `Health` vs `Support` and `Positive` vs `Freedom`, since their correlations are not as high as, say, 0.8 or 0.9, and we expect more predictors to improve model fit and provide potential interpretation to `Happiness`, so the advantage of their entering the model may outweigh the potential collinearity problem. Therefore, we don't remove any of them in the following models.

3 Methods

3.1 Train/test Splitting

Before the modeling, we split out data into a training set which includes the data of each country from 2011 to 2017 and a test set which includes the data of each country in 2018. After the split, there are a total of 805 training samples, and 115 testing samples. In the following sections, models will be fit on the training set and be tested on the test set, and prediction will be made on both sets to compare the training RMSE versus test RMSE.

3.2 Initial Modeling

At the very beginning, we build an intial linear regression model, with `Happiness` being the response and all other 13 variables as predictors. Note that in this initial model, `Country.Territory`, which includes 115 countries, creates 114 dummy variables.

Table 4: Anova table for the initial model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Country.Territory	114	925.535	8.119	82.589	0.000
Year	1	0.017	0.017	0.170	0.681
Population	1	0.505	0.505	5.140	0.024
GDP	1	0.167	0.167	1.700	0.193
Support	1	4.223	4.223	42.954	0.000
Health	1	1.382	1.382	14.058	0.000
Freedom	1	2.868	2.868	29.174	0.000
Generosity	1	0.043	0.043	0.437	0.509
Corruption	1	1.099	1.099	11.184	0.001
Positive	1	1.601	1.601	16.289	0.000
Negative	1	0.176	0.176	1.787	0.182
Government	1	1.140	1.140	11.594	0.001
Gini.Index	1	0.244	0.244	2.483	0.116
Residuals	678	66.649	0.098	NA	NA

Table 4 is the ANOVA table for this model and table 5 summarizes the R^2 and RMSE statistics. Unsurprisingly, the initial model, with 114 dummy variables for the countries, provides a high $R^2 = 0.934$, indicating an

Table 5: Statistics of the initial model

RMSE.train	RMSE.test	R2	R2adj
0.288	0.452	0.934	0.921

excellent model fit. However, there are several problems on `Country.Territory` being a predictor:

1. Potential over-fitting problem calls for attention. The test set RMSE is nearly 1.6 times higher than the training set RMSE. In fact, as shown in figure 2, different countries have clearly different baseline levels of `Happiness`, and within each country, there are only 8 records from 2011 to 2018. In other words, when `Country.Territory` enters the model as a strongly influential and deciding predictor on the response and splits the data into 115 tiny parts, we may arrive at quite inaccurate results on new data if, for example, a country for some reason has totally different values on other predictors in a year.
2. Another problem is that when the new data has different `Country.Territory` that doesn't exist in our training set, then we may completely fail to make a fair prediction. In other words, the model is hard to be generalized when meeting with other countries.

Due to these reasons, `Country.Territory` won't be included in the following models. A natural substitution for `Country.Territory` is `Region`. By including `Region` into the model, we consider the region-specific baseline levels of `Happiness` as well as avoid potential over-fitting as it only includes 10 levels so that within each region there are sufficient amount of representative samples (as shown in table 3).

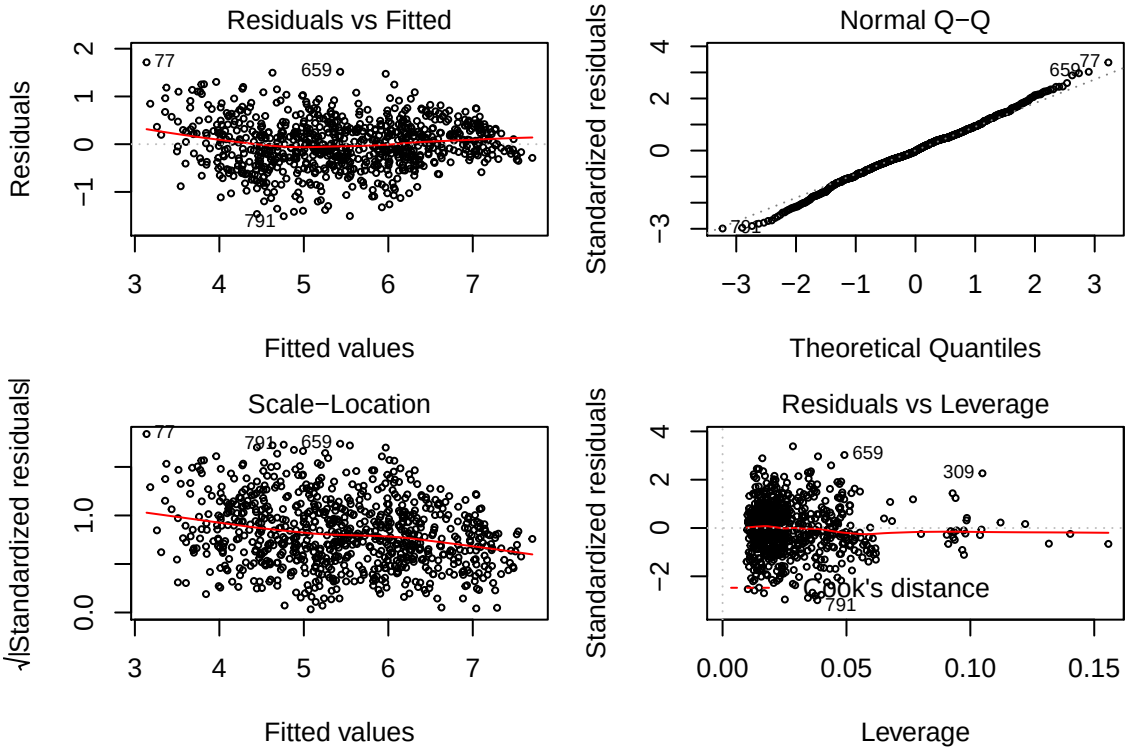

Figure 6: Model diagnosis for the model excluding country but including region

Table 6: Statistics of the model excluding country but including region

RMSE.train	RMSE.test	R2	R2adj
0.507	0.626	0.794	0.788

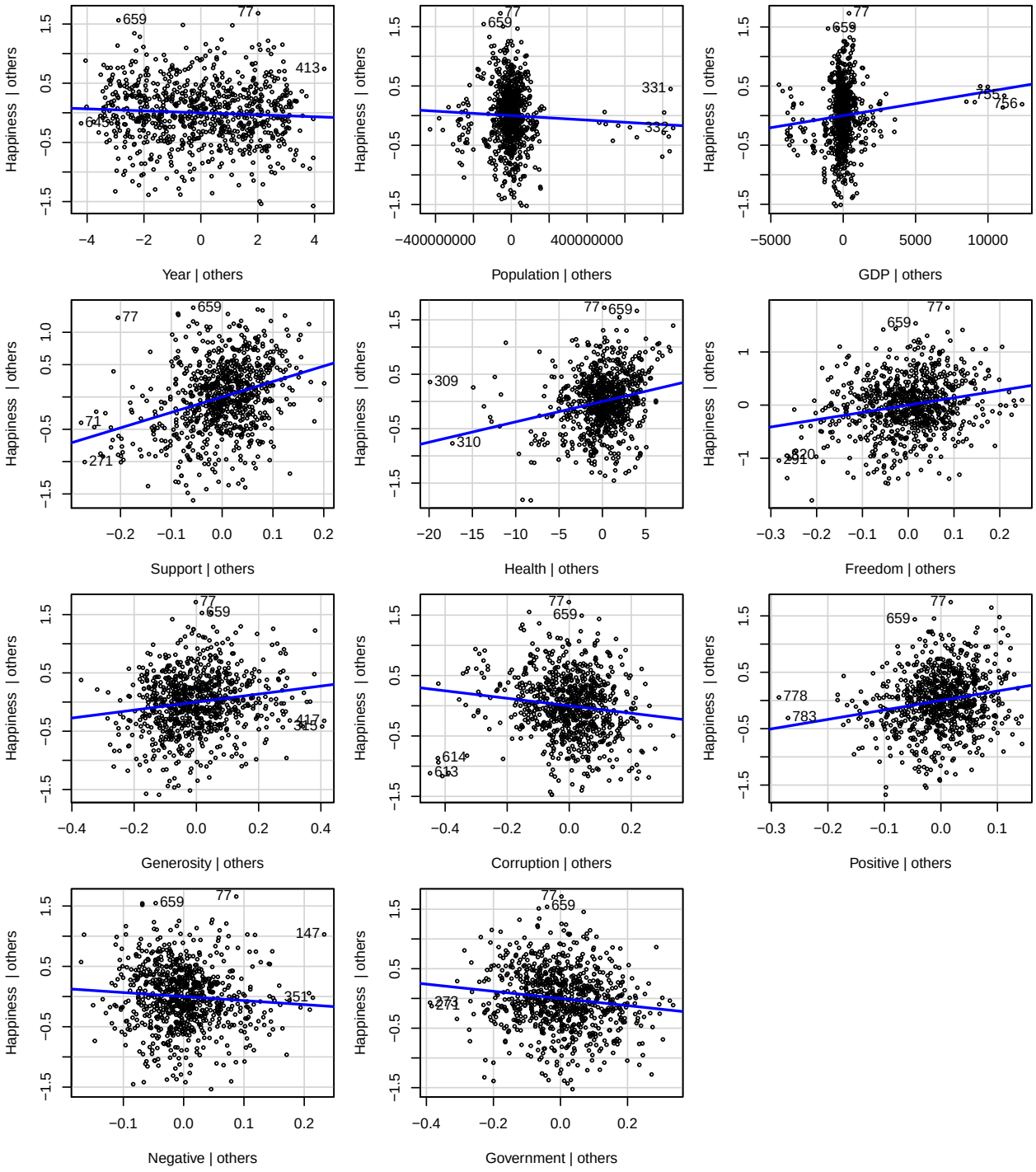

Figure 7: Added variable plots for the model without country intercept

Then we build a model excluding `Country.Territory` while including `Region`. Table 6 summarizes the statistics for this model. Unsurprisingly, we arrive at a much lower R^2 and higher RMSE after replacing `Country.Territory` with `Region`. But the potential overfitting problem is somehow corrected according to the RMSE.

The model diagnosis in figure 6 doesn't shows that there's no serious model assumptions violated. There is a slight downward trend in the variance versus the fitted values, indicating that the variance is somewhat related to the fitted values but in general, nothing worth concerning about. We don't see any influntial points based on the "Residual vs. Leverage" plot either.

However, the added-variable plots in Figure 7 imply some problems. The majority of points in the added-variable plots of `Population` and `GDP` cluster together, with a few points lying far away and dragging the partial-regression lines. This is because `Population` and `GDP` have large order of magnitudes compared to other predictors and are heavily right-tailed, as discussed before. Therfore, we will create a new predictor called `GpC` to denote the GDP per capita. Also, transformation will be considered, in the next section.

Table 7: Coefficients summary for the model with GDP per capita (GpC)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.897	18.273	0.979	0.328
Year	-0.009	0.009	-0.940	0.347
Support	2.084	0.242	8.621	0.000
Health	0.028	0.006	5.021	0.000
Freedom	1.108	0.208	5.328	0.000
Generosity	0.562	0.141	3.992	0.000
Corruption	-0.105	0.161	-0.649	0.517
Positive	1.567	0.294	5.339	0.000
Negative	-0.481	0.299	-1.608	0.108
Government	-0.389	0.148	-2.629	0.009
Gini.Index	-1.030	0.238	-4.328	0.000
RegionCommonwealth of Independent States	-0.008	0.078	-0.105	0.917
RegionEast Asia	-0.362	0.109	-3.326	0.001
RegionLatin America and Caribbean	0.382	0.083	4.614	0.000
RegionMiddle East and North Africa	0.192	0.081	2.368	0.018
RegionNorth America and ANZ	0.132	0.142	0.931	0.352
RegionSouth Asia	0.025	0.114	0.223	0.824
RegionSoutheast Asia	-0.285	0.119	-2.396	0.017
RegionSub-Saharan Africa	-0.408	0.101	-4.035	0.000
RegionWestern Europe	-0.146	0.092	-1.583	0.114
GpC	18699.004	2006.458	9.319	0.000

Replacing `Population` and `GDP` with `GpC`, we build a new model. Table 7 shows the coefficients estimates and the t-test for this model, and figure 8 shows the model diagnosis. Again, there are no identifiably serious problems in the model diagnosis. But a noticeable result is that the coefficient estimates of `GpC` is extremely large compared to others. We stop this section here at this model, and in the following section, tranformation will be considered to see if there can be further improvement.

3.3 Transformation

The recommended Box-Cox power transformation is shown in table 8, and the last column summarizes our *finally decided power transformation*. Here are our decided transformations:

1. There are two predictors `GpC` and `Negative` whose rounded recommended power are 0 and thus need log transformation.

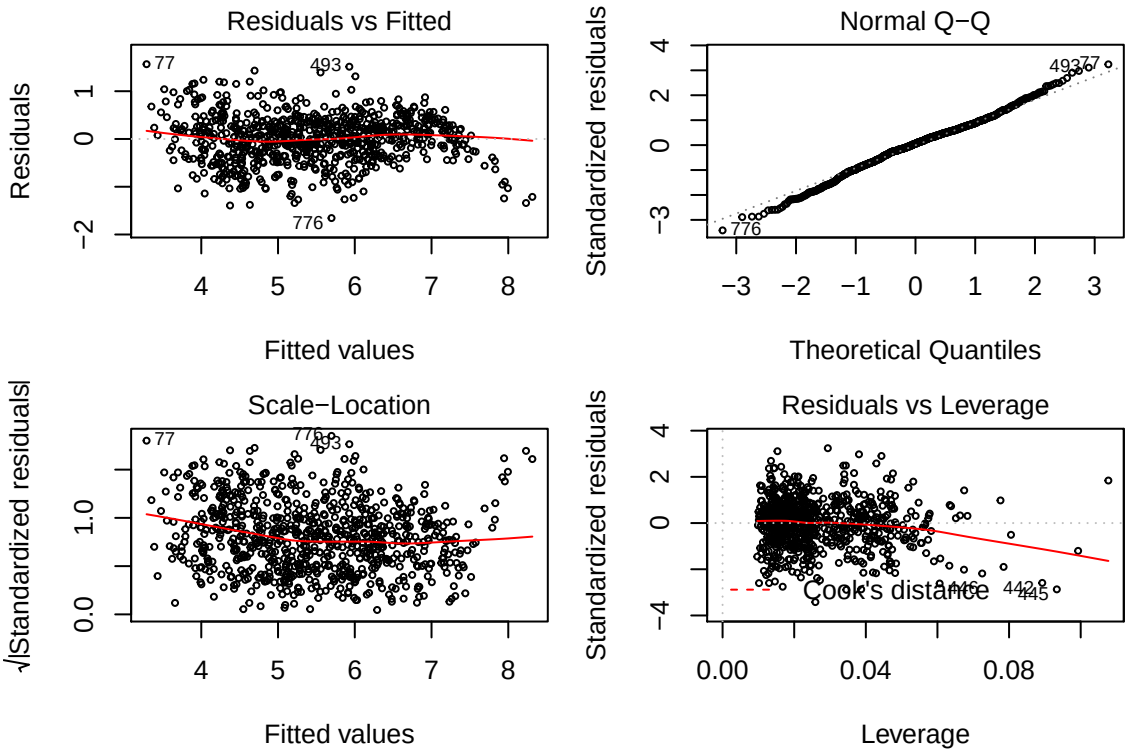

Figure 8: Model diagnosis for the model with GDP per capita (GpC)

Table 8: bcPower Transformations to Multinormality

	Est.Power	Rounded.Pwr	Wald.Lwr.Bnd	Wald.Upr.Bnd	Decided.Pwr
Happiness	1.64	1.6	1.41	1.87	2
Year	0.74	1.0	-60.11	61.58	1
Support	4.34	4.3	3.92	4.76	1
Health	4.67	4.7	4.19	5.14	1
Freedom	2.48	2.5	2.20	2.76	2
Generosity	-1.74	-2.0	-2.23	-1.24	-2
Corruption	2.76	2.8	2.53	2.99	1
Positive	1.87	2.0	1.50	2.25	2
Negative	0.10	0.0	-0.07	0.26	0
Government	0.53	0.5	0.42	0.65	0.5
Gini.Index	-0.31	-0.5	-0.52	-0.11	-0.5
GpC	0.04	0.0	0.00	0.07	0

2. `Government` and `Gini.Index` is recommended the power 0.5 and -0.5 respectively, meaning that square root and the inverse square root transformation will be helpful.

3. The recommended power for the response `Happiness` is 2, so square transformation will be applied.

4. There are three predictors `Support`, `Health` and `Corruption` having clearly larger powers compared to others. Look at the histograms in figure 1, then we can notice that these three variables are all left-skewed. In fact, it is commonly seen that variables with this kind of distributions will be recommended a large power by Box-Cox transformation in which indeed a maximim likelihood estimation is going on. But here, two facts support us not to do the recommended transformations on these variables: 1) we desire a more understandable and interpretable models, 2) the large power will significally scaling up or

down the range of the variables, may leading to completely non-comparable coefficients estimates with others.

5. For the rest of the variables, we refer to the recommended powers. Basically, they are all using square transformation.

Table 9: Coefficients summary for the model after transfomation

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	70.309	185.497	0.379	0.705
Year	-0.004	0.092	-0.045	0.964
Support	11.717	2.571	4.556	0.000
Health	0.078	0.061	1.273	0.203
Freedom	7.087	1.555	4.559	0.000
Generosity	-15.347	1.693	-9.067	0.000
Corruption	-4.738	1.573	-3.012	0.003
Positive	10.820	2.194	4.932	0.000
Negative	-2.820	0.842	-3.351	0.001
Government	-3.262	2.009	-1.624	0.105
Gini.Index	7.351	1.445	5.087	0.000
GpC	4.715	0.317	14.888	0.000
RegionCommonwealth of Independent States	1.163	0.805	1.445	0.149
RegionEast Asia	-3.682	1.112	-3.313	0.001
RegionLatin America and Caribbean	6.334	0.872	7.260	0.000
RegionMiddle East and North Africa	2.296	0.816	2.813	0.005
RegionNorth America and ANZ	3.726	1.369	2.721	0.007
RegionSouth Asia	3.992	1.181	3.380	0.001
RegionSoutheast Asia	-0.898	1.218	-0.737	0.461
RegionSub-Saharan Africa	-0.345	1.035	-0.334	0.739
RegionWestern Europe	0.564	0.863	0.653	0.514

Table 10: Comparison between models before and after transformation

	RMSE.train	RMSE.test	R2	R2adj
Model before transformation	0.484	0.595	0.812	0.808
Model after transformation	0.481	0.605	0.844	0.840

After the variable transformation, we build a linear regression model and see the following improvements:

1. The added-variable plots shown in figure 9 are much better than before. For each of the variables, the points are evenly scattered in general. Also, these plots can imply some potential outliers or influential points, such as the 493rd and the 776th observations; but overall, they are not quite outlying.

2. In addition, as shown in table 9, the coefficients estimates for all the predictors are basically of similar scales.

3. In terms of the RMSE and R^2 statistics, the improvement brought by transformation is slight, as shown in table 10. We somewhat increase the model fit on the training data, while making slightly worse prediction on the test data. Generally speaking, the transformed predictors have better explainatory ability on the response according to R^2.

3.4 Interactions and Model Selection

Now it's time to consider possible interactions among predictors. To guarantee the interpretability of the model, we limit the highest order of interaction to be 2.

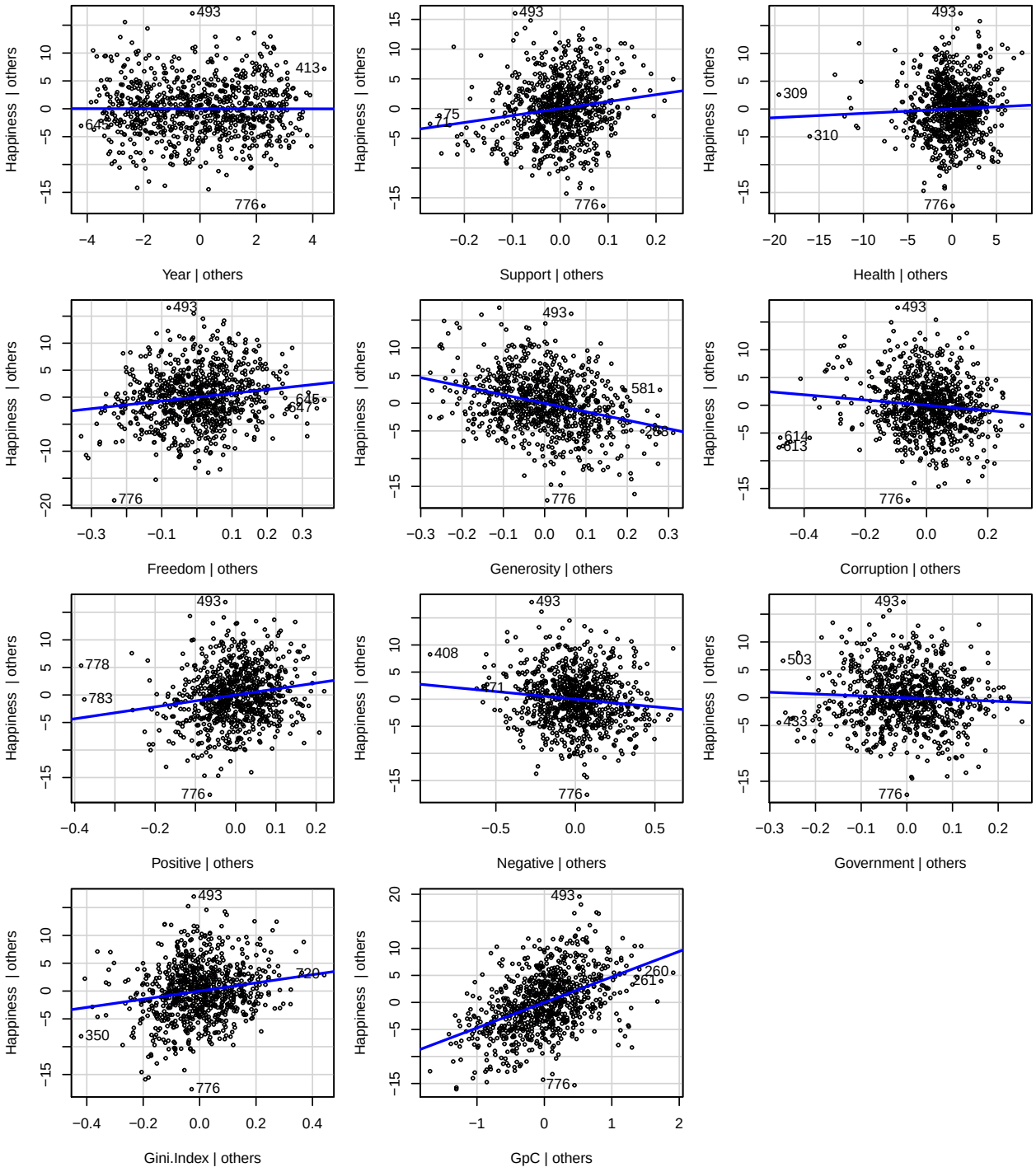

Figure 9: Added variable plots for the model after transfomation

Table 11: Comparison between model selection and shrinkage estimates

	Num of Params	RMSE.train	RMSE.test	R2	Max Dev	Bias
Model without interactions	21	0.481	0.605	0.844	2.251	-0.081
Backward AIC	99	0.353	0.461	0.913	1.859	-0.037
Backward BIC	47	0.392	0.515	0.892	2.151	-0.023
Lasso	74	0.388	0.491	0.893	1.726	-0.048

There are a total of 175 main effects and 2-order interactions in the model. We conduct backward AIC, BIC to do model selection and also do lasso shrinkage estimation. The statistics results are summarized in table 11.

As a result, the model selected by AIC has the highest R^2; this is an unsurprising result since AIC tends to select a large model to ensure the best predictive ability. On the contrary, BIC selects a much smaller model, involving 47 terms. Lasso provides a model of a middle size, smaller than AIC while larger than BIC.

Our final choice is the BIC model selection, due to the following reasons: 1) it provides a small model which can be easily interpreted; 2) comparing the results of Lasso and BIC, then we can note that the additional 27 terms from BIC to Lasso don't bring in too much improvement in R^2 and RMSE; 3) BIC results in a model with the smallest bias, meaning that on average, the prediction error is the lowest.

3.5 Final Model

Our final model is the model selected by backward BIC from the full models that include all the main effects and their 2-order interactions. Table 10 shows the coefficents estimates, the 95% confidence intervals as well as the p-value in the t-test. The final model includes 47 terms, where 20 are main effects and 27 are 2-order interactions.

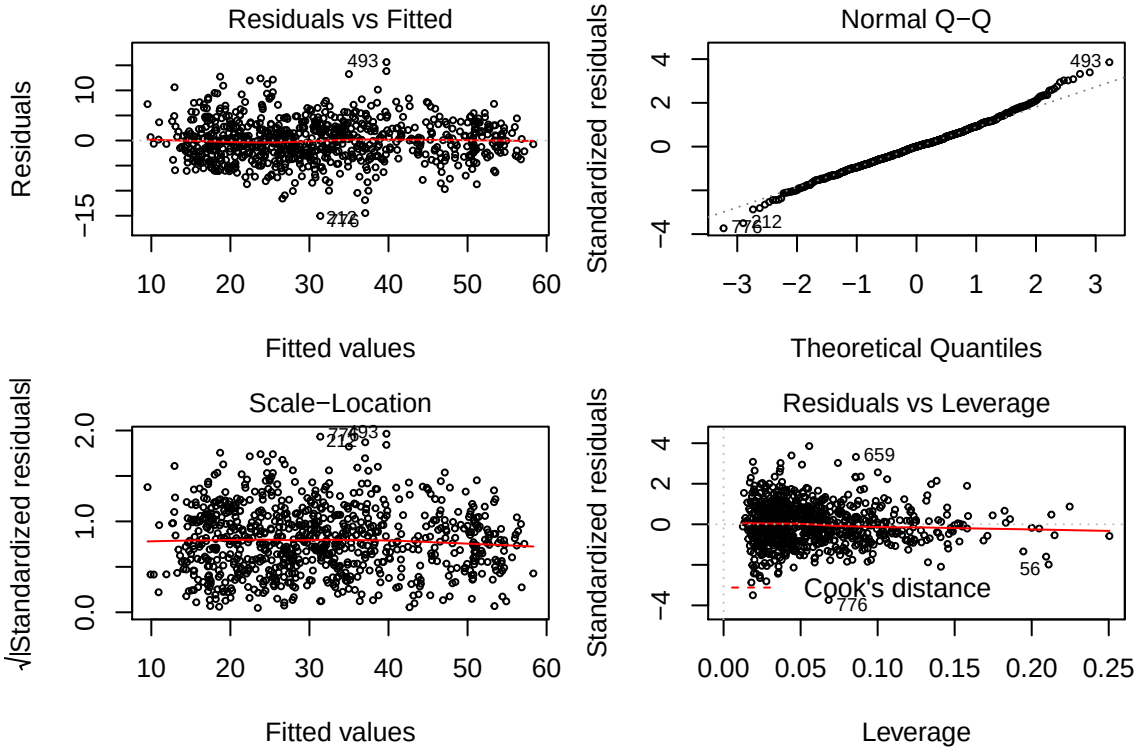

Figure 10: Model diagnosis for the final model

Table 12: Coefficients estimates of the final model

	Predictor	Estimates	2.5%	97.5%	Pr(>|t|)
1	(Intercept)	-255.269	-360.662	-149.876	0.000
2	Support	36.994	17.836	56.152	0.000
3	Health	1.349	0.075	2.623	0.038
4	Freedom	-24.303	-49.786	1.181	0.062
5	Generosity	-5.588	-9.027	-2.148	0.001
6	Corruption	203.068	141.321	264.816	0.000
7	Positive	21.780	5.366	38.193	0.009
8	Negative	-30.048	-43.745	-16.352	0.000
9	Government	-57.845	-95.390	-20.300	0.003
10	Gini.Index	4.682	1.967	7.397	0.001
11	GpC	-22.132	-30.082	-14.183	0.000
12	RegionCommonwealth of Independent States	1.966	-23.923	27.855	0.882
13	RegionEast Asia	10.772	-18.022	39.565	0.463
14	RegionLatin America and Caribbean	73.197	50.119	96.276	0.000
15	RegionMiddle East and North Africa	73.342	48.215	98.469	0.000
16	RegionNorth America and ANZ	63.470	-54.216	181.155	0.290
17	RegionSouth Asia	62.669	3.259	122.078	0.039
18	RegionSoutheast Asia	111.158	58.712	163.603	0.000
19	RegionSub-Saharan Africa	-19.412	-49.885	11.062	0.212
20	RegionWestern Europe	39.979	1.613	78.345	0.041
21	Support:Corruption	-32.901	-57.436	-8.367	0.009
22	Health:Government	1.341	0.750	1.932	0.000
23	Health:GpC	0.166	0.069	0.264	0.001
24	Freedom:Positive	27.386	12.073	42.700	0.000
25	Freedom:Government	-24.685	-40.056	-9.315	0.002
26	Freedom:GpC	-2.765	-4.710	-0.820	0.005
27	Corruption:GpC	14.535	9.690	19.381	0.000
28	Corruption:RegionCommonwealth of Independent States	6.071	-6.802	18.944	0.355
29	Corruption:RegionEast Asia	6.315	-13.502	26.131	0.532
30	Corruption:RegionLatin America and Caribbean	-15.213	-29.425	-1.001	0.036
31	Corruption:RegionMiddle East and North Africa	3.654	-13.539	20.846	0.677
32	Corruption:RegionNorth America and ANZ	-20.963	-37.770	-4.155	0.015
33	Corruption:RegionSouth Asia	2.634	-24.375	29.644	0.848
34	Corruption:RegionSoutheast Asia	-17.947	-43.804	7.909	0.173
35	Corruption:RegionSub-Saharan Africa	26.557	8.852	44.262	0.003
36	Corruption:RegionWestern Europe	-42.986	-55.236	-30.737	0.000
37	Positive:Government	-38.649	-63.385	-13.912	0.002
38	Negative:GpC	-2.221	-3.305	-1.136	0.000
39	GpC:RegionCommonwealth of Independent States	0.511	-1.613	2.635	0.637
40	GpC:RegionEast Asia	1.637	-0.947	4.220	0.214
41	GpC:RegionLatin America and Caribbean	4.709	2.807	6.611	0.000
42	GpC:RegionMiddle East and North Africa	6.287	4.431	8.143	0.000
43	GpC:RegionNorth America and ANZ	3.519	-7.930	14.968	0.546
44	GpC:RegionSouth Asia	4.811	0.170	9.453	0.042
45	GpC:RegionSoutheast Asia	7.601	4.295	10.907	0.000
46	GpC:RegionSub-Saharan Africa	0.522	-1.751	2.796	0.652
47	GpC:RegionWestern Europe	0.312	-3.269	3.894	0.864

The diagnosis plots look even better than previous plots, providing enough evidence that our linear model is suitable on our datset. The "Residual vs. Fitted" plot is a null plot and we do not see any heteroscedasticity in the residuals. The normal Q-Q plot shows that the residuals follow a normal distribution in general, except for the tail regions. The leverage plot shows that there are no influencial points.

All included variables show significance in the t-test, though few interactions are less significant. To explain our model, and demostrate how some of the variables affect the estimate of `happiness`, we will select a couple of the relatively more important variables to elaborate on, such as `Support`, `Negative`, `GpC`, `Curruption`, `RegionSoutheast Asia`, `RegionLatin America and Caribbean`, `RegionMiddle East and North Africa`, `Corruption:RegionWestern Europe`.

As mentioned in the introduction, the variables that measure the subjective feeling collected from the surveys are unitless, so we evaluate them based on their transformed values. The baseline `Happiness` score is -255.269 but has no practical meaning because there are no countries with all predictors being 0. However, as thee value of each variable increases in our final model, `Happiness` decreases or increases accordingly.

- For example, if a country's `Support` increases by 1 unit, the estimated squared happiness index will increases by 36.994 with a 95% confidence interval ranging from 17.84 to 56.152 units.
- The coefficient of `Negative` is -30.048, indicating that a country's squared happiness index will on average decrease by 2.86 with a with a 95% confidence interval ranging from 4.17 to 1.55 units if the average of three negative affect measures (worry, sadness and anger) increase by 10%.
- The coefficient of `GpC` is -22.132, indicating that a country's squared happiness index will on average decrease by 2.11 with a with a 95% confidence interval ranging from -2.86 to -1.35 units if the average of three negative affect measures increase by 10%.
- The coefficient of `Corruption` is 203.068, indicating that a country's squared happiness index will on average increase by 203.068 with a with a 95% confidence interval ranging from 141.321 to 264.816 units if the average of a country's corruption score increase by 1 unit. This is a bit counter-intuitive, and we will discuss this later.
- The coefficient of `RegionSoutheast Asia` is 111.158, indicating that a country's squared happiness index will on average increase by 111.158 with a with a 95% confidence interval ranging from 58.712 to 163.603 units if the country is in Southeast Asia.
- The coefficient of `RegionLatin America and Caribbean` is 73.197, indicating that a country's squared happiness index will on average increase by 73.197 with a with a 95% confidence interval ranging from 50.119 to 96.276 units if the country belongs to the Latin America and Caribbean region.
- The coefficient of `RegionMiddle East and North Africa` is 73.197, indicating that a country's squared happiness index will on average increase by 73.342 with a with a 95% confidence interval ranging from 48.215 to 98.469 units if the country belongs to the Middle East and North Africa region.
- The coefficient of `Corruption:RegionWestern Europe` is -42.986, indicating that a country's squared happiness index will on average decrease by 42.986 with a with a 95% confidence interval ranging from 55.236 to 30.737 units if the average of a country in Western Europe and its corruption score increases by 1 unit.

This model gives a good intuition to the important factors that contribute to the change of squared happiness index in a given country and is relatively easy to interpret. The corruption level affects the squared happiness index more negatively for a country in Western Europe than for a country in Southeast Asia. This could be interpreted from a few different levels. First, a country in Western Europe would likely to have a have higher living standards (higher GDP per Capital) and a more democratic government than a country from Southeast Asia, so with the basic material needs met and more political freedom, any increasee in the corruption level would be perceived as more intensely, and thus reflected in the relative larger decrease in the squared happiness index.

Nevertheless, one potential drawback is that some of the predictors have the opposite effect on squared happiness index because there are interaction terms that are highly correlated with them and hence they

need to balance their explainatory power by taking opposite signs from each other. For example, it makes no sense that the increasing in the level of corruption in a country indicates a significant increase in their squared happiness index. This is because there are many interaction terms taking `Corruption` into account. For instant, an increase in `Corruption:RegionWestern Europe` is associated with a decrease in squared happiness index. In addition, although it is obvious that a country's GDP per Capital should be positively correlated with the squared happiness index since a country's economy determines people's ability to have sufficient material goods, `GpC` has a negative effect on `Happiness` here, and that is also potentially due to the interaction terms that contain `GpC`, with similar reasons as `Corruption`. If we only look at the interaction terms of GpC with regions, we find that a 10% increase in `GpC` for a country that's in Western Europe has much less significant impact on `Happiness` than the same proprtion of increase in `GpC` for a country that's in the Southeast Asia region or the region containing Middle East annd North Afria. This is reasonable since there are much more developed countries in Western Europe and their GDP per Capita is generally higher than the countries in other two regions. Due to the concept of diminished marginal untility in economy, higher GDP or more material goods would only bring so much utility (which can be thought of as the enjoyment that something brings) upto a certain point. Hence, although a 10% increase in `GpC` for a country in Southeast Asia, Middle East, or North Afria would be less in absolute amont of value than the same percent increase for a country in Western Europe, the former one is associated with a much higher squared happiness index for its people.

4 Summary and Conclusion

4.1 Report for the United Nations

In making an effort to analyze the world's happiness index, we will first present you with the best model that we built to explain some of the underlying factors that affect countries' happiness index across the world, and then use 4 smaller models to further discover some factors' different impacts on a country's happiness index given its general geographical location.

4.1.0.1 General model

To build a model that would fit on any country in the world given relavant information, we grouped the countries by regions rather than using each individual country as separate predictors in the model. This helps us to achieve generalizability of the model on a broader range of countries that may not be included in our given dataset. We also created the variable `GpC` to better capture the information a country's GDP and population. We did transformations on a couple of the variables which can be found in section 3 with detailed explaination and procedures. In particular, we squared the happiness index, but since it doesn't have units in the first place, we can evaluate it based on its relative value after such transformation.

We fit a linear model using BIC model selection method, selecting the best predictors for a large pool of variables containing all possible interaction terms to in order to capture the combined effects of any given two original predictors. The best BIC model yields 47 significant predictors that affect the squared happiness index in different degrees.

Here are a couple of important findings that were also mentioned earlier: - If a country's score for `Support`, i.e., the national average response to the 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?"* –with 1 indicating *"yes"*, increases by 1 unit, the estimated increase of `Happiness` is about 6.08, and we are 95% confident that this increase is within the range from 4.223 to 7.49 units. - With a 10% increase in `Negative`, i.e., the average of measures of worry, sadness and anger, a country's squared happiness index will on average decrease by 5.48, and we are 95% confident that this decrease is within the range from 6.614 to 4.04 units. - If the average of a country in Western Europe and its corruption level increases by 1 unit, then its squared happiness index will on average decrease by 6.56, and we are 95% confident that this decrease is within the range from 7.43 to 5.54 units.

This finding supports the common assumption that countries' squared happiness index would be affected differently by the same predictors if they are located in different regions, and we want to examine regional differences on the biggest factors that contribute to happiness. Although the model is a good fit (see section

5 for model evaluation), it has one drawback makes the model a bit difficult to interpret if we just look at individual predictors. This is because with different interation terms including a combinationn of individual predictors many times, the explainatory power has to even out among the similar predictors to avoid inflating the prediction level. Therefore, we built a few smaller models given specified region and without interaction terms for further exploration of the factors. Again, since the aim of the following models is to provide a general idea of imporant factors to different regions, we roughly combined the regions according to continents, and only selected 4 major continents to evaluate. Note that the region group for `Africa` includes the region of `Middle East and North Africa`, so when we refer to Africa later, it includes countries in the Middle East as well.

4.1.0.2 Smaller regional models

Table 13: Coefficients estimates of Countries in Europe

	Estimates	2.5%	97.5%	p-value
(Intercept)	-510.609	-988.940	-32.277	0
Year	0.287	0.049	0.526	0
Support	15.181	4.384	25.978	0
Freedom	5.233	0.907	9.559	0
Generosity	-7.967	-12.816	-3.118	0
Corruption	-12.272	-16.493	-8.050	0
Positive	10.131	2.830	17.432	0
Negative	-4.079	-6.698	-1.461	0
GpC	4.040	2.734	5.347	0

Table 14: Coefficients estimates of Countries in Asia

	Estimates	2.5%	97.5%	p-value
(Intercept)	38.515	-1.737	78.768	0
Health	0.582	0.233	0.932	0
Generosity	-18.992	-26.506	-11.478	0
GpC	3.054	1.470	4.638	0

Table 15: Coefficients estimates of Countries in Africa

	Estimates	2.5%	97.5%	p-value
(Intercept)	79.612	65.858	93.367	0
Generosity	-22.823	-29.161	-16.485	0
Positive	15.340	9.417	21.263	0
Gini.Index	13.664	9.367	17.960	0
GpC	5.437	4.853	6.020	0

Table 16: Coefficients estimates of Countries in America

	Estimates	2.5%	97.5%	p-value
(Intercept)	63.998	46.742	81.255	0
Freedom	12.928	7.051	18.805	0
Positive	19.474	9.943	29.004	0
Negative	-8.066	-12.707	-3.424	0
GpC	4.798	3.841	5.755	0

In the models by region, we are using the BIC model selection method on the transformed predictors used for our final model in the previous part. However, we are not considering interaction terms for straight-forward interpretation since we want to provide you with an overall idea of what factors are more important for countries in different region groups.

- The most influncial factors on the squared happiness index for a country in Europe are `Corruption` and `GpC` while `Year` and `Freedom` are the least influctial ones. We see that with 10% increase in an European country's GDP per Capita, its squared happiness index will increase by 0.385 and we are 95% confident that the average increase will be between 0.26 and 0.51.
- The best BIC model for countries in Asia only has three signficiant factors left not counting the intercept: `Generosity`, `GpC`, and `Health.` For this particular model, intercept doesn't have a large influence at all on the squared happiness index comparing to other predictors. Note that `Generosity` here indicates how lack of donation this country is accordinng to its GDP per Capita level and `Health` measures the life expectancies in years.
- The best BIC model for countries in Africa has the predictors' importance level relatively evenly distributed amongthe intercept term and its 4 factors: `Generosity`, `GpC`, and `Gini.Index`, and `Positive.` The `Gini.Index` is the Gini Index of household income while `Positive` is the average of three positive affect measures: happiness, laugh and enjoyment.
- The model for countries in America also has the predictors' importance level relatively evenly distributed amongthe intercept term and its 4 factors: `Freedom`, `Positive`, and `Negative`, and `GpC`. Note that `Freedom` measures how satisfied people are with their freedom to choose what to do with their lives, and `Negative` is the average of three negative affect measures: worry, sadness and anger.

We see that the biggest factors on squared happiness index are different for countries in different regions yet all the models include GDP per Capita, which makes sense as the economy is critical to the country's well-bring as mentioned before. We will not dive into the detailed interpretations of the predictors since we are more interested in the overall difference and commonalities. Here are some interesting findings and conclusions:

- `Corruption` and `Year` are two unique factor for the European countries, and we interpreted that as those countries are more sensitive to the corruption in the institutions and their squared happiness index increases as time proceeds.
- Countries in Africa tend to correlate positivity (happiness, laugh and enjoyment) with more happiness while countries in America tend to correlate negativity (worry, sadness and anger) with less happiness.
- We also see here that countries in Asia and Africa tend have their squared happiness index highly correlated with with their generosity index.

In conclusion, the findings are very intriguing and it is true that countries in different regions have different factors that affect the people's happiness. This result could be improved by looking into the factors by even finer geographical partition.

5 Evaluation

In previous sections, brief evaluations are conducted along with the modeling procedure. In this section, we provide an final evaluation on our final model.

Table 17: Evaluation of the final model

	Num of Params	R2	RMSE	Max Dev	Bias	coverage
Final Model	47	0.892	0.515	2.151	-0.023	0.887

Table 17 summarizes the statistics for our final model.

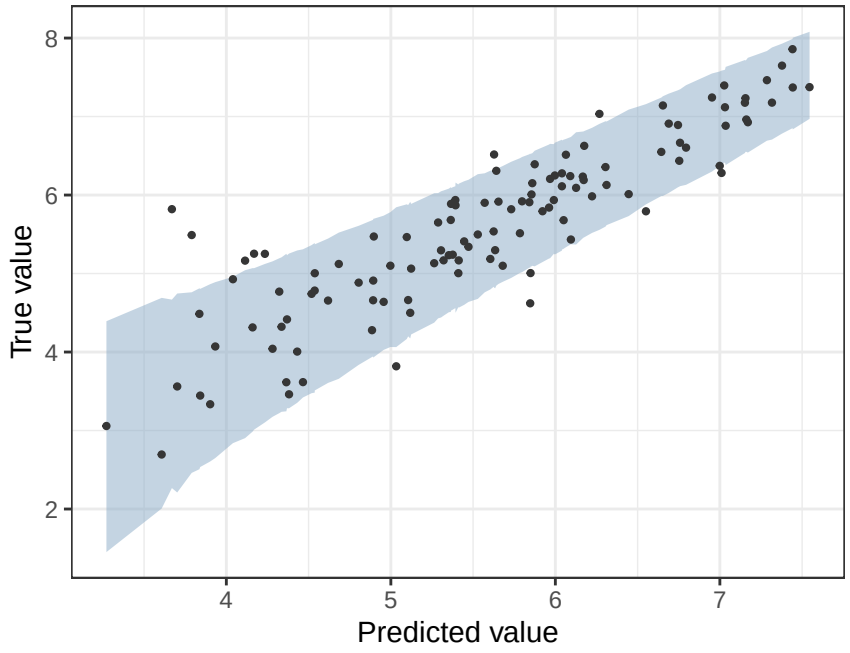

Figure 11: Model evaluation: coverage of the final model

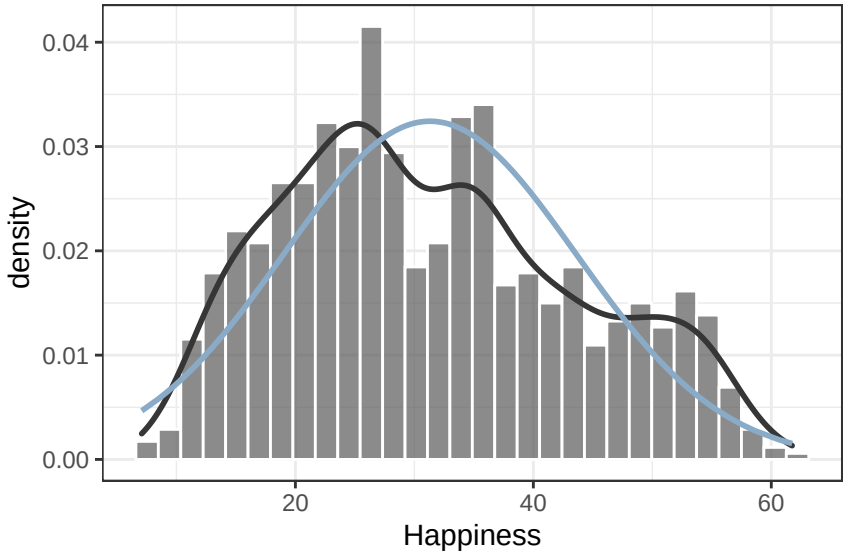

Figure 12: Model evaluation: response distribution with transformation

$R^2 = 0.892$ shows that our model has fairly good explanotory ability on the response `Happiness`. The bias is also low, indicating that on average, the prediction error is low.

However, the maximum deviation and coverage suggest some limitations of our model. In the worst prediction, we make an error greater than 2; also, our prediction intervals only capture 88.7% of the true values, lower than the desired 95% level. From figure 11 we can see that we make both overestimation and underestimation, while the underestimation cases are slightly more frequent than the overestimation, resulting in a negative bias.

Why this is the case? Remember a coverage ratio lower than the nominal confidence level implies that some model assumptions may be not suitable on our dataset. Since here we have pretty ideal residual plots (as shown in 10), then the only problem can be with the Q-Q plot, i.e., with the normality assumption. Indeed, this is true. In figure 12 we plot the empirical density and the theoretically normal density of the response after transformation. Note that there is a slight deviation from the empirical density to the normal, which is a potential reason for the coverage being slightly lower than the desired level. But in general, the empirical distribution is close to normal enough, so that our linear model works really well on the dataset and provides good predictions and interpretations.