Regression Analysis for Prediction: World Happiness in 2011

Prepared for

STA4164

University of Central Florida

By

Elaine Ng, Brandi Schilling, Charles Volpe & Timothy Simmons

November 18, 2019

Table of Contents

INTRODUCTION

The dataset of interest that will be analyzed is from the World Happiness Report published in 2011. The World Happiness Report ranks 146 countries by happiness levels as calculated by the national average responses to the Gallup World Poll and the Cantril ladder survey. The goal of this project is to determine which predictors have the greatest significance in determining a country's happiness.

An analysis of this dataset provides insight into a country's economic and social development which allows countries to better understand the variables that are most significant to their population's happiness. Countries can use this as a tool to improve their own economy and society by reviewing current aspects of life that have been taken into consideration in these evaluations, and by applying their understanding of the variables' effects on happiness levels. This project allows countries to see the successes and failures that their own population, as well as other countries, have experienced as a result of a change in variables; moreover, countries can emulate these methods used by the happiest and most successful countries, or they can use that information to help predict the state of their economy and society if they were to alter their own variables.

DESCRIPTION OF DATA

Each of the 146 countries within this dataset include the estimates of 6 variables that contribute to national happiness levels, and the extent to which they influence these quality of life evaluations. To measure the level of well-being of countries, the Gallup World Poll implements a questionnaire which evaluates the following 14 influential factors: business and economics, citizen engagement, communications and technology, education and families, environment and energy, food and shelter, government and policies, health, law and order, religion and ethics, social issues, well-being, work, and others. Alongside the Gallup World Poll, the Cantril ladder survey asks respondents to imagine a ladder with steps numbered 0-10, in which each step represents a quality of life with 0 being the worst possible life and 10 being the best possible life. Respondents are then asked to place themselves on a step in that ladder based on where they feel their current life falls in relation; this will be measured as the life ladder variable.

To provide a more accurate representation of a country's happiness level and to highlight any correlations and significance, the dataset utilizes 5 additional variables alongside the life ladder variable which are: GDP, social support, life expectancy, freedom, and corruption. The measurement of GDP is as determined by the World Development Indicators which have utilized purchasing power parity to provide a fair comparison between different currencies; this variable provides insight of a country's economy and growth rate. Social support measures the country's 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 the possible responses 0 if not true, and 1 if true. Life expectancy is calculated from previous data regarding country-specific healthy life expectancy and total life expectancy as reported by the World Health Organization and the World Development Indicators. Freedom to make life choices (denoted as freedom) measures the country's average response to the question, "Are you satisfied or dissatisfied with your freedom to choose what you do with your life?" with the possible responses 0 if not satisfied, and 1 if satisfied. Corruption perception (denoted as corruption) measures the country's average response to the following two questions: "Is corruption widespread throughout the government or not", and "Is corruption widespread within businesses or not?" with the possible responses 0 if not true and 1 if true.

All countries in this dataset are compared to a hypothetical country called Dystopia. This imaginary nation ranks overall lowest on the happiness score with each of its 5 variables equaling the lowest national averages. Due to Dystopia's low quality of life and happiness, it is able to serve as benchmark for comparing all countries to ensure that the variables of each country are positive (or nonzero).

RESEARCH QUESTIONS

Happiness is an ephemeral concept, and its measurement is equally challenging. However, broad statements which seem to correlate with happiness can be assessed and tied to more concrete measurements. The biggest interesting question this research will answer is 'What makes the world happy?' To understand the happiness levels among countries, we must first understand how each of the variables come into play when evaluating the average happiness of a country; in doing so, we will answer the questions, 'Of the five variables data has been collected, which has the biggest impact?' and 'Do any of the variables seem to not be significant in

assessing happiness?' The questions that our analysis will be answering will help identify ways in which countries can improve their own happiness levels, economies, and societies. In addition to helping countries identify areas of improvement in their own countries, it is beneficial to see how other variables of comparable countries can impact happiness levels; thus, we will be answering the question, 'Is there a reliable model that can be used to predict happiness?'

CHOICE OF METHODS TO ANALYZE DATA

The first step we have chosen to analyze the dataset is data editing. In this step, we ensured that the quality of the data is appropriate and identified any errors or inconsistencies by checking whether or not there are any outliers or influential points in the set that could affect the regression analysis. While data editing, we were able to locate missing values and correct them by using a method called mean imputation which allowed us to replace these missing values with the calculated mean of that column. We chose to implement the method of mean imputation after a test of normality was performed and was able to prove that our data is normally distributed and is therefore symmetric around the mean. To ensure that the model we will be using to predict happiness is reliable, we randomly split the dataset into two subsets, the Training group which contains 60% of the observations, and the holdout group which contains the other 40% of the observations.

Following the data editing step, we performed a descriptive analysis which provides the mean, median, standard deviation, minimum, and maximum for each of the 6 variables. Prior to searching for the best regression model, we examined the regression assumptions. The assumptions of linearity were checked by visual means such as the QQ-plot, normal distribution with the test of normality, constant variance with the test for homogeneity of variance, and collinearity by correlation matrices and the VIF (variance inflation factor).

After confirming that all of the regression assumptions have been met, we continued with finding the best regression model to predict happiness. Firstly, we selected the maximum model that contains all of the basic predictors, and we checked for any possible transformation and interaction terms. Then we used stepwise, backwards selection, and forward elimination to determine which variables should be included in the model. By comparing the coefficient of determination, F-statistics, P-values and Mallow's CP, we will be able to choose the best model. We will then use our best model to make a prediction for the training and holdout group, and

then compare their squared correlation coefficient in order to assess the reliability of our model. For the assessment of the reliability of our model, we follow the rule that: the smaller the difference, the higher the reliability.

Lastly, we applied our best and most reliable model to make a prediction for our whole dataset, and we calculated the difference between the actual and predicted values for happiness.

SUMMARY RESULTS

A) Data Editing

- Outliers, Leverage and Influential Points

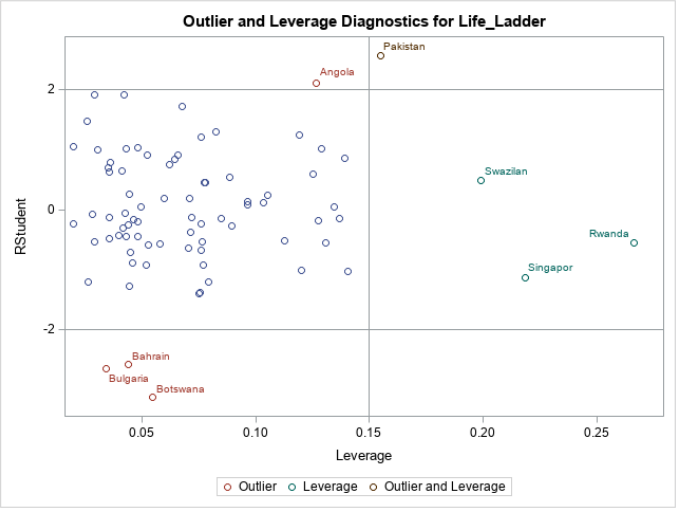

Figure 1: Outliers and Leverage Diagnostics

Figure 1 shows 4 observations as outliers; Angola, Bahrain, Botswana and Bulgaria, 3 leverages; Rwanda, Singapore and Swaziland, and one that is both an outlier and leverage point, Pakistan. But upon closer look at Table 1, their studentized residuals are all below 3, therefore they will not be counted as outliers. The 3 leverage points will be removed from the training set to eliminate their impacts on the fitted regression model.

Country	Angola	Bahrain	Bulgaria	Pakistan	Botswana
Studentized Residual	2.062	-2.487	-2.549	2.470	-2.951

Table 1: Country and their Studentized Residuals.

B) Descriptive Analysis

Variable	N	Mean	Median	Std Dev	Minimum	Maximum
Life_Ladder	76	5.5369399	5.3732431	1.1298660	3.5199211	7.7882319
GDP	76	9.3806041	9.4832211	1.1037075	6.6935630	11.8064709
Social_Support	76	0.8301147	0.8575269	0.0964288	0.5211036	0.9773776
Life_Expectancy	76	62.5052540	64.5800934	7.6485442	43.5934143	73.7421570
Freedom	76	0.7397751	0.7740757	0.1502786	0.3474140	0.9520344
Corruption	76	0.7700802	0.8089413	0.1658758	0.2200431	0.9769174

Figure 2: Descriptive Analysis.

Figure 2 shows the number of observations in our training set, the mean, median, standard deviation, minimum and maximum for each variable, including our dependent variable, Life Ladder.

C) Regression Assumptions

- Test for Normality

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.974969	Pr < W	0.1608
Kolmogorov-Smirnov	D	0.082263	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.082473	Pr > W-Sq	0.1968
Anderson-Darling	A-Sq	0.561283	Pr > A-Sq	0.1454

Figure 3: Test for Normality.

Figure 3 shows the test for Normality. Since our samples, even when broken into training/holdout data have N > 50 we use the Kolmogorov-Smirnov test. With a p-value >.15 our test fails to reject the null hypothesis and so we can conclude our data is normally distributed.

- Linearity

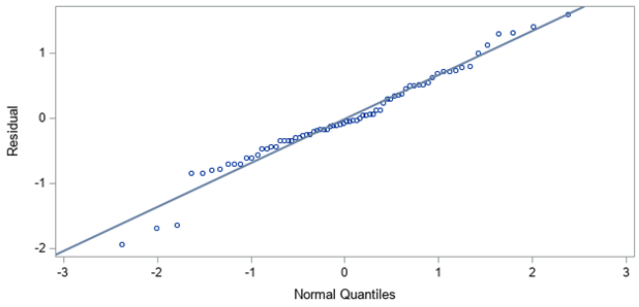

Figure 4: QQ-Plot

Figure 4 is the QQ-Plot for our residuals. It follows a linear line that is very close to a 45-degree line. From this we can conclude the data has a mostly linear relationship between dependent and independent variables.

- Heteroscedasticity

Test for Homogeneity of Variance

The REG Procedure
Model: MODEL1
Dependent Variable: Life_Ladder Life_Ladder

Test of First and Second Moment Specification		
DF	Chi-Square	Pr > ChiSq
20	16.48	0.6863

Figure 5: Test for Homogeneity of Variance.

We used the PROC REG option / spec to test the heteroscedasticity assumption. The test has a null hypothesis that the data is heteroscedastic, and with a p-value for the test .6863 we fail to reject the null hypothesis. Thus, our data satisfies the assumption of heteroscedasticity.

- Collinearity

To test for collinearity, we ran two tests. One uses the PROC REG / vif option, as seen on Figure 6 which checks for variance inflation due to collinearity. As a rule of thumb, a VIF > 10 presents a possible collinearity issue. None of our variables have a VIF value larger than 10, so we do not suspect collinearity will be a problem.

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	Intercept	1	-1.73619	0.87369	-1.99	0.0506	0
GDP	GDP	1	0.31023	0.11354	2.73	0.0079	2.71926
Social_support	Social support	1	2.72438	1.02434	2.66	0.0096	1.91602
Life_Expectancy	Life Expectancy	1	0.02870	0.01532	1.87	0.0650	2.51856
Freedom	Freedom	1	1.16944	0.68142	1.72	0.0903	1.80977
Corruption	Corruption	1	-0.70947	0.46018	-1.54	0.1274	1.28879

Figure 6: VIF

As a second test, we look at the correlation matrix for the data. GDP and Life Expectancy are the closest to possibly being an issue, but the rule of thumb we applied is $|r_{ij}| > 0.8$ which none crosses.

Given both tests, we conclude that we do not have any collinearity issues.

Having checked the assumptions for regression, we have seen nothing to suggest that the data is invalid for regression analysis.

Pearson Correlation Coefficients, N = 80 Prob > |r| under H0: Rho=0						
	Life_Ladder	GDP	Social_Support	Life_Expectancy	Freedom	Corruption
Life_Ladder	1.00000	0.69497 <.0001	0.65306 <.0001	0.65862 <.0001	0.55316 <.0001	-0.31179 0.0049
GDP	0.69497 <.0001	1.00000	0.57790 <.0001	0.76553 <.0001	0.39406 0.0003	-0.12417 0.2725
Social_Support	0.65306 <.0001	0.57790 <.0001	1.00000	0.52771 <.0001	0.55664 <.0001	-0.16655 0.1398
Life_Expectancy	0.65862 <.0001	0.76553 <.0001	0.52771 <.0001	1.00000	0.35882 0.0011	-0.18174 0.1067
Freedom	0.55316 <.0001	0.39406 0.0003	0.55664 <.0001	0.35882 0.0011	1.00000	-0.45017 <.0001
Corruption	-0.31179 0.0049	-0.12417 0.2725	-0.16655 0.1398	-0.18174 0.1067	-0.45017 <.0001	1.00000

Figure 7: Correlation Matrix

D) Selecting Best Regression Model

- Maximum Model

Basic Predictors: GDP, Social Support, Life Expectancy, Freedom and Corruption.

Possible Transformation: We continue by checking to see if we need to apply any transformations to our data. Our conclusion, by the end, is that no transformations need be applied. See below for the scatter plots we used to determine this. The oddest result, for corruption, still seems to show no transformation is required.

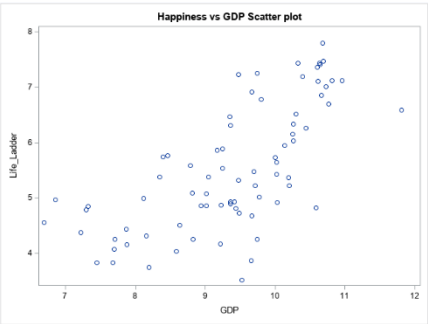

Figure 8: Life Ladder vs GDP

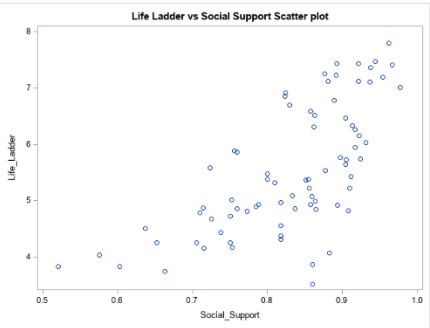

Figure 9: Life Ladder vs Social Support

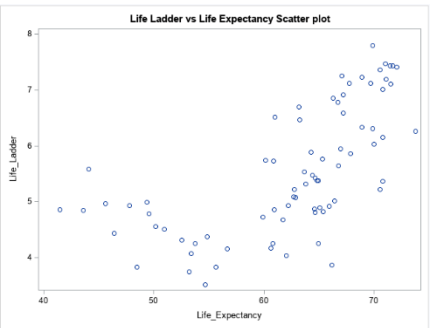

Figure 10: Life Ladder vs Life Expectancy

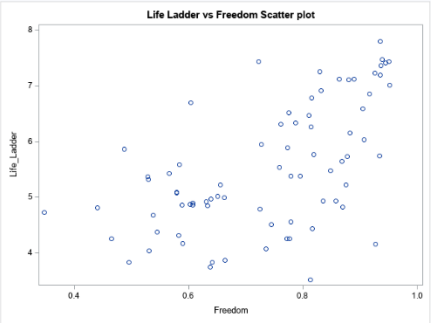

Figure 11: Life Ladder vs Freedom

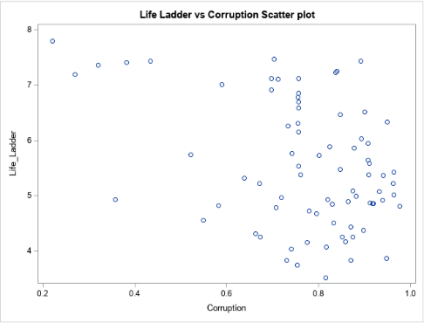

Figure 12: Life Ladder vs Corruption

Interaction Terms: GDP*Social Support, Social Support*Life Expectancy and Social Support*Freedom.

Using PROC GLM, we were able to generate scatter plots that show the effect of the change of a variable on another, which is the definition of interaction between factors. Below are the scatter plots (Figure 13 – Figure 16) for the variables we determined show interaction. In testing for interaction, we looked at whether the slope of the trend-lines on the scatter plots differ for different levels of a qualitative variable. Since all our variables are continuous quantitative variables it took a bit of finessing the data. The first step was breaking up the range of each of our variables into 4 sections. A dummy variable is added to the dataset representing the "level". For instance, the variable Life Expectancy has a range of about [41, 74]. The difference is 33, meaning we have 33/4 size categories. The lowest category (interaction dummy = 0) covers [41, 49.5), the next (interaction dummy = 1) covers [49.5, 58), and so on until 74.

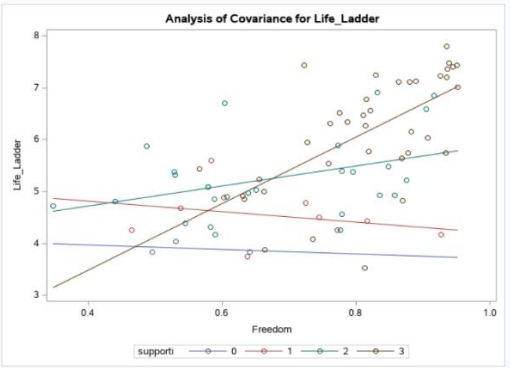

Figure 13: Analysis of Covariance Freedom # 1

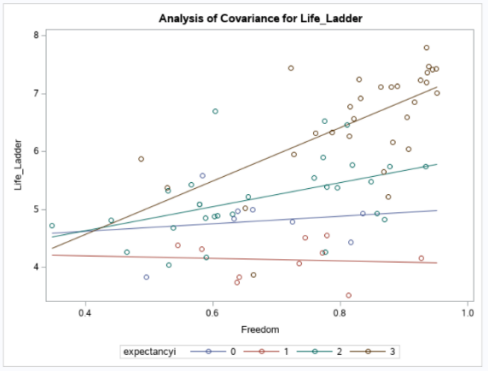

Figure 14: Analysis of Covariance Freedom # 2

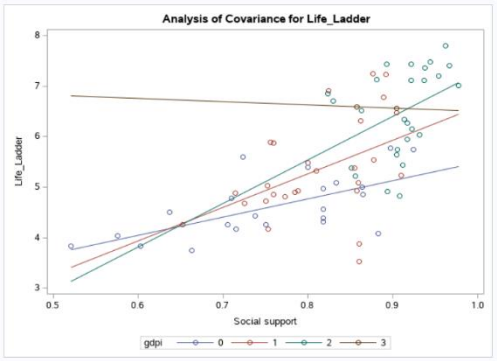

Figure 15: Analysis of Covariance Social Support

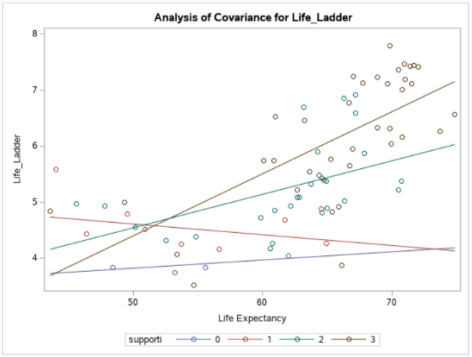

Figure 16: Analysis of Covariance Life Expectancy

- Searching Strategy

With transformations tested for as well as interactions accounted for, our next step is to find the best model. Due to there being some problems with stepwise variable selection such as, biased r-squared values and improper p-values we would not be using stepwise. A liberal alpha is difficult to define therefore, we will use backwards elimination where a smaller alpha will still yield good results. Using backwards elimination with $\alpha = 0.1$, the variables removed were the interaction term between social support and freedom, social support and GDP, social support, and finally, corruption. This leaves us with the model of $Y = \beta 0 + \beta 1 GDP + \beta 2 Life\_Expectancy + \beta 3 Freedom + \beta 4(Social\_Support * Life\_Expectancy) + \beta 5(Freedom * Life\_Expectancy)$.

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	11.16439	3.76216	3.61542	8.81	0.0042
GDP	0.23664	0.11283	1.80599	4.40	0.0397
Life_Expectancy	-0.18915	0.06034	4.03378	9.83	0.0026
Freedom	-14.02072	5.14753	3.04583	7.42	0.0082
social_support_life_expectancy	0.05404	0.01858	3.47271	8.46	0.0049
freedom_life_expectancy	0.24693	0.08080	3.83401	9.34	0.0032

Figure 17: Backwards Elimination Variables

This model has an R-square = 0.7043, F-statistics = 31.91, P-value = < 0.0001 and CP = 4.35.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	social_support_freedom	8	0.0003	0.7147	8.0583	0.06	0.8100
2	Social_Support	7	0.0007	0.7140	6.2038	0.15	0.7020
3	gdp_social_support	6	0.0030	0.7110	4.8699	0.69	0.4109
4	Corruption	5	0.0067	0.7043	4.3503	1.53	0.2205

Figure 18: Backwards Elimination Variables Removed

With Life Expectancy having the highest F-statistics and lowest P-value, this could be the most significant variable in predicting happiness.

E) Evaluating Reliability

By finding the squared correlation coefficient in both the training and holdout set, we can compare them to assess reliability. The smaller the difference the higher the reliability.

The squared correlation coefficient for the training set is 0.6991 and for the holdout set is 0.7803. The difference between those two is 0.0812. The difference is small therefore, we have a reliable model.

Root MSE	0.63745	R-Square	0.6991
Dependent Mean	5.51047	Adj R-Sq	0.6782
Coeff Var	11.56806		

Figure 19: R-Square of Training Group

Root MSE	0.51077	R-Square	0.7803
Dependent Mean	5.30871	Adj R-Sq	0.7613
Coeff Var	9.62142		

Figure 20: R-Square of Holdout Group

F) Prediction

After assessing reliability, we can apply our model to the whole data set and compare the actual values with our predicted value.

Obs	WP5	Life_Ladder	MODEL1	diff
1	Afghanis	3.83172	4.10230	-0.27058
2	Albania	5.86742	4.84803	1.01939
3	Algeria	5.31719	5.17818	0.13901
4	Angola	5.58900	4.78735	0.80165
5	Argentin	6.77581	6.08154	0.69426
6	Armenia	4.26049	4.59567	-0.33517
7	Australi	7.40562	7.17141	0.23421
8	Austria	7.47051	7.02708	0.44344
9	Azerbaij	4.68047	5.01753	-0.33706
10	Bahrain	4.82398	6.37918	-1.55521
11	Banglade	4.98565	4.56499	0.42066
12	Belarus	5.22531	5.69506	-0.46976
13	Belgium	7.11136	6.85333	0.25804
14	Benin	3.87028	3.82345	0.04683
15	Bolivia	5.77887	5.16243	0.61644
16	Bosnia a	4.99467	4.42089	0.57378
17	Botswana	3.51992	5.37097	-1.85105
18	Brazil	7.03782	5.98592	1.05190
19	Bulgaria	3.87538	5.63046	-1.75508
20	Burkina	4.78537	4.29139	0.49397
21	Burundi	3.70589	3.69752	0.00838
22	Cambodia	4.16123	4.73505	-0.57383
23	Cameroon	4.43389	4.31417	0.11972
24	Canada	7.42605	7.00677	0.41928
25	Central	3.67783	3.26697	0.41086
26	Chad	4.39348	4.70481	-0.31132
27	Chile	6.52633	5.76944	0.75689
28	China	5.03721	5.72432	-0.68712
29	Colombia	6.46395	5.83164	0.63231
30	Comoros	3.83849	4.40309	-0.56460
31	Congo Br	4.50982	4.51159	-0.00177
32	Congo (K	4.51696	4.18568	0.33129
33	Costa Ri	7.22889	6.37758	0.85131
34	Croatia	5.38537	5.20663	0.17874
35	Cyprus	6.68961	6.13632	0.55329
36	Czech Re	6.33149	6.30735	0.02414
37	Denmark	7.78823	6.98317	0.80507
38	Djibouti	4.36919	4.34039	0.02880
39	Dominica	5.39654	5.76258	-0.36605
40	Ecuador	5.79509	5.64334	0.15175
41	Egypt	4.17416	5.03585	-0.86169
42	El Salva	4.74129	5.14920	-0.40791
43	Estonia	5.48682	6.03012	-0.54330
44	Finland	7.35423	6.94555	0.40867
45	France	6.95919	6.88418	0.07501
46	Gabon	4.25540	4.91072	-0.65532
47	Georgia	4.20303	4.35547	-0.15244
48	Germany	6.62131	6.91979	-0.29847
49	Ghana	5.60820	4.61853	0.98967
50	Greece	5.37204	5.47393	-0.10189
51	Guatemal	5.74335	5.18195	0.56141
52	Guinea	4.04457	3.90698	0.13759
53	Haiti	4.84457	4.14160	0.70297
54	Honduras	4.96103	5.14706	-0.18603
55	Hong Kon	5.47401	6.88347	-1.40946
56	Hungary	4.91760	5.75291	-0.83531
57	India	4.63487	4.48421	0.15066
58	Indonesi	5.17261	5.45887	-0.28626
59	Iran	4.76751	.	.
60	Iraq	4.72537	4.85530	-0.12993
61	Ireland	7.00690	7.15782	-0.15092
62	Israel	7.43315	6.19632	1.23683
63	Italy	6.05709	5.86088	0.19621
64	Jamaica	5.37445	5.66672	-0.29227
65	Japan	6.26279	6.66468	-0.40189
66	Jordan	5.53933	5.66954	-0.13021
67	Kazakhst	5.73566	5.97227	-0.23660
68	Kenya	4.40531	4.80535	-0.40004
69	Kosovo	4.85950	4.99844	-0.13894
70	Kuwait	6.37770	6.26288	0.11482
71	Kyrgyzst	4.92105	5.27852	-0.35747
72	Laos	4.70375	4.74234	-0.03859

Figure 21: Actual Values, Predicted Values, and their differences # 1

Looking at the difference column, we can see that our model can predict happiness pretty accurately, with most differences being rather small.

73	Latvia	4.96681	5.41467	-0.44786
74	Lebanon	5.18757	5.28422	-0.09665
75	Lesotho	4.89751	4.61741	0.28010
76	Lithuani	5.43244	5.66434	-0.23190
77	Luxembou	7.10140	7.25904	-0.15764
78	Macedoni	4.89818	5.21564	-0.31746
79	Madagasc	4.38142	4.62592	-0.24451
80	Malawi	3.94606	3.99901	-0.05295
81	Malaysia	5.78637	5.73061	0.05576
82	Mali	4.66683	4.39007	0.27676
83	Malta	6.15472	6.67292	-0.51820
84	Mauritan	4.78480	4.70541	0.07939
85	Mauritiu	5.47707	5.75207	-0.27500
86	Mexico	6.90952	5.91342	0.99609
87	Moldova	5.79226	5.17714	0.61512
88	Mongolia	5.03117	5.64375	-0.61257
89	Monteneg	5.22312	5.24984	-0.02673
90	Morocco	5.08497	5.14097	-0.05599
91	Mozambiq	4.97111	4.39959	0.57153
92	Nepal	3.80944	4.51914	-0.70970
93	Netherla	7.56380	6.99774	0.56606
94	New Zeal	7.19064	6.96901	0.22163
95	Nicaragu	5.38571	5.29609	0.08961
96	Niger	4.55583	4.35087	0.20496
97	Oman	6.85298	.	.
98	Pakistan	5.26719	4.10966	1.15753
99	Palestin	4.75122	4.72353	0.02769
100	Panama	7.24808	6.07370	1.17438
101	Paraguay	5.67708	5.39353	0.28356
102	Peru	5.89246	5.38150	0.51096
103	Philippi	4.99396	5.22277	-0.22881
104	Poland	5.64620	6.29437	-0.64816
105	Portugal	5.22000	6.42685	-1.20685
106	Qatar	6.59160	6.73068	-0.13908
107	Romania	5.02276	5.33450	-0.31174
108	Russia	5.38877	5.65916	-0.27039
109	Rwanda	4.09744	4.02131	0.07612
110	Saudi Ar	6.69979	5.68436	1.01543
111	Senegal	3.83420	4.26138	-0.42717
112	Serbia	4.81519	4.90883	-0.09364
113	Sierra L	4.50164	4.12360	0.37804
114	Singapor	6.56104	6.90904	-0.34800
115	Slovakia	5.94505	6.07284	-0.12779
116	Slovenia	6.03596	6.72066	-0.68470
117	Somalila	4.93057	.	.
118	South Af	4.93051	5.00570	-0.07519
119	South Ko	6.94660	5.85937	1.08723
120	Spain	6.51825	6.73583	-0.21758
121	Sri Lank	4.18057	5.68185	-1.50128
122	Sudan	4.31446	4.86052	-0.54606
123	Swazilan	4.86709	4.97729	-0.11020
124	Sweden	7.38223	7.01587	0.36636
125	Syria	4.03789	4.32691	-0.28902
126	Taiwan	6.30892	.	.
127	Tajikist	4.26267	4.84651	-0.58384
128	Tanzania	4.07356	4.88868	-0.81512
129	Thailand	6.66361	6.11724	0.54637
130	Togo	2.93622	3.47247	-0.53625
131	Trinidad	6.51875	5.82669	0.69206
132	Tunisia	4.87648	4.97714	-0.10066
133	Turkey	5.27194	4.78724	0.48471
134	Turkmeni	5.79175	.	.
135	Uganda	4.82600	4.64721	0.17879
136	Ukraine	5.08313	5.26552	-0.18239
137	United A	7.11870	6.57904	0.53967
138	United K	6.86925	6.88087	-0.01162
139	United S	7.11514	6.70701	0.40813
140	Uruguay	6.55405	6.20665	0.34740
141	Uzbekist	5.73874	5.61123	0.12751
142	Venezuel	6.57979	5.97306	0.60673
143	Vietnam	5.76734	5.68908	0.07827
144	Yemen	3.74626	4.51790	-0.77165
145	Zambia	4.99911	4.87119	0.12792
146	Zimbabwe	4.84564	4.58887	0.25677

Figure 22: Actual Values, Predicted Values, and their differences # 2

CONCLUSION

After editing our data by removing any outliers, leverage, and influential points and then filling in all missing values, we were able to run some tests for regression assumptions. We conducted tests for Normality, Linearity, Heteroscedasticity, as well as Collinearity. The tests confirm that a regression analysis is appropriate for our data set. We found the maximum model by including all basic predictors, possible transformations and interaction terms. Using the backwards elimination strategy, the best model was chosen and applied to both training and holdout group to test for reliability. Since, our model appears to be reliable for predicting happiness, it was implemented for the entire data set and compared to the actual values. Our analysis suggests that GDP, Life Expectancy, Freedom, and the interaction between Social Support and Life Expectancy are important predictors for estimating happiness. This finding can help countries see the importance that these variables have on their population's happiness.

Resources

Sustainable Development Solutions Network. (2017, June 14). World Happiness Report. Retrieved from https://www.kaggle.com/unsdsn/world-happiness.

Code

```
data Happiness;
infile "\\Client\C$\NewDataCSV.csv" firstobs = 2 missover dlm = "," dsd;
input WP5$ Country$ Year Life_Ladder GDP Social_Support Life_Expectancy
Freedom Corruption;
run;

data Happiness2;
set Happiness (drop = Country);
where Year = 2011;
gdp_social_support = GDP * Social_Support;
social_support_life_expectancy = Social_Support * Life_Expectancy;
social_support_freedom = Social_Support * Freedom;
freedom_life_expectancy = Freedom * Life_Expectancy;
run;

proc print data = Happiness2;
title 'All observations for 2011';
run;

%let prop_Model = 0.6;
%let prop_Holdout = 0.4;
data Model Holdout;
array p[2] _temporary_ (&prop_Model, &prop_Holdout);
set Happiness2;
call streaminit(123);
_k = rand("Table", of p[*]);
if _k = 1 then output Model;
else output Holdout;
drop _k;
run;

proc print data = Model;
var WP5 Life_Ladder GDP Social_Support Life_Expectancy Freedom Corruption;
title 'Training Group';
run;

proc print data = Holdout;
var WP5 Life_Ladder GDP Social_Support Life_Expectancy Freedom Corruption;
title 'Holdout Group';
run;

ods graphics on;
proc reg data = Model;
model Life_Ladder = GDP Social_Support Life_Expectancy Freedom Corruption;
output out = normalityTest r = resid;
title 'Test of Normality';
run;
proc univariate data = normalityTest plot normal;
var resid;
run;
ods graphics off;

proc stdize data = Model reponly method = mean out = CompleteModel;
var GDP Social_Support Life_Expectancy Freedom Corruption;
```

```
run;

proc print data = CompleteModel;
var WP5 Life_Ladder GDP Social_Support Life_Expectancy Freedom Corruption;
title 'Complete Model';
run;

proc stdize data = Holdout reponly method = mean out = CompleteHoldout;
var GDP Social_Support Life_Expectancy Freedom Corruption;
run;

proc print data = CompleteHoldout;
var WP5 Life_Ladder GDP Social_Support Life_Expectancy Freedom Corruption;
title 'Complete Holdout';
run;

proc reg data = CompleteModel;
model Life_Ladder = GDP Social_Support Life_Expectancy Freedom Corruption / r
influence;
title 'Outliers and Influential Points';
run;

ods graphics on;
proc reg data = CompleteModel
plots(label)=(CooksD RStudentByLeverage RStudentByPredicted DFFITS DFBETAS);
id WP5;
model Life_Ladder = GDP Social_Support Life_Expectancy Freedom Corruption;
title 'Outliers and Influential Points Plots';
run;
ods graphics off;

data noLeverage;
set CompleteModel;
if WP5 = 'Swazilan' THEN DELETE;
if WP5 = 'Rwanda' THEN DELETE;
if WP5 = 'Singapor' THEN DELETE;
if WP5 = 'Pakistan' THEN DELETE;
run;

proc means data = noLeverage n mean median std min max;
var Life_Ladder GDP Social_Support Life_Expectancy Freedom Corruption;
title 'Descriptive Statistics';
run;

proc corr data = CompleteModel;
var Life_Ladder GDP Social_Support Life_Expectancy Freedom Corruption;
title 'Correlation Matrix';
run;

proc reg data = CompleteModel ;
model Life_Ladder = GDP Social_Support Life_Expectancy Freedom Corruption /
vif;
ods graphics off;
run;

proc reg data = CompleteModel;
```

```
model Life_Ladder = GDP Social_Support Life_Expectancy Freedom Corruption /
spec;
title 'Test for Homogeneity of Variance';
run;

proc sgplot data = CompleteModel;
scatter y = Life_Ladder x = GDP;
title 'Happiness vs GDP Scatter plot';
run;

proc sgplot data = CompleteModel;
scatter y = Life_Ladder x = Social_Support;
title 'Life Ladder vs Social Support Scatter plot';
run;

proc sgplot data = CompleteModel;
scatter y = Life_Ladder x = Life_Expectancy;
title 'Life Ladder vs Life Expectancy Scatter plot';
run;

proc sgplot data = CompleteModel;
scatter y = Life_Ladder x = Freedom;
title 'Life Ladder vs Freedom Scatter plot';
run;

proc sgplot data = CompleteModel;
scatter y = Life_Ladder x = Corruption;
title 'Life Ladder vs Corruption Scatter plot';
run;

data gdpInteraction;
set NoLeverage;
if gdp > 11 then gdpi = 3;
else if gdp > 10 then gdpi = 2;
else if gdp > 9 then gdpi = 1;
else gdpi = 0;
run;

data supportInteraction;
set NoLeverage;
if social_support > .86 then supporti = 3;
else if social_support > .74 then supporti = 2;
else if social_support > .62 then supporti = 1;
else supporti = 0;
run;

data expectancyInteraction;
set NoLeverage;
if life_expectancy > 66 then expectancyi = 3;
else if life_expectancy > 58 then expectancyi = 2;
else if life_expectancy > 50 then expectancyi = 1;
else expectancyi = 0;
run;

data freedomInteraction;
set NoLeverage;
if freedom > .80 then freedomi = 3;
```

```
else if freedom > .65 then freedomi = 2;
else if freedom > .5 then freedomi = 1;
else freedomi = 0;
run;

data corruptionInteraction;
set NoLeverage;
if corruption > .75 then corruptioni = 3;
else if corruption > .54 then corruptioni = 2;
else if corruption > .33 then corruptioni = 1;
else corruptioni = 0;
run;

ods graphics on;
proc glm data=gdpInteraction;
class gdpi;
model Life_Ladder = Social_Support | gdpi / solution;
run;

proc glm data=gdpInteraction;
class gdpi;
model Life_Ladder = Life_Expectancy | gdpi / solution;
run;

proc glm data=gdpInteraction;
class gdpi;
model Life_Ladder = Freedom | gdpi / solution;
run;

proc glm data=gdpInteraction;
class gdpi;
model Life_Ladder = corruption | gdpi / solution;
run;

proc glm data=supportInteraction;
class supporti;
model Life_Ladder = Life_Expectancy | supporti / solution;
run;

proc glm data=supportInteraction;
class supporti;
model Life_Ladder = Freedom | supporti / solution;
run;

proc glm data=supportInteraction;
class supporti;
model Life_Ladder = corruption | supporti / solution;
run;

proc glm data=expectancyInteraction;
class expectancyi;
model Life_Ladder = Freedom | expectancyi / solution;
run;

proc glm data=expectancyInteraction;
class expectancyi;
model Life_Ladder = corruption | expectancyi / solution;
```

```
run;

proc glm data=freedomInteraction;
class freedomi;
model Life_Ladder = corruption | freedomi / solution;
run;
ods graphics off;

proc print data = noLeverage;
var WP5 Life_Ladder GDP Social_Support Life_Expectancy Freedom Corruption
gdp_social_support social_support_life_expectancy social_support_freedom
freedom_life_expectancy;
title 'Data without Leverage';
run;

proc reg data = noLeverage;
model Life_Ladder = GDP Social_Support Life_Expectancy Freedom Corruption
gdp_social_support social_support_life_expectancy social_support_freedom
freedom_life_expectancy/ selection = backward SLS = 0.1;
title 'Backwards Elimination without Leverage';
run;

proc reg data = CompleteModel;
model Life_Ladder = GDP Life_Expectancy Freedom
social_support_life_expectancy freedom_life_expectancy;
title 'Squared Correlation Coefficient Complete Model';
run;

proc reg data = CompleteHoldout;
model Life_Ladder = GDP Life_Expectancy Freedom
social_support_life_expectancy freedom_life_expectancy;
title 'Squared Correlation Coefficient Complete Holdout';
run;

proc reg data = Happiness2 outest = prediction;
model Life_Ladder = GDP Life_Expectancy Freedom
social_support_life_expectancy freedom_life_expectancy;
title 'Squared Correlation Coefficient All Observations';
run;

proc print data = prediction;
title 'Model 1';
run;

proc score data = Happiness2 score = prediction out = RScoreP type = parms;
var GDP Life_Expectancy Freedom social_support_life_expectancy
freedom_life_expectancy;
run;

proc print data = RScoreP;
title 'Predictions with Model 1';
run;

data compare;
set RScoreP;
diff = Life_Ladder - MODEL1;
run;
```

```
proc print data = compare;
var WP5 Life_Ladder MODEL1 diff;
title 'Difference between Actual and Predicted values';
run;
```