World Happiness Report 2019 - Multiple linear regression analysis

Fanni Kiss

2021-01-03

Introduction

Current report attempts to show how six key variables contribute to explaining the full sample of national annual average happiness scores in the World Happiness Report 2019. These variables are GDP per capita, social support, healthy life expectancy, freedom to make life choices, generosity and perceptions of corruption. The values are scores by country, which are based on individuals' own assessments of their lives. The six variables are used to explain the variation of happiness across countries.

GitHub repository

Data

The World Happiness Report is a publication of the Sustainable Development Solutions Network, powered by data from the Gallup World Poll. The cleaned data set is downloaded from Kaggle. Descriptive statistics and histograms of explanatory variables are shown in the appendix.

Model

The correlations between the Happiness Score and GDP per capita, Social support and Heathy life expectancy were strong and positive, so first, I use a simple linear regression model to predict the Happiness Score based on one variable. Then we can observe, how accurate the predicted Happiness Score compared to the true values. After that, I use multiple linear regression model to predict the happiness score more precisely using the three variables simultaneously and then using all variables simultaneously.

Simple Linear Regression

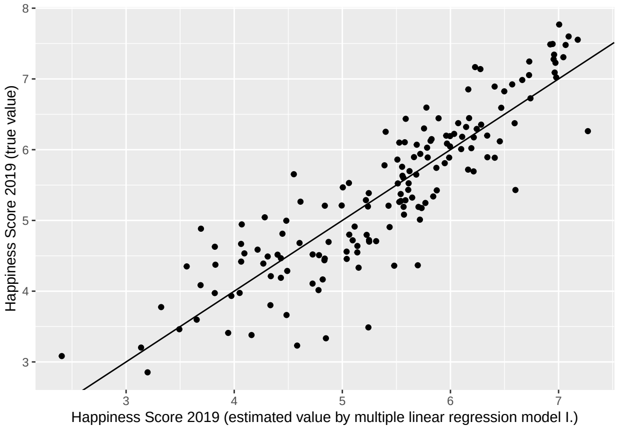
Firstly, I regress Happiness Score on GDP per capita. The multiple R-squared is 0.63, which means that GDP per capita explains 63% of the Happiness Score value. Also, the p-value of the coefficient of the GPD_per_capita variable is 0, so we can consider the predictor variable statistically significant. The regression table is shown in the appendix. The simple linear regression model is the next: Happiness Score (estimated) = 3.4 + 2.22 * GDP_per_capita The GDP per capita alone explains 63% of the Happiness Score. In the next section, I complete the regression model with further variables to predict more precisely the Happiness Score.

Multiple Linear Regression Model I.

The World Happiness Report data set allows us to check, how do the six different variables explain the Happiness Score. I apply a multiple linear regression using the GDP per capita, social support, life expectancy, freedom to make life choices, generosity and corruption to see, how do these variables explain the Happiness Score. The result of the regression is a model, which tells us, how does the Happiness Score changes on average with one unit larger value of one variable but the same vaue of the rest of the variables. The multiple linear regression model for the World Happiness Report 2019 is the next: Happinnes Score (estimated) = 1.79 + 0.77 * GDP__per__capita + 1.12 * social__support + 1.08 * life__expectancy + 1.45 * freedom__of__choice + 0.49 * generosity + 0.97 * corruption The multiple R-squared values of the model is 0.78, which means that the six explanatory variables explain 78% of the Happinness Score. There are two variables, generosity and corruption, which confidence interval ranges involve zero. The interval that contains the true value of the regression coefficients can be both negative and positive by 95% probability. This means, that generosity and corruption explain weakly the Happiness Score in my analysis. Also, the p-values of generosity and corruption are high (0.41 and 0.16 consecutively), which means that generosity and corruption are not statistically significant predictor variables.

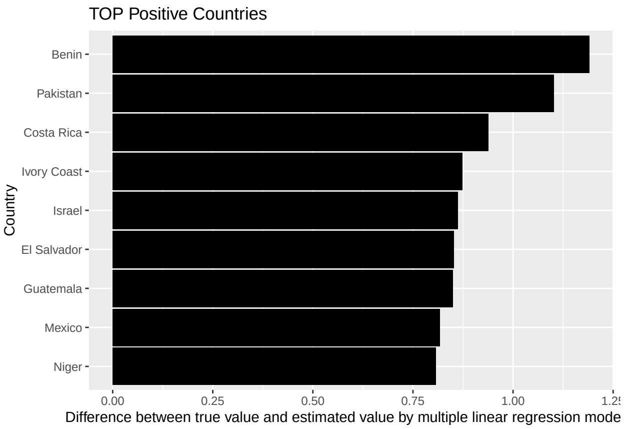
Comparison of estimated and true value of Happiness Score (I.)

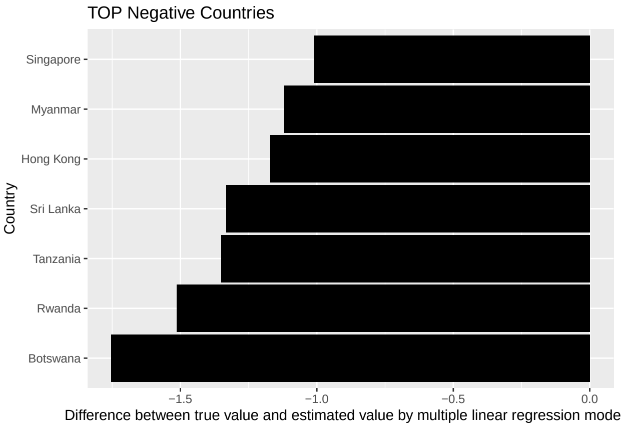
Due to the multiple linear regression model, we have estimated values for the Happiness Score. The plot below shows the estimated values and the true values. The smaller is the difference between the estimated value and the true value of the Happiness Score of a give country, the closer is the point to the 45 degree line. If the model would explain perfectly the Happiness Score, all the point would lie on the 45 degree line. However, in our case, the model does not explain all the Happiness Scores perfectly. There are countries, which are above the predicted happiness level and there are ones, which are below. Let us see, which ones are the top countries, that are the least well explained by the multiple linear regression model.



I set two threshold values to pick those countries, which shows a great difference between true Happiness Score value (y) and the estimated value (yE) based on my multiple linear regression model. Those countries, which shows higher than 0.8 difference between the true and the estimated value, are the “TOP Positive Countries”. The scatterplot below shows the countries which are “happier than they supposed to be” according to our multiple linear regression model. Benin, Pakistan, Costa Rica, Ivory Coast, Israel, El Salvador, Guatemala, Mexico and Niger have higher Happiness Score in the World Happiness Report than predicted by my multiple linear regression model. To understand the true value of the Happiness Score, we should consider to involve further explanatory variables into the multiple linear regression model, such as weather or connection to sea/ocean, as all of the countries have warm climate and most of them have connection to a sea or ocean.

Similarly, those countries, which shows lower than -1 difference between the true and the estimated value, are the “TOP Negative Countries”. The scatterplot below shows countries which are “less happier than they supposed to be" according to our multiple linear regression model. Botswana, Rwanda, Tanzania, Sri Lanka, Hong Kong, Myanmar and Singapore have lower Happiness Score in the World Happiness Report than the predicted value by my multiple linear regression model. The mentioned countries used to be part of a colony in the colonial period of the history. We might involve a dummy variable for further analysis to have a more precise model.



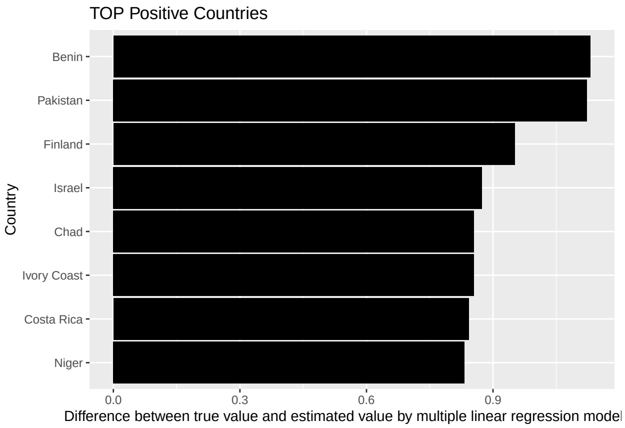


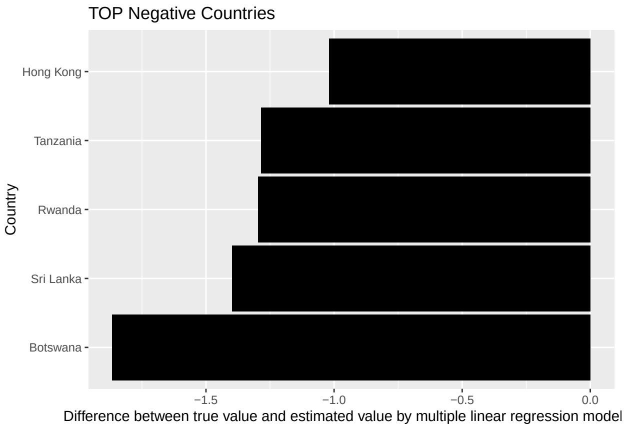
Multiple Linear Regression Model II.

The second multiple linear regression model involves four explanatory variables to predict the Happiness Score: GDP_per_capita, Social_support, Life_expectancy and Freedom_of_choice. The R-squared value is 0.77, which means that these explanatory variables explain 77% of the Happiness Score. The result of the second model is the next: Happiness Score = 1.89 + 0.81 * GDP_per_capita + 1.02 * Social_support + 1.14 * Life_expectancy + 1.85 * Freedom_of_choice This model exclude the Generosity and Corruption, which were not statistically significant based on their p-value.

Comparison of estimated and true value of Happiness Score (II.)

The true value of the Happiness Score and the estimated value by the second model are shown in the appendix by the scatterplot. I used the second model to see wether the TOP Negative and Positive Countries are changing. Finland and Chad have appeared as "happier countries than they supposed to be" according to the second model. However, Singapore and Myanmar have disappeared from the TOP Negative countries, which means that the difference between the true value of the Happiness Score and the estimated value by the second model was smaller, thus, more precise by the second model, which ignored two variables: Generosity and Corruption.





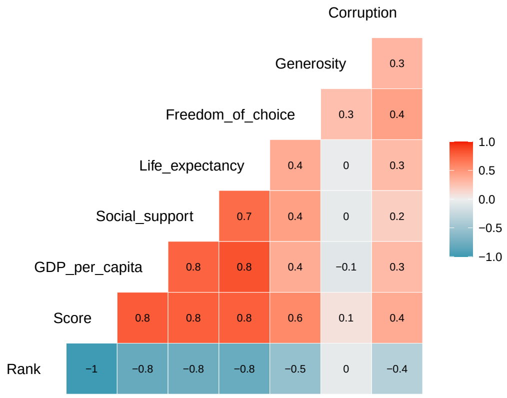
Collinearity

In this section, I look at the pair-wise correlation among the explanatory variables. The correlation graph below illustrates the dependence between the variables at the same time. The table contains the correlation coefficients between each variable and the others. In the correlation graph, it can be seen that all variables have a positive influence on the Happiness Score. There are three variables (GDP per capita, Social support, Healthy life expectancy), which have a strong correlation coefficient (0.8) with Happiness Score. Also, there are strong but imperfect correlations among three explanatory variables (GDP per capita, Social support and Healthy life expectancy), which means that multicollinearity exists in the model. It means that there are not many observations that are the same in one explanatory variable but different in the other explanatory variables. In the World Happiness Report it means, that countries with high GDP per capita tend to associate with high social support and high healthy life expectancy.

```
## [1] FALSE
```

```
##
## FALSE
##  1404
```

```
## Warning in ggcorr(g, label = TRUE, label_size = 2.9, hjust = 1, layout.exp = 2):
## data in column(s) 'Country' are not numeric and were ignored
```



Robustness analysis

My analysis assumes that the six different variables in the World Happiness Report can predict the Happiness Score in a certain level. I used two different multiple linear regression models. One of them involves all the variables contained by the World Happiness Report. However, two of the explanatory variables were not statistically significant (Generosity and Corruption). Thus, I made a second multiple linear regression model, which ignores the Generosity and Corruption to see if it predicts the Happiness Score better than the first model. The table below compares the two multiple linear regression models. To compare the two multiple linear regression models, I use the value of residual standard error. In model I., residual SE is lower (0.534) than in model II. (0.54). The residual SE is zero, if the model predicts perfectly. Consequently, the model I. is able to predict more precisely.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Szo, jan. 02, 2021 - 22:56:20 % Requires LaTeX packages: dcolumn

External validity

The World Happiness data set consists of variables, which are representative for all countries. Therefore, the model could predict the ranking of the World Happiness Report for the future globally. However, the prediction could be more precise if further variables would be involved into the analysis. As the residual analysis has shown, a dummy variable for each country if it used to be a colony or not, would have an additional value for the prediction.

Summary

Current report attempted to predict the Happiness Score for each country based on the data set of the World Happiness Report 2019. The analytics aims to compare the true values of the Happiness Score and the estimated values. The estimation was made by two different models. One of them involved all the six variables contained by the data set. Two variables were not considered statistically significant, therefore, the second multiple linear regression model ignored these two variables (Generosity and Corruption). Both of the models have shown a high R-squared value (77-78%), which allows us to conclude that the used variables explain the Happiness Score on a high level. However, at the beginning, the GDP per capita was regressed on the Happiness Score, which also explains 63% of the Happiness Score. In the analysis, we explored a multicollinearity among three explanatory variables (GDP per capita, Social support and Healthy life expectancy), which means that in most of the countries, these variables are tend to be associated and do not vary from each other.

Table 1: Estimation results of the two multiple linear regression models

	Dependent variable:	
	Score	
	(1)	(2)
GDP_per_capita	0.775^{***}	0.811^{***}
	(0.218)	(0.216)
Social_support	1.124^{***}	1.017^{***}
	(0.237)	(0.235)
Life_expectancy	1.078^{***}	1.141^{***}
	(0.335)	(0.337)
Freedom_of_choice	1.455^{***}	1.846^{***}
	(0.375)	(0.340)
Generosity	0.490	
	(0.498)	
Corruption	0.972^{*}	
	(0.542)	
Constant	1.795^{***}	1.892^{***}
	(0.211)	(0.199)
Observations	156	156
R^2	0.779	0.771
Adjusted R^2	0.770	0.765
Residual Std. Error	0.534 (df = 149)	0.540 (df = 151)
F Statistic	87.618^{***} (df = 6; 149)	127.048^{***} (df = 4; 151)
Note:		*p<0.1; **p<0.05; ***p<0.01

Appendix

Descriptive statistics of each variable

Table 2: Summary statistics for Happiness Score 2019

mean	median	std	iq_range	min	max	skew	numObs
5.407096	5.3795	1.11312	1.64	2.853	7.769	0.0113396	156

Table 3: Summary statistics for GDP per capita

mean	median	std	iq_range	min	max	skew	numObs
5.407096	5.3795	1.11312	1.64	2.853	7.769	0.0113396	156

Table 4: Summary stat for social support

mean	median	std	iq_range	min	max	skew	numObs
5.407096	5.3795	1.11312	1.64	2.853	7.769	0.0113396	156

Table 5: Summary stat for healthy life expectancy

mean	median	std	iq_range	min	max	skew	numObs
5.407096	5.3795	1.11312	1.64	2.853	7.769	0.0113396	156

Table 6: Summary stat for freedom to make life choices

mean	median	std	iq_range	min	max	skew	numObs
5.407096	5.3795	1.11312	1.64	2.853	7.769	0.0113396	156

Table 7: Summary stat for perceptions of corruption

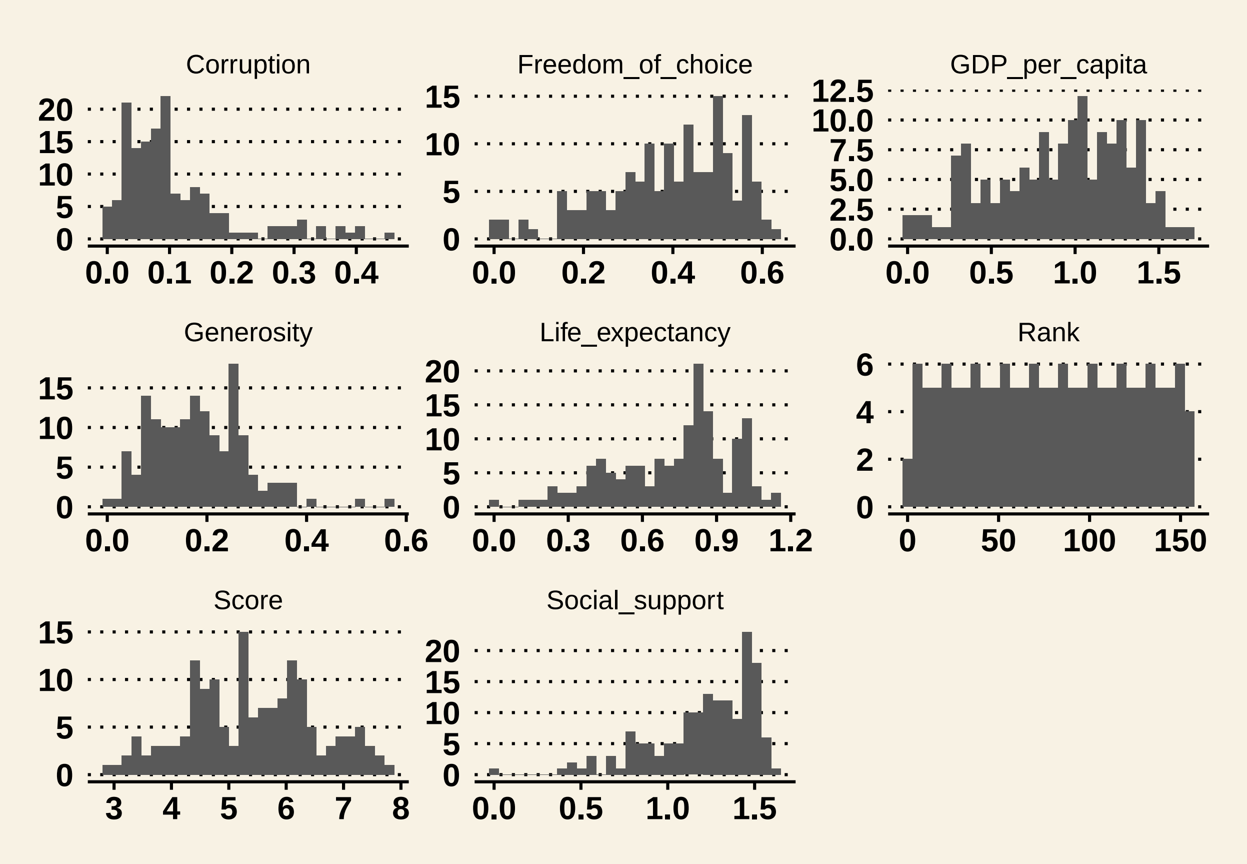
mean	median	std	iq_range	min	max	skew	numObs
5.407096	5.3795	1.11312	1.64	2.853	7.769	0.0113396	156

Table 8: Summary stat for generosity

mean	median	std	iq_range	min	max	skew	numObs
5.407096	5.3795	1.11312	1.64	2.853	7.769	0.0113396	156

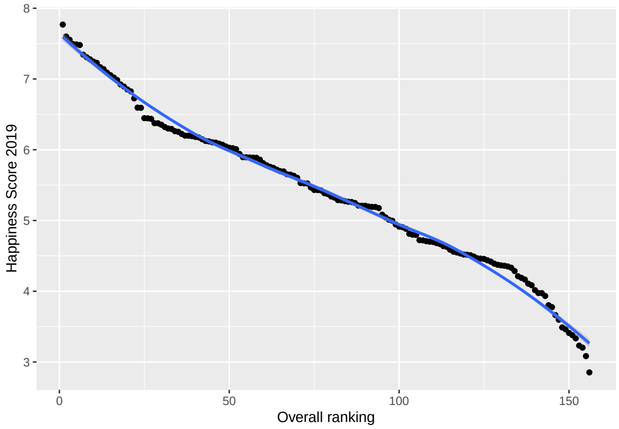
Histograms of each variable

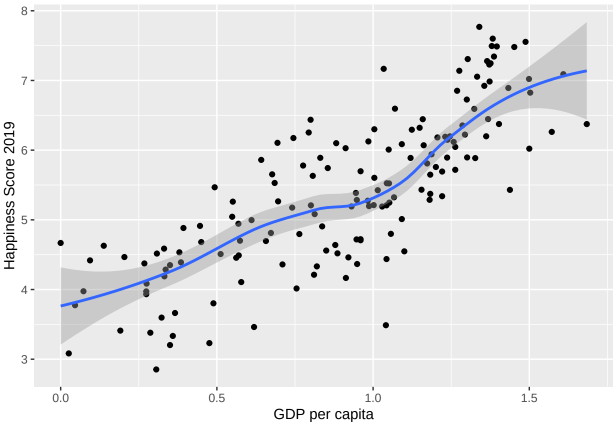
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

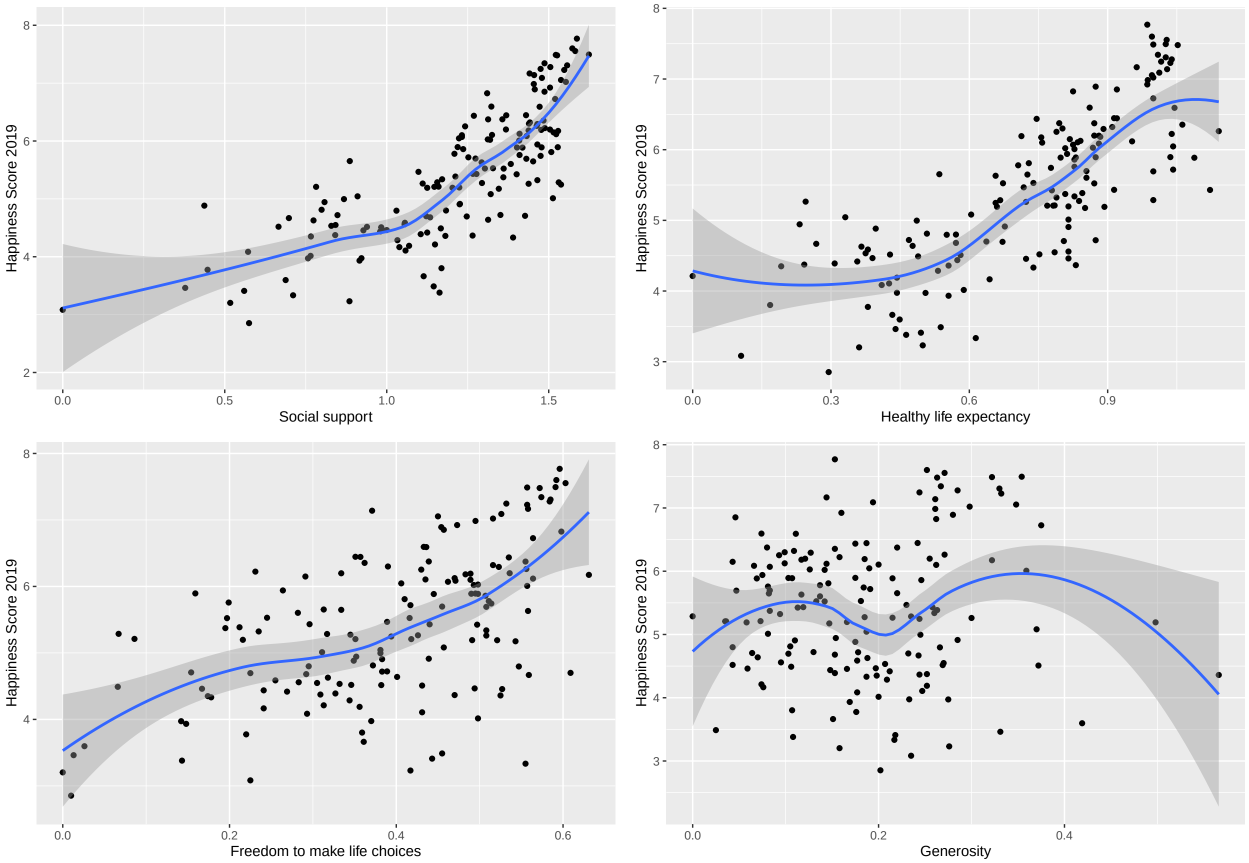


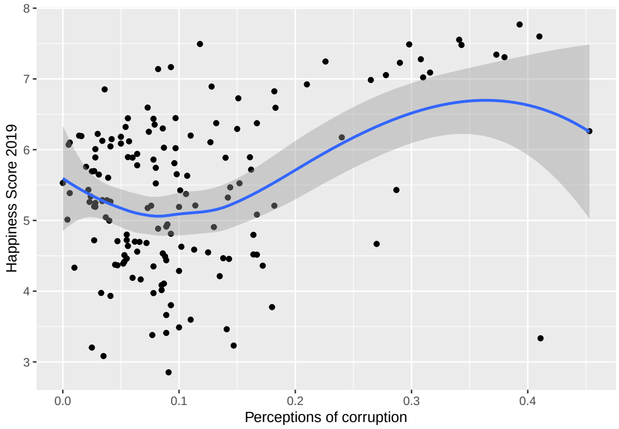
Scatterplots

The scatterplots below shows the association between the Happines Score and each explanatory variable with a lowess curve.









## Regression tables

The regression table shows the result of the simple linear regression for Happiness Score on GDP per capita.

```
## Warning: package 'kableExtra' was built under R version 4.0.3

##
## Attaching package: 'kableExtra'

## The following object is masked from 'package:dplyr':
##
##     group_rows
```

The regression table below shows the result of the multiple linear regression model I.

Table 9: Result of the simple linear regression on Happiness Score by GDP per capita

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
(Intercept)	3.40	0.13	26.10	0	3.14	3.66	154	Score
GDP__per__capita	2.22	0.13	16.78	0	1.96	2.48	154	Score

Table 10: Result of the multiple linear regression on Happiness Score

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
(Intercept)	1.80	0.24	7.49	0.00	1.32	2.27	149	Score
GDP__per__capita	0.78	0.22	3.55	0.00	0.34	1.21	149	Score
Social__support	1.12	0.25	4.44	0.00	0.62	1.62	149	Score
Life__expectancy	1.08	0.35	3.07	0.00	0.39	1.77	149	Score
Freedom__of__choice	1.45	0.34	4.26	0.00	0.78	2.13	149	Score
Generosity	0.49	0.59	0.83	0.41	-0.68	1.66	149	Score
Corruption	0.97	0.69	1.41	0.16	-0.39	2.34	149	Score

The regression table below shows the result of the multiple linear regression model II.

The scatterplot below shows the true values and the estimated values by the model II.

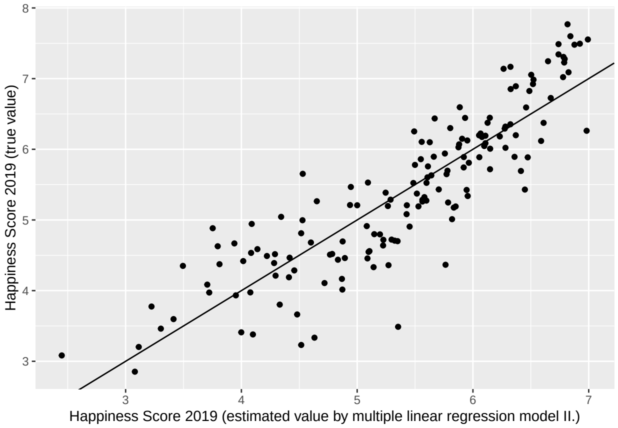


Table 11: Result of the multiple linear regression on Happiness Score

term	estimate	std.error	statistic	p.value	conf.low	conf.high	df	outcome
(Intercept)	1.89	0.21	8.86	0	1.47	2.31	151	Score
GDP__per__capita	0.81	0.21	3.87	0	0.40	1.22	151	Score
Social__support	1.02	0.25	4.07	0	0.52	1.51	151	Score
Life__expectancy	1.14	0.36	3.21	0	0.44	1.84	151	Score
Freedom__of__choice	1.85	0.33	5.67	0	1.20	2.49	151	Score