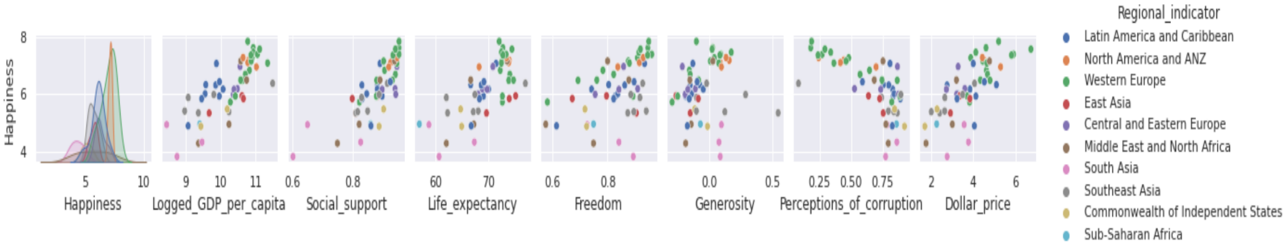
Project Part Three: Multiple Regression

Michael Ortiz, Joshua Head, Alexander Zommer

July 11, 2021

1 Data Preparation

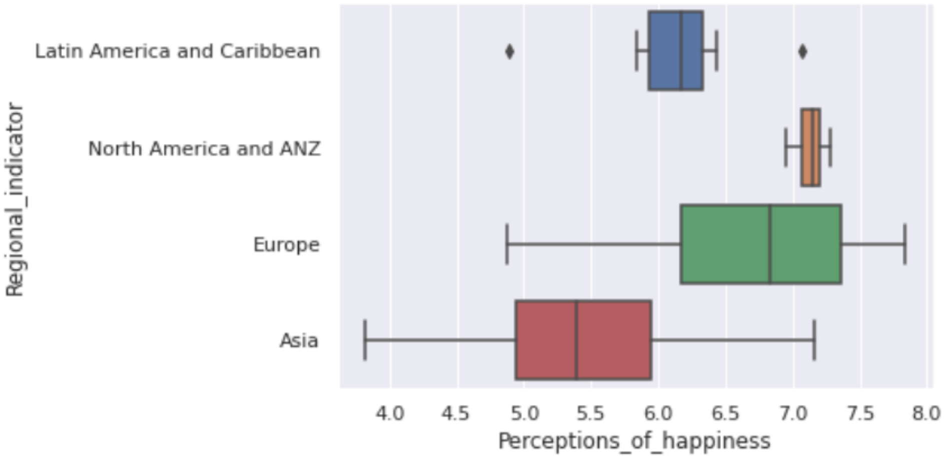
Our dataset contains 51 observations, a maximum of 7 continuous predictor variables, 1 categorical variable with 8 categories, and 1 response variable (Perceptions of Happiness). The data is visualized below.



Before we can build an appropriate model, we need to perform some minor data cleaning.

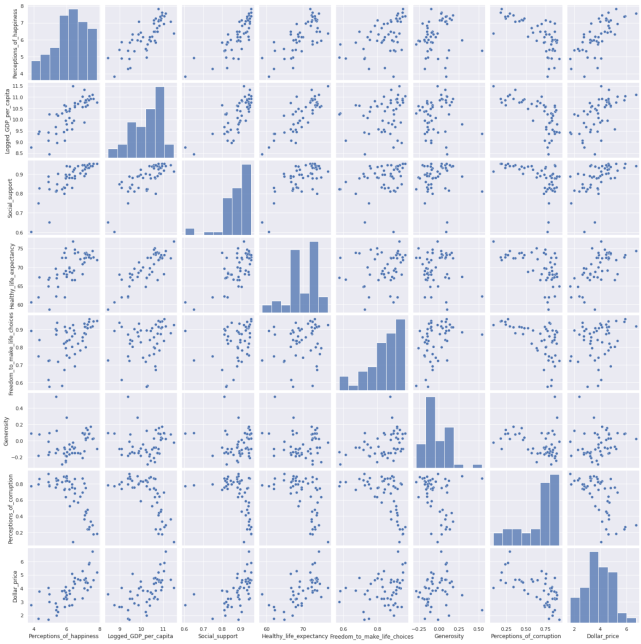
1.1 Simplifying the categorical variable

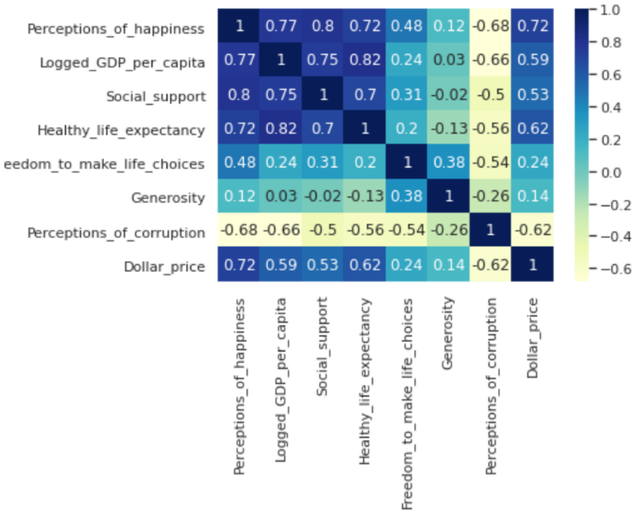
We have too many categories for Regional Indicator, and one of those categories (Sub-Saharan Africa) contains only one observation, which will not bode well for our model. Therefore, we remove that observation from the set; our total number of observations is now 50. We also consolidate the remaining categories into 4 larger ones represented by dummy variables: North America, Latin America and Caribbean, Europe, and Asia. The box plots of these categories given their happiness scores are shown below.



2 Checking Collinearity

Before we build our model, it is a good idea to get an idea of how our predictors relate to each other. We visualize the scatter plots of each predictor plotted against each other, and we produce a correlation matrix to get a numerical idea of the collinearity.

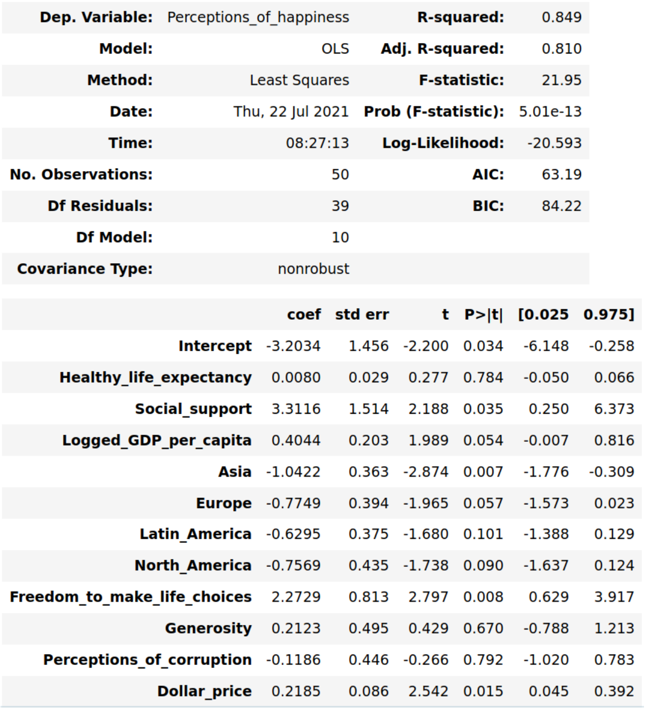




Our plots imply that there may be strong collinearity between our predictors Logged GDP, Social Support, and Life Expectancy. We will need to account for this when building our model.

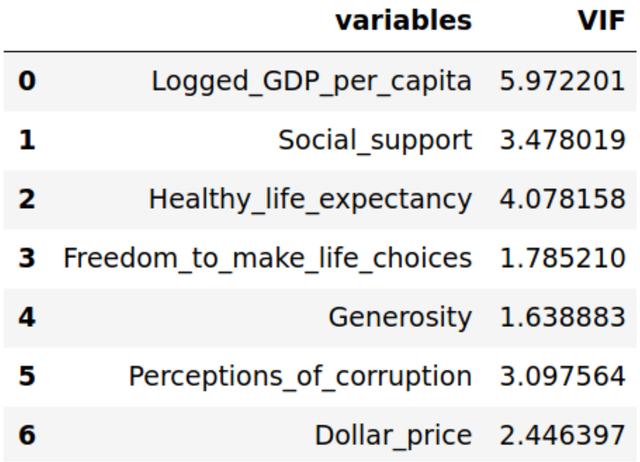
3 Selecting a Model

3.1 Model using all Predictors



Dep. Variable:	Perceptions_of_happiness	**R-squared:**	0.849
Model:	OLS	**Adj. R-squared:**	0.810
Method:	Least Squares	**F-statistic:**	21.95
Date:	Thu, 22 Jul 2021	**Prob (F-statistic):**	5.01e-13
Time:	08:27:13	**Log-Likelihood:**	-20.593
No. Observations:	50	**AIC:**	63.19
Df Residuals:	39	**BIC:**	84.22
Df Model:	10		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-3.2034	1.456	-2.200	0.034	-6.148	-0.258
Healthy_life_expectancy	0.0080	0.029	0.277	0.784	-0.050	0.066
Social_support	3.3116	1.514	2.188	0.035	0.250	6.373
Logged_GDP_per_capita	0.4044	0.203	1.989	0.054	-0.007	0.816
Asia	-1.0422	0.363	-2.874	0.007	-1.776	-0.309
Europe	-0.7749	0.394	-1.965	0.057	-1.573	0.023
Latin_America	-0.6295	0.375	-1.680	0.101	-1.388	0.129
North_America	-0.7569	0.435	-1.738	0.090	-1.637	0.124
Freedom_to_make_life_choices	2.2729	0.813	2.797	0.008	0.629	3.917
Generosity	0.2123	0.495	0.429	0.670	-0.788	1.213
Perceptions_of_corruption	-0.1186	0.446	-0.266	0.792	-1.020	0.783
Dollar_price	0.2185	0.086	2.542	0.015	0.045	0.392



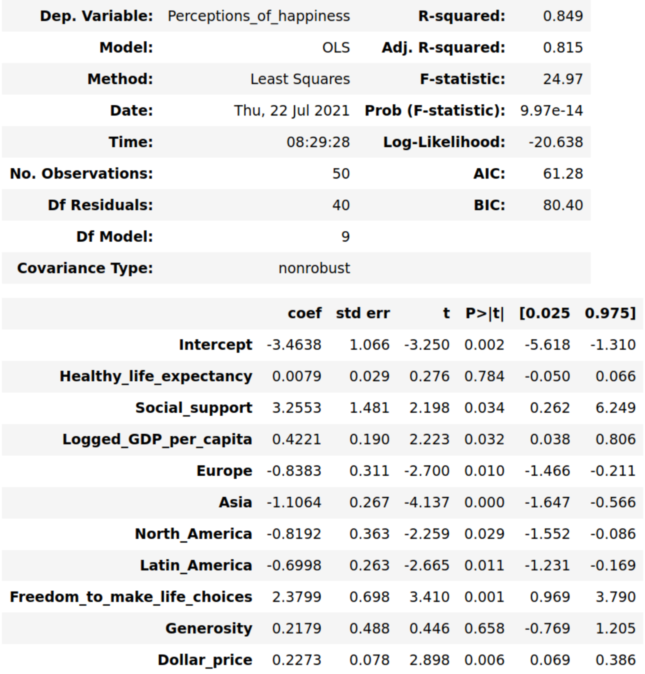
	variables	**VIF**
0	Logged_GDP_per_capita	5.972201
1	Social_support	3.478019
2	Healthy_life_expectancy	4.078158
3	Freedom_to_make_life_choices	1.785210
4	Generosity	1.638883
5	Perceptions_of_corruption	3.097564
6	Dollar_price	2.446397

Surprisingly, our model constructed using all of our predictors performed very well, with an $R^2 = 0.849$ and an adjusted $R^2 = 0.810$, but there are a handful of issues which we will address shortly.

Our F-Statistic of 21.95 and P-Value of $5.01e - 13$ indicate that, for any reasonable value of α, our model is significant for predicting values of Perceptions of happiness. We calculated the Variance Inflation Factor (VIF) for each of our continuous variables, and as we suspected, GDP, Social Support, and Life Expectancy are prime suspects of collinearity. We also performed t-tests using $\alpha = 0.05$ on each of our predictors to test for their significance in predicting happiness.

Despite the implication of collinearity, our VIF values are nonetheless relatively small, yet some of our predictors featured extremely large P-Values when performing the t-test. Therefore, our strategy is to address that first by removing the variable with the largest P-Value, one at a time, until all predictors return $p < 0.05$.

3.2 Remove Corruption Predictor

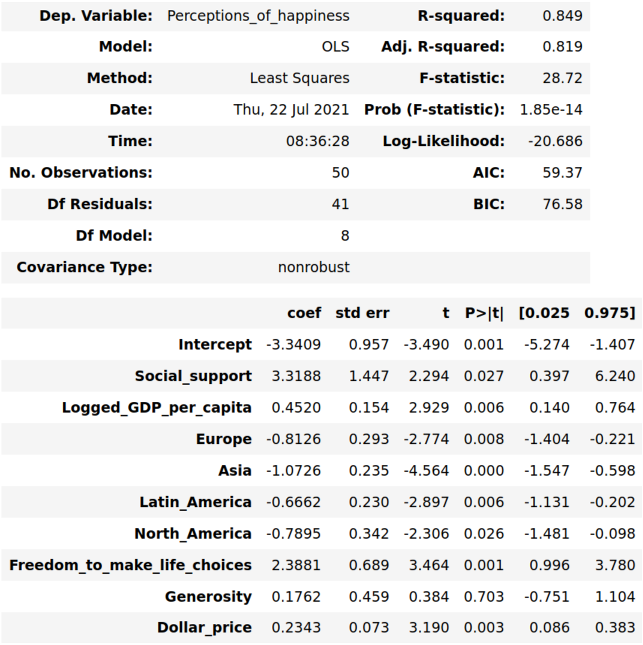


Dep. Variable:	Perceptions_of_happiness	**R-squared:**	0.849
Model:	OLS	**Adj. R-squared:**	0.815
Method:	Least Squares	**F-statistic:**	24.97
Date:	Thu, 22 Jul 2021	**Prob (F-statistic):**	9.97e-14
Time:	08:29:28	**Log-Likelihood:**	-20.638
No. Observations:	50	**AIC:**	61.28
Df Residuals:	40	**BIC:**	80.40
Df Model:	9		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-3.4638	1.066	-3.250	0.002	-5.618	-1.310
Healthy_life_expectancy	0.0079	0.029	0.276	0.784	-0.050	0.066
Social_support	3.2553	1.481	2.198	0.034	0.262	6.249
Logged_GDP_per_capita	0.4221	0.190	2.223	0.032	0.038	0.806
Europe	-0.8383	0.311	-2.700	0.010	-1.466	-0.211
Asia	-1.1064	0.267	-4.137	0.000	-1.647	-0.566
North_America	-0.8192	0.363	-2.259	0.029	-1.552	-0.086
Latin_America	-0.6998	0.263	-2.665	0.011	-1.231	-0.169
Freedom_to_make_life_choices	2.3799	0.698	3.410	0.001	0.969	3.790
Generosity	0.2179	0.488	0.446	0.658	-0.769	1.205
Dollar_price	0.2273	0.078	2.898	0.006	0.069	0.386

Our model from section [3.1] computed a terrible P-Value of 0.792 for Perceptions of Corruption, so we removed it first. $R^2 = 0.849$ did not change at all, yet adjusted $R^2 = 0.815$ increased. This is a great sign that this predictor was not contributing much of anything to our model.

3.3 Removing Life Expectancy

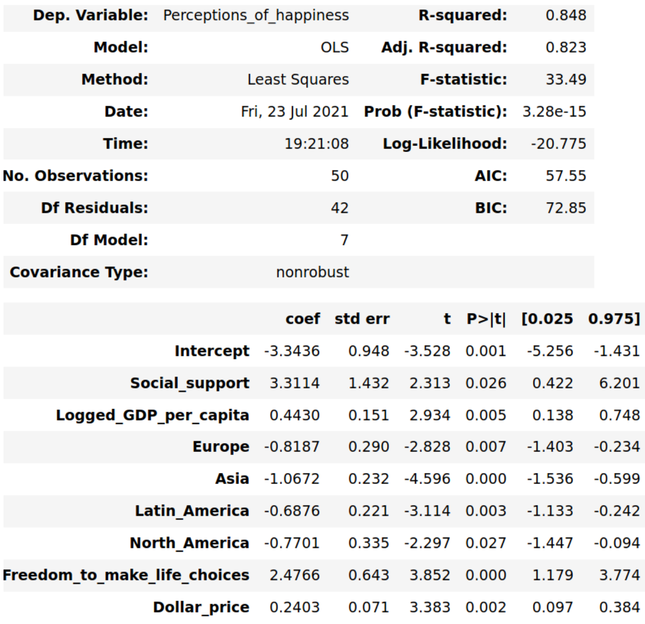


Dep. Variable:	Perceptions_of_happiness	**R-squared:**	0.849
Model:	OLS	**Adj. R-squared:**	0.819
Method:	Least Squares	**F-statistic:**	28.72
Date:	Thu, 22 Jul 2021	**Prob (F-statistic):**	1.85e-14
Time:	08:36:28	**Log-Likelihood:**	-20.686
No. Observations:	50	**AIC:**	59.37
Df Residuals:	41	**BIC:**	76.58
Df Model:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-3.3409	0.957	-3.490	0.001	-5.274	-1.407
Social_support	3.3188	1.447	2.294	0.027	0.397	6.240
Logged_GDP_per_capita	0.4520	0.154	2.929	0.006	0.140	0.764
Europe	-0.8126	0.293	-2.774	0.008	-1.404	-0.221
Asia	-1.0726	0.235	-4.564	0.000	-1.547	-0.598
Latin_America	-0.6662	0.230	-2.897	0.006	-1.131	-0.202
North_America	-0.7895	0.342	-2.306	0.026	-1.481	-0.098
Freedom_to_make_life_choices	2.3881	0.689	3.464	0.001	0.996	3.780
Generosity	0.1762	0.459	0.384	0.703	-0.751	1.104
Dollar_price	0.2343	0.073	3.190	0.003	0.086	0.383

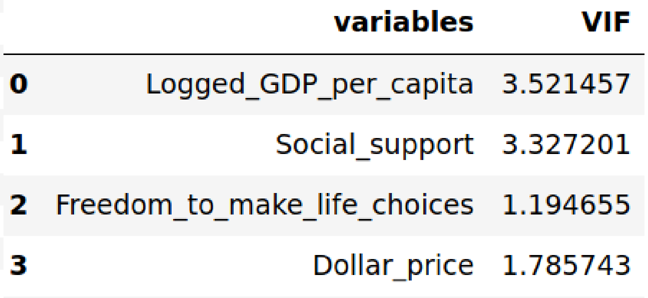
Our next predictor to go was Healthy Life Expectancy, with model [3.2] showing a P-value of 0.784. Yet again R^2 remained unchanged at 0.849, and adjusted R^2 increased to 0.819.

3.4 Final Model



Dep. Variable:	Perceptions_of_happiness	**R-squared:**	0.848
Model:	OLS	**Adj. R-squared:**	0.823
Method:	Least Squares	**F-statistic:**	33.49
Date:	Fri, 23 Jul 2021	**Prob (F-statistic):**	3.28e-15
Time:	19:21:08	**Log-Likelihood:**	-20.775
No. Observations:	50	**AIC:**	57.55
Df Residuals:	42	**BIC:**	72.85
Df Model:	7		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	-3.3436	0.948	-3.528	0.001	-5.256	-1.431
Social_support	3.3114	1.432	2.313	0.026	0.422	6.201
Logged_GDP_per_capita	0.4430	0.151	2.934	0.005	0.138	0.748
Europe	-0.8187	0.290	-2.828	0.007	-1.403	-0.234
Asia	-1.0672	0.232	-4.596	0.000	-1.536	-0.599
Latin_America	-0.6876	0.221	-3.114	0.003	-1.133	-0.242
North_America	-0.7701	0.335	-2.297	0.027	-1.447	-0.094
Freedom_to_make_life_choices	2.4766	0.643	3.852	0.000	1.179	3.774
Dollar_price	0.2403	0.071	3.383	0.002	0.097	0.384



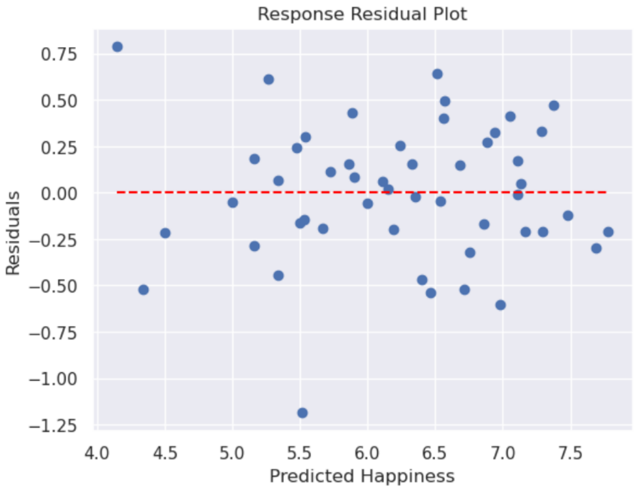
	variables	**VIF**
0	Logged_GDP_per_capita	3.521457
1	Social_support	3.327201
2	Freedom_to_make_life_choices	1.194655
3	Dollar_price	1.785743

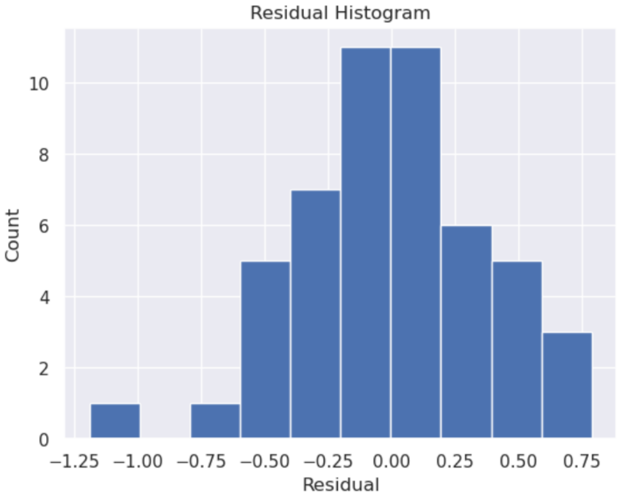
The final predictor we removed was Generosity, with model [3.3] reporting a P-Value of 0.703. This model insignificantly reduced $R^2 = 0.848$ from 0.849, and increased adjusted R^2 to 0.823. At this point, for $\alpha = 0.05$, all predictors are significant. With that information, we recalculated the VIF scores for the remaining variables and we are no longer seeing signs of collinearity using VIF > 5 as an indication. Any further attempts at adjusting the model by either removing terms, adding interaction terms, or transforming predictors resulted in significantly worse performance.

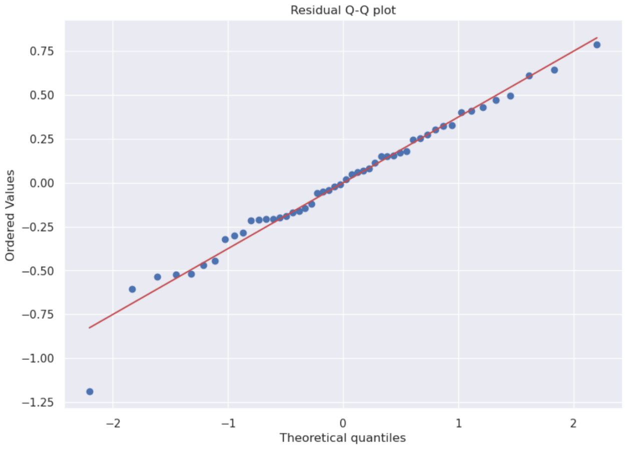
To summarize, between models [3.1] and [3.4] our R^2 reduced by 0.001 from 0.849 to 0.848, adjusted R^2 increased to 0.823 from 0.810, our F-Statistic increased from 21.95 to 33.49, we removed all insignificant predictors via t-tests, and our VIF scores no longer hint at collinearity.

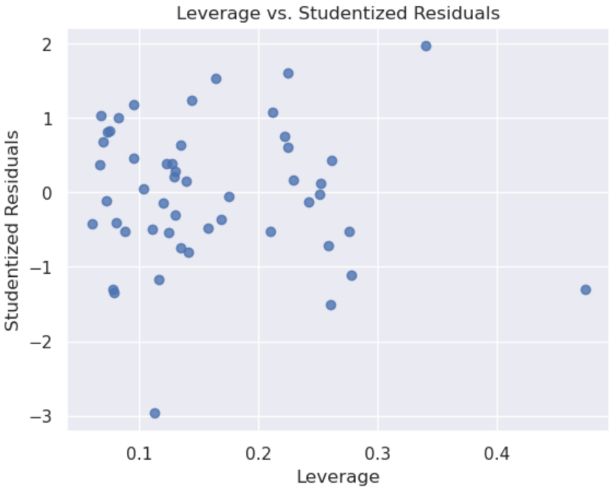
4 Analyzing the Model

4.1 Diagnostic Plots



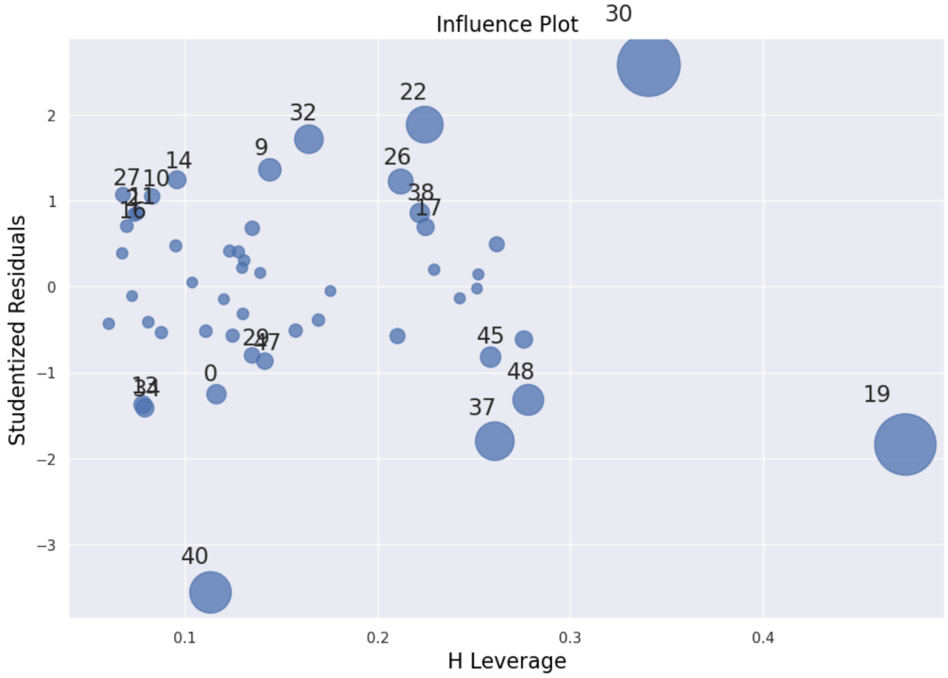




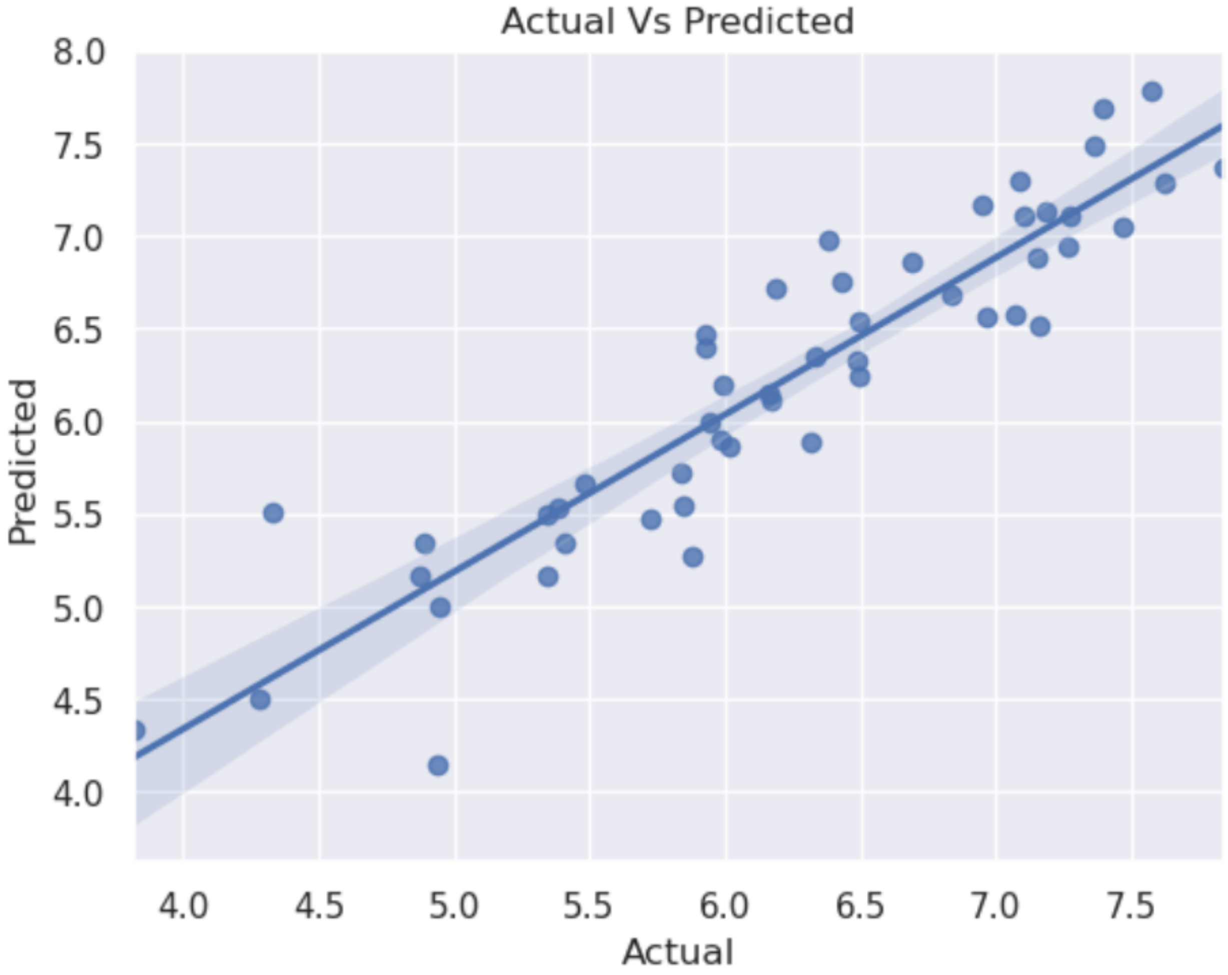


4.2 Interpreting the Plots

The random scatter in the residual plot, combined with the shape of the residual histogram and linearity in the Q-Q plot strongly indicate that our residuals are normally distributed and we not in violation of the homoscedasticity assumption. We confirm this by performing the shapiro test which resulted as ($statistic = 0.9799081087112427, pvalue = 0.5480934381484985$). We now have overwhelming evidence of normality in our residuals. We also suspect, however, that there are anywhere from 1 to 3 outliers in our data. Our first 3 plots indicate there is an outlier with an extreme negative residual, and our leverage plot indicates that there are 2 points with high leverage. We analyze further by visualizing Cook's Distance via an influence plot. In this plot, the larger an observation is, the more influential (higher Cook's Distance) it is.



We see that our extreme negative residual does not end up being that influential, but our observations with high leverage (30 and 19) are significantly more influential than the others and may in fact be outliers. The model may benefit from their removal. Nonetheless, our model has shown significant evidence that it is a strong predictor of happiness. We further confirm this by plotting our predicted vs actual values of happiness and observe a strongly positive correlation.


Actual Vs Predicted
Predicted
Actual
8.0
7.5
7.0
6.5
6.0
5.5
5.0
4.5
4.0
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5