Ethan Engel

Probability & Statistics 4442 – Spring 2021

Final Project – Response Surface Regression & Methodology

The research question to be investigated is two tiered. First, can a meaningful model to predict the conditions for famine within a nation can be developed from the dataset? Second, can a model developed from Response Surface Methodology (RSM) be more informative and/or fit the data better than Multiple Linear Regression (MLR) alone? In the end, while RSM did lead to models which fit the data significantly better than MLR alone, the utility of the models was left in question.

The data set was pieced together from four different sources. The continuous outcome variable was Global Hunger Index Score (GHI\_score), from dataverse.harvard.edu. This is an index computed from four statistics of a nation. The first is total proportion of the population that is undernourished. Second is the total proportion of the child population that is experiencing “wasting”, which means developmentally underweight for their height. Thirdly is the proportion of the child population that is experiencing growth stunting, and finally, the child mortality rate. There were three predictor variables in the dataset, all of which were continuous. The first was Fragile State Index (FSI\_score) from fragilestatesindex.org, which is a measure of the degree of conflict, corruption, and ongoing human rights violations for a nation. The second predictor was GDP/capita from data.un.org. The final predictor to be used was Temperature Change (curr\_temp\_change) from fao.org. This is a measure, from the most recent meteorological years, of a nation’s current temperature relative to the baseline climatological years of 1950-1980. There were a total of 117 nations represented in the set. Nations were left out of the set if data was missing. In addition, most developed nations were left out of the set, as they would minimally inform building the model.

Regarding exploratory analysis, the maximum GHI score was 46.9 for the Central African Republic. while there were several nations in the set who had GHI scores of <5, which were interpolated as 2.5. The distribution for GHI scores was skewed to the right, showing evidence of a normality violation. As for the predictor variables, the FSI scores showed a far more symmetric distribution, although there was a low-end outlier, Uruguay, whose FSI score was 33.4. Yemen had the highest FSI score of the set, at 112.4. The distribution for GDP/capita was heavily skewed to the right, with six upper end outliers, and clustered at the lower end. The range was from $434.77 (Malawi) to $31999.27 (Kuwait). The non-outlier portion of the temperature change distribution was symmetric, but there were six upper end outliers and one on the lower end. The highest temperature change of the set was Russia at 3.699 degrees over baseline climatology and the lowest was Nepal at .02 degrees below. Nepal was the only nation of the set with a negative temperature change. The potential normality violations were not terribly concerning, as RSM is resistant to normality violations.

Using the RSM library and function call in R, a series of three models would be generated, each nested in the next. The first would be a first order model, which is essentially linear. The second model would have pure quadratic terms added as predictors. The third has all two-way interactions of the linear terms added as predictor coefficients. Since the first model is nested in the second, and the second in the third, anova nested model comparisons would be run to test for fit on the data, checking against other summary statistic values for consistency.

RSM is essentially an extension and enhancement of Multiple Linear Regression, much like quadratic functions and cubic functions are a progression beyond linear functions within the overarching set of polynomial functions. The data requirements then are the same for RSM as they are for MLR. Factor variables can be used as well. The RSM function call uses multivariable calculus to test for curvature within the data which may make the addition of polynomial terms generate a better fit. This progression of polynomial terms is a short one, as the addition of cubic terms to a model immediately presents a danger of overfitting and therefore is infrequently employed. Cubic terms and beyond are not eligible as options for arguments to pass into the rsm() function call.

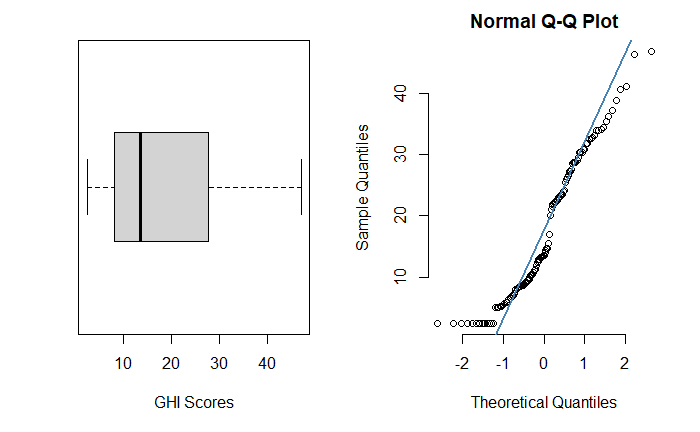
The three models generated the following summary statistics:

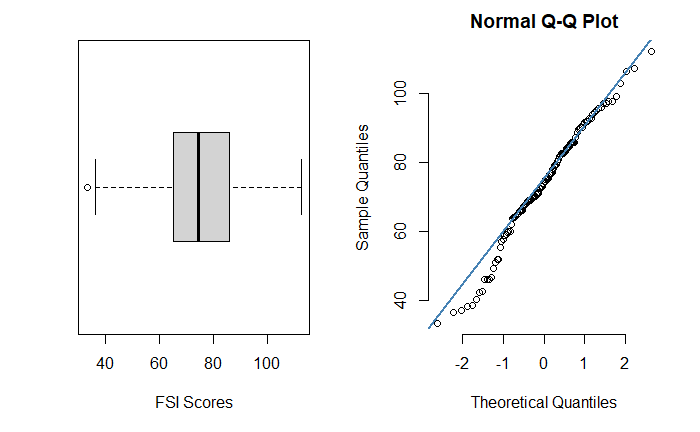
|  |  |  |  |
| --- | --- | --- | --- |
|  | Adjusted R-squared | Lack of Fit | AIC |
| First order Model (FO) | .61 | 6085 | 804 |
| First order + Pure Quadratic Terms (FO+PQ) | .73 | 4051 | 763 |
| Full Second Order Model with interactions of the Linear Terms | .74 | 3808 | 762 |

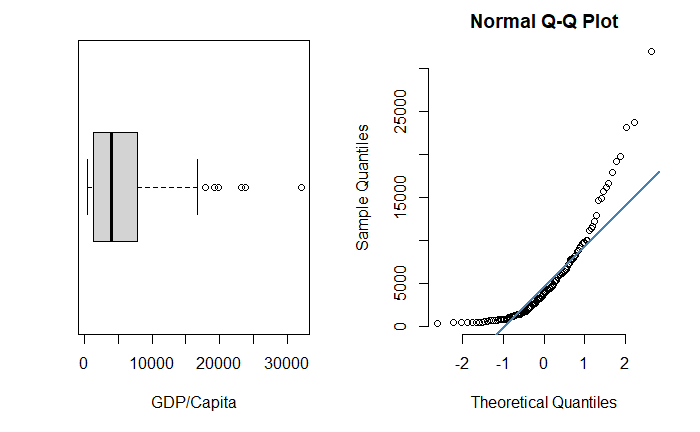
The anova comparison of the first model to the second, generated an F statistic > 18 on 3 degrees of freedom and p value of 9.46e-10, so adding the pure quadratic terms generated a model which fit the data far better. This was consistent with improvements across all three summary statistics in the table. While the three summary statistics improved slightly from the second to the third model, the anova comparison of the two generated an F statistic of 2.27 on the df=3 which rendered a p value of .08, not less than any meaningful alpha threshold.

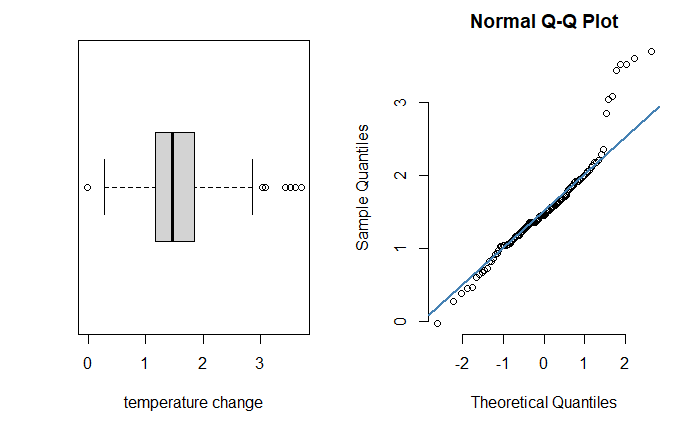
Going forward, the second model, with linear and pure quadratic terms, but without the linear interaction terms, is the one which will be evaluated and have its coefficients inspected. For this model, the coefficients for the intercept, the FSI scores (both linear and quadratic terms), and the GDP per capita (also both linear and quadratic terms) were all statistically significant. However, the temperature change coefficients, both linear and quadratic, were not statistically significant. This was maybe the most compelling component of the model. The meaningfulness and applicability of the model is rather tenuous. There is a stationary, or optimal, point when FSI score =61.7, GDP per capita=0 and temperature change=4.18. This is not especially informative, as a maximum value for GHI score would intuitively be expected at a high FSI score and at the minimum possible GDP per capita. In addition, the temperature change value is extrapolating beyond the current maximum temperature change of the dataset. Furthermore, the output for the model mentions that a “near stationary ridge situation” is being detected. Since the linear and quadratic coefficients for GDP per capita were the most statistically significant, it seems likely that this is occurring at or near the GDP per capita value of the stationary point value. Further inspection of a contour plot when GDP per capita=0 sheds little light upon the scenario. The GHI scores are generally greater at this slice than at the mean GDP per capita, but that was to be expected. The overall result was a model that fits reasonably well, but of little predictive use.

Exploratory Analysis Visuals









Response Surface Graphs

