Erica Wertel ID# 112528977

Vaina Vincent ID# 112166555

Alyssa Yuan ID# 112814476

How Factors Influence Happiness Score through Nonlinear Regressions

Introduction:

In Project 1, to find an answer to our research question, we ran a linear regression model corrected for heteroskedastic errors for Happiness score with Generosity, computed the hypothesis tests for 5% and 10% significance levels, and calculated the 90% confidence interval all of which confirmed that Generosity did have an effect on Happiness score, when the Generosity index increases by one unit, the Happiness Score increases by about 0.886 units. We realized that when the summary statistics are computed for the linear regression of Happiness Score and Generosity, the multiple R-squared is 0.005749 which means that the linear regression computed with only one regressor being Generosity can explain about 0.57% of the data. Therefore, we ran a linear regression with two regressors to try to explain more of the data that affects Happiness score. We found that the coefficient of the variable GDP.per.capita is greater than the coefficient of the variable Generosity. When we increase the GDP.per.capita index by one unit, Happiness Score will increase by about 2.25 units, but this regression also made us realize that the variable Generosity has more of an effect on Happiness Score than previously thought in the first linear regression model with only Generosity as a variable. By adding the variable GDP.per.capita, the coefficient of Generosity rose from about 0.886 to 1.635. Also, the multiple R-squared increased from 0.005749 to 0.6497 which shows that by controlling for GDP.per.capita, we can explain more of the data than by running a regression only for the variable Generosity.

In this first project we assumed that the regression functions are linear, which means that the effects on our dependent variable, Happiness score, by change in our independent variables Generosity and GDP.per.capita are always constant but it is a very strong assumption that is not necessarily true. Assuming a linear regression may result in wrong estimates, so the estimator of the effect on the dependent variable would be biased and we would need to take into account the non-linearity. It is possible that the relationship between the independent and dependent variables is non-linear which means that the effect of a change in the independent variables on the dependent variable is not going to be constant for different values of the independent variables. In this first project, we also notice that our regression model suffers from omitted variable bias and simultaneous causality bias (X causes Y, Y causes X) which are threats to the internal validity of the regression which is why we could use instrumental variables to help eliminate this bias.

We are doing non-linear regressions; the general non-linear population regression function equation is: *Yi = f(X1i , X2i ,…, Xki) + ui , i = 1,…, n*

The change in Y associated with a change in X1 , holding X2 ,…, Xk constant is:  
 *ΔY = f(X1 + ΔX1, X2,…, Xk) – f(X1, X2,…, Xk)*It is the difference between the value of the population regression function before and after changing X1, holding X2, …., Xk constant.

More specifically we are using a quadratic nonlinear regression model which is defined by:

*Yi = β0 + β1Xi + β2Xi2* It is similar to the linear multiple regression model, the difference is that the regressors are powers of X. The coefficients in this regression function are difficult to interpret but not the function itself. To interpret the estimated regression function we have to plot predicted values as a function of the independent variable X and compute the predicted change in the dependent variable from a change in the independent variable ( ΔY/ΔX ) at different values of X.  
We are also using the log-linear population regression function defined by:

*ln(Yi ) = β0 + β1Xi + ui*Logarithmic functions transform permit modeling relations in “percentage” terms, rather than linearly. The interpretation is as follows: a change in X by one unit (ΔX = 1) is associated with a 100% change in Y.  
Lastly, we will be using interacting variables which helps us measure if the change of the dependent variable Y when we change the independent variable X1 depends on another independent variable X2. Specifically we will calculate the interactions between two continuous variables, characterized by:  *Yi = β0+ β1X1i+ β2X2i + β3(X1i* ***×*** *X2i ) + ui*.  
In a linear equation where X1 and X2 are continuous, the effect of X1 doesn’t depend on X2 and vice versa, which is why we added the interaction term X1i **×** X2i as a regressor

Our econometric models are the Quadratic and Log-Linear models. We didn't do LogLog models because this model is often used to estimate supply and demand equations. Since we’re talking about the effects of independent variables on the happiness score, LogLog was not appropriate.

Although Linear regression is helpful with predicting the value of the dependent variable using the independent variables, Nonlinear regression can be a strong alternative to linear regression. This is because it allows for malleable curve-fitting functionality. We used the Quadratic and Log-Linear nonlinear regression for our dataset to better describe the nature of the data. They could provide a fitted line plot that could show the regression line following the data closely without any deviations. Although it's not possible to calculate the R-squared, the average absolute distance from the data points to the regression line lowers and you want a lower value because it means the data points are closer to the fit line.

I’ll be talking about what the quadratic,log linear and interacting variables can do that the linear one couldn’t. The Quadratic nonlinear regression model accounts for the direction change between the dependent and independent variable. The Log Linear regression models the rate change at a constant percent per year, while the linear regression only models the rate change at a constant fixed amount per year. The interacting variables adds to our understanding of the relationship among the variables in the model and allows for more hypotheses to be tested. It tells us how two or more independent variables work together to impact the dependent variable.

Data:

After conducting a linear regression to find the effect Generosity has on Happiness Score in project one, we decided to perform a quadratic regression to hopefully get a more accurate correlation between the variables. By performing an estimation of the quadratic function *Scorei = β0 + β1Generosityi + β2Generosityi2 + μi*  , we can interpret the relationship between the variables Happiness Score and and Generosity and compare whether the estimation of the quadratic was more applicable to the data than the linear model whose equation is *Scorei = β0 + β1Generosityi + μi* . Initially, we made a variable named gensq which is the square of all the values in the column Generosity. Next, we estimated the quadratic equation *Scorei = β0 + β1Generosityi + β2Generosityi2 + μi* and corrected the estimation for heteroskedastic standard errors in Figure 1 below.

Figure 1

Quadratic

|  | Estimate | Std. Errors | t-value | p-value |
| --- | --- | --- | --- | --- |
| Intercept | 5.09449 | 0.24924 | 20.4403 | <2e-16 |
| Generosity | 2.60331 | 2.52577 | 1.0307 | 0.3043 |
| gensq | -3.90431 | 5.36761 | -0.7274 | 0.4681 |

According to the estimates computed, the variable Generosity will have a positive effect on Happiness Score by about 2.60 units. However, once a maximum point is reached, the variable Generosity will have a negative effect on Happiness Score of about -3.90 units. After conducting the quadratic estimation, we summarized the regression and found the data in Figure 2.

Figure 2

Quadratic

| Multiple R-squared | 0.008317 |
| --- | --- |
| Adjusted R-squared | -0.004646 |
| P-value | 0.5279 |

While the quadratic estimation appears to offer some helpful information about the relationship between Happiness Score and Generosity, the Multiple R-squared is a very low value at 0.008317. Therefore, the quadratic regression conducted in Figure 1 can only explain approximately 0.83% of the data. This extremely low Multiple R-squared value is primarily because we are only taking into account Generosity’s impact on Happiness Score and not any other variable. Based on the p-value for the whole regression, we can conclude that there is a high probability that these results could have come up in a random distribution. Therefore, these estimates are not statistically significant and not a good fit for the data. When comparing the quadratic regression of Generosity to the linear regression of Generosity in project 1, the linear regression determined that an increase in the Generosity index by one unit would increase the Happiness Score by about 0.886 units. However, the quadratic estimation gives more detail since Generosity increases Happiness Score by a greater amount at first, but will start to have a negative effect on Happiness Score at a certain maximum point. Also, in the linear regression, the Multiple R-squared was a value of 0.005749. Since the Multiple R-squared increased from 0.005749 in the linear regression to a value of 0.008317 in the quadratic regression, the quadratic regression explains more of the data than the linear model. However, the p-value of the whole linear regression is lower so the linear regression is comparatively more statistically significant than the quadratic regression. Even though the quadratic regression model can explain more of the data than the linear model, neither regression fits the data well primarily because there are other factors that influence Happiness Score such as GDP per capita and healthy life expectancy. We can infer that these models suffer from omitted variable bias.

Next, an F-test was conducted to determine whether a linear or quadratic model is more appropriate to estimate the data. The null hypothesis is that the population coefficient of gensq equals zero while the alternative hypothesis is that the population coefficient gensq is not equal to zero. The results give a p-value for the F-statistic of 0.4681. Since the p-value is greater than an alpha equal to 0.01, we would accept the null hypothesis at the 1% significance level. Therefore, we would assume that the linear model is a better fit than the quadratic model for the data on Happiness Score.

These results have serious implications that affect economic policy. If governments decide to follow the quadratic model, they will realize that generosity only increases happiness up to a specific point. So, governments will want to maximize the well-being of their citizens to become more generous and hopefully more happy as a result. However, if there is too much of an incentive to be generous, people may not have enough money for their own lifestyle and their happiness level may decrease as a result. So, economic policy would take into account that generosity increases happiness, but only until a certain level.

After comparing a linear model to a quadratic model, we estimated a log linear model with the equation *lScorei = β0 + β1GDP.per.capitai + μi* where *lScore* is the log of all the values in the column Score. The log of Happiness Score allows us to normalize the data in the variable to fit a normal distribution better than the original data. The results of the log linear regression in Figure 3 are corrected for heteroskedastic standard errors below.

Figure 3

Log Linear

|  | Estimate | Std. Errors | t-value | p-value |
| --- | --- | --- | --- | --- |
| Intercept | 1.279293 | 0.029257 | 43.726 | <2.2e-16 |
| GDP.per.capita | 0.426705 | 0.027117 | 15.736 | <2.2e-16 |

| Multiple R-squared | 0.6254 |
| --- | --- |
| Adjusted R-squared | 0.623 |
| P-value | <2.2e-16 |

According to the table, a change in GDP.per.capita by one unit is associated with a 42.671% change in Happiness Score. Therefore, changing the GDP.per.capita by a small amount will increase the Happiness Score by almost 50%. The variable GDP is important to increasing a person’s happiness level as opposed to the variable Generosity. Also, looking at the Multiple R-squared in Figure 3, the variable GDP.per.capita can explain about 62.5% of the data on the log of Happiness Score. Comparing the log linear model to the linear model with the equation *Scorei = β0 + β1GDP.per.capitai + μi* , a change of one unit in the variable GDP.per.capita will result in an increase in the Happiness Score by approximately 2.22 units in the linear model. Also, the linear model is able to explain more of the data than the log linear model because the Multiple R-squared value of the linear model is about 0.63. Therefore, the linear model can explain about 0.5% more of the data than the log linear model can. When comparing the p-values of the linear and log linear regression models, both models have an overall p-value of 2.2e-16 so there is a low probability that the results came from a random distribution. Consequently, both the linear and log linear models are statistically significant and are a good fit for the data.

Next, we performed an F-test to decide whether a linear model or a log linear model is more appropriate to represent the data. The null hypothesis is that a linear model best represents the data while the alternative hypothesis is that the log linear model best represents the data. The p-value estimated for the F-statistic is 2.2e-16 which is less than an alpha equal to 0.01. We can conclude that because the p-value is less than the alpha value, we can reject the null hypothesis that the population is better modeled linearly at the 1% significance level and assume that the log linear model fits the data best.

When discussing economic policy in relation to the results found above, economists will find that GDP per person impacts happiness more than the level of generosity of a person. Therefore, policy by governments should help promote financial welfare of their citizens to maximize the people’s happiness. Governments must find ways to increase jobs and encourage trade which promotes GDP for the country and increases GDP per capita so the citizens can lead a happy life.

Lastly, we estimated a regression with two continuous interacting variables. The variables chosen were Generosity and GDP.per.capita in the model equation *Scorei = β0 + β1Generosityi + β2GDP.per.capita + β3(Generosityi \* GDP.per.capitai)+ μi* . Since the variables Generosity and GDP.per.capita have continuous values, we would conduct a regression for interacting variables between two continuous variables as opposed to one or both variables being binary. The results to the model equation above are depicted in Figure 4.

Figure 4

Interacting Continuous Variables

|  | Estimate | Std. Errors | t-value | p-value |
| --- | --- | --- | --- | --- |
| Intercept | 3.86307 | 0.39158 | 9.8653 | <2.2e-16 |
| Generosity | -2.04919 | 1.78476 | -1.1482 | 0.252073 |
| GDP.per.capita | 1.43309 | 0.36697 | 3.9052 | 0.0001412 |
| Generosity\*GDP.per.capita | 3.80776 | 1.64351 | 2.3168 | 0.0218472 |

| Multiple R-squared | 0.659 |
| --- | --- |
| Adjusted R-squared | 0.6523 |
| P-value | <2.2e-16 |

By conducting a regression with the interacting variables Generosity\*GDP.per.capita the marginal effect of Generosity on Happiness Score will not be constant and will be a function of the GDP.per.capita. The interaction of the terms indicate that the effect of Generosity on Happiness Score is different for different values of GDP.per.capita. The coefficient of the interaction term being positive indicates that the higher the GDP.per.capita , the greater the effect Generosity has on Happiness Score. Since the p-value of the whole regression is extremely low, we can infer that the results from the regression are statistically significant and not the product of a random distribution. Additionally, the high multiple R-squared value specifies that the regression can explain 65.9% of the data. This value is likely the result of including many variables in the regression. Therefore, an F-test is needed to decide whether a linear regression model or a regression with interacting variables is more appropriate. When comparing the multiple R-squared value above with the value from the linear model with equation *Scorei = β0 + β1Generosityi + β2GDP.per.capita + μi* from project one, the linear model has a multiple R-squared value of 0.6497 that is slightly lower than the model with interacting variables, so the linear model cannot explain as much of the data than the nonlinear regression. Also, the p-values of the overall regression models for both linear and nonlinear are the same and thus, both models are statistically significant.

The F-test conducted tested the null hypothesis that the coefficient of I(Generosity \* GDP.per.capita) equals 0 versus the alternative hypothesis that the coefficient of I(Generosity \* GDP.per.capita) does not equal 0. Since the p-value for the F-statistic is 0.02185 which is slightly greater than 0.01, we determined that we would accept the null hypothesis that the coefficient of I(Generosity \* GDP.per.capita) equals 0 at the 1% significance level. Therefore, according to the F-test, the linear model would provide a more accurate representation of the data than the regression model with interacting variables.

When taking into consideration the regression model including interacting variables. Economic policy will be influenced since the effect that Generosity has on Happiness Score depends on GDP.per.capita. Economists would want to encourage the growth of a country’s GDP while also increasing people’s level of generosity. Therefore, economists may advise governments to focus on increasing the number of jobs since it will raise GDP and make the citizens wealthy so they will be inclined to be more generous.

While the nonlinear regressions above help to interpret the data in many different ways, there is still a threat to internal validity with omitted variable bias, simultaneity, and measurement error. The above regressions suffer from omitted variable bias because there are an infinite amount of factors that can impact Happiness Score. Happiness is likely influenced by factors such as positive family relations and unemployment level in a particular country. Also, simultaneity occurs because the explanatory variable is determined jointly with the dependent variable. Therefore, it is possible that while Generosity and GDP.per.capita impact Happiness Score, Happiness Score can also impact Generosity and GDP.per.capita. If people are happy, they might work harder at their jobs and raise the GDP while also giving more money to charity. Lastly, the models may suffer from measurement error simply because of random error which occurs in every experiment and system error because of a possible miscalibration of the software. Fortunately, a regression with instrumental variables can help eliminate the bias. We can conduct an instrumental variable estimator with a single regressor, X, and single instrument, Z, by isolating the part of the independent variable that is not correlated with the error term. The instrument must be relevant which means that its correlation with the independent variable is not equal to zero and the instrument must be exogenous where its correlation with the error term is zero. Once these two conditions are met, we will have a valid instrument for our regression. When performing the regression, we would first want to regress X on Z using OLS to obtain the uncorrelated part of X. Next, we would regress Y on X hat where the coefficient of X hat is the two stage least squares estimator. Regressions with instrumental variables can help provide consistent estimates specifically when the explanatory variable is correlated with the error term.

Conclusion:

Overall, nonlinear regressions prove to aid us in finding relationships among variables we wouldn’t have known with a linear regression. In the quadratic regression, we saw that Generosity has a positive effect on Happiness Score, which we knew in the linear model, but Generosity will start to negatively affect Happiness Score once a maximum point is reached, which we didn’t know in a linear model. Similarly, by conducting a log linear regression, we found out that an increase in one unit of GDP.per.capita, will increase Happiness Score by almost 50% while the linear model could not provide information on percentage increases. The regression with interacting variables proved that the marginal effect of Generosity is not constant and will be impacted by GDP.per.capita. These nonlinear regression help to expand our knowledge about factors that affect happiness and could be used to justify changes in economic policy to promote well-being.

References:

Stock, James H. and Watson Mark W. Introduction to Economics. Pearson Addison-Wesley.

2011

Sustainable Development Solutions Network. 2012. “World Happiness Report.”

Kaggle.com. 2012. https://www.kaggle.com/unsdsn/world-happiness.

‌