ECO 321 - Econometrics

Project 2 - Non-Linear Regression with Multiple Regressors

Due Date: 12/11/2020

Group 8

Group Member Names:

Emirhan Akkaya (SBU ID#: 112768575)

Parv Joshi (SBU ID#: 112169570)

Shail Shah (SBU ID#: 112315115)

A. Our Findings From Project-1

In Project One, we found a high correlation (causality) between two independent variables and our dependent variable. The other two independent variables we had used did not correlate with the dependent variable that we used. The two variables that did have causality were GDP per capita and Military Expenditure, and the two variables that did not have causality were Government Savings and Central Government Debt. The final model that we used, using heteroskedasticity robust standard errors, and a level of significance of 1% was:

$$GHS = -3.9607e+02 + 7.0514e-02 \times GDPpc + 9.7850e-09 \times ME$$

This model had a multiple R-squared of 0.7219 (it explains 72.19% of the variance) and an Adjusted R-squared value of 0.7128 (it explains 71.28% of the variance). It has a residual standard error of 1084 (on 122 degrees of freedom).

We faced a limitation when running the regressions for project one because some of our data were not perfectly linear, so our regressions were not wholly accurate. Our group will be running regressions that are not substantially affected by non-linear regressors to address this limitation. For project one, we also did not have any interaction terms. Interaction terms would have given us the ability to multiply two variables to see how they would affect each other. We could not use interaction terms for project one because doing so would have caused our data to be non-linear. We plan to use interaction terms for project two since we can take non-linear data into account.

We plan to use the OLS estimation for project two, as we did in project one, because this estimation helps us minimize possible errors that we may not see in our data. We also plan to use non-linear regressions, including the linear-log and log-linear population regression, the cubic regression, the log-log regression, interactions between non-linear regressors, and finally, the IV regressions. For project two, we also added the log of the histograms because the log gave us a better picture and understanding of the data that we used in our project. The log of the histogram gave us a better picture since the histogram without the log function was approximately normal.

B. Econometric Models we Plan to Use

As we did in Project 1, for this Project as well, we are looking to use Ordinary Least Squares (OLS) as one of our essential econometric models. The OLS estimates the parameters in our regression model and minimizes the sum of the squared residuals. The OLS also has three least squares (LSA) assumptions. These assumptions are E(u|X = x) = 0, which implies that for any given x, the mean of the error given x is equal to zero. The second LSA assumption is that each distribution is independently and identically distributed. This means that each distribution is chosen from the same population or set of data and that the chosen entities are randomly selected from the distribution. Randomness is fundamental in statistical analysis because it allows us to assume no bias in the data we are working with. Finally, the third LSA assumption states that large outliers are very rare. This is important for an OLS fit because an outlier can dramatically affect the digital data set's distribution or population. In our R code, we used the government healthcare as the dependent variable and all other independent variables as the regressors.

Also, for our project, we assume homoscedasticity but used heteroskedasticity because it was a better fit for our project since we had multiple dependent variables. Heteroskedasticity was a better fit for our data because heteroskedasticity takes into account unequal variances, and we had variances that were not equal to each other in our data.

We might use the IV regression model at some point in our project because the IV regression enables us to break our regression into two differing parts. The act of restructuring our model will allow us to consider a diverse set of circumstances that will give us the best result that we see in our regressions. The IV regression model may also take endogenous and exogenous variables into account, which will separate variables that directly affect the dependent variable and variables that do not affect the dependent variable.

C. Explanation of how our Econometric Models are related to Chapter-6

As we mentioned in our introduction, we believe that substantial non-linearities in our data would be better addressed by obtaining a non-linear regression model that may better explain the relationship between our dependent and independent variables. Hence this project is going to include nonlinear functions of single independent variables. Say we have a regression function $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i + \ldots + \beta_r + u_i$, this regression function contains linear terms and is a linear regression model that depicts a linear relationship between the dependent variable (Y) and the independent variables (X_i). However, as we have seen in Chapter 6, we can introduce polynomial terms and logarithmic terms to see if the nonlinear regressors give us a more accurate prediction. We are going to do the same in this project. With the same regression model that we see above, we introduce polynomial terms, $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 \ldots + \beta_r + u_i$, like in this equation where we have introduced squared and cubic values of the independent variables. We can also include logarithmic values of our variables like in, $Y_i = \beta_0 + \beta_1 ln(X_i) + u_i$ or $ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$. Once we run a regression model with an equation that contains polynomial terms of logarithmic terms, we can use the OLS Econometric model as explained in the section above for estimating coefficients, hypothesis testing, and finding a more accurate regression equation that is better able to explain the relationship between the dependent and independent variables. It is important to note that interpreting the OLS model's coefficients might be a little tricky when including logarithmic terms or polynomials. Some manual calculations may be involved later on to get the correct values of the estimated coefficients. In line with our variables, we may see how the square or cubic functions of each of our independent variables, GDPpc, ME, GS, and CGD, affect our dependent variable Government Healthcare Spending (GHS) and if the log of any of the independent variables has an impact on Government Healthcare Spending and vice versa.

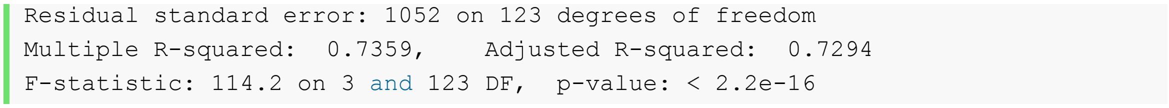
We have also seen in chapter 6 the use and importance of interaction terms and how they may be able to better explain the relationship between the variables in question. We will explore different interaction terms to see if they have a substantial effect on Government Healthcare Spending. Since all of our data variables are continuous, we will only be focusing on the interaction terms of two continuous variables. In our case, we concluded in Project 1 that Central Government Debt and Government Savings do not affect Government Healthcare Spending. The interaction between these two continuous variables may impact Government Healthcare Spending, which we will look to explore in this project, among other things. Say we have a regression function like, $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i$ which has the interaction terms $(X_{1i} \times X_{2i})$ as a regressor. We can again use our OLS econometric model to interpret the coefficients in the most appropriate way to see if the interaction term has a substantial effect on the dependent variable.

D. Summary from Non-Linear Regression Model

Transformations for GDP per capita:

We ran linear, squared, cubic, linear-log, log-linear, and log-log regressions for all of our independent variables. We found that the cubic function fit the best for the GDP per capita independent variable when running our R tests. The tables that we conducted for all of the models that we tested for but did not see best can be found in the appendix at the end of this essay document.

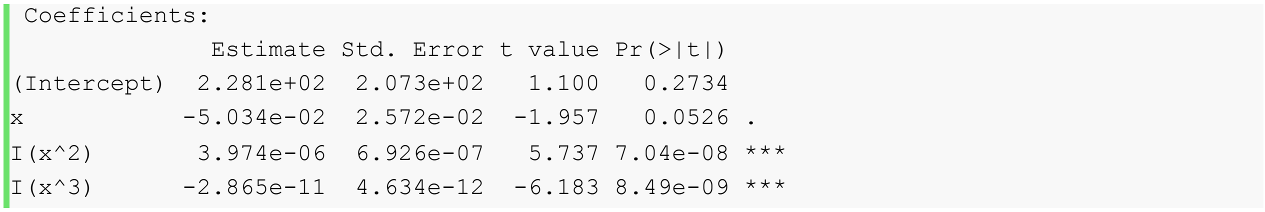
The quadratic model that we decided to use for this variable was x^3. We decided to use this over the other models because it gave the best picture for the GDP per capita independent variable. We assumed homoscedasticity for the cubic function when we ran the model in R. Assuming homoscedasticity meant that the errors have equal variance.



```
Residual standard error: 1052 on 123 degrees of freedom
Multiple R-squared:  0.7359,    Adjusted R-squared:  0.7294
F-statistic: 114.2 on 3 and 123 DF,  p-value: < 2.2e-16
```

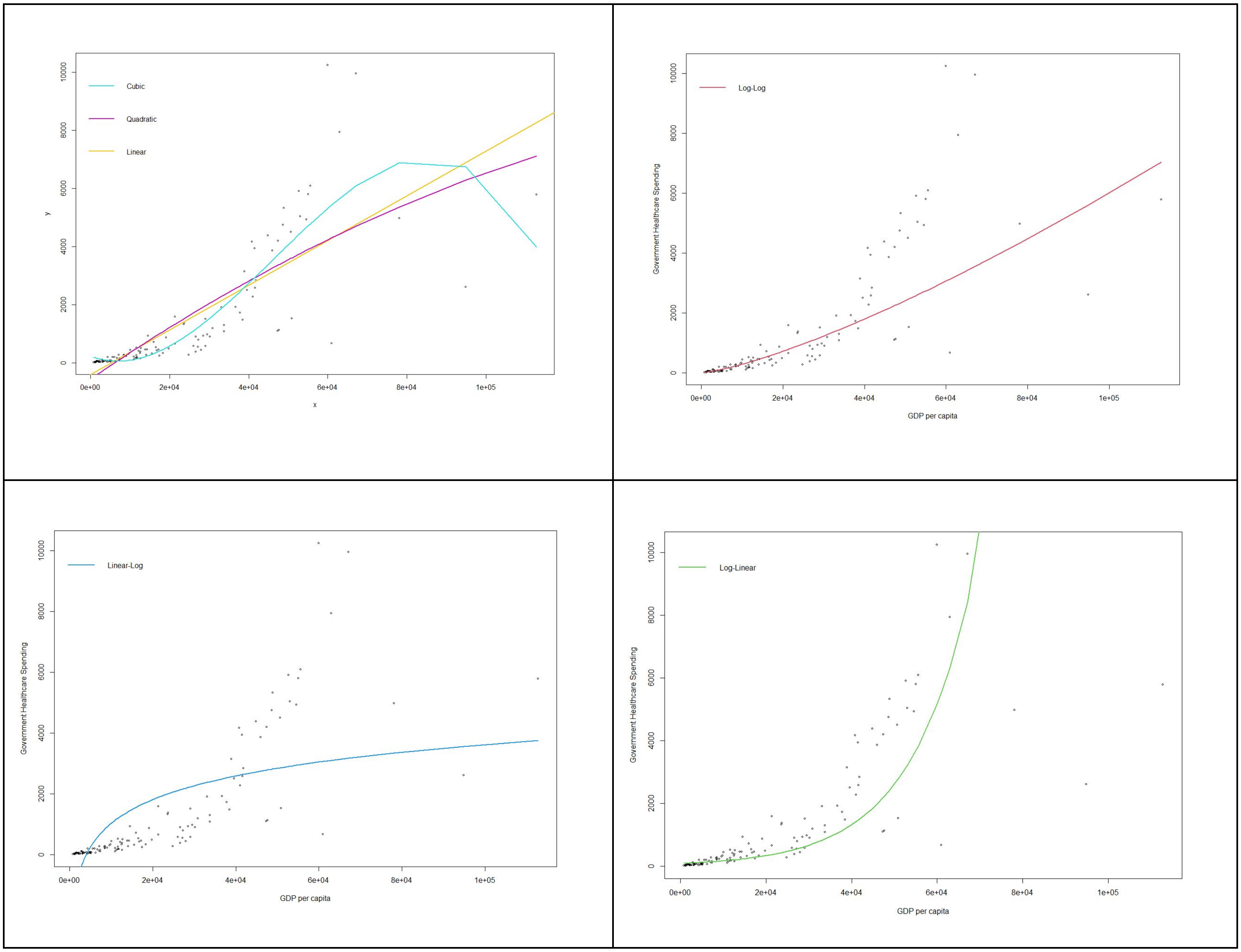
The adjusted R-squared value here is 0.7294, which means that our model can explain 72.94% of the variance in our model. Our model cannot explain 1 - 0.7294 or 0.2706 of the variance in our model.

We also assumed heteroskedasticity for our project. When we presume heteroskedasticity, our error terms will not have the same variance. Thus, our model has a more significant number of possibilities since the error term is not uniform throughout our data. The error terms not being uniform is also a better fit for the model. Here is our output table assuming heteroskedasticity:



```
 Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.281e+02  2.073e+02   1.100   0.2734
x           -5.034e-02  2.572e-02  -1.957   0.0526 .
I(x^2)       3.974e-06  6.926e-07   5.737 7.04e-08 ***
I(x^3)      -2.865e-11  4.634e-12  -6.183 8.49e-09 ***
```

These are the p-value, standard errors, and t-values for the cubic model, all of our other models are defined in the appendix. We then compared our transformations.



The log-log model is the best fit from our log transformations since the graph of this transformation is in between what the log-linear and linear-log graphs present. The slope of our linear-log model is to steep at the beginning of the graph, discounting all of the data-points at the beginning of the graph, and the slope is too small at the end of the graph, which implies that the model does not sufficiently take the points near the top of the graph into account. Our log-linear graph is fine up until the point where it gets past 4e+4 for the x-axis. The graph is sufficient until this point since the lower half of the graph's data points are disregarded because of the quadratic slope the line presents. The log-log model delivers the best mix of both previous graphs.

Considering the polynomial functions.

```
Wald test

Model 1: y ~ x + I(x^2)
Model 2: y ~ x
  Res.Df Df      F Pr(>F)
1    124
2    125 -1 1.0144 0.3158
```

When we compare the linear model with the quadratic model using the waldtest, we get a p-value of 0.3158 > 0.05 (our chosen level of significance). Thus, model 2 (linear model) is better than our quadratic model. Comparing the linear model with the cubic one, we get the following results from the waldtest:

```
Wald test

Model 1: y ~ x + I(x^2) + I(x^3)
Model 2: y ~ x
  Res.Df Df      F  Pr(>F)
1    123
2    125 -2 5.8476 0.00375 **
```

Here, since our p-value is 0.00375 < 0.05 (our chosen level of significance), our model 1 (cubic) is better. This means that now we need to compare the cubic model and the log-log model to find our best transformation for the GDP per capita variable. We can do so by considering the graphs.

Comparing the log-log and cubic curves, we can see that the cubic transformation is better than the log-log transformation. We came to this conclusion because looking at the graph. We can see that the cubic polynomial's slope, after 8e+04, comes down a little, and this is important since there are data-points on the lower half of the graph, which could potentially change the outcome of our conclusion. The log-log model did not fully take this into account. Moreover, for points between 4e+04 and 6e+04, the log-log transformation is too low and ignores most points above it. These are the main reasons to claim that the cubic transformation is better among the polynomials. Here is the output we got when we ran our test for a cubic transformation:

Our group also found the estimates and standard errors for this variable:

Transformations for Military Expenditure:

Our next independent variable is Military Expenditure (ME). This was an important variable in our first project as it was part of our final model, but that was a linear function. Hence, we were interested to see how this variable might change when the aspects of chapter 6 were taken into account. Like we did for GDP per capita, for Military Expenditure, we ran linear, square, cubic, log-linear, linear-log, and log-log regressions to see which was the best fit. The findings for all of these can be found in Section B of the appendix. When running all the regressions, comparing the results using the waldtest method, and closely looking at the graph, we concluded that the cubic model for Military Expenditure is better among the polynomials. Our graphs and reasoning for this are as follows.

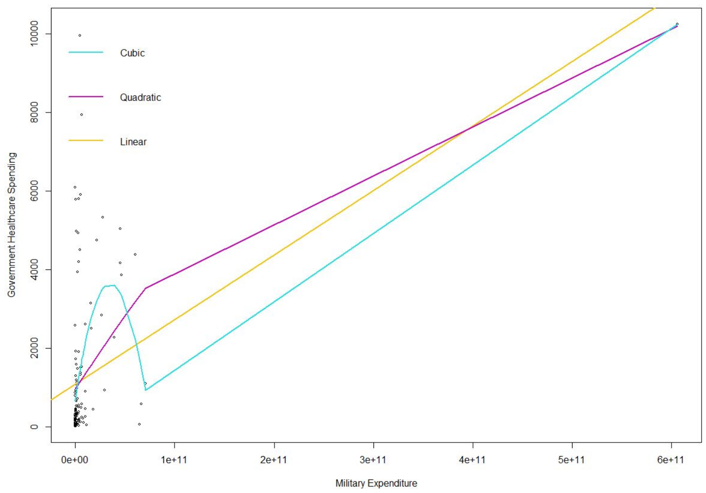
Considering our log models, we immediately received an error while plotting our log-log and linear-log curves. The error said: “Error in lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) : NA/NaN/Inf in 'x'.” This is because we have some countries with zero military expenditure, and the log of zero is not defined. This is why we could not create these regression models for them at all. We see this as our log-log and linear-log curves being non-steady and non-continuous. Hence, the only log model we would consider is our log-linear model, the graph for which can be seen below.

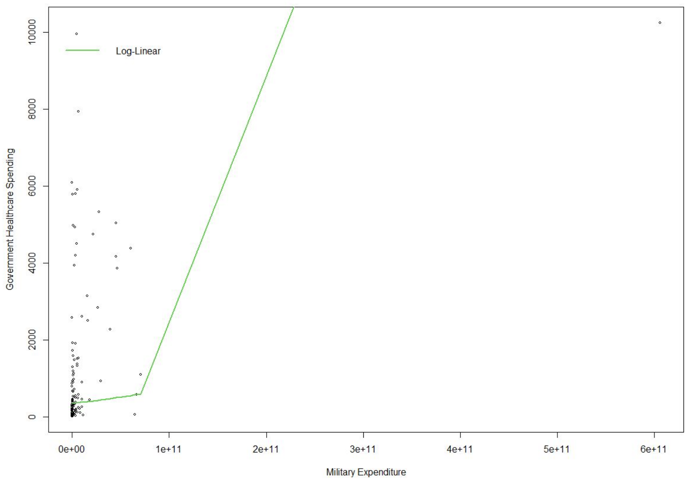
Now considering the polynomial functions.

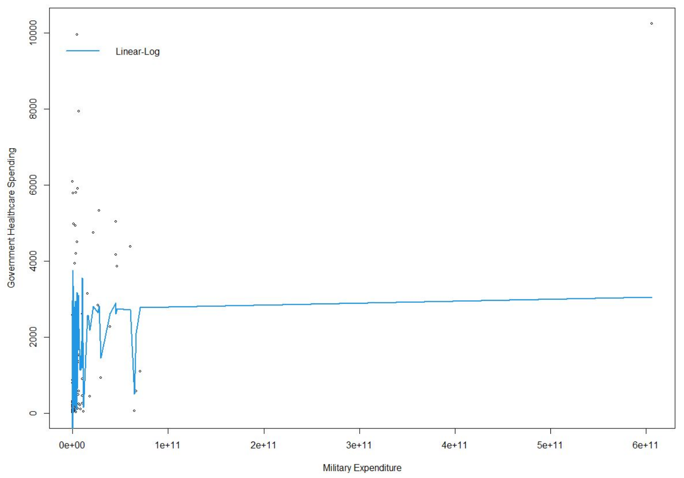
```
Wald test

Model 1: y ~ x + I(x^2)
Model 2: y ~ x
  Res.Df Df      F Pr(>F)
1    124
2    125 -1 2.4227 0.1221
```

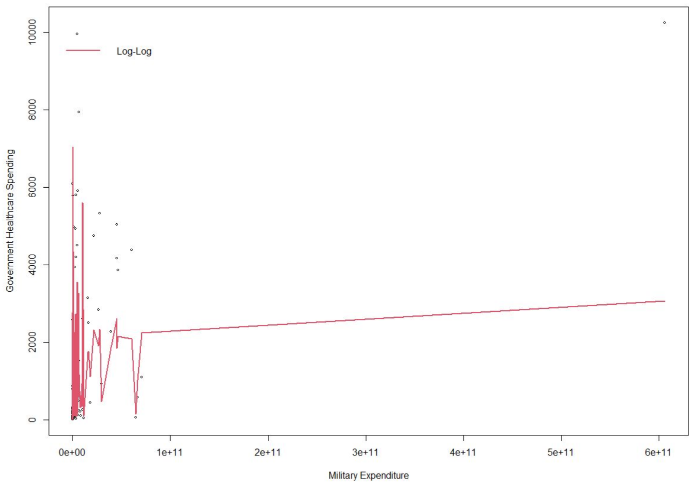
When we compare the linear model with the quadratic model using the waldtest, we get a p-value of 0.1221 > 0.05 (our chosen level of significance). Thus, model 2 (linear model) is better than our quadratic model.

When we compare our linear model to our cubic model using the waldtest, we get an error that reads, “Error: computationally singular system: reciprocal condition number = 1.2889e-27”. This means that the matrix is invertible. Hence we will have to use the graph to find the better one.









When we look at the graphs for the linear model and the cubic model, we see that the cubic model has more points under it and can cover more data values. The graph's initial loop tells us that it accounts for more values and is a better fit than the linear model. Hence the cubic model is the better of all the polynomial models.

We now have to compare the cubic model to the log-linear model. We would make this comparison based on the graphs we have obtained. We see that the log-linear graph curve is pretty low and does not consider many points above. Whereas for the cubic model, we see a significant number of points under the curve and the loop in the curve in a way compensates for the date points above it. Hence for Military Expenditure, we think that the cubic model is best.

Our values for the estimated coefficients, adjusted R-squared, standard errors, t-test coefficients, etc. for the cubic model can be found below (similar values for the other models of ME can be found in section B of the appendix):

Assuming homoscedasticity:

```
Residual standard error: 1711 on 123 degrees of freedom
Multiple R-squared:  0.3007,    Adjusted R-squared:  0.2836
F-statistic: 17.63 on 3 and 123 DF,  p-value: 1.393e-09
```

Assuming heteroskedasticity:

```
t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)  6.6193e+02  1.5556e+02  4.2552 4.100e-05 ***
x            1.7570e-07  3.5793e-08  4.9088 2.846e-06 ***
I(x^2)      -2.7270e-18  6.5651e-19 -4.1537 6.069e-05 ***
I(x^3)       4.0646e-30  9.8923e-31  4.1088 7.204e-05 ***
```

<u>Transformations for Government Savings:</u>

We also ran tests for all of the possible functions that we were planning to use for the Government Savings independent variable. Our results when assuming and not assuming homoscedasticity are given in part C of the appendix found at the end of this essay. For this variable, we found that the cubic model was also the best for transforming the data.

We decided to use the cubic model over the other models because it gave us the best picture for the Government Savings independent variable. We assumed homoscedasticity for the cubic function when we ran the model in R, and these are the results that we obtained when running the first test:

```
Residual standard error: 1800 on 123 degrees of freedom
Multiple R-squared:  0.2258,    Adjusted R-squared:  0.2069 
F-statistic: 11.96 on 3 and 123 DF,  p-value: 6.335e-07
```

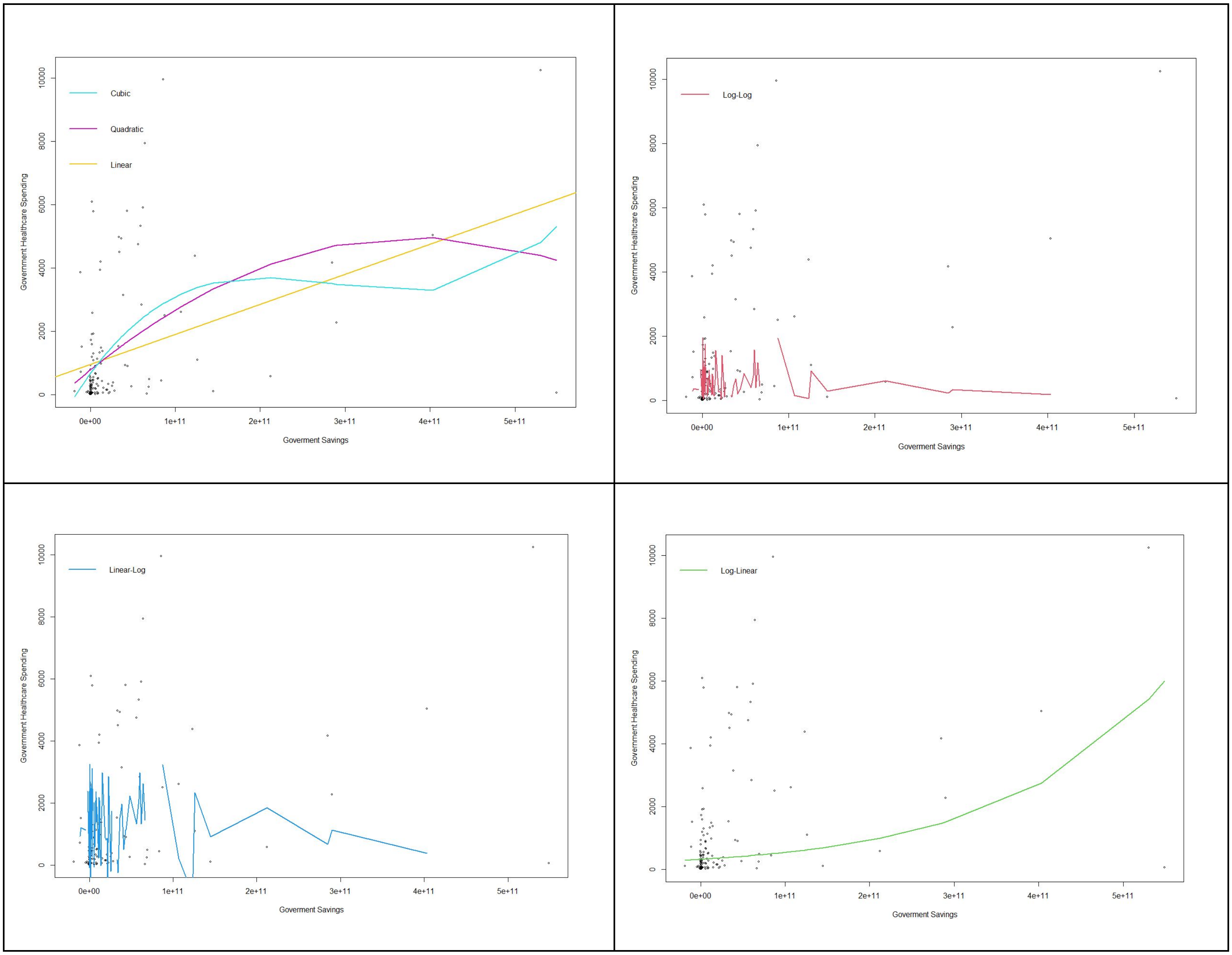
The adjusted R squared value here is .2069 which, as said before, implies that our model can predict 20.69% of the data. Our group also assumed heteroskedasticity for this model, and here is out output when we assume that the error terms are not uniform:

```
t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)  6.6497e+02  1.4005e+02  4.7481 5.608e-06 ***
x            3.6565e-08  1.3800e-08  2.6496  0.009117 ** 
I(x^2)      -1.3882e-19  8.2552e-20 -1.6816  0.095189 .  
I(x^3)       1.5964e-31  1.2429e-31  1.2844  0.201405    
```

These are the values of the p-value, standard errors, and t-statistics for the cubic model for the cubic model. Our group concluded that the cubic model was the best transformation for our data for various reasons. Firstly, as we did for the GDP per capita independent variable, we ran a waldtest for the models here.

```
Model 1: y ~ x + I(x^2)
Model 2: y ~ x
  Res.Df Df      F Pr(>F)
1    124
2    125 -1 0.1154 0.7346
```

The first wald test that we conducted was between the linear and quadratic transformations. We concluded that the linear model was a better transformation between these two models since it had a p-value of 0.7346. When we tried to compare linear and cubic transformations, an error message was thrown at us saying: “Error in solve.default(vc[ovar, ovar]) : system is computationally singular: reciprocal condition number = 1.74503e-25.” It implies that our data for both models form a variance-covariance matrix with linearly dependent columns, and hence our variance-covariance matrix is not invertible. This is why we could not compare them using the waldtest and hence tried to compare them graphically.



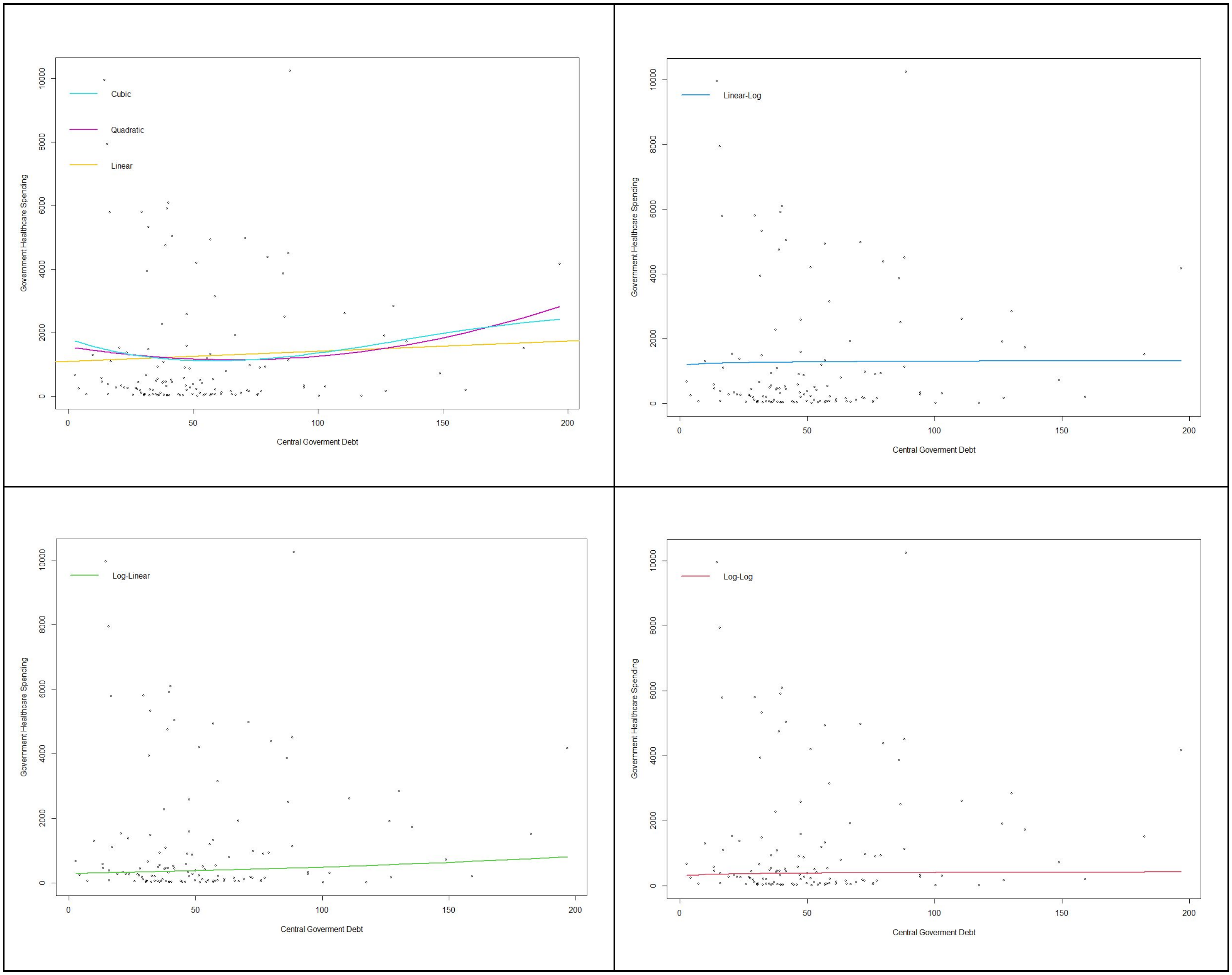
This time also we received an error while plotting our log-log and linear-log functions saying "Warning message: In log(x): NaNs produced." This error arose since some of the countries we interpreted had negative government savings values (those countries had budget deficits in 2017). Log of a negative value is not defined, which caused a problem while running our tests. The log-linear line showed no error, but again it was not a steady curve. This time, this curve was not even continuous since it has brakes in the middle. So at this point, we had concluded that the log models were not useful for this variable. Hence we only considered the log-linear graph for further analysis.

After all of these tests were conducted to find the best transformation among polynomials, we had to compare the linear and cubic transformations graphically. We found that the cubic model was a lot more accurate near the graph's extremes, especially before 1e+11. Moreover, the linear line just kept going up for x above 1e+11, not painting an accurate picture of how the data should be transformed since extreme points are better explained by the cubic curve.

Finally, we can conclude that the cubic model was the best transformation for the government savings independent variable. Comparing the log-linear and cubic transformations, the better model that should be used here is the cubic model since the log-linear one is too low for x below 1e+11.

Transformations for Central Government Debt:

Our final independent variable is the Central Government Debt (CGD). In Project 1, we concluded that CGD was not significant, and we did not include it in our final model. For this variable, we ran regressions against the polynomial and log functions, as we did for the other three variables, and the results for all of those can be found in Section D of the Appendix. For Central Government Debt, we concluded that the Linear-Log Model is the better transformation, i.e., we could conclude for the following reasons.



Firstly, we compare our polynomial models, i.e., the linear, quadratic, and cubic models.

We used the waldtest to make the comparisons.

Results for comparing the linear and quadratic models using the waldtest, we got:

```
Wald test

Model 1: y ~ x + I(x^2)
Model 2: y ~ x
  Res.Df Df      F Pr(>F)
1    124
2    125 -1 0.0432 0.8358
```

The results we obtain give us a p-value of 0.8358. Since the p-value of 0.8358 > 0.05 (our chosen level of significance), we can say that the linear model is better than the quadratic model. Furthermore, when we compare the linear model with the cubic model using the waldtest, we get the following results.

```
Wald test

Model 1: y ~ x + I(x^2) + I(x^3)
Model 2: y ~ x
  Res.Df Df     F Pr(>F)
1    123
2    125 -2 0.549 0.5789
```

This gives us a p-value of 0.5789. Again, since the p-value of 0.5789 > 0.05 (our chosen level of significance), we can say that our linear model again proves to be better than the cubic model and thus the better of the polynomial models.

Next, to compare all the log models, we are going to look at the graphs. We concluded that the graph that was the best fit was the Linear-log graph. This graph curve is much higher, with many data points under it and relatively fewer data points above it. The Log-Linear and Log-Log graphs seem to be discounting many data points, deeming them not the best fit. Comparing the linear model with the linear-log model, we would again look at the graphs for both those models. Both the curves seem to be somewhat horizontal and are relatively flat. However, we also think that the linear-log model is a better fit. The linear model starts low and increases in slope in its incline slightly but only towards the end. The Linear-Log model graph begins at a higher point, and although it slopes down a little bit, it is higher than most date values initially, thus taking into account more data points. It decreases in its slope by a bit that is only towards the end and is relatively close to its endpoint. Hence we think that the linear-log model is the best for the Central Government Debt variable.

Our values for the estimated coefficients, adjusted R-squared, standard errors, t-test coefficients, etc. for the Linear-Log model can be found below (similar values for the other models of CGD can be found in section D of the appendix):

```
Residual standard error: 2030 on 125 degrees of freedom
Multiple R-squared:  0.0001105, Adjusted R-squared:  -0.007889
F-statistic: 0.01382 on 1 and 125 DF,  p-value: 0.9066
```

Here is out output when we assume that the error terms are not uniform:

```
t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1157.601   1162.596  0.9957   0.3213
I(log(x))     31.515    293.791  0.1073   0.9147
```

Results of our Non-Linear model

Compared to Project 1, there has been a significant change to our model in this project. We were able to consider nonlinearities, conducting all the tests, and plotting the graphs showed us that the non-linear components of our independent variables were a much better fit for our model. Furthermore, we noticed that the interaction between our non-linear terms was better, and the interaction terms were significant at the 5% level (our chosen level of significance). Thus using the concepts we learned in Chapter 6, we were able to improve on our model.

The regression model we considered in Project 1 was :

$$GHS \;=\; \beta_0 \;+\; \beta_1 \times GDPpc \;+ \beta_2 \times ME \;+ \beta_3 \times GS + \beta_4 \times CGD$$

The final model we concluded was :

$$GHS \;=\; -3.9607e+02 \;+\; 7.0514e-02 \;\times GDPpc \;+\; 9.7850e-09 \times ME$$

We ran a regression with the transformations we saw as the best fit from the section above for this project. For GDP per capita, Military Expenditure, and Government Savings, the cubic models were the best fit, and for Central Government Debt, the linear-log mode was the best. At the same time, we also included, in the regression, interaction terms between these transformations. Interaction terms helped us see the cause and effect relationship of certain variables we used for our project. Without interaction terms in project one, we really could not grasp the full picture of how one variable affects another. Here are our model results:

Our R-squared, F-statistic, p-values and residual standard error values are:

```
Residual standard error: 673.4 on 105 degrees of freedom
Multiple R-squared:  0.9076,    Adjusted R-squared:  0.8891
F-statistic: 49.09 on 21 and 105 DF,  p-value: < 2.2e-16
```

Here is our output when we assume that the error terms are not uniform:

```
                                       Estimate  Std. Error t value  Pr(>|t|)
(Intercept)                         -3.8947e+02  5.0476e+02 -0.7716 0.4420866
GDPpc                               -1.2882e-02  2.7975e-02 -0.4605 0.6461146
I(GDPpc^2)                           2.1485e-06  1.0862e-06  1.9779 0.0505574 .
I(GDPpc^3)                          -2.0089e-11  8.5587e-12 -2.3472 0.0207895 *
ME                                  -1.6853e-08  4.3371e-08 -0.3886 0.6983794
I(ME^2)                              2.5970e-18  2.8787e-18  0.9022 0.3690362
I(ME^3)                              2.4700e-28  3.9822e-28  0.6203 0.5364342
GS                                  -1.7582e-09  8.6480e-09 -0.2033 0.8392886
I(GS^2)                              2.1267e-19  1.4812e-19  1.4358 0.1540183
I(GS^3)                             -3.8393e-30  3.2176e-30 -1.1932 0.2354641
I(log(CGD))                          1.2867e+02  1.1297e+02  1.1390 0.2572911
I(GDPpc^3 * ME^3)                   -2.3230e-42  3.6449e-42 -0.6373 0.5253033
I(GDPpc^3 * GS^3)                    8.5049e-44  2.0649e-44  4.1188 7.612e-05 ***
I(GDPpc^3 * log(CGD))                2.0892e-12  3.0979e-12  0.6744 0.5015492
I(ME^3 * GS^3)                      -2.0953e-62  6.1090e-62 -0.3430 0.7322933
I(ME^3 * log(CGD))                  -7.3409e-29  9.9507e-29 -0.7377 0.4623275
I(GS^3 * log(CGD))                   6.3918e-31  8.4787e-31  0.7539 0.4526183
I(GDPpc^3 * ME^3 * GS^3)            -2.5060e-76  3.9680e-76 -0.6316 0.5290511
I(GDPpc^3 * ME^3 * log(CGD))         6.1234e-43  8.5259e-43  0.7182 0.4742251
I(GDPpc^3 * GS^3 * log(CGD))        -1.9819e-44  5.1430e-45 -3.8536 0.0002007 ***
I(GS^3 * ME^3 * log(CGD))            6.5377e-63  1.6167e-62  0.4044 0.6867446
I(GDPpc^3 * GS^3 * ME^3 * log(CGD))  4.7108e-77  9.9316e-77  0.4743 0.6362539
```

Based on the results, there were only three terms that we found were significant at the 5% level of significance. Those are $GDPpc^3,\ GDPpc^3 \times GS^3,\ and\ GDPpc^3 \times GS^3 \times log(CGD)$.

The goodness of fit and how our results have changed compared to the previous model:

To compare the goodness of fit between our linear regressions and our non-linear regressions, we decided that it would be best to compare our adjusted R-squared values and standard error values we found in R. For our linear model (from project 1), we discovered that our adjusted R-squared value here was 0.7128, while for the non-linear model (project 2 model) it was 0.8891. A 71.28 percentage for our adjusted R-squared value shows us that only 71.28% of the variance in our model was explained, while a 72.3% shows that 88.9% of the variance was accounted for in our models. It is a very significant increase. For our standard errors, we also had slightly better data here as well. In our first linear regression model, we analyzed that our model had a 1084 residual standard error (with 122 degrees of freedom). For our non-linear regression model, our standard error had decreased to 673.4, with 105 degrees of freedom. This shows that our model's standard error decreased and consequently improves our second model's goodness of fit. The second model's goodness of fit was better because of a higher R-squared value and lower standard error.

We also used the F-test to compare our data points compared to the regression model and our hypotheses. Our linear model (from project 1) found that our F-statistic was 230.2, while for our non-linear model, it was 49.09. The F-statistic is lower for the non-linear model. Still, it did not show a clear and precise picture of which model would be poised as the better one for us, so we decided to compare the p-values since comparing using F-test and p-values is equivalent. The p-values that we obtained for both models were the same, which was less than $2.2\,e-16$. We decided to use the alpha level of 5%, and our p-value was much smaller than the alpha level. Based on the value we obtained from our p-value, we concluded that both models were good enough at a 5% level of significance.

Implications of our results on Economic Policy:

To find the implications of our economic policy results, we need to understand the real-world economic meaning behind each of our model's significant terms. The explanation for those is as follows:

$GDPpc^3$:

Our first term $GDPpc^3$ had a significance level of 0.0207, which is less than our 5% alpha. $GDPpc^3$ is correlated with Government Healthcare Spending because when the GDPpc increases, government healthcare spending also increases. After a point, as the people get wealthier and live better lifestyles, their spending on healthcare can decrease as they are not as prone to illnesses and have better access to healthier nutrition and water. In this period, healthcare spending can reduce, but later on, healthcare spending can increase again as people tend to get older and sicker and suffer from illnesses. Furthermore, as people get richer, they tend to spend more on elective surgeries, which increases in GDPpc increases healthcare spending again. This behavior of the GDPpc depicts a cubic model and explains why $GDPpc^3$ is an essential term for our model.

$(GS)^3 : (GDPpc)^3$:

In our final tests that we ran in R, we found that the interaction between the cubic functions of GDP per capita and government savings is significant. We know that these two interactions are significant because the p-value we obtained in our test was less than the critical alpha level of 5%. Government Savings can be correlated with the government healthcare spending variable that we have since the more a country saves, the greater probability there is that the government would utilize towards healthcare. Also, GDP per capita is correlated with government healthcare spending because a higher GDP per capita implies that a country's residents are earning more annual income, which means more savings for the government since a country can more heavily tax its residents. A wealthier government may spend less on its healthcare system and focus on other things like defense and infrastructure. Finally, GDP per capita and government savings are also related to each other. As mentioned before, a higher GDP per capita implies that the government has more access to money. The cubic interaction between government spending and GDP per capita suggests that as GDP per capita increases, the government savings would strongly follow its path. Still, later on, there would be a point where there is a negative slope between their interactions. After the graph point where there was a negative relationship between GDP per capita and government spending, the graph would eventually have a positive correlation and continue to slope upwards.

$(GS)^3 : log(CGD) : (GDPpc)^3$:

After conducting our final test for our project, we found that the third and final significant interaction was between three variables. The three variables were the cubic transformation of GDP per capita, the cubic transformation of government savings, and the log transformation of central government debt. We arrived at this non-linear model since the p-value obtained from this test was less than the alpha level of .05. The p-value for this interaction term was .0002. GDP per capita may be correlated with government healthcare spending since if a country's population is making more money, they can spend more on their healthcare systems. However, we predict that the government decreases its total spending on the healthcare system, even though it has more money, most probably because since a more significant percentage of the population is more wealthy, people generally would be more healthy.

Since this is a cubic function, though, there is a point in the correlation between these two variables where it starts to decrease, and this may be because since more people are making more money, everyone is healthier. However, after another point in the graph, the line's slope on the graph increases again. We assume that it also increases because now people are making even more money than before. We believe that they work longer hours, and because of this, more people are stressed, which may lead some to get to the hospital, increasing government healthcare spending once again. For the log of central government debt and government healthcare spending, we could imply that the greater government debt a country has, the less they would spend on healthcare since they do not have as many resources available. For instance, with a 1% increase in the total central government debt a country has, we would expect a decrease of one unit in the total healthcare spending a country puts forth. Finally, for our cubic transformation of the government savings, this implies that as government savings goes up, the total healthcare spending of a nation goes up, until a point on the graph where it comes down, most likely due to a mini 'golden-age.'

After this mini "golden-age" is over, healthcare spending and government savings continue in the same direction. Taking the interaction term into account, including all of the variables, we found that these three terms have an inverse relationship with government spending on healthcare. This is most likely because of all of the different possible situations possible in a country and how much demand there is for healthcare in a given year. For instance, if the world were to be going through a pandemic, healthcare demand would most probably be very high since everyone is concerned for their health, while in a typical year, healthcare demand may not be so high.

The overall effect on economic policy:

Since none of our significant terms included the Military Expenditure (ME) variable, we can assume that we do not require incorporating the Military Expenditure variable in our final model.

Since we know that Military Expenditure is insignificant, we know that change in Military Expenditure will not affect economic policies relating to government healthcare spending. We found that some of our estimates were negative from the results obtained after running all of our tests, implying that these variables increased as government healthcare spending decreased. We also found that some of our other variables' estimation values were positively correlated, indicating that government healthcare spending variable increases as these variables increase. Economists could then conclude that the positive estimations have a greater impact when trying to estimate which variables affect government healthcare spending the most and which affect healthcare spending the least. To name a few variables that had negative estimation values, they were GDP per capita, Government savings, the cubic function of GDP per capita, and so on. Some positive estimation values were the log of central government debt, the squared function of military expenditure, the cubic function of military expenditure interacted with the log of central government debt, etc. These key and important values show Economists how and where to allocate useful resources and with what costs.

E. <u>How our Econometric models are related to Chapter 7</u>

Our econometric models are related to the contents of chapter seven in many ways. Chapter seven talks about the conditions for instrumental variables and how it helps to avoid biases. Our model does suffer from endogeneity bias. The reasons for the bias are mainly omitted variables and simultaneous causality among our dependent and independent variables. Regarding the omitted variable bias, our model did not include all of the possible variables that may affect government healthcare spending. For instance, government spending on education may affect GDP per capita (one of your independent variables), affecting government healthcare spending.

We did not take into account any variable relating to education in our project. Government spending on education is a dependent variable to GDP per capita since a country's wealth may determine how much money they allocate to education. Other omitted variables include population size, population demographics, number of medical institutions/facilities, cost of medical goods, and medicines. All of these affected government healthcare spending and were not part of our model. As much as we would like to include them, their data is not as readily available.

At the same time, we do think that our model suffers from simultaneous causality bias as well. Simultaneity essentially means when our independent variable causes our dependent variable to change, and at the same time, our dependent variable can cause our independent variable to change. In our model, we believe that the independent variable Government Savings can cause this bias. A decrease or increase in Government Saving can affect how much money the government can spend on Healthcare. Simultaneously, when the government is forced to spend more on healthcare, it can affect their savings. A prime example is the Covid-19 pandemic. An increase in healthcare spending is taking away from the government's savings, causing it to decrease.

We believe that our measurements do not contain significant errors-in-variables bias when we talk about the measurement error. This is because all of our data was collected from the World Bank and International Monetary Fund databases, and their data is considered the most reliable for economic variables. Hence, we assume that our information is consistent and accurate. Therefore, we do not think that this error is present in our model.

After knowing that we have endogeneity bias, we wanted to see how to resolve it. Our best option was using instrumental variable regression. The conditions for a valid instrument are:

(i) Instrument Relevance (i.e., the correlation between the independent variable and instrumental variable should be significant), and

(ii) Instrument Exogeneity (i.e., the instrumental variable should not be correlated with the error term)

Moreover, there is another condition that the instrumental variable should not directly affect our GHS variable significantly. If it does, it would be an independent variable and not an instrumental variable.

For having an instrumental variable, we had two choices. One is to check if any of our variables included work as an instrument or not. The other was to find such a variable that satisfies the conditions mentioned above. However, we did not go for the second option since finding such a variable is difficult. Even if we find it, we cannot guarantee if we will find the data for all the 127 countries we considered. Hence, finding and including a new variable would reduce our number of observations. Since 127 is already low, a lower number of observations would make our model too specific for our considered countries and not a general model. Hene, we went for the first option.

For our final model, the one variable that we did not include was Military Expenditure (ME). This is because none of the ME terms were significant at the 5% level (our chosen level of significance). The inclusion of other variables means that they directly affect our dependent variable and hence cannot be used as an instrumental variable. Instead, they must be used as independent variables. Now, ME was one variable that we did have the data available and could be considered to see if it satisfies the conditions for an Instrumental Variable (IV). The first condition for an IV is Instrument Relevance. This sees if our IV is correlated to any independent variables ($corr\,(Z_i, X_i) \neq 0$). It is intuitive that Military Expenditure is related to GDP, Government Savings, and Central Government Debt and hence satisfies the first condition. This is because the more the GDP per capita of a country, the more they can spend on Military Expenditure. The more they spend on Military Expenditure, the less they will be their savings. Even more the central government debt of a country, less they will be able to accommodate towards Military Expenditure. Therefore, ME satisfies the first condition. This can be proved using the following first stage regressions:

First Stage Regression (using heteroskedasticity robust standard errors) for GDP per capita on Military Expenditure:

```
t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 2.0876e+04 1.8760e+03 11.1278 < 2.2e-16 ***
ME          8.7160e-08 2.3152e-08  3.7647 0.0002554 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

First Stage Regression (using heteroskedasticity robust standard errors) for Government Savings on Military Expenditure:

```
t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 2.2329e+10 5.1046e+09  4.3742 2.542e-05 ***
ME          1.0419e+00 2.1001e-01  4.9614 2.235e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

First Stage Regression (using heteroskedasticity robust standard errors) for Government Savings on Military Expenditure:

```
t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 5.3313e+01 3.0564e+00 17.4430 < 2.2e-16 ***
ME          6.6718e-11 2.0050e-11  3.3276  0.001151 ** 
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The second condition for an IV is Instrument Exogeneity. This means that our IV should not be related to our error term ($corr\,(Z_i, u_i) \;= 0$). The error term captures any of the omitted variables in our model that affect our dependent variable, Government Healthcare Spending. However, ME fails to satisfy this condition. This is because the error term captures variables like population size and corruption index directly correlate to Healthcare Spending. They also affect Military Expenditure because a country's population determines how much a country spends on its military and how much funds and manpower they have. Corruption also affects ME because government officials can take bribes from weapons manufacturers to direct taxpayers to their companies and earn government contracts. In a way, this also redirects money away from healthcare spending. Thus, the Military Expenditure does not satisfy this condition and hence is not a good instrumental variable for our model.

Since we do not have any instrumental variable, let us explain the process of overcoming our endogeneity bias if we had an instrumental variable. Let us call the instrumental variable 'IV' for our purposes. We could overcome the bias by using Two-Stage Least Squares (TSLS or 2SLS). The first thing we do here is isolate the independent variable from the original OLS model (assuming IV is correlated to that independent variable, say 'X,' not correlated to the error term, and does not directly affect the dependent variable). Performing a regression of X on IV will give us a first-stage regression that estimated IV's dependence on X. We can then use the predicted values of this regression as the independent variable in our second stage.

Our second stage of regression would include the predicted estimator of the regression ($\widehat{X}$) as the independent variable. We know that $\widehat{X}$ will be uncorrelated with error and hence will make the second stage valid as an OLS regression. This can be done in R using the ‘ivreg’ function, which is a part of the ‘AER.’ This is how the syntax would look in that case:

```
ivreg_model = ivreg(Y ~ X | IV, data = data)
summary(ivreg_model, vcov = sandwich, df = Inf)
```

Conclusion

Using non-linear transformations and interaction terms, we found our final model to be:

$$\boxed{GHS = (-3.8947e+02) + (-2.0089e-11) \times GDPpc^3 + (8.5049e-44) \times GDPpc^3 \times GS^3 + (-1.9819e-44) \times GDPpc^3 \times GS^3 \times log(CGD)}$$

It has a multiple R-squared of 0.9076 (90.76% of variance explained) and an adjusted R-squared of 0.8891(88.91% of variance explained). The residual standard error: 673.4 on 105 df. This is a lot better than our linear model in the first project. This could have been improved if we could find enough data on a valid instrument.

Appendix

A. GDPpc (GDP per capita)

A.1 : Linear Model:

Assuming homoscedasticity:

```
Residual standard error: 1204 on 125 degrees of freedom
Multiple R-squared:  0.6481,    Adjusted R-squared:  0.6452
F-statistic: 230.2 on 1 and 125 DF,  p-value: < 2.2e-16
```

Assuming heteroskedasticity:

```
t test of coefficients:

                Estimate  Std. Error t value  Pr(>|t|)
 (Intercept) -401.032086  141.658804 -2.8310   0.00541 **
 x              0.076794    0.010337  7.4293 1.497e-11 ***
```

A.2 : Square Model

Assuming homoscedasticity:

```
Residual standard error: 1199 on 124 degrees of freedom
Multiple R-squared:  0.6538,    Adjusted R-squared:  0.6482
F-statistic: 117.1 on 2 and 124 DF,  p-value: < 2.2e-16
```

Assuming heteroskedasticity:

```
t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept) -5.5517e+02  1.2335e+02 -4.5008 1.539e-05 ***
x            9.3315e-02  1.5074e-02  6.1903 8.061e-09 ***
I(x^2)      -2.2410e-07  2.2250e-07 -1.0072    0.3158
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

A.3 : Linear-Log Model

Assuming homoscedasticity:

```
Residual standard error: 1521 on 125 degrees of freedom
```

```
Multiple R-squared:  0.4383,    Adjusted R-squared:  0.4338
F-statistic: 97.53 on 1 and 125 DF,  p-value: < 2.2e-16
```

Assuming heteroskedasticity:

```
t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) -9298.95    1250.70 -7.4350 1.453e-11 ***
I(log(x))    1122.25     142.24  7.8897 1.307e-12 ***
```

A.4 : Log-Linear Model

Assuming homoscedasticity

```
Residual standard error: 0.8364 on 125 degrees of freedom
Multiple R-squared:  0.7503,    Adjusted R-squared:  0.7483
F-statistic: 375.6 on 1 and 125 DF,  p-value: < 2.2e-16
```

Assuming heteroskedasticity:

```
t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) 4.4622e+00 1.4984e-01  29.781 < 2.2e-16 ***
x           6.8142e-05 7.5773e-06   8.993 3.216e-15 ***
```

A.5 : Log-Log Model

Assuming homoscedasticity:

```
Residual standard error: 0.564 on 125 degrees of freedom
Multiple R-squared:  0.8865,    Adjusted R-squared:  0.8856
F-statistic:   976 on 1 and 125 DF,  p-value: < 2.2e-16
```

Assuming heteroskedasticity:

```
t test of coefficients:

             Estimate Std. Error t value  Pr(>|t|)
(Intercept) -6.452556   0.430241 -14.998 < 2.2e-16 ***
I(log(x))    1.316201   0.046356  28.393 < 2.2e-16 ***
```

B. ME (Military Expenditure)

B.1 : Linear Model

Assuming homoscedasticity:

```
Residual standard error: 1816 on 125 degrees of freedom
Multiple R-squared:  0.1993,    Adjusted R-squared:  0.1929
F-statistic: 31.11 on 1 and 125 DF,  p-value: 1.434e-07
```

Assuming heteroskedasticity:

```
t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) 1.0932e+03 1.6279e+02  6.7156 5.887e-10 ***
x           1.6411e-08 1.4771e-09 11.1102 < 2.2e-16 ***
```

B.2 : Square Model

Assuming homoscedasticity:

```
Residual standard error: 1799 on 124 degrees of freedom
Multiple R-squared:  0.221, Adjusted R-squared:  0.2084
F-statistic: 17.59 on 2 and 124 DF,  p-value: 1.883e-07
```

Assuming heteroskedasticity:

```
t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)
(Intercept)  9.6110e+02  1.6970e+02  5.6634 9.797e-08 ***
x            3.9264e-08  1.5550e-08  2.5250   0.01283 *
I(x^2)      -3.9653e-20  2.5476e-20 -1.5565   0.12214
```

B.3 : Log-Linear

Assuming homoscedasticity:

```
Residual standard error: 1.621 on 125 degrees of freedom
Multiple R-squared:  0.06244,   Adjusted R-squared:  0.05494
F-statistic: 8.325 on 1 and 125 DF,  p-value: 0.004608
```

Assuming heteroskedasticity:

```
t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 5.8664e+00 1.4628e-01 40.1049 < 2.2e-16 ***
x           7.5746e-12 2.0490e-12  3.6967 0.0003255 ***
```

C. GS (Government Savings)

C.1 : Linear Model

Assuming homoscedasticity:

```
Residual standard error: 1854 on 125 degrees of freedom
Multiple R-squared:  0.1658,	Adjusted R-squared:  0.1591 
F-statistic: 24.84 on 1 and 125 DF,  p-value: 2.026e-06
```

Assuming heteroskedasticity:

```
              Estimate Std. Error t value  Pr(>|t|)    
(Intercept) 9.5437e+02 1.6650e+02  5.7320 7.029e-08 ***
x           9.4919e-09 4.1411e-09  2.2921   0.02357 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
```

C.2 : Square Model

Assuming homoscedasticity:

```
Residual standard error: 1820 on 124 degrees of freedom
Multiple R-squared:  0.2029,	Adjusted R-squared:   0.19 
F-statistic: 15.78 on 2 and 124 DF,  p-value: 7.841e-07
```

Assuming heteroskedasticity:

```
t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)    
(Intercept)  7.8106e+02  1.3880e+02  5.6274 1.157e-07 ***
x            2.1697e-08  8.3364e-09  2.6027   0.01038 *  
I(x^2)      -2.8048e-20  2.3057e-20 -1.2165   0.22612    
```

C.3 : Log-Linear

Assuming homoscedasticity:

```
Residual standard error: 1.608 on 125 degrees of freedom
Multiple R-squared:  0.07703,   Adjusted R-squared:  0.06965
F-statistic: 10.43 on 1 and 125 DF,  p-value: 0.001582
```

Assuming heteroskedasticity:

```
t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.7699e+00 1.5461e-01 37.3200  < 2e-16 ***
x           5.3356e-12 2.5762e-12  2.0711  0.04041 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

D. CGD (Central Government Debt)

D.1 : Linear Model

Assuming homoscedasticity:

```
Residual standard error: 2027 on 125 degrees of freedom
Multiple R-squared:  0.003006,  Adjusted R-squared:  -0.00497
F-statistic: 0.3769 on 1 and 125 DF,  p-value: 0.5404
```

Assuming heteroskedasticity:

```
t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1103.5690   357.8942  3.0835 0.002518 **
x              3.2108     5.2285  0.6141 0.540272
```

D.2 : Square Model

Assuming homoscedasticity:

```
Residual standard error: 2026 on 124 degrees of freedom
Multiple R-squared:  0.01179,   Adjusted R-squared:  -0.004147
F-statistic: 0.7398 on 2 and 124 DF,  p-value: 0.4793
```

Assuming heteroskedasticity:

```
t test of coefficients:

               Estimate  Std. Error t value Pr(>|t|)  
(Intercept) 1568.939752  689.982558  2.2739  0.02469 *
x            -12.796957   18.646944 -0.6863  0.49382  
I(x^2)         0.097446    0.093362  1.0437  0.29863  
```

D.3 : Cubic Model

Assuming homoscedasticity:

```
Residual standard error: 2033 on 123 degrees of freedom
Multiple R-squared:  0.01316,   Adjusted R-squared:  -0.01091 
F-statistic: 0.5467 on 3 and 123 DF,  p-value: 0.6513
```

Assuming heteroskedasticity:

```
t test of coefficients:

               Estimate  Std. Error t value Pr(>|t|)  
(Intercept)  1.8172e+03  9.9824e+02  1.8204  0.07113 .
x           -2.6826e+01  3.9492e+01 -0.6793  0.49824  
I(x^2)       2.9591e-01  4.6909e-01  0.6308  0.52933  
I(x^3)      -7.3172e-04  1.6424e-03 -0.4455  0.65674  
```

D.4 : Log-Linear Model

Assuming homoscedasticity:

```
Residual standard error: 1.664 on 125 degrees of freedom
Multiple R-squared:  0.01191,   Adjusted R-squared:  0.004004 
F-statistic: 1.507 on 1 and 125 DF,  p-value: 0.222
```

Assuming heteroskedasticity:

```
t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 5.6664297  0.2610009 21.7104   <2e-16 ***
x           0.0052702  0.0039444  1.3361   0.1839    
```

D.5 : Log-Log Model

Assuming homoscedasticity:

```
Residual standard error: 1.673 on 125 degrees of freedom
Multiple R-squared:  0.000828,  Adjusted R-squared:  -0.007165
F-statistic: 0.1036 on 1 and 125 DF,  p-value: 0.7481
```

Assuming heteroskedasticity:

```
t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 5.681512   0.767751  7.4002 1.743e-11 ***
I(log(x))   0.071132   0.201593  0.3529    0.7248
```