**Introduction**

Gross domestic product (GDP) is the overall market value of final goods and services produced within a country in a given period. Forecasting the GDP in the future because people need to understand and need to know the health of the economy. Also, we want to understand and see the difference or similarly of independent and dependent variables. We want to know if the independent and dependent variables have relationship. In the data of GPD, it has 10 independent variables such as U.S. population, personal savings rate, average, unemployment rate, inflation rate, housing start, non-farm payrolls, national debt, gold average price, U.S. imports, U.S. Exports, and the political party of the U.S. President (dummy variables). The dummy variables is to take account of qualitative factors in regression because there some things that has categories rather than numbers associated with them. Since the U.S. President is a dummy variable, I introduce a dummy variable which takes the value of 0 if the sampled individual is a Democratic and the value 1 if is a Republican. These variables have relationship with independent variables and they impact the GPD. There are 6 stages of revising the multiple regression models. Stage 1 is evaluation of the first model, stage 2 is evaluation of the multicollinearity and fixing for multicollinearity, stage 3 is add some explanatory variables based on your evaluations, stage 4 is test for individual significance (t test) and drop one variable at a time, and stage 5 is autocorrelation and autoregressive models for time series data. The final stage 6 is special revisions of time series data. This project is to get the highest value of “Adjusted R-Squared” for the best model that satisfies all the requirements of F Test, t Test, and is the simplest model as per the Parsimony criteria.

**Model #1**

In the first stage is to analyze the Multiple Linear Regression Model with all the explanatory variables and evaluates the Regression output for all the parts of the first model. In “Stage 1: Evaluation of the First Model,” the Model #1, R square is 99.90 percent and the model is near 100 percent perfect. R-Square is the most important number of the output and it tells how well the regression line approximates the real data. Therefore, the number of R-square tells how much of the output variable’s variance is explained by the input variable's variance. The equation of R-square:

**= =**

The adjusted R square is explained the accuracy of the regression equation. However, the standard error is 161.0497 which is the overall forecast for the model. The significance F is 5.303207E-57; therefore, is lower than F when F is 4147.5607. Significance of F is the probability that the Regression output could have been obtained by chance. The equation for F = MSR/MSE. For P-value, there are seven independent variable is less than 5%. X3, X7, and X10 is more than 5%. The individual significant are X1, X2, X4, X5, X6, X8, and X9. The equation of Multiple Linear Regression in Model #1 is = -10136.2126 +62.4612 (X1) – 83.2489 (X2) + 23.5805 (X3) – 58.2224 (X4) – 0.4515 (X5) + 108.0037 (X6) – 0.1458 (X7) + 3.6246 (X8) – 2.1721 (X9) – 57.6445 (X10). Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data.

|  |  |  |
| --- | --- | --- |
| SUMMARY OUTPUT (**Model #1**) |  | |
|  |  | |
| *Regression Statistics* | | |
| Multiple R | 0.99951814 | |
| **R Square** | **0.9990** | |
| **Adjusted R Square** | **0.9988** | |
| Standard Error | 161.0497 | |
| Observations | 51 | |
| ANOVA |  |  | |  |  |  |  |
|  | *df* | *SS* | | *MS* | *F* | *Significance F* |  |
| Regression | 10 | 1075752692 | | 107575269.2 | 4147.5607 | 5.03207E-57 |  |
| Residual | 40 | 1037479.885 | | 25936.99711 |  |  |  |
| Total | 50 | 1076790172 | |  |  |  |  |
|  |  |  | |  |  |  |  |
|  | *Coefficients* | *Standard Error* | | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | -10136.2126 | 690.9372 | | -14.6702 | 0.00000000000000001087 | -11532.6487 | -8739.7765 |
| U.S. Population (in million) (X1) | 62.4612 | 3.0349 | | 20.5808 | 0.00000000000000000000006850 | 56.3274 | 68.5950 |
| Personal Savings Rate (X2) | -83.2489 | 31.8647 | | -2.6126 | 0.0126 | -147.6499 | -18.8479 |
| Average unemployment rate (X3) | 23.5805 | 25.0880 | | 0.9399 | 0.3529 | -27.1242 | 74.2853 |
| Inflation Rates (X4) | -58.2224 | 13.0896 | | -4.4480 | 0.00006742 | -84.6775 | -31.7674 |
| Housing Start (in thousands) (X5) | -0.4515 | 0.0994 | | -4.5402 | 0.00005058 | -0.6524 | -0.2505 |
| National Debt (in Trillions) (X6) | 108.0037 | 49.4677 | | 2.1833 | 0.03494 | 8.0257 | 207.9818 |
| Gold (Average Price) (X7) | -0.1458 | 0.2668 | | -0.5464 | 0.5878 | -0.6850 | 0.3935 |
| U.S. Imports (in Billions) (X8) | 3.6246 | 0.3541 | | 10.2359 | 0.0000000000009817 | 2.9089 | 4.3403 |
| U.S. Exports (in Billions) (X9) | -2.1721 | 0.6323 | | -3.4352 | 0.001393 | -3.4501 | -0.8942 |
| The political party of the US President (X10) | -57.6445 | 66.1753 | | -0.8711 | 0.3889 | -191.3897 | 76.1008 |

**Multicollinearity**

Multicollinearity is used to determine several different independent variables in multiple regression model are closely correlated to one another. In “Stage 2: Evaluation of the multicollinearity and fixing for multicollinearity,” I planned to revise the model by dropping one independent variable each time and find which model has the highest R Squared for multicollinearity. In multicollinearity, all the variables have higher than 80 percent in R-Square.

Multicollinearity is the overall P value is very low and all the individual P values are high. We use the independent variable in regression analysis to introduce to any variable being used to predict the value of the dependent variable. Therefore, the model fits the data well, even though none of the X variables have a statistically significant impact on predicting Y (Anderson, p. 693). All the independent variables in multicollinearity are larger than 80 percent.

|  |  |  |  |
| --- | --- | --- | --- |
| **Table of R Square for Multicollinearity** | | | |
| **Variables** | **R Square for Multicollinearity** | **Multcollinearity (More than 80%?)** | **Variance Inflationary Factory** |
| U.S. Population (in million) (X1) | 0.9888 | Yes | 89.5560 |
| Personal Saving Rate (X2) | 0.9989 | Yes | 886.6019 |
| Average unemployment rate (X3) | 0.99902 | Yes | 1015.4628 |
| Inflation Rates (X4) | 0.9986 | Yes | 694.4186 |
| Housing Start (in thousands) (X5) | 0.9985 | Yes | 684.9241 |
| National Debt (in Trillions) (X6) | 0.9989 | Yes | 927.3734 |
| Gold (Average Price) (X7) | 0.99903 | Yes | 1030.2016 |
| U.S. Imports (in Billions) (X8) | 0.9965 | Yes | 286.7632 |
| U.S. Imports (in Billions) (X9) | 0.9988 | Yes | 801.4526 |
| The political party of the US President (X10) | 0.99902 | Yes | 1018.5681 |

The coefficients are in the multiple linear regression is the size of the coefficient for each independent variable that gives the size of the effect on the dependent variable. Also, the size of the coefficient is positive or negative that gives the direction of the effect. When the coefficient is positive, it means how much the dependent variable is expected to increase. On the other hand, if the coefficient is negative, it mean that it decrease when the independent variable increases by one. However, in regression with multiple independent variables, the coefficient shows how much the dependent variable is expected to increase when the independent variable increases by one, holding all the other independent variables constant (Princeton University, 2007).

**Table of Coefficients for each 10 Variables**

|  |  |
| --- | --- |
|  | *Coefficients* |
| Intercept | -10136.2126 |
| U.S. Population (in million) (X1) | 62.4612 |
| Personal Savings Rate (X2) | -83.2489 |
| Average unemployment rate (X3) | 23.5805 |
| Inflation Rates (X4) | -58.2224 |
| Housing Start (in thousands) (X5) | -0.4515 |
| National Debt (in Trillions) (X6) | 108.0037 |
| Gold (Average Price) (X7) | -0.1458 |
| U.S. Imports (in Billions) (X8) | 3.6246 |
| U.S. Exports (in Billions) (X9) | -2.1721 |
| The political party of the US President (X10) | -57.6445 |

I dropped the ones at a time with the highest R-Square and revised the model again. As a result, I dropped 5 independent variables maximum. Although some of the independent variables are still more than 80% in multicollinearity R-Square, I cannot drop anymore independent variables. The rules for the project are to drop maximum of 5 independent variables. I had dropped 5 independent variables. If I drop more variables, there will be one or no variables left. In the real world, most the independent variables in multicollinearity are less than 80%. Also, the value in coefficient would be positive.

Table for R-Square for Independent Variables. Drop the highest R-Square that is more than 80 percent.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variables | Model #1 | Model #2 | Model #3 | Model #4 | Model #5 |
| U.S. Population (in million) (X1) | 0.9888 | 0.9883 | 0.9883 | 0.9857 | 0.9821 |
| Personal Saving Rate (X2) | 0.9989 | 0.99881 | 0.9988 | 0.9987 | 0.9986 |
| Average unemployment rate (X3) | 0.99902 | 0.998895  **Drop** |  |  |  |
| Inflation Rates (X4) | 0.9986 | 0.9977 | 0.9974 | 0.9974 | 0.9986 |
| Housing Start (in thousands) (X5) | 0.9985 | 0.9984 | 0.9983 | 0.9983 | 0.9982 |
| National Debt (in Trillions) (X6) | 0.9989 | 0.99839 | 0.99889  **Drop** |  |  |
| Gold (Average Price) (X7) | 0.99903  **Drop** |  |  |  |  |
| U.S. Imports (in Billions) (X8) | 0.9965 | 0.9962 | 0.9962 | 0.9960 | 0.9954 |
| U.S. Imports (in Billions) (X9) | 0.9988 | 0.9987 | 0.9987 | 0.9987 | 0.9987  **Drop** |
| The political party of the US President (X10) | 0.99902 | 0.99888 | 0.99886 | 0.9988  **Drop** |  |

|  |  |  |
| --- | --- | --- |
| Name | R-Square | Adjusted R-Square |
| Model #1 | 0.9990 | 0.9988 |
| Model #2 | 0.9989 | 0.9987 |
| Model #3 | 0.9989 | 0.9987 |
| Model #4 | 0.9989 | 0.9987 |
| Model #5 | 0.9988 | 0.9987 |
| Final Model | 0.9987 | 0.9985 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT  (**Final Model for Stage 2** ) |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.9993 |  |  |  |  |  |
| **R Square** | **0.9987** |  |  |  |  |  |
| **Adjusted R Square** | **0.9985** |  |  |  |  |  |
| Standard Error | 177.7850 |  |  |  |  |  |
| Observations | 51 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 5 | 1.075E+09 | 2.15E+08 | 6804.506274 | 1.44076E-63 |  |
| Residual | 45 | 1422338.4 | 31607.52 |  |  |  |
| Total | 50 | 1.077E+09 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | -9674.9766 | 518.0240 | -18.6767 | 0.00000000000000000000007965 | -10718.3305 | -8631.6227 |
| U.S. Population (in million) (X1) | 58.6615 | 2.0363 | 28.8084 | 0.000000000000000000000000000001221 | 54.5602 | 62.7627 |
| Personal Savings Rate (X2) | -62.2391 | 19.4444 | -3.2009 | 0.002514 | -101.4021 | -23.0761 |
| Inflation Rates (X4) | -71.7403 | 10.6461 | -6.7386 | 0.00000002490 | -93.1827 | -50.2979 |
| Housing Start (in thousands) (X5) | -0.3145 | 0.0799 | -3.9371 | 0.0002835 | -0.4754 | -0.1536 |
| U.S. Imports (in Billions) (X8) | 2.7812 | 0.1124 | 24.7532 | 0.0000000000000000000000000007568 | 2.5549 | 3.0076 |

**Quadratic Terms**

In stage 3, I added quadratic terms for the explanatory variables that show a quadratic pattern. Three independent variables show a quadratic pattern. Anderson said, “Residual plot is against the independent variable x is a graph in which the values of the independent variables are represented by the horizontal axis and the corresponding residual values are represented by the vertical axis” (p. 34) Each of the independent variable, I added a trend line in each of the plot and display the R-squared value. Therefore, three of the variables is more than 2% and square each of them. Durbin-Watson is a test statistic used to discover the presence of autocorrelation in the residuals from a regression analysis. In addition, there a relationship between values separated from each other by a given time lag in Durbin-Watson (p. 61). The model for stage 3, Durbin-Watson statistic is 1.04886; as a result, is more than 5% and it passes the test. However, the “P” in Durbin Watson, the risk level is low and is 0.0000. The “P” in Durbin-Watson means the measure of the risk level that we are wrong in our conclusions.

|  |  |  |
| --- | --- | --- |
| **Table of R-Square for Residual Plot** | | |
| Variables | R-Square for Residual Plot | Comments |
| U.S. Population (in million) (X1) | 0.0255 | Square (Is more than 2%) |
| Personal Saving Rates (X2) | 0.0069 |  |
| Inflation Rates (X4) | 0.2754 | Square (Is more than 2%) |
| Housing Start (in thousands) (X5) | 0.0044 |  |
| U.S. Imports (in Billions) (X8) | 0.0236 | Square (Is more than 2%) |

|  |  |
| --- | --- |
| **R-Square of the Final model for Stage 3** | |
| R Square | Adjusted R Square |
| 0.9991 | 0.9990 |

SUMMARY OUTPUT

(**Final Model for Stage 3**)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.9997 |  |  |  |  |  |
| **R Square** | **0.9994** |  |  |  |  |  |
| Adjusted R Square | 0.9992 |  |  |  |  |  |
| Standard Error | 127.1144 |  |  |  |  |  |
| Observations | 51 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 8 | 1076111533 | 134513942 | 8324.8788 | 1.24476E-64 |  |
| Residual | 42 | 678638.7709 | 16158.066 |  |  |  |
| Total | 50 | 1076790172 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | -1264.0444 | 2411.3900 | -0.5242 | 0.6029 | -6130.4264 | 3602.3377 |
| U.S. Population (in million) (X1) | -13.1430 | 22.2783 | -0.5899 | 0.5584 | -58.1025 | 31.8164 |
| Personal Savings Rate (X2) | 8.5807 | 20.5805 | 0.4169 | 0.6789 | -32.9524 | 50.1138 |
| Inflation Rates (X4) | -90.5321 | 33.8073 | -2.6779 | 0.01053 | -158.7581 | -22.3061 |
| Housing Start (in thousands) (X5) | -0.1958 | 0.0608 | -3.2198 | 0.002476 | -0.3185 | -0.0731 |
| U.S. Imports (in Billions) (X8) | 4.6509 | 0.6078 | 7.6515 | 0.000000001726 | 3.4242 | 5.8776 |
| U.S. Population (in million) Square | 0.1264 | 0.0499 | 2.5302 | 0.01524 | 0.0256 | 0.2271 |
| Inflation Rates Square | 3.0493 | 2.2703 | 1.3431 | 0.1865 | -1.5324 | 7.6310 |
| U.S. Imports (in Billions) Square | -0.0005 | 0.0001 | -3.9567 | 0.0002870 | -0.0008 | -0.0003 |

**P-Value**

In stage 4, all the independent variables must pass the p-value. According to Anderson, The p-value is used to test the hypothesis of equal population means (p.540). Backward elimination is a procedure that starts with a model that includes all the independent variables. After that, delete one independent variable at a time using the same procedure as step-wise regression. The final step is to remove the variables in the model that have a p-value greater than p-value (p. 754). The Durbin-Watson statistic had reduced when I revised it. The value is 1.04026 and it still more than 5%.

|  |  |  |
| --- | --- | --- |
| **Table 1 for p-value** | | |
| Variables | p-value | Comment |
| U.S. Population (in million) (X1) | 0.5584 |  |
| Personal Saving Rates (X2) | 0.6789 | **Drop** |
| Inflation Rates (X4) | 0.01053 |  |
| Housing Start (in thousands) (X5) | 0.002476 |  |
| U.S. Imports (in Billions) (X8) | 0.000000001726 |  |
| U.S. Population (in million) Square | 0.01524 |  |
| Inflation Rates Square | 0.1865 |  |
| U.S. Imports (in Billions) Square | 0.0002870 |  |

R Square: 0.9994

Adjusted R Square: 0.9992

|  |  |  |
| --- | --- | --- |
| **Table 2 for p-value** | | |
| Variables | p-value | Comment |
| U.S. Population (in million) (X1) | 0.6215 | **Drop** |
| Inflation Rates (X4) | 0.006773 |  |
| Housing Start (in thousands) (X5) | 0.0006818 |  |
| U.S. Imports (in Billions) (X8) | 0.0000000000001328 |  |
| U.S. Population (in million) Square | 0.01549 |  |
| Inflation Rates Square | 0.14060 |  |
| U.S. Imports (in Billions) Square | 0.000002710 |  |

R Square: 0.9994

Adjusted R Square: 0.9993

|  |  |  |
| --- | --- | --- |
| **Table 3 for p-value** | | |
| Variables | p-value | Comment |
| Inflation Rates (X4) | 0.00006817 |  |
| Housing Start (in thousands) (X5) | 0.0004476 |  |
| U.S. Imports (in Billions) (X8) | 0.00000000000002408 |  |
| U.S. Population (in million) Square | 0.00000000000001442 |  |
| Inflation Rates Square | 0.03051 | **Drop** |
| U.S. Imports (in Billions) Square | 0.000001870 |  |

R Square: 0.9994

Adjusted R Square: 0.9993

|  |  |  |
| --- | --- | --- |
| **Table 4 for p-value** | | |
| Variables | p-value | Comment |
| Inflation Rates (X4) | 0.000000005356 | Drop none because each independent variable is less than 5% |
| Housing Start (in thousands) (X5) | 0.00004256 |  |
| U.S. Imports (in Billions) (X8) | 0.000000000000003086 |  |
| U.S. Population (in million) Square | 0.00000000000007573 |  |
| U.S. Imports (in Billions) Square | 0.0000002296 |  |

R Square: 0.9993

Adjusted R Square: 0.9992

**Conclusion**

Stage 5 and 6 was not used because the model was already best fit. The final model of R-Square is 99.93 percent and adjusted R-Square is 99.92%. In Statgraphics, there is a statistically significant relationship between the variables at the 95.0% confidence level. All the independent variables are less than 0.05, it means the term is statistically significant at the 95% confidence level. Also, in Statgraphics, it says not to drop any more variables from the model. Therefore, the model passes the t-Test and p-value. Also, the model passes the F test and Durbin-Watson statistic. The equation for the final model:

= -2319.56 - 47.0561 (Inflation Rates) - 0.200315 (Housing) + 5.09971 (U.S. Imports) + 0.0862252 (U.S. Population Square) - 0.000605914 (U.S. Imports Square).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SUMMARY OUTPUT (**Final Model**) |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.9996 |  |  |  |  |  |
| **R Square** | **0.9993** |  |  |  |  |  |
| **Adjusted R Square** | **0.9992** |  |  |  |  |  |
| Standard Error | 130.2316 |  |  |  |  |  |
| Observations | 51 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | ***F*** | ***Significance F*** |  |
| Regression | 5 | 1.076E+09 | 2.15E+08 | **12688.7921** | **1.191E-69** |  |
| Residual | 45 | 763212.34 | 16960.27 |  |  |  |
| Total | 50 | 1.077E+09 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | ***P-value*** | *Lower 95%* | *Upper 95%* |
| Intercept | -2535.4628 | 368.6916 | -6.8769 | **0.00000001552** | -3278.0457 | -1792.8799 |
| Inflation Rates (X4) | -53.8712 | 7.4940 | -7.1886 | **0.000000005356** | -68.9649 | -38.7776 |
| Housing Start (in thousands) (X5) | -0.2473 | 0.0545 | -4.5345 | **0.00004256** | -0.3572 | -0.1375 |
| U.S. Imports (in Billions) (X8) | 4.7684 | 0.4077 | 11.6951 | **0.000000000000003086** | 3.9472 | 5.5896 |
| U.S. Population (in million) Square | 0.0943 | 0.0089 | 10.6228 | **0.00000000000007573** | 0.0764 | 0.1121 |
| U.S. Imports (in Billions) Square | -0.0005 | 0.0001 | -6.0899 | **0.0000002296** | -0.0007 | -0.0004 |

References

Anderson, D., Sweeney, D., & Williams, T. (2012). *Modern Business Statistics with Microsoft Office Excel*. Mason, OH: South-Western Cengage Learning.

Cameron, Colin A. (2009). Excel 2007: Multiple Regression. Retrieved April 20, 2012, from http://cameron.econ.ucdavis.edu/excel/ex61multipleregression.html

Princeton University. (2007). Interpreting Regression Output. Retrieved April 20, 2012, from http://dss.princeton.edu/online\_help/analysis/interpreting\_regression.htm