CKME136: Predicting GDP To Understand Interest Rate Effects

# Introduction

The context of this problem is a historically low (near-zero) interest rate environment (see image below).

Figure 1: US Fed Funds 1954 to Current



Source: macrotrends.com

Interest rates are near historic lows. In fact, interest rates are near zero. Let us set aside negative interest rate conditions for a moment. In most economic environments, negative interest rates are improbable, though not impossible. In this scenario, zero interest rates become a floor. When we have a floor for a key variable in economic modeling variable, we are arguably provided a higher-than-usual level of predictive certainty. In short, we “know” interest rates are going up. If direction is known, then only the rate of increase remains to be determined; namely the amount of interest rate increase and speed at which those increases are achieved. Arguably, this provides a more certain backdrop for predictive economic and financial modeling. This is the backdrop for this predictive modeling project.

Economic modeling is an important tool for both public and private sectors, institutions and individuals. Arguably, any operating forecast for a government, an enterprise or an individual making an investment decision is at least, in part, affected by economic conditions. This project proposes to tackle the problem of developing a reliable predictive modeling tool, for economic forecasting in the USA. It will do so by analyzing multiple economic and capital markets variables in the US.

Research Problem: How can we predict GDP to understand how interest rate increases might affect GDP?

# Literature Review

Below is a summary of some of the Literature Review available for predictive economic and capital markets modeling:

[CBO Budget & Economic Outlook - 2015 to 2025](https://www.cbo.gov/publication/49892)

This document provides a long-term outlook of US Federal Fiscal balances. It includes an economic forecast model including GDP projections with sector (Y = G + C + I + [X-M]) projections, labour, inflation and interest rate projections among others. I think it will provide a useful guide for economic modeling both in terms of methodology and as a benchmark for gauging the reasonability of any empirical results generated during this project.

[IMF - World Economic Outlook - April 2015](http://www.imf.org/external/pubs/ft/weo/2015/01/index.htm)

This document provides another sample of economic predictive modeling. It identifies some non-traditional economic measures such as output potential, output gap and probability of recession distributions that might be worthy of further revenue when I am doing data analysis.

[World Bank - Global Economic Prospects](http://www.worldbank.org/gep)

Similar to the IMF citation, this document provides guidance as to what the final output of a predictive economic model should look like. It provides some unique data points, such as federal reserve policy projections, in terms of interest rate ranges. Data presentation in terms of medians overlaid onto projection ranges are helpful. May try it in R.

http://www.fanniemae.com/portal/research-insights/forecast.html

This site contains current and archived monthly projections. It is especially helpful in the sense that it projects interest-sensitive housing investment.

[OECD Economic Outlook](http://www.oecd.org/eco/economicoutlook.htm)

This site contains several unique (to OECD) aggregated economic data points, such as Composite Leading Indicator (CLI). These will be included in the data analysis. As with many of the data points, there are some disparities between what is available for Canada vs. USA, if only in the lengths of the data series. Still … it should be a useful data source that was not contemplated in the original project abstract.

[FRB St. Louis-Kevin Kliesen-A Guide to Tracking the U.S. Economy-Q1 2015](https://research.stlouisfed.org/publications/review/2014/q1/kliesen.pdf)

This document provides guidance on how to forecast US economy and what variables are important. It references other forecast models.

<https://www.imf.org/external/pubs/ft/fandd/basics/models.htm>

This article distinguishes between theoretical and empirical economic models. It talks about evaluation models based on mean errors and volatility. The objective is to reduce mean errors to zero while minimizing volatility.

<https://voxeu.org/article/how-should-we-make-economic-forecasts>

This article is a bit brief, but it raises the question as to adding and removing variables and how it affects the accuracy of economic models. Which variables to include or not include and how they might affect Type II errors will be applied in this project.

<https://en.wikipedia.org/wiki/Economic_forecasting>

Wikipedia provides some clear guidance on forecast method steps and references multiple sources of economic projection models.

**BOOK ABOUT CANADA ECONOMIC FORECASTING:**

<https://books.google.ca/books?hl=en&lr=&id=fSKYAAAAQBAJ&oi=fnd&pg=PR1&dq=CANADA+ECONOMIC+FORECASTING&ots=GCQpioKwGA&sig=w2prjVsifPNU7Yxafob7oTe964Q#v=onepage&q=CANADA%20ECONOMIC%20FORECASTING&f=false>

This online book provides detail on how to forecast for particular asset classes. It will be useful for the capital markets predictive modeling component of this project.

**For future reference only – no comments:**

<https://www.bdc.ca/en/blog/pages/2018-economic-outlook-global-growth-brings-good-news-canadian-entrepreneurs.aspx>

<http://www.oecd.org/eco/outlook/forecastingmethodsandanalyticaltools.htm>

# Dataset

For the purposes of this capstone, data is pulled from several websites include Bureau of Economic Analysis, Bureau of Labor Statistics, Federal Reserve and Investing.com.

## Data Sourcing

A summary of the data sources is presented in the table below:

Figure 2: Data Collected for Capstone

|  | **Data Item** | **Source URL** | **Source Provider** | **Raw Interval** | **Start Date** | **End Date** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | US Qtrly Import Index Capital Goods | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 1979Q4 | 2018Q3 |
| 2 | US Income and Spending | https://apps.bea.gov/iTable/iTable.cfm?reqid=19&step=2#reqid=19&step=2&isuri=1&1921=survey | BEA | Monthly | 01-Jan-59 | 01-Sep-18 |
| 3 | US Qtrly Import Price Index (ex fuels) | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 1984Q4 | 2018Q3 |
| 4 | US Qtrly Import Fuel Price Index | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 1984Q4 | 2018Q3 |
| 5 | US Import Price Index | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 1982Q3 | 2018Q3 |
| 6 | US NonfaRM Labor Cost Qtrly Change | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 1947Q2 | 2018Q3 |
| 7 | US Nonfarm Real Hourly Cost Qtrly Change | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 1947Q2 | 2018Q3 |
| 8 | US Nonfarm Output Qtrly Change | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 1947Q2 | 2018Q3 |
| 9 | US Real Hourly Cost Qtrly Change | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 1947Q2 | 2018Q3 |
| 10 | US Unit Labor Cost Qtrly Change | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 1947Q2 | 2018Q3 |
| 11 | US Output Per Hour Qtrly Change | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 1947Q2 | 2018Q3 |
| 12 | US Wkly Earnings Real | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 1979Q1 | 2018Q3 |
| 13 | US Wkly Earnings | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 1979Q1 | 2018Q3 |
| 14 | US Hrly Lab Cost Benefits | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 2008Q1 | 2018Q2 |
| 15 | US Hrly Lab Cost Wages | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 2008Q1 | 2018Q2 |
| 16 | US Hrly Lab Cost Combined | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 2008Q1 | 2018Q2 |
| 17 | US ECI Benefits | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 2001Q1 | 2018Q3 |
| 18 | US ECI Wages | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 2001Q1 | 2018Q3 |
| 19 | US ECI Combined | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Quarterly | 2001Q1 | 2018Q3 |
| 20 | US Not in Labor Force Discouraged | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-94 | 01-Oct-18 |
| 21 | US Not in Labor Force Searching | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-94 | 01-Oct-18 |
| 22 | US Not in Labor Force | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-75 | 01-Oct-18 |
| 23 | US New Entrants Unemployed | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-67 | 01-Oct-18 |
| 24 | US Median Weeks Unemployed | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jul-67 | 01-Oct-18 |
| 25 | US Avg Weeks Unemployed | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-48 | 01-Oct-18 |
| 26 | US Unemployment Rate | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-48 | 01-Oct-18 |
| 27 | US Unemployment | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-48 | 01-Oct-18 |
| 28 | US Part Time Employed | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-68 | 01-Oct-18 |
| 29 | US Full Time Employed | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-68 | 01-Oct-18 |
| 30 | US Employment Ratio | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-48 | 01-Oct-18 |
| 31 | US Total Employment | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-48 | 01-Oct-18 |
| 32 | US Labor Participation | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-48 | 01-Oct-18 |
| 33 | US Civilian Labor Force | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-48 | 01-Oct-18 |
| 34 | US Non-farm employment | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-48 | 01-Oct-18 |
| 35 | US CCI |  | OECD |  |  |  |
| 36 | US CLI |  | OECD |  |  |  |
| 37 | US BCI |  | OECD |  |  |  |
| 38 | US Industrial Production and Capacity Utilization | https://www.federalreserve.gov/datadownload/Download.aspx?rel=G17&series=b2f049239f5cf183f4725cecbda0e4bd&filetype=csv&label=include&layout=seriescolumn&from=01/01/1919&to=12/31/1920 | Fed Reserve | Monthly | 01-Jan-19 | 01-Nov-18 |
| 39 | US Non-revolving | https://www.federalreserve.gov/datadownload/Download.aspx?rel=G19&series=8bf9516c0580cb3eb79502c041a63da1&filetype=csv&label=include&layout=seriescolumn&lastObs=&type=package | Fed Reserve | Monthly | 01-Jan-43 | 01-Nov-18 |
| 40 | US Revolving Credit | https://www.federalreserve.gov/datadownload/Download.aspx?rel=G19&series=6122591a0c6ce5e44eba1275f1360517&filetype=csv&label=include&layout=seriescolumn&lastObs=&type=package | Fed Reserve | Monthly | 01-Jan-68 | 01-Nov-18 |
| 41 | US Major Holders | https://www.federalreserve.gov/datadownload/Download.aspx?rel=G19&series=b0a993954f179e866ccda94309fd7bef&filetype=csv&label=include&layout=seriescolumn&lastObs=&type=package | Fed Reserve | Monthly | 01-Jan-43 | 01-Nov-18 |
| 42 | US Terms of Credit | https://www.federalreserve.gov/datadownload/Download.aspx?rel=G19&series=72a2e956fb1c6b8dbff0a1b6a4b32324&filetype=csv&label=include&layout=seriescolumn&lastObs=&type=package | Fed Reserve | Monthly | 01-Jan-76 | 01-Nov-18 |
| 43 | US Consumer Credit SA | https://www.federalreserve.gov/datadownload/Download.aspx?rel=G19&series=88c443018ff95e5f2b2bae86fbd226cd&filetype=csv&label=include&layout=seriescolumn&lastObs=&type=package | Fed Reserve | Monthly | 01-Feb-43 | 01-Nov-18 |
| 44 | US Consumer Credit No SA | https://www.federalreserve.gov/datadownload/Download.aspx?rel=G19&series=47b3133fcba3957706678b2a55cb5a97&filetype=csv&label=include&layout=seriescolumn&lastObs=&type=package | Fed Reserve | Monthly | 01-Jan-43 | 01-Nov-18 |
| 45 | US Mortgage Debt | https://www.federalreserve.gov/data/mortoutstand/current.htm | Fed Reserve |  |  |  |
| 46 | US Owned Receivables | https://www.federalreserve.gov/data.htm | Fed Reserve |  |  |  |
| 47 | US Finance Company Receivables | https://www.federalreserve.gov/data.htm | Fed Reserve |  |  |  |
| 48 | US Deposit Reserves | https://www.federalreserve.gov/data.htm | Fed Reserve |  |  |  |
| 49 | US FX Rates | https://www.federalreserve.gov/data.htm | Fed Reserve |  |  |  |
| 50 | US Commercial Paper Rates | https://www.federalreserve.gov/datadownload/Download.aspx?rel=CP&series=593ce926936cbd64b3c79b960a792b85&filetype=csv&label=include&layout=seriescolumn&from=01/01/1998&to=12/31/2018 | Fed Reserve | Daily | 1998 | 2018 |
| 51 | US Assets and Liabilities of Banks | https://www.federalreserve.gov/releases/h8/current/default.htm | Fed Reserve | Quarterly |  |  |
| 52 | US Charge Off Rates | https://www.federalreserve.gov/releases/chargeoff/chgallsa.htm | Fed Reserve | Quarterly | 1985Q1 | 2018Q2 |
| 53 | US Delinquency Rates | https://www.federalreserve.gov/releases/chargeoff/delallsa.htm | Fed Reserve | Quarterly | 1985Q1 | 2018Q2 |
| 54 | US Monetary Base | https://fred.stlouisfed.org/series/BASE | Fed Reserve | Biweekly | 15-Feb-84 | current |
| 55 | US Housing Prices | tps://fred.stlouisfed.org/series/USSTHPI | Fed Reserve | Quarterly | 1975Q1 | 2018Q1 |
| 56 | US Libor | https://fred.stlouisfed.org/series/USD3MTD156N | Fed Reserve | Daily | 02-Jan-86 | 31-Oct-18 |
| 57 | US Interest Rates | https://www.federalreserve.gov/datadownload/ | Fed Reserve | Monthly | 01-Jan-34 | 01-Sep-18 |
| 58 | US Consumer Debt | https://www.federalreserve.gov/releases/g19/hist/cc\_hist\_sa\_levels.html | Fed Reserve | Monthly | 01-Jan-43 | 01-Aug-18 |
| 59 | US GDP | https://www.bea.gov/national/xls/gdplev.xlsx | BEA | Quarterly | 1947Q1 | 2018Q2 |
| 60 | US CPI base 82 84 | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-13 | 01-Sep-18 |
| 61 | US CPI base 67 | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-13 | 01-Sep-18 |
| 62 | US CPI F&B | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-47 | 01-Sep-18 |
| 63 | US CPI Housing | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-47 | 01-Sep-18 |
| 64 | US CPI Transpo | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-67 | 01-Sep-18 |
| 65 | US Total Employment | https://data.bls.gov/pdq/SurveyOutputServlet | BLS | Monthly | 01-Jan-48 | 01-Oct-18 |
| 66 | US S&P 500 Daily | http://www.investing.com | Investing.com | Daily | 10-Sep-97 | 16-Oct-18 |
| 67 | US S&P 500 Monthly | http://www.investing.com | Investing.com | Monthly | 01-Feb-70 | 01-Oct-18 |
| 68 | US DJIA Daily | http://www.investing.com | Investing.com | Daily | 01-Jan-07 | 16-Oct-18 |
| 69 | US DJIA Monthly | http://www.investing.com | Investing.com | Monthly | 01-Feb-85 | 01-Oct-18 |

## Data Cleaning and Normalizing

The data format varied by sources. These variations included data layout and value formatting.

For layout, some data was downloaded in columns and some in rows. Some data is reported in days, or weeks or months or quarters. To accommodate these layout variations, the following data cleaning had to be done:

1. organize data into columns for upload into R
2. regroup data to a common quarterly reporting period and label it properly (for example, for monthly data, all data is labeled as YYYYQN or YYYYQ1 or YYYYQ2 or YYYYQ3 or YYYYQ4, then only the integer ended values are pulled in via EXCEL INDEX-MATCH fomulas)
3. convert to quarterly average values from daily reporting

### Organize Data Into Columns For Upload Into R

Use offset functionality to convert rows to columns

Monthly Reporting

=OFFSET($A$1,MATCH(VALUE(LEFT(R20,4)),A:A,0)-1,12)

Quarterly Reporting

=OFFSET($A$1,MATCH(VALUE(LEFT(R49,4)),B:B,0)-1,VALUE(RIGHT(R49,1))+1)

### Regroup Data To A Common Quarterly Reporting Period and Label It Properly

For MMM-YY (date formatted) in a single column, use IFS functionality to apply quarterly labels

=IFS(MONTH(A8)=3,YEAR(A8)&"Q1",MONTH(A8)=6,YEAR(A8)&"Q2",MONTH(A8)=9,YEAR(A8)&"Q3",MONTH(A8)=12,YEAR(A8)&"Q4",TRUE,YEAR(A8)&"QN")

For MMM-YY (text formatted) in a single column, use IFS functionality to apply quarterly labels

=IFS(VALUE(RIGHT(G4,1))=3,LEFT(G4,4)&"Q1",VALUE(RIGHT(G4,1))=6,LEFT(G4,4)&"Q2",VALUE(RIGHT(G4,1))=9,LEFT(G4,4)&"Q3",VALUE(RIGHT(G4,2))=12,LEFT(G4,4)&"Q4",TRUE,"")

For YYYY:MM:00 (format not sure) in a single column, use text extract and concatenation

=LEFT(C5,4)&"Q"&MID(C5,7,1)

For YYYY-MM-DD (date formatted; monthly) use date and concatenation functionality:

=IFS(MONTH(A20)=1,YEAR(A20)-1&"Q4",MONTH(A20)=4,YEAR(A20)&"Q1",MONTH(A20)=7,YEAR(A20)&"Q2",MONTH(A20)=10,YEAR(A20)&"Q3",TRUE,"")

For YYYY-MM-DD (date formatted; daily) use date and concatenation functionality:

=IFS(AND(MONTH(B13)=3,MONTH(B14)=4),YEAR(B13)&"Q1",AND(MONTH(B13)=6,MONTH(B14)=7),YEAR(B13)&"Q2",AND(MONTH(B13)=9,MONTH(B14)=10),YEAR(B13)&"Q3",AND(MONTH(B13)=12,MONTH(B14)=1),YEAR(B13)&"Q4",TRUE,YEAR(B13)&"QN")

For YYYY & QTR# (text formatted) in two columns, use concatenate

=B30&"Q"&RIGHT(C30,1)

### Recalculate Quarterly Values From Daily Reporting

Equity capital markets data, converting daily data to quarterly averages is as below:

Whenever daily date changes from membership in one quarter to another, calc the number of data points for the quarter just completed, identified as follows …

=IF(RIGHT(B80,2)<>"QN",MAX(C$2:C79)+1,"")

… then take the daily average based on the number of days (counted above) for that quarter

=IF(C80<>"",AVERAGE(OFFSET(K80,0,0,-(ROW(C80)-MATCH(C80-1,C:C,0)),1)),"")

NOTE – that the day count here is variable because it is trading days so it varies across quarters with timing of weekends and holidays

The formulas above are sampled from the XLS attached. Relative references are thus subject to change.

Data was pulled from many sources. It is pulled together in this XLS. For each data source (and topic) there is an worksheet (69 in total). From those work sheets, the values are sent to a single worksheet which can be saves as a CSV for upload into R. (please see XLS embedded below).



On each worksheet there may be several variables. To control size, only certain data points were selected. That reduced the number of variables to 114.

# Approach (Planning Level)

## Step 1: Assemble Data From Sources

The first step will be pulling in data from sources. There are many data sources for economic and capital markets. Moreover, the data points are spread out across many sources. Further, often, there are many data sources for the same data point. Controlling scope will be important part of completing this step.

## Step 2: Filter and Clean Data

After data is pulled from sources data will have to be filtered. For example, it might not be correct to evaluate seasonally-adjusted values with those that are not seasonally-adjusted. Data will have to be filtered to ensure that issues of logical compatibility are addressed before analyzing the data.

The data will be time series but sourced based on different intervals. Some data might be available daily, others monthly and some quarterly. Data will have to be manipulated so that the time intervals are consistent. Based on data review to this point, quarterly is likely to be the relevant interval.

## Step 3: Organize and Group Data

Data will be grouped by category, for example economic versus capital markets. Data will be tabulated into a common time series presentation. Data will be formatted so that it can be loaded into R.

## Step 4: Preliminary Analysis

Preliminary analysis will review of time series graphs. It will include correlation analysis of all variables (based on groupings in Step 3). Output is to include correlation matrix and time series visualizations. Highly correlated values (positive or negative) with the dependent variable will be the first variables used in the next step.

## Step 5: Model and Vet

In R, will use regression (including both linear and random forest) and other modeling to build out predive models. Initial analysis will be based on the variables identified in step 4. Other variables, more specifically, interest data may be introduced to see if there is any material affect on results.

# Implementation (Level) Steps

## Step 1: Load Normalized Data Into R Via CSV Upload

**Data is extracted from XLS, into a CSV and uploaded.**

Steps are as follows:

USData=read.csv("C:/Users/nmccl/Desktop/CKME project/USData.csv", header = TRUE, na.strings="" )

(format date data as date type)

USData$Date = as.Date(USData$Date, "%m/%d/%Y")



## Step 2: Generate Time Series Plots

**Generate the time series plots:**

plot\_list = list()

for(i in 3:116) {

GraphTitle = (colnames(USData)[i])

p = ggplot(data = USData[!is.na(USData[,i]),], aes\_string(x=colnames(USData)[1], y=colnames(USData)[i])) + geom\_line() + theme(axis.title.y=element\_blank(), panel.border = element\_rect("black", fill = NA, size =2))+ggtitle(GraphTitle)

plot\_list[[i]] = p

}

**Write plots to PDF file:**

pdf("C:/Users/nmccl/Desktop/CKME project/USDataTimeSeriesPlots.pdf")

for(i in 3:116)

{

print(plot\_list[[i]])

}

dev.off()

The time series plots are included in the file attached:



## Step 3: Calculate Correlations

**Calculate the correlations:**

USDataCor = rcorr(as.matrix(USData\_NoDates))

**Output the correlations to a CSV:**

(use the following function to flatten correlations so it can be put in CSV file:)

#function below puts all correlations into a table

flattenCorrMatrix <- function(cormat, pmat) {

ut <- upper.tri(cormat)

data.frame(

row = rownames(cormat)[row(cormat)[ut]],

column = rownames(cormat)[col(cormat)[ut]],

cor =(cormat)[ut],

p = pmat[ut]

)

}

(apply the function:)

flattenCorrMatrix(USDataCor$r, USDataCor$P)

(write to CSV:)

write.csv(USCorrOutput,"C:/Users/nmccl/Desktop/CKME project/USDataCors.csv")



## Step 4: Filter The Correlations For Dependent Variable

In order to predict overall GDP, the number of data points was reduced from 114 to 12. By filtering the data in an XLS for correlation greater than |0.75| AND p values less than 0, the number of variables was reduced to 12. Five interest rates variables were also carried forward to enable further interest rate analysis.

It should be noted that many GDP components are included in the CSV upload data. For example, Consumer GDP or Government Spending GDP. These were all removed from consideration as they are part of overall GDP.

Figure 3: Variables Extracted For Regression Analysis

| **VARIABLE NAME** | **REGRESSION LABEL** |
| --- | --- |
| US\_GDP\_Y\_\_\_\_\_\_Gross\_domestic\_product\_\_USD\_M\_ | Y |
| Personal\_Income\_Real\_\_2012\_\_\_USD\_B\_2012\_ | X1 |
| US\_Population\_\_000s\_ | X2 |
| USCPI\_UrbanB8284\_\_Index\_8284\_ | X3 |
| Ttl\_IndCapUtil\_\_USD\_M\_ | X4 |
| Total\_Consumer\_Revolving\_Debt\_\_USD\_M\_ | X5 |
| US\_HousingPrices\_Base80\_\_Index\_80\_ | X6 |
| US\_Employed\_\_000s\_ | X7 |
| All\_Mtgs\_\_USD\_M\_ | X8 |
| Import\_Export\_Price\_All\_Commodities\_\_Index\_2000\_ | X9 |
| US\_MonetaryBase\_\_USD\_M\_ | X10 |
| SP500\_\_Index\_ | X11 |
| LoanDel\_C.I\_\_Perc\_ | X12 |
| X13Mos\_Libor\_USD\_\_Perc\_ | XI1 |
| X10Year\_Treasury\_Rate\_\_Perc\_ | XI2 |
| Federal\_funds\_effective\_rate\_\_Index\_8284\_ | XI3 |
| X3Month\_Tbill\_\_Perc\_ | XI4 |
| Prime\_Rate\_\_Perc\_ | XI5 |

As a reminder, the interest variables did not satisfy correlation AND p-value filters. They are included here for supplemental analysis only.



## Step 5: Regression Analysis Data Preparation

To prepare data frame for regression analysis, we remove dates and subset for variables noted in Figure 3.

#remove non regression columns

USData\_NoDates = USData[,3:length(USData)] #to rid the data frame of non-data columns for correlation analysis

USDataRegr = subset(USData\_NoDates, select = c(US\_GDP\_Y\_\_\_\_\_\_Gross\_domestic\_product\_\_USD\_M\_,

Personal\_Income\_Real\_\_2012\_\_\_USD\_B\_2012\_,

US\_Population\_\_000s\_,

USCPI\_UrbanB8284\_\_Index\_8284\_,

Ttl\_IndCapUtil\_\_USD\_M\_,

Total\_Consumer\_Revolving\_Debt\_\_USD\_M\_,

US\_HousingPrices\_Base80\_\_Index\_80\_,

US\_Employed\_\_000s\_,

All\_Mtgs\_\_USD\_M\_,

Import\_Export\_Price\_All\_Commodities\_\_Index\_2000\_,

US\_MonetaryBase\_\_USD\_M\_,

SP500\_\_Index\_,

LoanDel\_C.I\_\_Perc\_,

X13Mos\_Libor\_USD\_\_Perc\_,

X10Year\_Treasury\_Rate\_\_Perc\_,

Federal\_funds\_effective\_rate\_\_Index\_8284\_,

X3Month\_Tbill\_\_Perc\_,

Prime\_Rate\_\_Perc\_))

To prepare data frame for regression analysis, we remove incomplete rows. Regression analysis breaks down if missing values. Applying means or medians did not seem appropriate for this data.

#keep only complete data for rows for regression analysis

USDataRegrClean = USDataRegr[complete.cases(USDataRegr),]

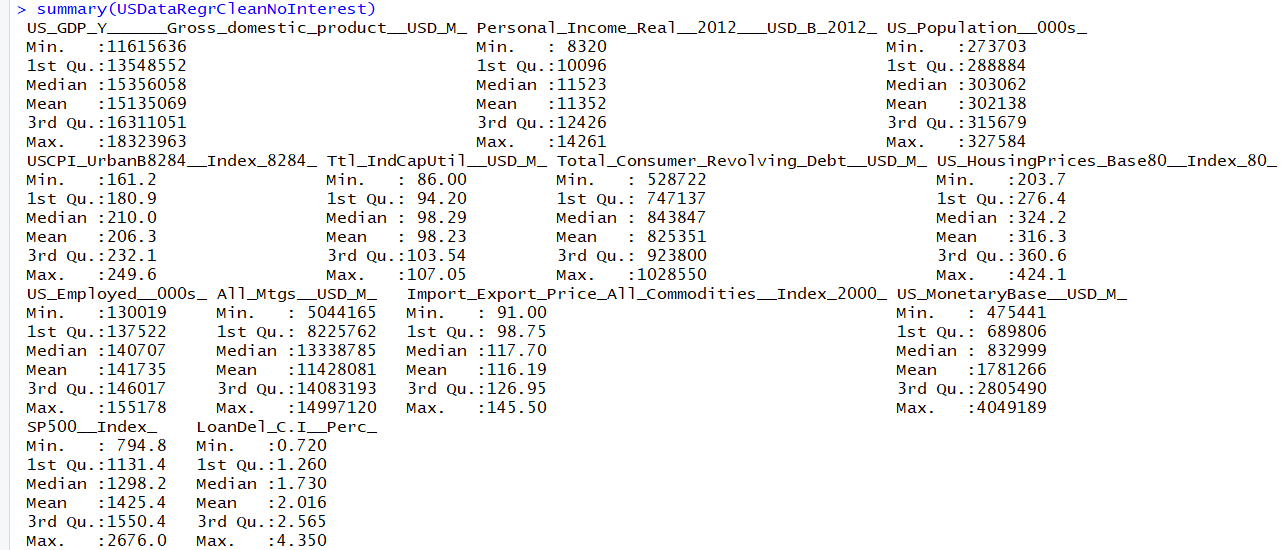
For initial regression analysis, we use only the non-interest data. This is the data that passed our filters (correlation > |0.75| and p < 0.05).

Here is the prepped data.

#review data without interest

summary(USDataRegrCleanNoInterest)

Figure 4: Variables Extracted For Regression Analysis - Summary

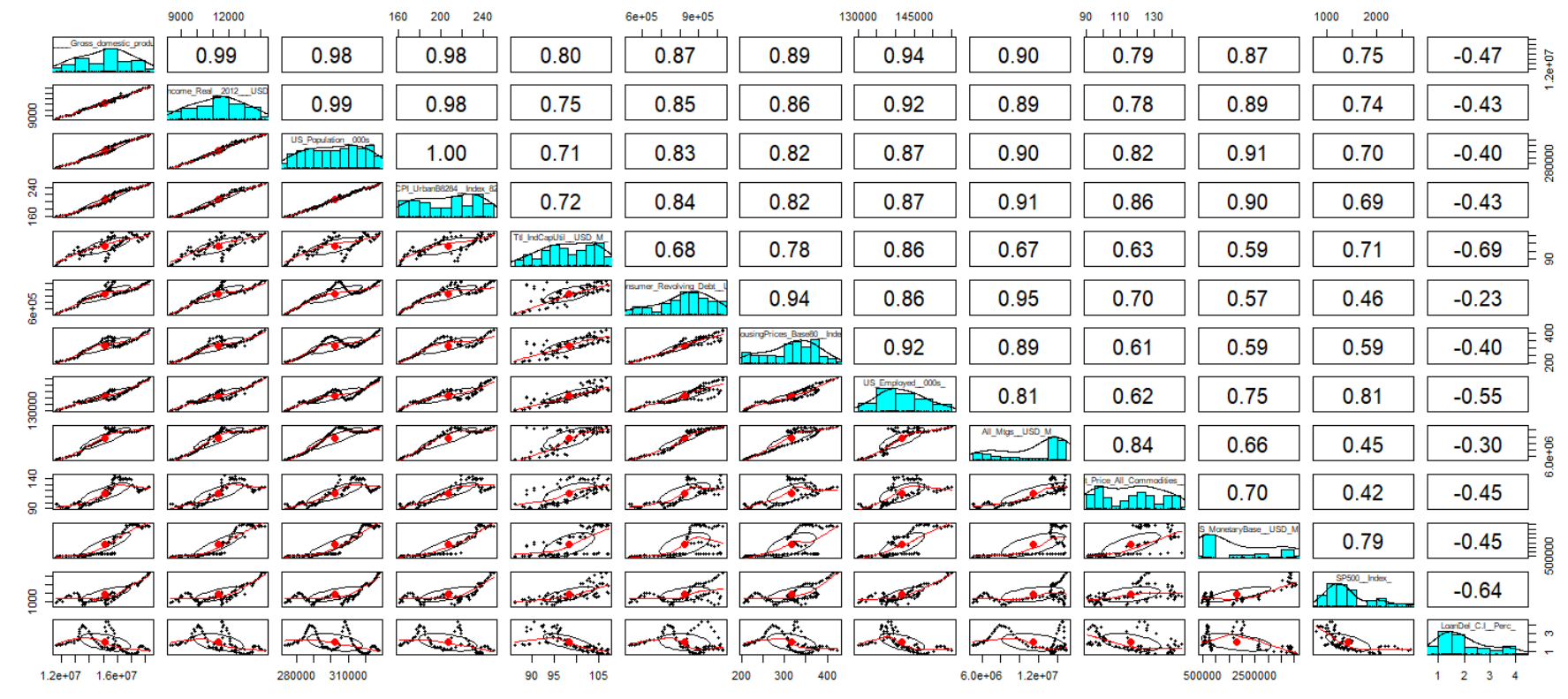


pairs.panels(USDataRegrCleanNoInterest, col = "red", cex.cor = 2)

Below is the pairs panel for selected data (interest rate data not included).

There are some highly correlated values in the pairs panel. Given the nature of the subject matter, this is probably inherent to the data. Leaps library’s regsubset will pick the optimal variables for our model – see later in this document.

Figure 5: Pairs Panels For Variables Extracted For Regression Analysis



## Step 6: Create Training and Validation Sets

#divide data into training and validation

set.seed(2018)

training.size = 0.8

train.index=sample.int(nrow(USDataRegrCleanNoInterest), round(nrow(USDataRegrCleanNoInterest)\* training.size))

#training data

train.USDataRegrCleanNoInterest = USDataRegrCleanNoInterest[train.index,]

View(train.USDataRegrCleanNoInterest)

#validation data

validate.USDataRegrCleanNoInterest = USDataRegrCleanNoInterest[-train.index,]

## Step 7: Select Optimized Models

Use regsubsets from the leaps package to identify the optimized model, based on adjusted R-squared.

#select dependent variables for optimal model

USRegResults = regsubsets(US\_GDP\_Y\_\_\_\_\_\_Gross\_domestic\_product\_\_USD\_M\_ ~

Personal\_Income\_Real\_\_2012\_\_\_USD\_B\_2012\_ +

USCPI\_UrbanB8284\_\_Index\_8284\_ +

Ttl\_IndCapUtil\_\_USD\_M\_ +

Total\_Consumer\_Revolving\_Debt\_\_USD\_M\_ +

US\_HousingPrices\_Base80\_\_Index\_80\_ +

US\_Employed\_\_000s\_ +

All\_Mtgs\_\_USD\_M\_ +

Import\_Export\_Price\_All\_Commodities\_\_Index\_2000\_ +

US\_MonetaryBase\_\_USD\_M\_ +

SP500\_\_Index\_ +

LoanDel\_C.I\_\_Perc\_,

data = train.USDataRegrCleanNoInterest,

nbest =4)

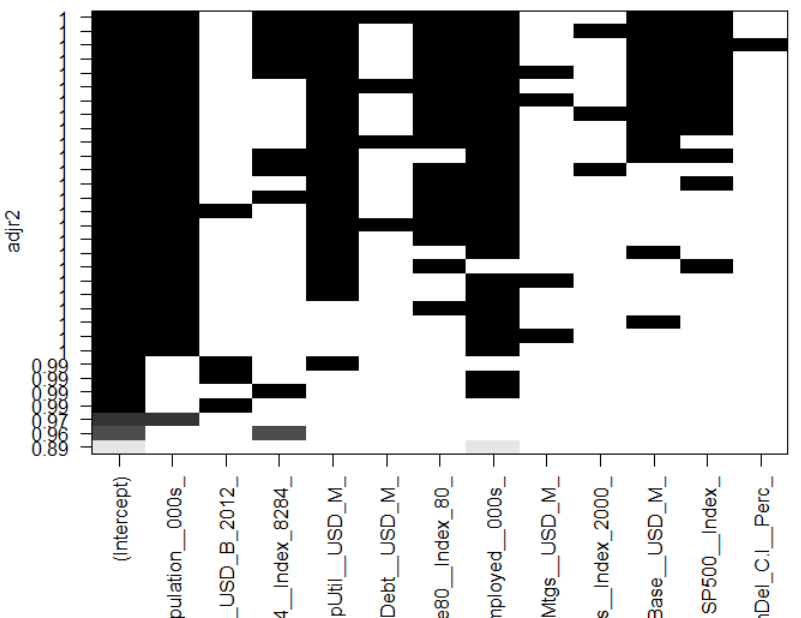
View the results:

#see summary and plot for optimal model

summary(USRegResults)

plot(USRegResults, scale="adjr2")

Figure 6: Optimal Regression Models From (leaps) Package



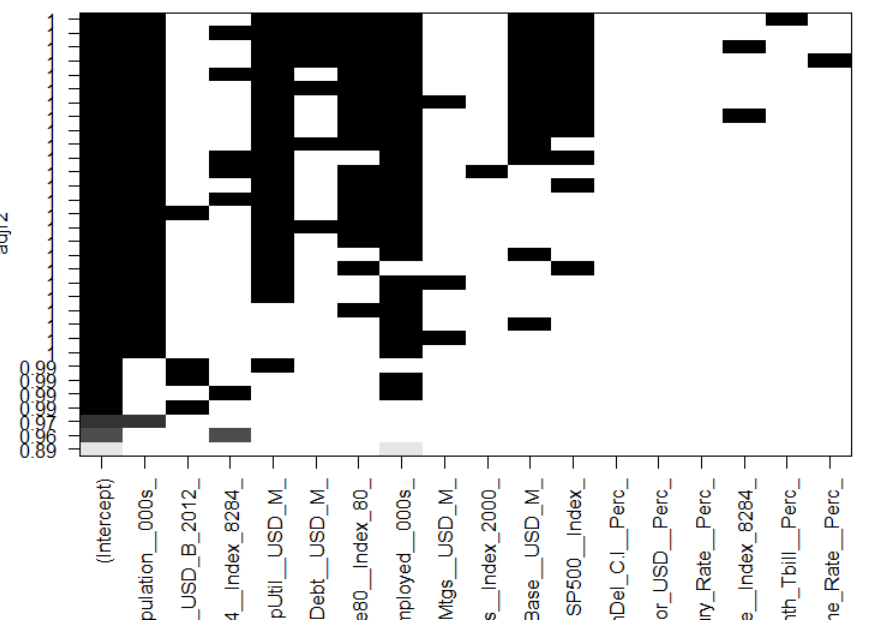
A summary of the models selected by regsubsets is below:

Figure 7: Top Five Models For Regression Analysis

| VARIABLE NAME | MODEL1 | MODEL2 | MODEL3 | MODEL4 | MODEL5 |
| --- | --- | --- | --- | --- | --- |
| US\_GDP\_Y\_\_\_\_\_\_Gross\_domestic\_product\_\_USD\_M\_ | Y | Y | Y | Y | Y |
| Personal\_Income\_Real\_\_2012\_\_\_USD\_B\_2012\_ |  |  |  |  |  |
| US\_Population\_\_000s\_ | X | X | X | X | X |
| USCPI\_UrbanB8284\_\_Index\_8284\_ | X | X | X | X | X |
| Ttl\_IndCapUtil\_\_USD\_M\_ |  | X | X | X | X |
| Total\_Consumer\_Revolving\_Debt\_\_USD\_M\_ | X |  |  |  |  |
| US\_HousingPrices\_Base80\_\_Index\_80\_ | X | X | X | X | X |
| US\_Employed\_\_000s\_ | X | X | X | X | X |
| All\_Mtgs\_\_USD\_M\_ |  |  |  |  | X |
| Import\_Export\_Price\_All\_Commodities\_\_Index\_2000\_ |  | X |  |  |  |
| US\_MonetaryBase\_\_USD\_M\_ | X | X | X | X | X |
| SP500\_\_Index\_ | X | X | X | X | X |
| LoanDel\_C.I\_\_Perc\_ |  |  | X |  |  |

Later we ran the same function, but included interest data points.

Figure 8: Optimal Regression Models From (leaps) Package – Interest Data Included



A summary of the models with interest data selected by regsubsets is below:

Figure 9: Top Five (With Interest Data) Models For Regression Analysis

| VARIABLE NAME | MODEL1 | MODEL2 | MODEL3 | MODEL4 | MODEL5 |
| --- | --- | --- | --- | --- | --- |
| US\_GDP\_Y\_\_\_\_\_\_Gross\_domestic\_product\_\_USD\_M\_ | Y | Y | Y | Y | Y |
| Personal\_Income\_Real\_\_2012\_\_\_USD\_B\_2012\_ |  |  |  |  |  |
| US\_Population\_\_000s\_ | X | X | X | X | X |
| USCPI\_UrbanB8284\_\_Index\_8284\_ |  | X |  |  | X |
| Ttl\_IndCapUtil\_\_USD\_M\_ | X | X | X | X | X |
| Total\_Consumer\_Revolving\_Debt\_\_USD\_M\_ | X | X | X | X |  |
| US\_HousingPrices\_Base80\_\_Index\_80\_ |  | X | X | X | X |
| US\_Employed\_\_000s\_ | X | X | X | X |  |
| All\_Mtgs\_\_USD\_M\_ |  |  |  |  | X |
| Import\_Export\_Price\_All\_Commodities\_\_Index\_2000\_ |  |  |  |  |  |
| US\_MonetaryBase\_\_USD\_M\_ | X | X | X | X | X |
| SP500\_\_Index\_ | X | X |  | X | X |
| LoanDel\_C.I\_\_Perc\_ |  |  |  |  |  |
| X13Mos\_Libor\_USD\_\_Perc\_ |  |  |  |  |  |
| X10Year\_Treasury\_Rate\_\_Perc\_ |  |  |  |  |  |
| Federal\_funds\_effective\_rate\_\_Index\_8284\_ |  |  | X |  |  |
| X3Month\_Tbill\_\_Perc\_ | X |  |  |  |  |
| Prime\_Rate\_\_Perc\_ |  |  |  | X |  |

## Step 7: Run Regression Models

The models above are presented in order, from best to worst, based on the attribute we chose to measure – in our case, Adjusted R-squared. So we ran all five models, first as linear regression and then as random forest regressions. As such, we have twenty regressions comprised of:

1. Five linear regressions without interest data
2. Five linear regressions with interest data forced in
3. Five random forest regressions without interest data
4. Five random forest regressions with interest data forced in

Below is sample code for one of the linear regressions without interest data:

USRegrNoInt1 = lm(US\_GDP\_Y\_\_\_\_\_\_Gross\_domestic\_product\_\_USD\_M\_ ~

US\_Population\_\_000s\_ +

USCPI\_UrbanB8284\_\_Index\_8284\_ +

Ttl\_IndCapUtil\_\_USD\_M\_ +

Total\_Consumer\_Revolving\_Debt\_\_USD\_M\_ +

US\_HousingPrices\_Base80\_\_Index\_80\_ +

US\_Employed\_\_000s\_ +

US\_MonetaryBase\_\_USD\_M\_ +

SP500\_\_Index\_,

data = train.USDataRegrCleanNoInterest)

After the regressions are run, the models are evaluated by running predictions and then evaluating the predicted results.

Below is sample code for one of the regression models without interest data prediction and evaluation:

train.USDataRegrCleanNoInterest$PredGDP1 = predict(USRegrNoInt1, newdata = subset(train.USDataRegrCleanNoInterest,

select = c(

US\_Population\_\_000s\_,

USCPI\_UrbanB8284\_\_Index\_8284\_,

Ttl\_IndCapUtil\_\_USD\_M\_,

Total\_Consumer\_Revolving\_Debt\_\_USD\_M\_,

US\_HousingPrices\_Base80\_\_Index\_80\_,

US\_Employed\_\_000s\_,

US\_MonetaryBase\_\_USD\_M\_,

SP500\_\_Index\_) ) )

validate.USDataRegrCleanNoInterest$PredGDP1 = predict(USRegrNoInt1, newdata = subset(validate.USDataRegrCleanNoInterest,

select = c(

US\_Population\_\_000s\_,

USCPI\_UrbanB8284\_\_Index\_8284\_,

Ttl\_IndCapUtil\_\_USD\_M\_,

Total\_Consumer\_Revolving\_Debt\_\_USD\_M\_,

US\_HousingPrices\_Base80\_\_Index\_80\_,

US\_Employed\_\_000s\_,

US\_MonetaryBase\_\_USD\_M\_,

SP500\_\_Index\_) ) )

summary(USRegrNoInt1)

train1.corr = round(cor(train.USDataRegrCleanNoInterest$PredGDP1, train.USDataRegrCleanNoInterest$US\_GDP\_Y\_\_\_\_\_\_Gross\_domestic\_product\_\_USD\_M\_),2)

train1.RMSE = round(sqrt(mean((train.USDataRegrCleanNoInterest$PredGDP1 - train.USDataRegrCleanNoInterest$US\_GDP\_Y\_\_\_\_\_\_Gross\_domestic\_product\_\_USD\_M\_)^2)))

train1.MAE = round(mean(abs(train.USDataRegrCleanNoInterest$PredGDP1 - train.USDataRegrCleanNoInterest$US\_GDP\_Y\_\_\_\_\_\_Gross\_domestic\_product\_\_USD\_M\_)))

c(train1.corr^2, train1.RMSE, train1.MAE)

validate1.corr = round(cor(validate.USDataRegrCleanNoInterest$PredGDP1, validate.USDataRegrCleanNoInterest$US\_GDP\_Y\_\_\_\_\_\_Gross\_domestic\_product\_\_USD\_M\_),2)

validate1.RMSE = round(sqrt(mean((validate.USDataRegrCleanNoInterest$PredGDP1 - validate.USDataRegrCleanNoInterest$US\_GDP\_Y\_\_\_\_\_\_Gross\_domestic\_product\_\_USD\_M\_)^2)))

validate1.MAE = round(mean(abs(validate.USDataRegrCleanNoInterest$PredGDP1 - validate.USDataRegrCleanNoInterest$US\_GDP\_Y\_\_\_\_\_\_Gross\_domestic\_product\_\_USD\_M\_)))

c(validate1.corr^2, validate1.RMSE, validate1.MAE)

# Results

All the 20 models generated strong adjusted R^2 near 1 and with p values well below 0.05. The linear regression models were slightly strong than random forest on the basis of adjusted R^2.

If we compute the correlation between predicted price and actual price, it should match the R-squared of the model and it does. See training adjusted R^2 versus validation adjusted R^2.

If we compute the root mean square error between predicted price and actual price and the mean absolute error for the same values, it provides a range of error.

Generally the validation set error ranges were a bit higher than the training set ranges. For linear regression models the values averaged 26% to 40% higher, whilst for random forest the differences averaged 56% to 64% more. On the basis of expected error ranges, linear regression models fared better.

Forcing in interest rate data did not seem to have much impact on the models. For linear regression models there is no difference in adjusted R^2. For random forest, the difference is only 70 bps.

Figure 10: Regression Model Results

| Type | Int Data | Model # | Model: Adj R^2 | Train: Adj R^2 | Train: RMSE | Train: MAE | Valid: Adj R^2 | Valid: RMSE | Valid: MAE |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| regr | n | model1 | 0.9988 | 1 | 58628 | 48267 | 1 | 79779 | 57577 |
| regr | n | model2 | 0.9988 | 1 | 58866 | 48123 | 1 | 80366 | 59231 |
| regr | n | model3 | 0.9988 | 1 | 59078 | 48980 | 1 | 77258 | 54731 |
| regr | n | model4 | 0.9988 | 1 | 59820 | 48999 | 1 | 84447 | 61996 |
| regr | n | model5 | 0.9988 | 1 | 59426 | 48928 | 1 | 83862 | 61453 |
| regr | y | model1 | 0.9988 | 1 | 58617 | 48359 | 1 | 85725 | 66116 |
| regr | y | model2 | 0.9988 | 1 | 58628 | 48267 | 1 | 79779 | 57577 |
| regr | y | model3 | 0.9988 | 1 | 58662 | 48288 | 1 | 85078 | 65771 |
| regr | y | model4 | 0.9988 | 1 | 58857 | 48379 | 1 | 84900 | 65475 |
| regr | y | model5 | 0.9988 | 1 | 59426 | 48928 | 1 | 83862 | 61453 |
| rand forst | n | model1 | 0.9962 | 1 | 57023 | 42651 | 1 | 83879 | 73509 |
| rand forst | n | model2 | 0.9957 | 1 | 59115 | 43314 | 1 | 94342 | 80634 |
| rand forst | n | model3 | 0.9952 | 1 | 59916 | 46216 | 1 | 101423 | 78104 |
| rand forst | n | model4 | 0.9958 | 1 | 57267 | 43153 | 1 | 91350 | 74727 |
| rand forst | n | model5 | 0.9961 | 1 | 56333 | 42303 | 1 | 83743 | 65271 |
| rand forst | y | model1 | 0.9500 | 1 | 61288 | 44266 | 1 | 85725 | 66116 |
| rand forst | y | model2 | 0.9961 | 1 | 54735 | 41599 | 1 | 88247 | 75293 |
| rand forst | y | model3 | 0.9951 | 1 | 61788 | 46071 | 1 | 96638 | 74473 |
| rand forst | y | model4 | 0.9952 | 1 | 59553 | 45854 | 1 | 98720 | 79312 |
| rand forst | y | model5 | 0.9961 | 1 | 55647 | 41110 | 1 | 86292 | 67595 |

As the graphs below show, there is little difference between actual US GDP and predicted GDP using either linear or random forest regression. Moreover, introducing interest data, also does not seem to change prediction accuracy markedly.

Figure 11: GDP Predicted Values (no interest data – training data)

US GDP = actual

PredGDP# = linear regression predicted model#

PredGDPRF# = random forest regression predicted model#

Figure 12: GDP Predicted Values (with interest data – training data)

US GDP = actual

PredGDP# = linear regression predicted model#

PredGDPRF# = random forest regression predicted model#

# Conclusions

## KEY CONCLUSIONS

This study produced some conclusions and generated many questions.

There seems to be a handful of variables that can be used to predict GDP quite accurately. These are:

1. population
2. CPI
3. capacity utilization
4. number employed
5. money supply
6. equity markets.

If we force in interest rates as an independent variable, the model’s ability to predict is only slightly diminished but the predictor variables change.

1. CPI becomes slightly less important
2. Revolving debt becomes more important
3. Mortgages outstanding become only slightly more important.

Overall, it is possible that interest levels may not be so impactful on TOTAL GDP, but moreso on the MIX of GDP.

## KEY QUESTIONS

This model could probably be developed further. For example, it might be useful to just force in economic, fiscal and monetary tools measured values. This would mean removing population, which grows steady state over time (see time series graph) and is less controllable. It might mean forcing in values that are more controllable by public and central policy makers. How these predictors dive GDP could adjust their decision-making accordingly.

The definition of GDP might be changed to include debt-adjusted GDP. Afterall, spending-measured GDP can be generated anytime with more debt, but at some point debt-funded could become more destructive than constructive. Using GROSS GDP – net of total debt might also be a productive exercise. Gross GDP net of debt could be a better measure to be predicted and eventually targeted.