Final Project

Andrew McGuirk

October 24, 2018

Section 1: Abstract

When I started this project I wanted to be able to build a model that would effectively predict US GDP. When I finished it I was able to reduce my test MSE of my original model by a factor of six by the time I got to my final model, and Generalized Additive Model that has a .99 adjusted R-squared to Real GDP. To build this model I found data for various economic variables of the website for the St. Louis Federal Reserve. I aimed to predict Real GDP (which means it is inflation adjusted) by using variables such as the price of oil, the unemployment rate, the yield on the 10 year US Treasury Note, the Consumer Price Index, Personal Consumer Expenditures, and Real Disposable Personal Income. The earliest I was able to find data points for all of these variables was January of 1964 so that is where my data will start and my data ends with the Q1 GDP numbers for 2018. To create my GAM I used a combination of basis splines for the variables. For each individual variable I used the LOOCV method to estimate how many knots to put in my basis spline. Then I used each individual basis spline and put them together to create a Generalized Additive Model. Something I noticed in the summary output for my GAM is that the coefficients increase in magnitude as the date gets closer and closer to the present. This makes a lot of sense because the Real GDP numbers have increased six-fold since the beginning of the data set and it would make sense that each variables affect would be magnified as time has gone on.

Section 2: Motivation and Background

For my project, I wanted to be able to build a model that would accurately predict the United States Gross Domestic Product (GDP), which is a measure of total economic activity in a given country. The formula for GDP is as follows:

$$GDP = C + I + G + NX$$

Where C is the total sum of all private consumption in the United States, I is the sum of all the country's investment, G is the sum of all government spending, and NX are the net exports (exports - imports).

I wanted to be able to put a model like this together because I wanted to have a statistically significant model that could help predict whether or not a recession is looming in the near future. A recession is a business cycle contraction which results in the slowing of economic growth and is defined by two straight quarters (6 months) of negative GDP growth. Because I wanted my model to predict whether a recession is *coming* rather than if it is *here* I am looking to predict GDP with a 6 month lead time, meaning if I have all the economic data today I can make a prediction of what GDP will be in 6 months. I also wanted to have a lead time between my predictors and GDP because when talking about the economy sometimes it takes a fair amount of time for changes in economic variables like monetary or fiscal policy to ripple through to the entire economy and make meaningful impacts on GDP numbers.

When thinking about what sorts of variables I wanted to look at when trying to predict GDP I thought about what specifically would have an effect on or be a good way of tracking each of those individual variables in the GDP formula. So the variables I initially decided to look at are PCE (personal consumer expenditures), the Unemployment Rate, the 10 Year US Treasury note rate, CPI (consumer price index) less food and energy which is a measure of inflation, the price of a barrel of oil, and finally RDPI which is real disposable personal income.

Summary of article on 10 year treasury note: https://www.thebalance.com/10-year-treasury-note-3305795. The 10 year treasury note is the benchmark of the United States government debt. It is the rate that the Fed keeps its eyes on the most and uses as a reference point of whether or not to alter interest rates. As yields on the 10 year rise, so do various other yields in the United States such as the fixed-mortgage rate and corporate debt yield because investors are expecting a higher return. Since the note is backed by the US treasury and there is essentially no chance of the US defaulting on this debt (and if they do there will be bigger problems to worry about) this is considered to be a risk less asset, as such returns are expected to be lower than equities or corporate debt because there is much more risk associated with those types of securities. The 10 year treasury note is in the middle of what is commonly referred to as the "yield curve". Simply explained, the yield curve is a graphical representation of government bonds of different maturities and their respective yields. On the X-axis is the maturity date, ranging from 1 month maturity date to 30 years. Then Y-axis is just the yield to maturity if you were to hold the bond up until its maturity date. Typically, the yield curve is upward trending as the longer you hold the bond the more you are expecting to get paid by the government. However, since these bonds are bought and sold on the secondary open market, they have the ability to fluctuate based on the markets expectations of future growth. When future growth is uncertain, people tend to sell short-term US bonds causing the yields to rise (yields move inversely to prices). When certain short-term yields are higher than long-term yields this is known as an "inversion" of the yield curve and is often seen as a sign that a recession is coming in the near future.

Summary of article on CPI less food and energy: https://www.bls.gov/opub/btn/archive/the-so-called-core-index-history-and-uses-of-the-index-for-all- items-less-food-and-energy.pdf CPI is generally an index that measures a basket of goods and how the prices of those goods change over time. If the price of this basket of goods goes up by, say 2% over the course of a year, then it is generally safe to say that we have had 2% inflation in the past year (which is a pretty standard number and the inflation rate that the Fed targets on a yearly basis). The reason a lot of economists (and myself) use CPI less food and energy is not only because of how necessary the two are for a society to function and thrive, but also because of how volatile they have been in recent years especially in comparison to everything else in the CPI basket. Additionally, food and energy prices are more subject to significant "shocks" than prices of other goods and services.

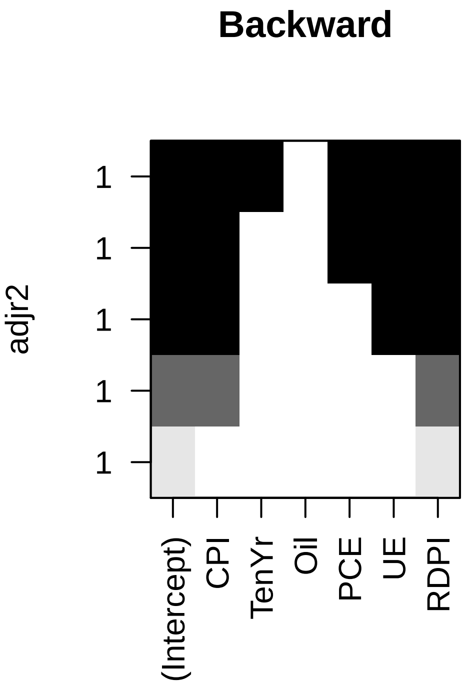
Section 3: Exploratory Data Analysis

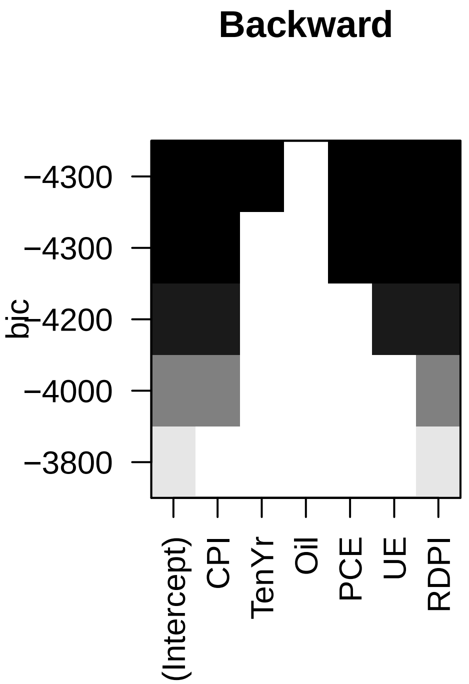
With the data in this project I had to do some manipulation considering that a lot of these variables are given at different rates. For instance, the 10 year US treasury note rate is constantly changing from second to second because it is being traded on the secondary open-market, while GDP is only measured quarterly. So in order to get my data to match up with dates I had to manipulate my data set a little bit. Since GDP is quarterly and most everything else I had was changing either daily, weekly, or monthly, I had to find a consistent time period to analyze these variables with each other. For GDP, say GDP was 2000 for January (Q1) of 1964, I then used that GDP number monthly for the full quarter, so January, February, and March of 1964 all have the same reading for GDP. For other variables that were measured either daily or weekly, I needed to convert them into monthly numbers. By using a few built in excel functions I was able to average some of these other variables like the 10 year treasury rate or the price of oil over the course of each month and obtained a monthly reading for those variables to go with the GDP numbers. Historically the GDP data on the St. Louis Federal Reserve website goes back to the 1940s, but some of these other variables aren't as historically backed, so my data set begins on January of 1964, which was the earliest date I could find data on **all** of my variables. The data is monthly and I have a total of 10 variables with 675 observations. Not all of these variables will be used in my analysis because they either don't carry much weight (i.e. date), or are other response variables to see which response will yield the best predictive model. The possible response variables I am looking at are GDP with a 2Q lead, *Real* GDP with a 2Q lead (i.e. inflation adjusted), and GDP growth rate. I began with all three possible responses but in the end only looked at Real GDP.

Variable Selection

PCE seemed like an obvious choice because simply put the more people are willing to use their money for consumption the higher the total consumption in the GDP calculation will be. In economic booms people will consume more, in busts, less. The unemployment rate was another fairly obvious choice. The higher the level of employment the more overall consumption there will be in both the private and public sectors. The private sector because fewer people are without work and they can use their income to stimulate the economy. Then for the public sector tax revenue will be higher due to higher income and more people paying taxes. I included the consumer price index as a way to account for inflation over the years considering the data goes back to 1964. RDPI seemed like a necessary choice as well for the higher real disposable incomes are the more money is circulating through the economy which leads to job growth which directly affects GDP. I included the price of oil because one of the standard measures of CPI does not include food or energy considering how necessary they are for a society to survive and thrive and thus will always be in demand.

The 10 year treasury note rate is essentially the cost of borrowing, the higher the 10 year rate the more it costs to borrow money and vice versa. It is considered the benchmark US debt rate and reflects the current level of monetary policy dictated by the Fed. The Fed has the power to raise and lower interest rates given the current climate of the economy. When we are in a bull market and at high employment, the Fed will raise interest rates to curb spending so as to lower the risk of rampant inflation, this is known as the tightening of monetary policy. Now, when we are in bad times economically the Fed will lower interest rates, decreasing the cost of borrowing and increasing the total amount of credit available. When debts rise at a faster rate than incomes this causes there to be debt bubbles, and when these bubbles finally burst it leads to defaults and restructurings of debts. Defaults and restructurings are deflationary in nature and so to offset that the Fed will then print money and use that money to purchase US government debt which ultimately affects the open-market rate of treasury bonds, this is called quantitative easing. When interest rates are lower, then that is a sign that we are in good economic times. Businesses will be able to borrow more at a lower rate without too much risk of default or restructuring and that money will help create jobs and strengthen the economy. When interest rates are higher it means that the Fed is much more cautious about future US growth and doesn’t want to run the risk of the economy overheating.



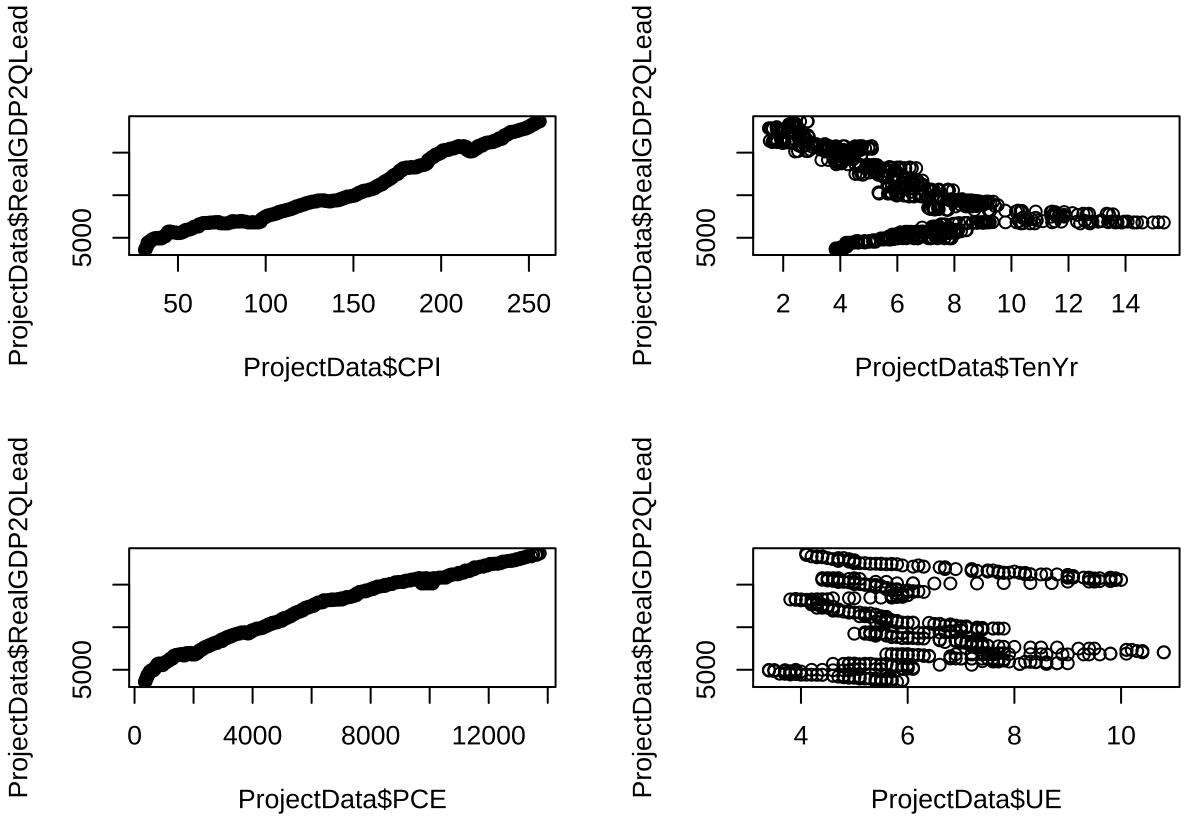


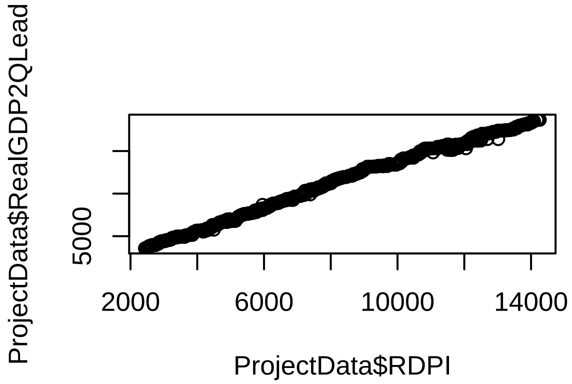
Here I used simple backwards elimination to look at the affect the variables I had picked would have on Real GDP. It makes some sense as to why the adjusted R-squared of this model is rounding up to 1, and that's because some of these variables, like Real GDP, are in a near constant growth period as the economy has grown over the past 60 years. Looking only at two variables, CPI and RDPI, they already give us an adjusted R-squared of something close to 1, so adding variables is only going to increase it from there. Based on the above plot, though it looks like the variable for the price of a barrel of Oil is not very statistically significant when it comes to predicting Real GDP, so I will not use it in my final model.

Section 4: Methods

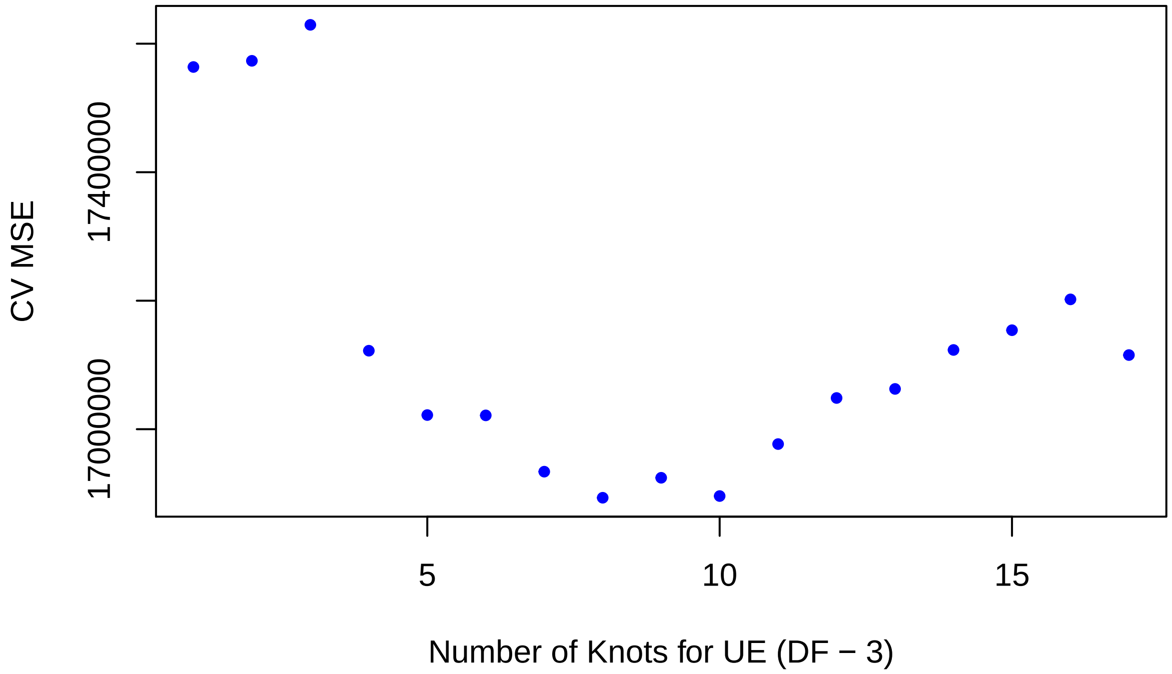
I decided to use two different methods for building a model to predict Real GDP. The first method I decided to use was a classic cross validation technique to try and make a good predictive model. I used the LOOCV method, which is an iterative process that removes one data point at a time and builds a model with the n - 1 data points, this process then repeats itself n times until every data point has been left out and used to build the predictive model. I wanted to use a fairly simple technique to start this off because the adjusted R-squared for the model without Oil found from best subset selection was already near 1. I thought that maybe a simple cross validation technique would work well with this data set considering how close to perfect the relationship between these variables already were. However, when running this technique I got a test MSE of 32,855.75, which as I will show later, is less than what I got when I ended up running a Generalized Additive Model (GAM) of basis splines. A GAM is a multivariate extension of a spline, meaning I can use different splines on different predictors in the model. A spline is a stat learning model that will fit a high order polynomial between the predictor and the response variable. I used the LOOCV method to test exactly how many knots I should put in my spline for each variable. The plots below show the number of knots to use compared with the test MSE given we use that many knots. To build my final GAM I used the number of knots for each variable that would help minimize the test MSE and combined the splines into one generalized additive model.

```
## [1] 32855.75
```

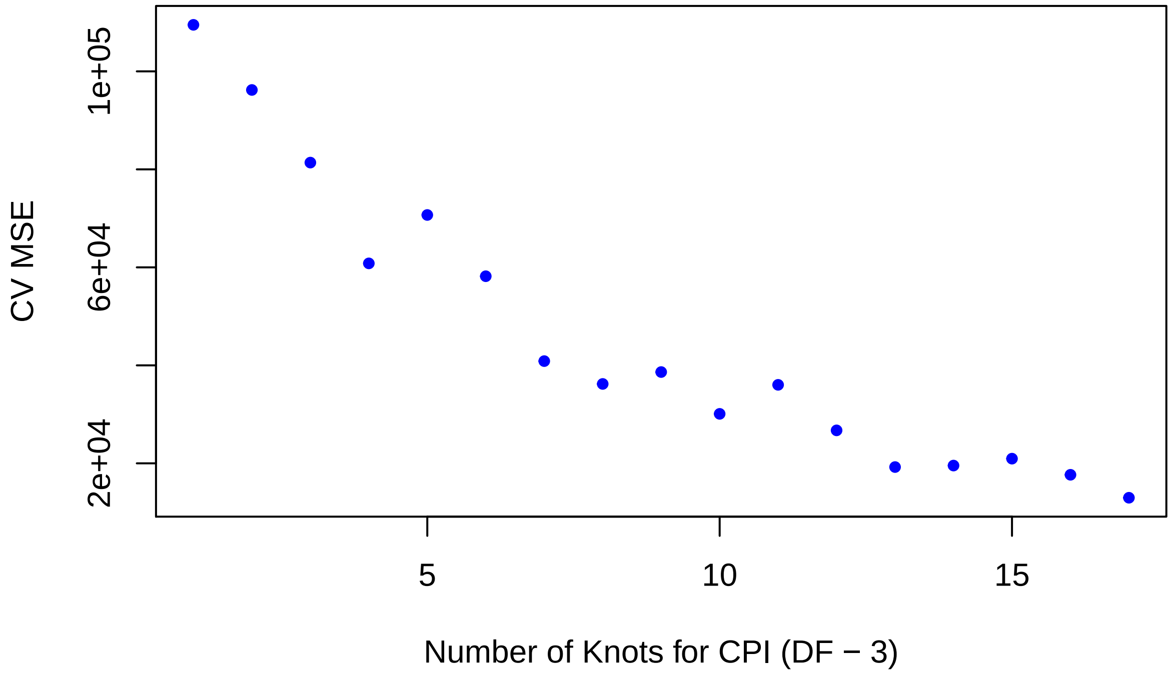




A lot of these variables seem to have pretty complicated relationships with Real GDP considering how cyclical they can be (like the TenYr rate and unemployment rate). This just further validates the thought that I had that these variables require a high order polynomial to compare to Real GDP. I will explore these relationships as I build a generalized additive model

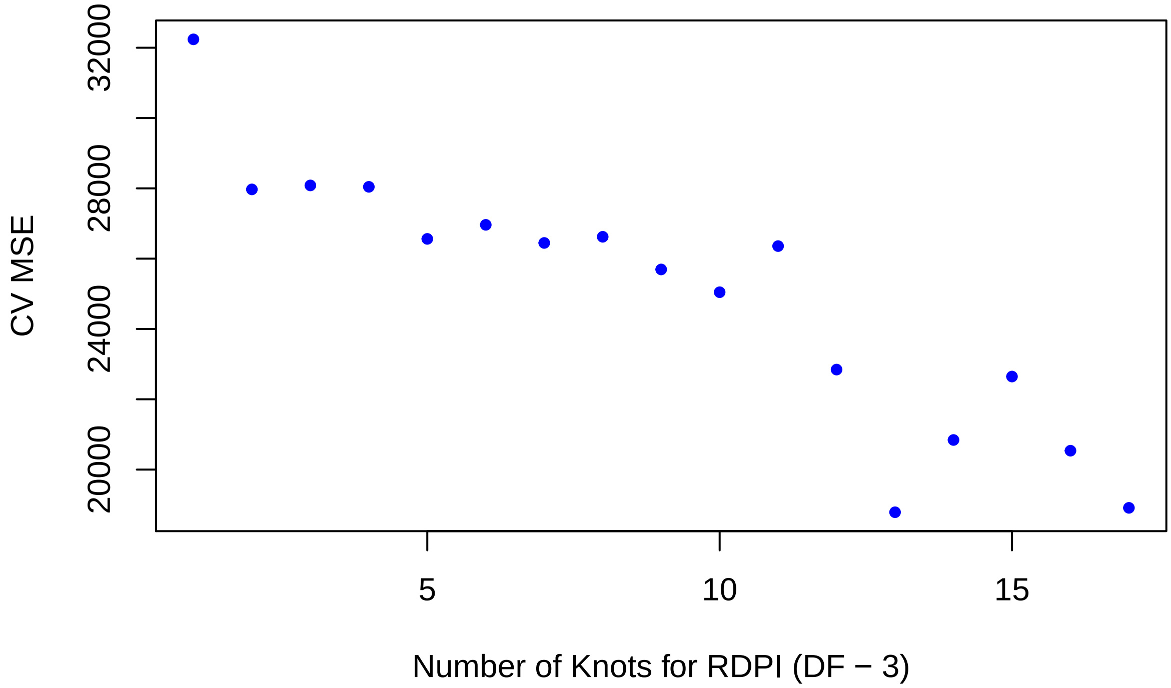


With a basis spline for the Unemployment rate I was only able to get an adjusted R-squared of .12. This isn't that surprising considering the plot between UE and the Real GDP.

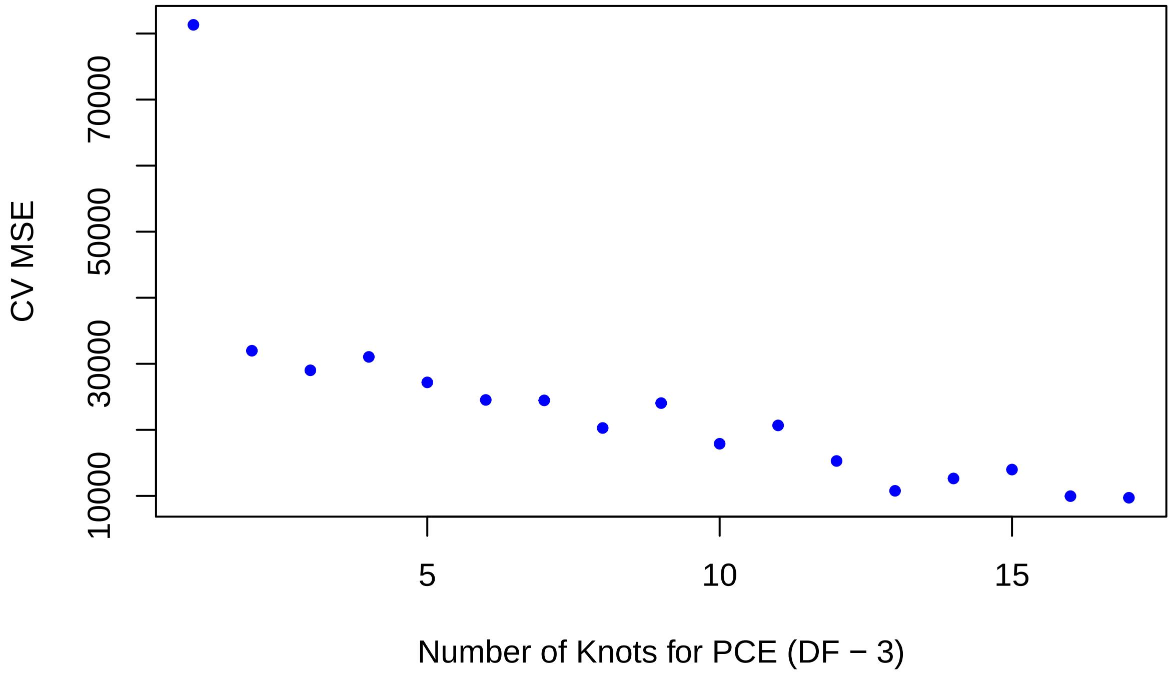


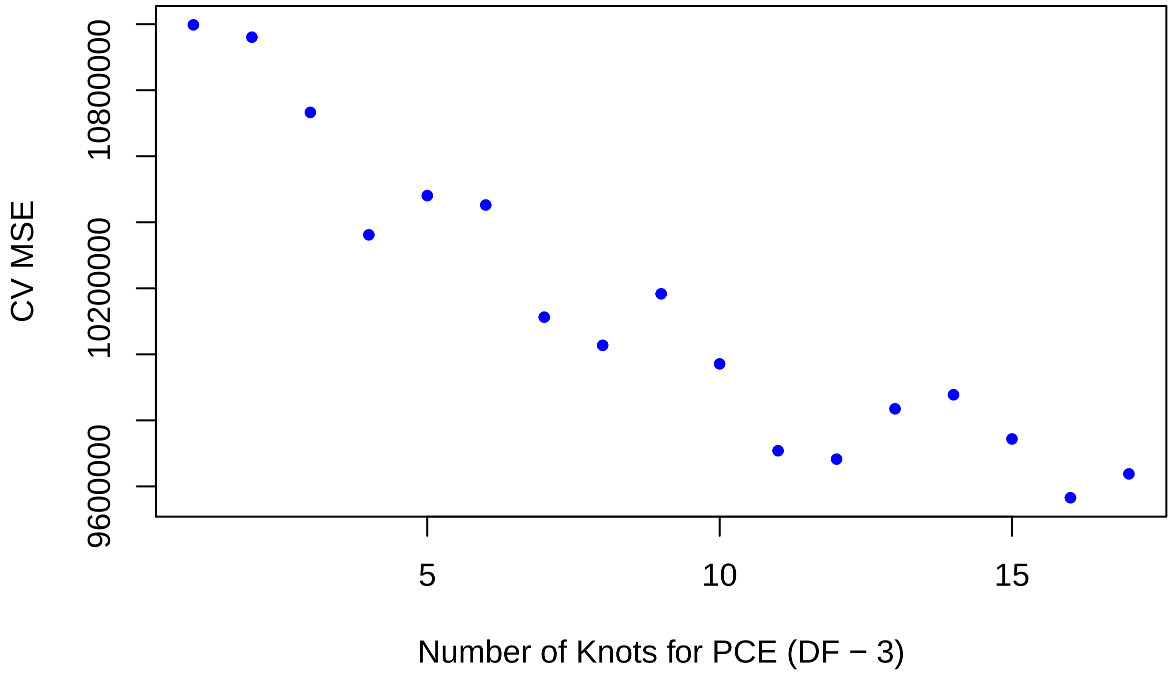
```
## [1] 17
```

Running a basis spline to find a relationship between CPI and Real GDP the basis spline gives an adjusted R-squared of .999. This is great because this relationship looked close to linear but there were a few distinct variations in the pattern.


CV MSE
20000
24000
28000
32000
5
10
15
Number of Knots for RDPI (DF − 3)

```
## [1] 13
```


CV MSE
10000
30000
50000
70000
5
10
15
Number of Knots for PCE (DF − 3)



With a basis spline for explaining Real GDP with the 10 year treasury note, the 10 year note was able to explain 53.3% of the variability.

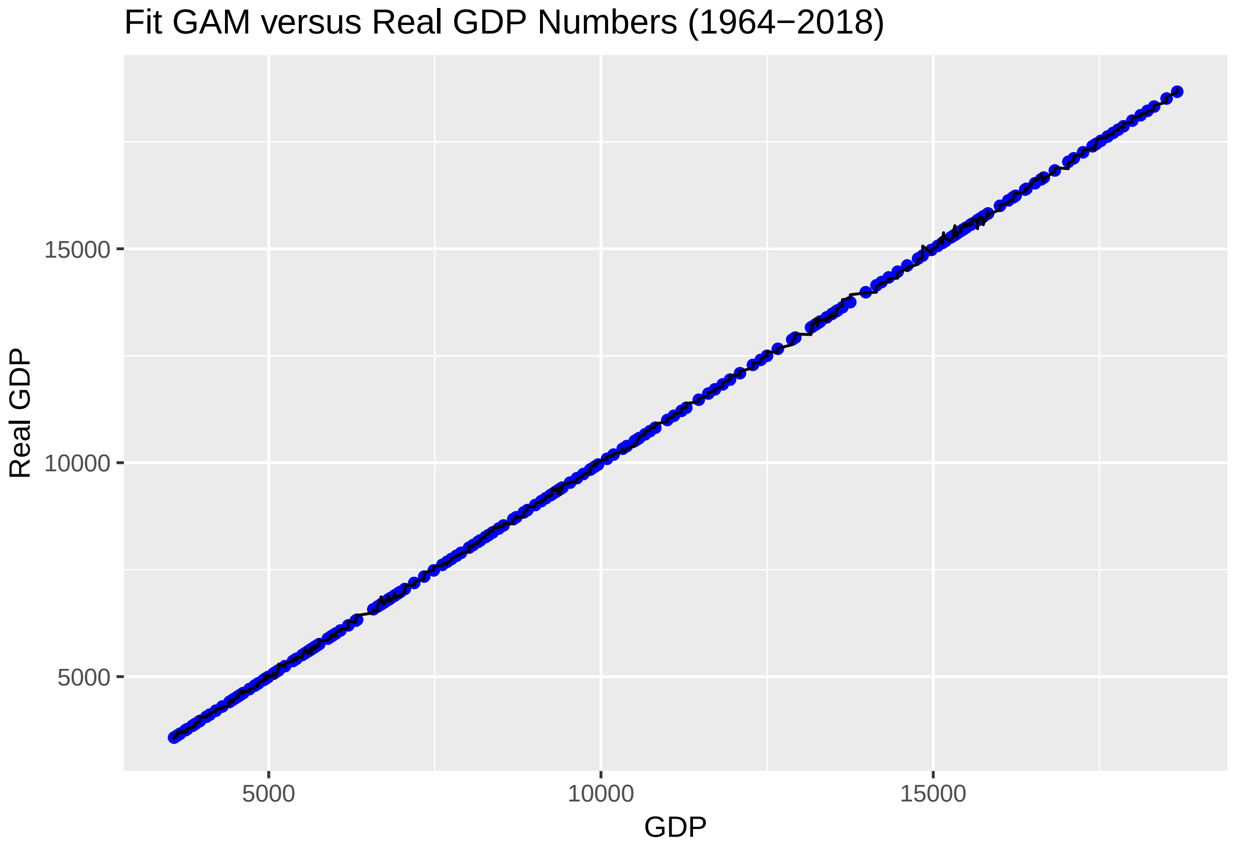
Once I ran LOOCV to figure out how many knots my basis spline should have for each variable, I decided to build a GAM full of different basis splines. I am going to use the GAM as my final model for a variety of reasons. The first and clearest reason is the more than 80% reduction in the test MSE. Another reason is that a lot of these variables do not have linear or linear-like relationships with Real GDP, so it would not make sense to use a method like LOOCV which was only creating a linear model for me. Once I was able to fit the GAM it had an adjusted R-squared of .999 and a test MSE of 5310. Overall I'm very happy how much my model has improved.

```
## Loading required package: foreach
```

```
## Loaded gam 1.16
```

```
## [1] 5310.018
```

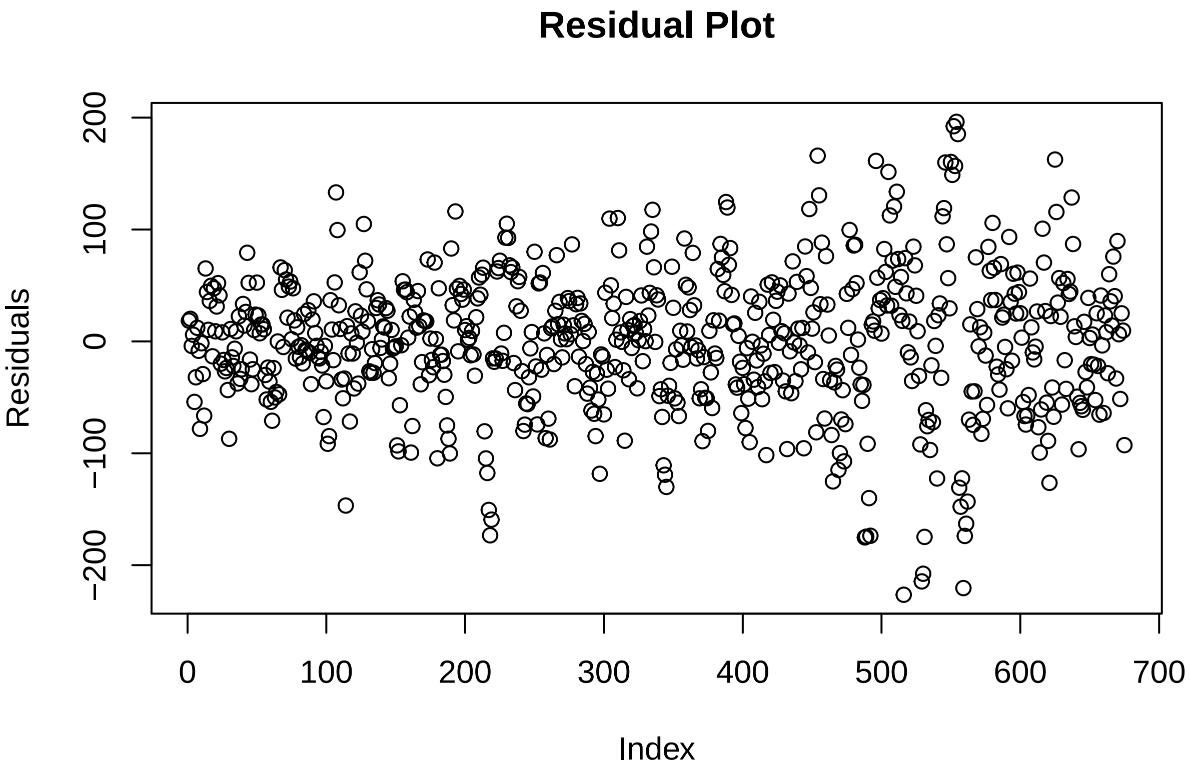
Running a GAM with the number of knots determined by LOOCV for each of the variables, I got a test MSE using LOOCV of only 5310. This is much better than my original LOOCV model or my spline model that I fit so I will stick with this model.



This is a plot of the US GDP over time compared to the GAM I used for UE, PCE, CPI, and RDPI. This model looks as if it is able to capture almost all of the variability in the Real GDP data and looks like it could potentially be a very good model to predict future GDP numbers.

Section 5: Results

After running my GAM and fitting it against the GDP Data, it appears that the model I have built is pretty robust. Normally a test MSE as high as this one at 5310 would worry me but considering the order of magnitude of our response variable a small percent error could result in a very high squared residual. Not only that, but considering my first model using LOOCV, the test MSE for the spline is nearly a sixth, meaning my model has made significant improvements coming from a model that had near perfect correlation to begin with. Despite the significant improvements in the test MSE of the model, the model leaves me a little disappointed because it's not the most easy to interpret. Since my model is a Generalized Additive Model with a number of basis splines it has over 80 different knots and as such over 80 coefficients. While this model may be robust and good at finding the relationship between my predictors and Real GDP, it isn't the easiest to actually go about calculating what the next GDP number would be if I were to have all the economic variables at the moment, which is really what I wanted to build with my model.



Despite that setback though I'd like to bring attention to the above residual plot. The residuals generally look like they have zero mean and the highest residuals go from -226.4 to 196.23 with a Median of 3.13. So on average this model comes within 3 points of calculating the Real GDP and at most it overpredicted by 226 and underpredicted by 196. I'd say that's a pretty strong model.

Section 6: Conclusion and Future Work

I am pretty happy with how well I was able to fit Real GDP with the variables I had selected. What I was really hoping for when this project started though was a concrete way to predict future GDP and in turn GDP growth of the United States. What I think I've ended up with is just an explanation of the relationship between GDP and the UE, CPI, RDPI, 10 Year rate, and PCE. Once this project was finished I really wanted to be able to give this model current economic data and based off of the 675 or so previous data points, would have liked it to just spit out an estimate of what the US GDP would be in two quarters. Unfortunately this is not the type of model that I have built and given more time and a little more knowledge on the subject I would have liked to build a model just like how I'm describing. I'm not disappointed with the results of my model, moreso with the functionality of it. Maybe in the future I will be able to build a model more like I was hoping to predict. Otherwise I'm happy with how well my model was able to explain almost all of the variability in Real GDP.

Going forward if I had a little more knowledge about GAMs and how to interpret them and their various knots I would have faired much better with my analysis. I would have liked to had a formula that I could just plug and chug current economic data. It's also important to keep this model updated by updating the economic data in my data set as time goes on as the economy is ever changing and these variables are constantly changing as well.

For completions sake I will include the summary output of my GAM below.

```
##
## Call:
## lm(formula = RealGDP2QLead ~ bs(UE, df = 11) + bs(RDPI, df = 16) +
##     bs(TenYr, df = 19) + bs(PCE, df = 20) + bs(CPI, df = 20),
##     data = ProjectData)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -226.442  -35.472    3.134   37.934  196.231
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3460.810     96.644  35.810  < 2e-16 ***
## bs(UE, df = 11)1       -71.076     63.990  -1.111 0.267137
## bs(UE, df = 11)2       117.569     62.928   1.868 0.062215 .
## bs(UE, df = 11)3         8.035     65.799   0.122 0.902846
## bs(UE, df = 11)4        80.077     64.795   1.236 0.217007
## bs(UE, df = 11)5        58.300     67.988   0.858 0.391515
## bs(UE, df = 11)6        35.957     69.875   0.515 0.607034
## bs(UE, df = 11)7        22.780     75.899   0.300 0.764178
## bs(UE, df = 11)8        62.106     77.492   0.801 0.423190
## bs(UE, df = 11)9       239.198     88.863   2.692 0.007310 **
## bs(UE, df = 11)10      204.349     97.808   2.089 0.037112 *
## bs(UE, df = 11)11      241.386    102.425   2.357 0.018765 *
## bs(RDPI, df = 16)1    -508.327    304.862  -1.667 0.095967 .
## bs(RDPI, df = 16)2     112.718    453.363   0.249 0.803737
## bs(RDPI, df = 16)3   -1085.857    582.434  -1.864 0.062772 .
## bs(RDPI, df = 16)4      43.497    591.505   0.074 0.941405
## bs(RDPI, df = 16)5    -182.080    578.371  -0.315 0.753014
## bs(RDPI, df = 16)6     629.788    633.010   0.995 0.320189
## bs(RDPI, df = 16)7    -210.329    646.048  -0.326 0.744872
## bs(RDPI, df = 16)8     626.341    666.150   0.940 0.347481
## bs(RDPI, df = 16)9      60.926    694.716   0.088 0.930146
## bs(RDPI, df = 16)10    157.788    756.686   0.209 0.834891
## bs(RDPI, df = 16)11   -235.080    782.118  -0.301 0.763849
## bs(RDPI, df = 16)12    638.797    771.564   0.828 0.408049
## bs(RDPI, df = 16)13    809.750    780.094   1.038 0.299690
## bs(RDPI, df = 16)14    591.724    788.602   0.750 0.453347
## bs(RDPI, df = 16)15    434.898    807.708   0.538 0.590480
## bs(RDPI, df = 16)16    548.804    855.445   0.642 0.521421
## bs(TenYr, df = 19)1     35.280     77.529   0.455 0.649232
## bs(TenYr, df = 19)2     21.095     53.666   0.393 0.694405
## bs(TenYr, df = 19)3    125.557     73.167   1.716 0.086681 .
## bs(TenYr, df = 19)4     63.428     63.586   0.998 0.318925
## bs(TenYr, df = 19)5     39.342     69.275   0.568 0.570309
## bs(TenYr, df = 19)6     95.921     70.894   1.353 0.176570
## bs(TenYr, df = 19)7    -83.371     73.233  -1.138 0.255402
## bs(TenYr, df = 19)8     -7.377     71.724  -0.103 0.918118
## bs(TenYr, df = 19)9     33.989     74.402   0.457 0.647964
## bs(TenYr, df = 19)10   -17.175     74.428  -0.231 0.817584
## bs(TenYr, df = 19)11    -5.963     76.148  -0.078 0.937612
## bs(TenYr, df = 19)12    18.832     77.369   0.243 0.807778
## bs(TenYr, df = 19)13     4.943     76.054   0.065 0.948204
## bs(TenYr, df = 19)14    42.441     77.808   0.545 0.585640
## bs(TenYr, df = 19)15    32.173     79.605   0.404 0.686239
## bs(TenYr, df = 19)16    16.867     86.165   0.196 0.844870
## bs(TenYr, df = 19)17    48.122    101.284   0.475 0.634882
## bs(TenYr, df = 19)18   140.715    100.629   1.398 0.162534
## bs(TenYr, df = 19)19   132.312    104.047   1.272 0.203997
## bs(PCE, df = 20)1      378.560    316.674   1.195 0.232403
## bs(PCE, df = 20)2      574.511    461.501   1.245 0.213673
## bs(PCE, df = 20)3     1536.875    571.673   2.688 0.007383 **
## bs(PCE, df = 20)4     3562.470    714.447   4.986 8.11e-07 ***
## bs(PCE, df = 20)5     2224.131    899.031   2.474 0.013645 *
## bs(PCE, df = 20)6     5610.749    889.672   6.307 5.61e-10 ***
## bs(PCE, df = 20)7     3826.996   1063.559   3.598 0.000347 ***
## bs(PCE, df = 20)8     6772.061   1077.285   6.286 6.34e-10 ***
## bs(PCE, df = 20)9     6149.479   1164.066   5.283 1.79e-07 ***
## bs(PCE, df = 20)10    8246.738   1227.195   6.720 4.30e-11 ***
## bs(PCE, df = 20)11    8914.288   1245.823   7.155 2.50e-12 ***
## bs(PCE, df = 20)12    8866.784   1361.621   6.512 1.59e-10 ***
## bs(PCE, df = 20)13   10419.240   1469.962   7.088 3.91e-12 ***
## bs(PCE, df = 20)14   12291.999   1522.931   8.071 3.94e-15 ***
## bs(PCE, df = 20)15   10509.425   1549.171   6.784 2.86e-11 ***
## bs(PCE, df = 20)16   13112.510   1612.510   8.132 2.52e-15 ***
## bs(PCE, df = 20)17   12438.007   1607.806   7.736 4.48e-14 ***
## bs(PCE, df = 20)18   14911.347   1666.648   8.947  < 2e-16 ***
## bs(PCE, df = 20)19   14434.171   1726.388   8.361 4.52e-16 ***
## bs(PCE, df = 20)20   15030.390   1733.122   8.672  < 2e-16 ***
## bs(CPI, df = 20)1       92.977    241.244   0.385 0.700077
## bs(CPI, df = 20)2      552.538    276.413   1.999 0.046073 *
## bs(CPI, df = 20)3      409.118    493.072   0.830 0.407026
## bs(CPI, df = 20)4     -879.396    563.613  -1.560 0.119231
## bs(CPI, df = 20)5     -372.316    736.657  -0.505 0.613458
## bs(CPI, df = 20)6    -2314.262    805.072  -2.875 0.004192 **
## bs(CPI, df = 20)7     -707.893    936.805  -0.756 0.450164
## bs(CPI, df = 20)8    -3318.007    978.964  -3.389 0.000748 ***
## bs(CPI, df = 20)9     -985.921   1084.343  -0.909 0.363600
## bs(CPI, df = 20)10   -3176.820   1116.470  -2.845 0.004590 **
## bs(CPI, df = 20)11   -2779.419   1142.682  -2.432 0.015298 *
## bs(CPI, df = 20)12   -1504.982   1201.882  -1.252 0.210999
## bs(CPI, df = 20)13    -632.132   1410.196  -0.448 0.654132
## bs(CPI, df = 20)14   -2331.011   1429.375  -1.631 0.103470
## bs(CPI, df = 20)15    1019.506   1492.933   0.683 0.494946
## bs(CPI, df = 20)16   -2641.253   1533.027  -1.723 0.085432 .
## bs(CPI, df = 20)17    -612.277   1572.074  -0.389 0.697069
## bs(CPI, df = 20)18   -1268.550   1572.195  -0.807 0.420070
## bs(CPI, df = 20)19    -500.072   1635.478  -0.306 0.759892
## bs(CPI, df = 20)20    -412.143   1709.853  -0.241 0.809608
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 66.28 on 588 degrees of freedom
## Multiple R-squared:  0.9998, Adjusted R-squared:  0.9998
## F-statistic: 3.593e+04 on 86 and 588 DF,  p-value: < 2.2e-16
```