MA 574

Regression Project

Kaiduo Zheng

Sahil Shahani

1. Give a brief description of gross domestic product(GDP).

   Gross Domestic Product is the broad measurement of nation's economic health. It is the monetary value of all goods and services produced within a country's border in a specific time period.[1]

   Difference between Nominal and Real GDP:

   Nominal GDP of a country is ambiguous, we cannot tell the growth of the economy. In a period of time, the prices of goods and services increases typically leading to increased GDP. We cannot distinguish if the increase in GDP is because of escalation in prices or result of production increase. Real GDP makes adjustment to the prices of goods and services with respect to a base year and then is usually smaller than Nominal GDP, indicating real growth of economy, excluding inflation. [1]

   There are two main methods in calculating GDP:

   1) Expenditure Approach: The Expenditure Approach/Spending Approach is the mula that is spent by different entities participating in the economy. This method takes into account totals consumption, private domestic investment, government spending and net exports. This is a most popular way to estimate GDP. [2]

   $$GDP = Personal\ Consumption + Investments + Govt.\ Spending + Net\ Exports$$

   2) Income Approach: "The income approach to measuring gross domestic product (GDP) is based on the accounting reality that all expenditures in an economy should equal the total income generated by the production of all economic goods and services." [3]

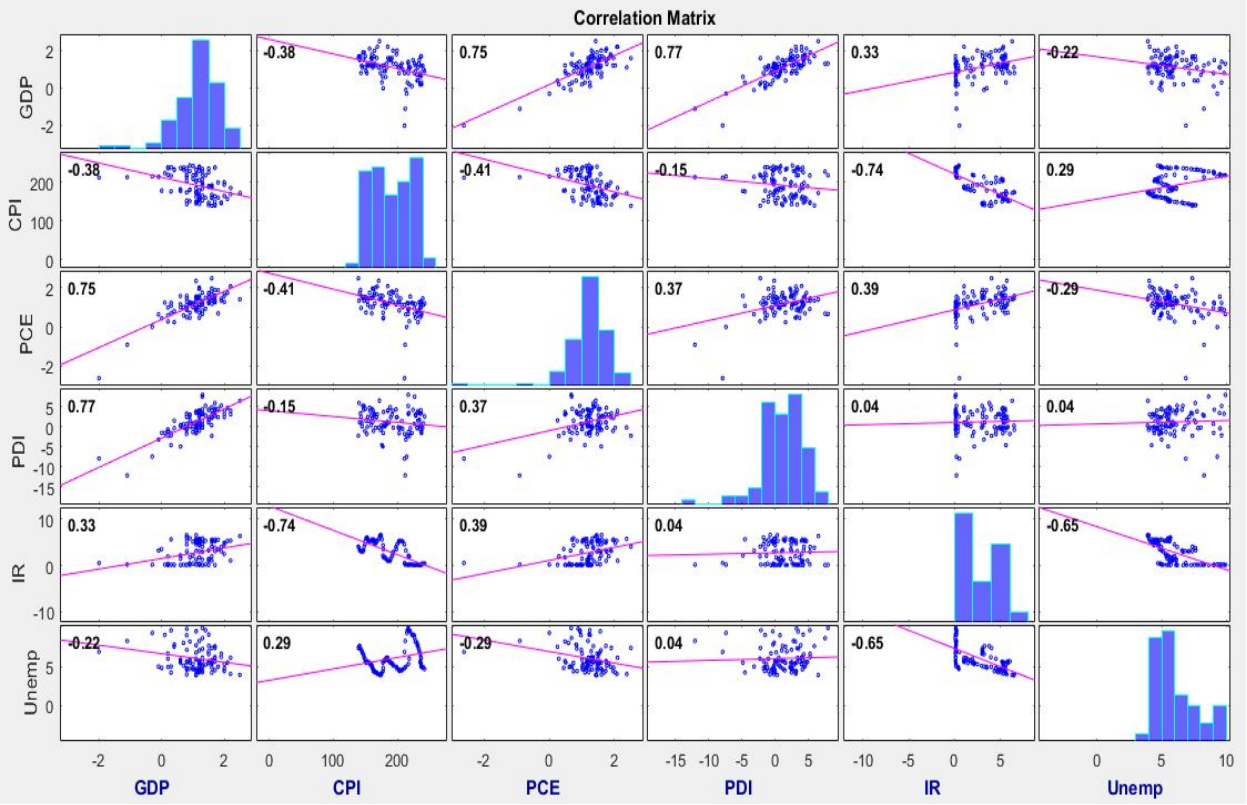
   $$GDP = Total\ National\ Income + Sales\ Tax + Depreciation + Net\ Foreign\ Factor$$

2. Use either MATLAB or Python to run a multiple linear regression of GDP onto a set of 4 to 5 predictor variables. Two of these predictors should be measures of inflation.

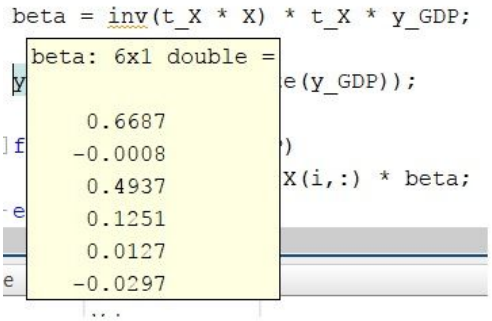
   We are using Nominal GDP quarterly data in our project from 1997-2016 i.e 100 data points. We have downloaded the data in Billion dollars from Bloomberg and calculated the period over period growth of Nominal GDP for 3 of our predictors.We downloaded the Consumer Price Index(CPI) and Interest rate (IR) as our measure of inflation in the economy from FRED (Federal Reserve Bank of Economic Data - St Louis) website.

   Following is the list of predictors we are using to predict Nominal GDP over next quarters.

   1) Personal Consumer Expenditures(PCE)
   2) Private Domestic Investment (PDI)
   3) Interest Rate (IR)
   4) Unemployment Rate (UER)
   5) Consumer Price Index(CPI)



Correlation Plot



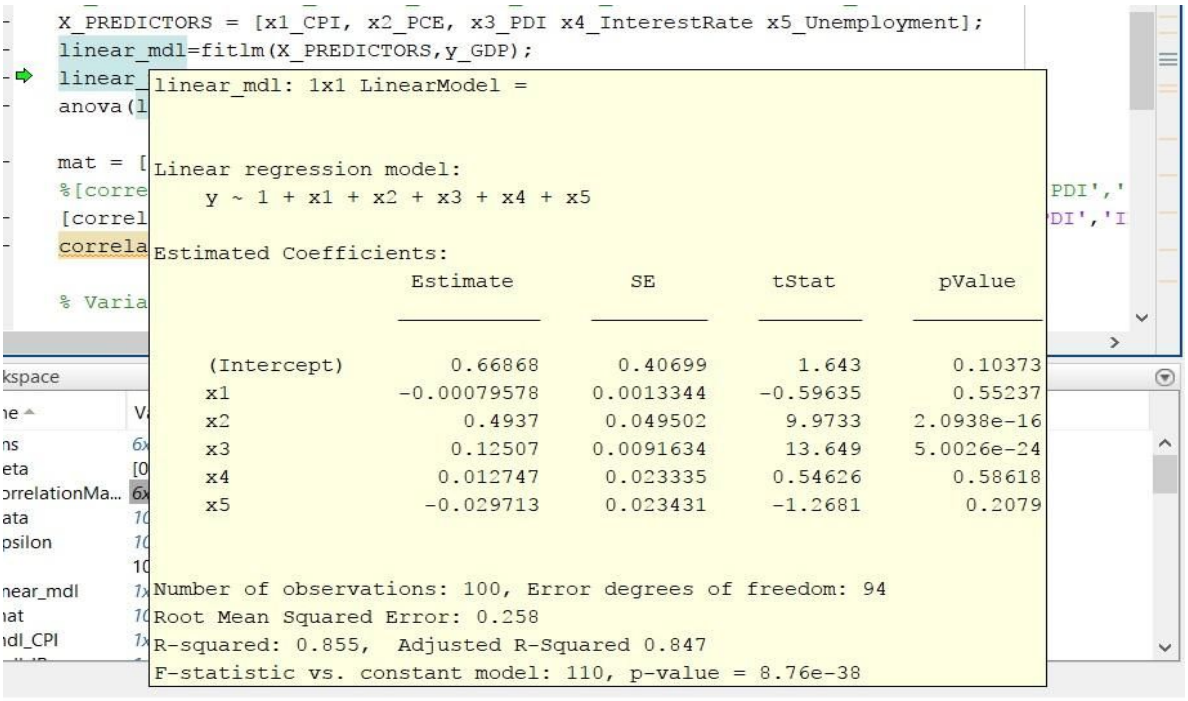
Using the Linear Algebra taught in class, we calculate the Beta vector. As we have 5 predictors in our original model, we have (p+1) predictors, that includes an intercept.

We have written our function *performRegression* which takes input arguments as X predictor matrix and response vector and uses the equation (1) to find predicted values. It calculates the estimators of beta using equation (2), along with intercept $\beta_0 = \mathrm{E}[\mathrm{Y}|\mathrm{X}=0]$.

```
function [y_GDP_hat, epsilon] = performRegression(X, y_GDP)
```

$$\hat{Y}_i = \hat{\beta}_i . X_i$$ ..... ……..... Eq.(1)

$$\hat{\beta} = (X^T . X)^{-1} X^T Y$$ …….. ……...Eq (2)



In the above picture, we also fit pre-programmed regression model to the dataset and verified out beta coefficients. The P-value associated with each predictor in the model tells us if the predictor is relevant.

As you can see the picture, x2 & x3 predictors (PCE and PDI) have insignificant p-values, thereby rejecting null hypothesis and concluding that PCE and PDI contribute in predicting Y - our target variable.

$\beta_2 = 0\ \beta_3 = 0$ ---------- ----------Null Hypothesis

We also verify the estimators of beta and fitted values in the model and they are exactly the same as we found using our function *performRegression.*

Root Mean Squared Error: 0.258. R-squared: 0.855. Adjusted R-Squared: 0.847

The R-squared value and the adjusted R-Square value are relatively large enough to say the model is good and useful.

3. Run an analysis of variance (ANOVA) on your regression. Comment on the values in your ANOVA. What can you conclude from this analysis? Is this regression useful?

ANOVA is partition of the variance in Y into 2 parts.

$$(Y_i - \bar{Y}_i) = (\hat{Y}_i - \bar{Y}_i) + (\hat{Y}_i - Y_i)$$ .....Equation (3)

Total Variance = Sum of squares of Variance Explained by the Regression Model + Sum of squares of Residual Variance.

TSS = RSS + SSE .....Abbreviated Form

$(Y_i - \bar{Y}_i)$ .....Total Variance

$(\hat{Y}_i - \bar{Y}_i)$ .....Regression Variance

$(\hat{Y}_i - Y_i)$ .....Residual Variance

ANOVA TABLE:

```
sum_RSS =

   36.7563


sum_SSE =

    6.2537


total_variance =

   43.0100
```

The sum of square of variance explained by the model is 36.7563, giving R-squared 0.855.

R-squared = RSS/TSS = 36.7563/43.01 = 0.855

As we know higher the R-squared and lower the residual variance, the better is the model.

From the adjacent picture, we can see that we have unexplained(Residual variance) very low (6.2537).

So we can say that our regression is useful.

4. For each of the predictor variables, compute the variance inflation factor, VIFj . Based on these values, should you eliminate any predictor variables ?

"Variance Inflation factor for each predictor tells us how much the variance of its $\hat{\beta}_i$. is increased by having other predictor variables in the model" [4].

It basically tells us the multicollinearity between variables. It removes redundancy, if two predictors are collinear and have very high VIF values, means that there is redundancy i.e VIF >=4, calls for elimination of those variables.

Methodology:

1) We regress each predictor on other predictor variables and fit a linear regression model to this data.

2) The we calculate R-Squared of the model.

3) VIF = 1/(1-R-Squared)

4) Heuristic Rule VIF >= 4. Eliminate it

Results:

```
vif_CPI =
    2.7380

vif_PCE =
    1.4624

vif_PDI =
    1.2044

vif_InterestRate =
    4.1185

vif_Unemployment =
    2.0832
```

Here we have created 5 functions:

VIF_CPI(X_CPI_Regress, x1_CPI) - Accept

VIF_PCE(X_PCE_Regress, x2_PCE) - Accept

VIF_PDI(X_PDI_Regress, x3_PDI) - Accept

VIF_InterestRate(X_Interest_Rate_Regress, x4_InterestRate) - Eliminate

VIF_Unemployment(X_Unemployment_Regress, x5_Unemployment) - Accept

5) For all the different combinations of the remaining predictor variables,compute the Cp of each regression. Comparing these values, what suggestions would you make about which predictor variables to use in your multiple linear regression.

Remaining Predictors:

1) Consumer Price Index (CPI)

2) Personal Consumer Expenditure (PCE)

3) Private Domestic Investment(PDI)

4) Unemployment Rate.

Cp statistic is a model selection criteria just like AIC, BIC . Basically, it tells us how well a model predicts. A model with smallest Cp value is preferred.[4]

Methodology

1) We use different combinations of the above 4 predictors and obtain 15 models and their Cp Statistic.
2) Cp = SSE(p) / Var(epsilon_M) - n + 2*(p+1) where epsilon_M $\sigma^2_{\epsilon,M}$ =

Results:

	A	B
1	p=1	
2	Cp_stat_CPI	488.2408
3	Cp_stat_PCE	203.0006
4	Cp_stat_PDI	180.2552
5	Cp_stat_UE	550.413
6		
7	p=2	
8	Cp_stat_CPI_PDI	134.5913
9	Cp_stat_CPI_UE	480.4067
10	Cp_stat_PDI_UE	136.6033
11	Cp_stat_PCE_UE	204.9776
12	Cp_stat_CPI_PCE	201.3926
13	Cp_stat_PCE_PDI	14.2445
14		
15	p=3	
16	Cp_stat_CPI_PCE_Unemploymer	203.342
17	Cp_stat_CPI_PDI_Unemploymen	112.5911
18	Cp_stat_PCE_PDI_Unemploymer	9.6001
19	Cp_stat_CPI_PCE_PDI	12.276
20		
21	p=4	
22	Cp_statistic_Final	9.3143

SSE(p) : Sum of squares of residual variance.

p - no of predictors (diff. subsets)

n - number of data points.

M - all predictors

We want SSE(p) to be small as possible.

In our 4- predictor model, SSE(4) seems to be very low, enabling Cp to be 9.3143 which is close to (p+1) predictors [4].

6) For each observation in the data set, compute the leverage Hii. For the n total observations plot the leverage values ? Are there any observations that have concerning leverage values ?

Definition:

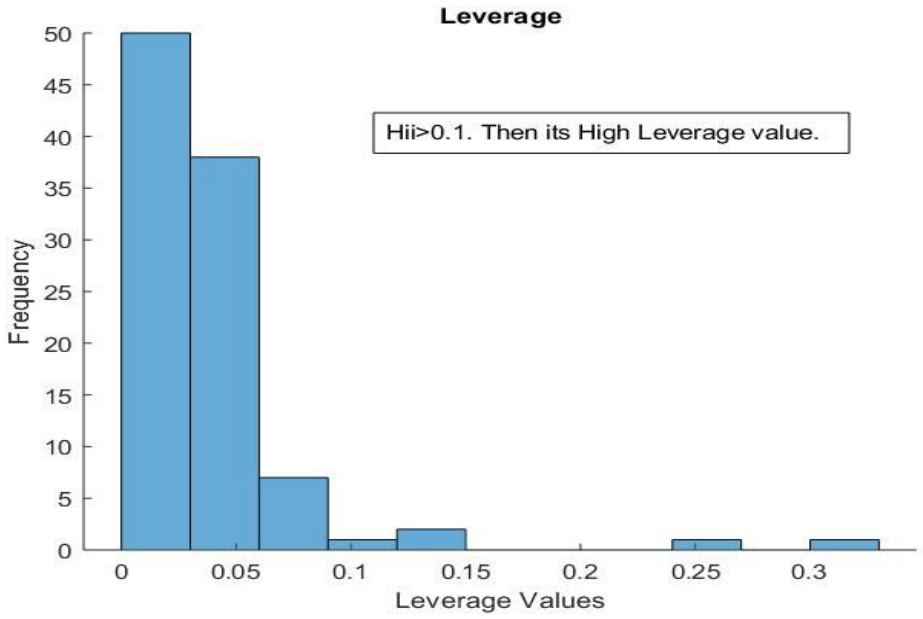
"Leverage, of i^{th} observation denoted by H_{ii} is a measure of how much influence Y_i has on its own fitted value"[4]. The higher the leverage value, the more accurately Y_i is predicted. Hence we will have low residual (e_i) and almost unbiased prediction.

Methodology: *function leveragePoints()*

1) Find the hat matrix, using this $X(X^TX)^{-1}X^T$ formula.

2) Find the leverage vector (H_{ii}) which is the diagonal of hat-matrix.

3) Heuristic Value: 2*(p+1)/n



C	D	E
Leverage	**L>0.1**	**Index**
0.144746	1	1
0.301445	1	68
0.241539	1	69
0.137147	1	70
0.118439	1	72

Results

These are the leverage values for data points 1,68, 69,70 and 72.

7) For each observation compute the raw residual, the studentized residual,and the externally studentized residual. Plot these over time. Compare and discuss your results. Are there any observations that are causes for concern ?

Raw Residual: Raw residual is just actual minus forecasted value.[4]

$$\hat{\epsilon} = Y_i - \hat{Y}_i \quad \text{. ……..,..... Equation (4)}$$

Studentized Residual: This residual overcomes the shortcomings of standardized residual of fixed standard error. [4]

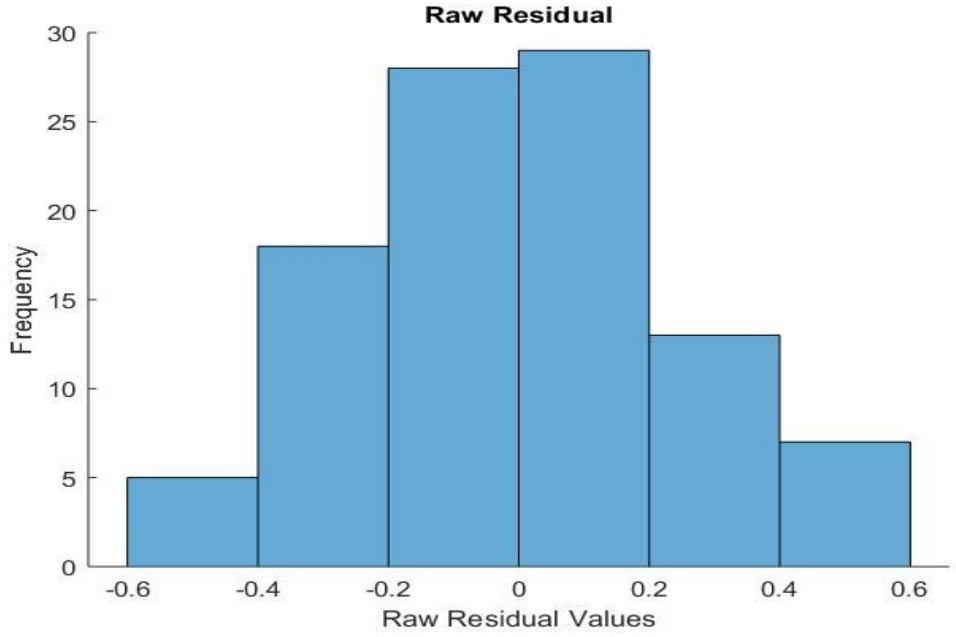
$$\epsilon_i / \; s\sqrt{\hat{1 - H_{ii}}} \quad \text{……..Equation (5)}$$

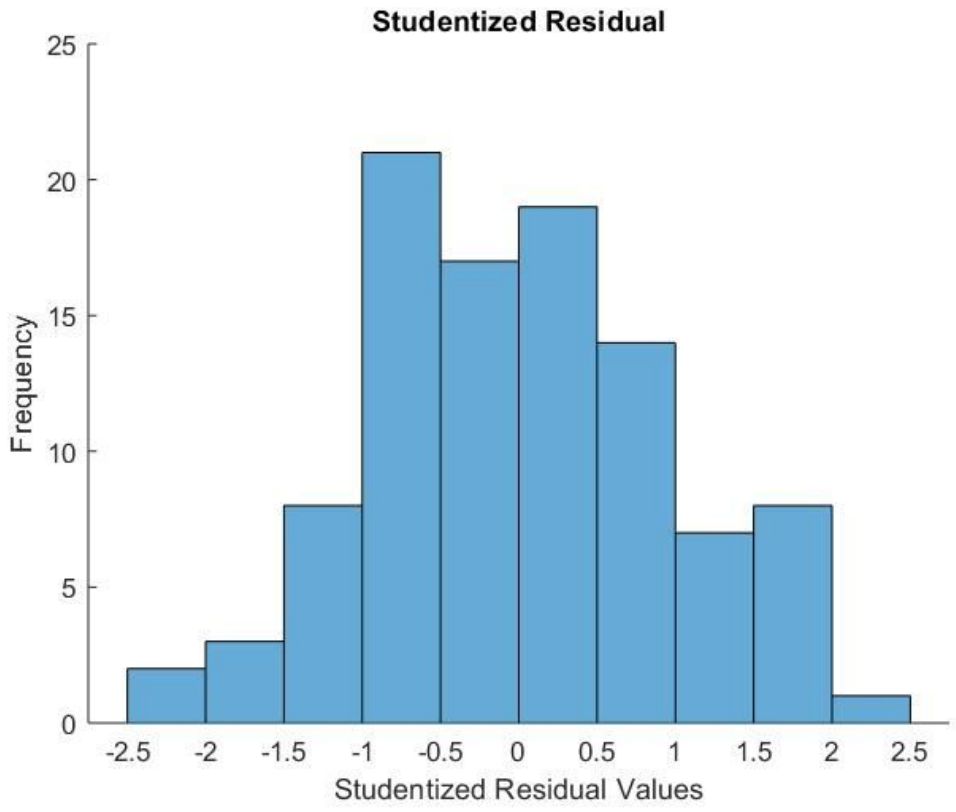
External Studentized Residual / RStudent: "The studentized residual of ith data point that does not use the ith data point" [4]

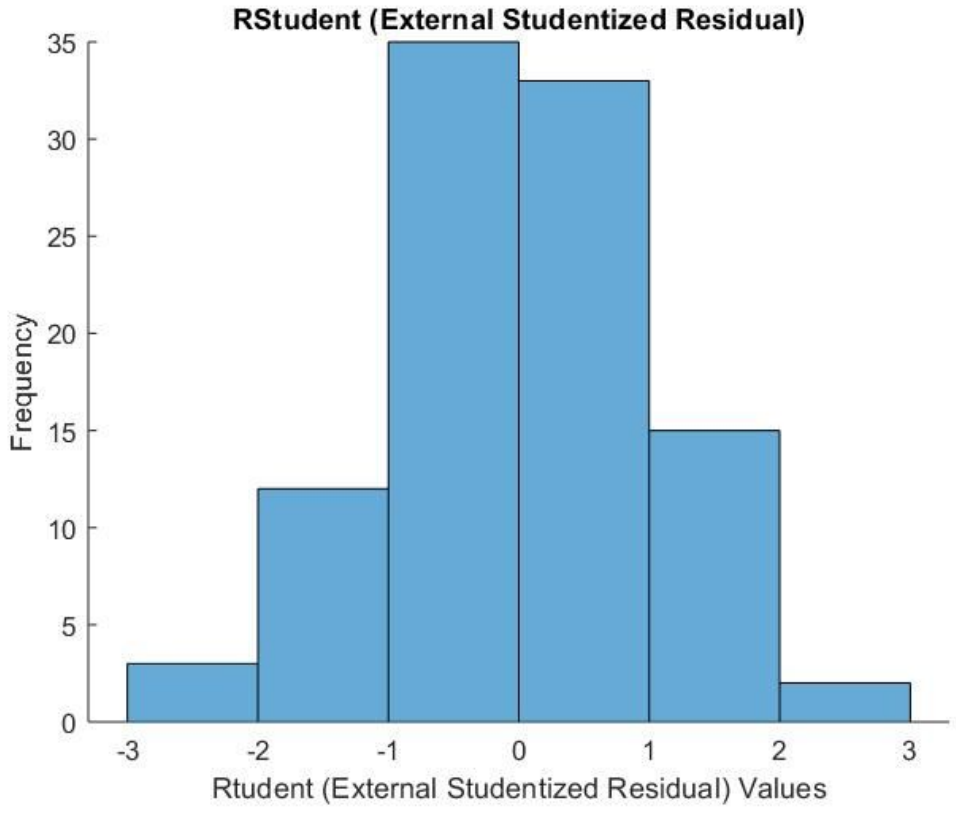
$$\hat{\epsilon_i} / \; s(-i)\sqrt{1 - H_{ii}} \quad \text{.....Equation (6)}$$

Methodology :

1) We remove the ith data point and fit model to the remaining data.
2) We find the residual vector of fitted values against the ith removed dataset.
3) Using Equation (6) we find the RStudent Value




Studentized Residual
Frequency
25
20
15
10
5
0
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5
Studentized Residual Values


RStudent (External Studentized Residual)
Frequency
35
30
25
20
15
10
5
0
-3
-2
-1
0
1
2
3
Rtudent (External Studentized Residual) Values

	A	B	C	D
1	Rstude	InternalSRR	Levera	L>0.1
2	1.098209	0.93309046	0.144746	1
69	-0.09406	-0.100478688	0.301445	1
70	2.835159	2.072371436	0.241539	1
71	1.731861	1.508193021	0.137147	1
73	-0.94453	-0.778863835	0.118439	1
102				

RStudent values for leverage points

Results:

The raw residuals have no absolute value greater than 0.6.
Data Point 69 could be outlying as it has a absolute value of greater than 2 for Internal Studentized Residual.
The external studentized residual also indicates potential problem with data point 69.

Below is the picture of predictor and response variables of leverage points. We can see that data point 69 is outlier as it has high predictor and response values

	A	B	C	D	E	F	G
1	Date	Interest rate	CPI	PCE	PDI	GDP	Unemployment
2	31-03-1992	4.02	139.1	2.5	2.25	1.6	7.37
68	30-09-2008	1.94	218.877	0.29	-2.3	0.2	6
69	31-12-2008	0.51	211.398	-2.62	-7.91	-2	6.87
70	31-03-2009	0.18	212.495	-0.91	-12.08	-1.1	8.27
71	30-06-2009	0.18	214.79	0	-7.41	-0.3	9.3
72	30-09-2009	0.16	215.861	1.24	-2.16	0.3	9.63
74							

The blue highlight is our model and red (Interest Rate) is predictor we have eliminated.

Residual Analysis, reflect the **Global Financial Crisis of 2008 .** In 2008, 4th quarter, the economy suffered huge losses because of the housing market bubble, the stock markets collapsed, recession started (Unemployment rate spiked to 6.87 and then 8.27), banks and other big financial institutions were bailed out. As you can see reducing the interest rates also did not work and could save the economy from collapsing. [5]

8) For each observation, compute the Cook's Distance. Plot these over time. Are there any cases that are a cause for concern?

"A high leverage value and high RStudent value indicate potential problems with the data set but not how much influence a data point has on its estimates. Cook's Distance tells how much the fitted value changes if ith observation is deleted"

$$\frac{\sum_{j=1}^{n}(\hat{Y}_j - Y_j(\hat{-i}))}{(p+1)s^2}$$ …….. Equation (7)

Methodology:

1) We remove the ith data point and fit model to the remaining data.
2) We find the residual vector of fitted values against the ith removed dataset.
3) Using Equation (7) we find the Cook's Distance Value

Result:

C	D	E
Leverage	L>0.1	Cook's Distance
0.144745716	1	0.5048511
0.301444536	1	37.71099834
0.241539456	1	42.48468287
0.137146643	1	44.89311683
0.118439433	1	47.17512502

Yes, data point 69, 70 and 72 appears to be cause of concern as they have high cook's distance implying that deleting this data point from data set changes the fitted values of other points.

9) Based on your results and your ANOVA what conclusions and recommendations can you make?

Based on ANOVA results, we can conclude that the regression is useful. As R-squared is 0.854, we can say that most of the variance is explained by the model. Adjusted R-squared value is 0.844.

b) Based on your work choose a multiple linear regression model to run and compare it to your regression from Question 2.

We selected a model with these predictors:

1) Consumer Price Index (CPI)

2) Personal Consumer Expenditure (PCE)

3) Private Domestic Investment(PDI)

4) Unemployment Rate.

We eliminated the following predictor:

1) Interest Rate

We selected the predictors based on Variance Inflation factor which describes the multicollinearity between predictors. $VIF_{Interest\ Rate} >= 4$. There was certainly redundancy as Consumer Price Index (CPI) and Interest Rate are correlated.

In the economy, when inflation is high, governments increase the interest rate to curb the inflation and when inflation is low, governments decrease the interest rates to stimulate the economy.

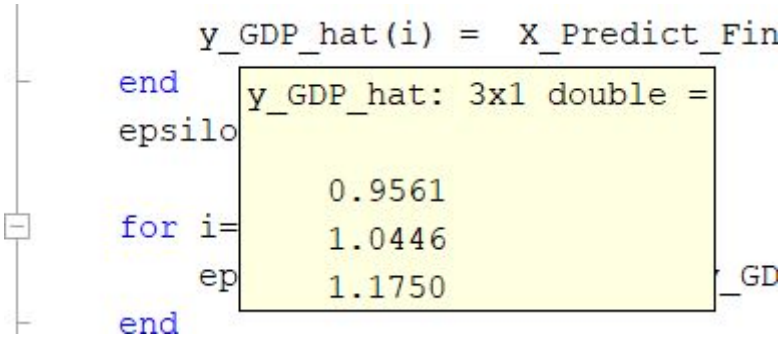
c) Make sure to explicitly justify your choice and explicitly give the regression coefficients.

Regression Coefficients:
Intercept: 0.8567
X1 = 0.00133
X2 = 0.495
X3 = 0.125
X4 = -0.0382

Although our regression coefficients show that X1 (Consumer Price Index(CPI)) and X4 (Unemployment) have there estimators close to 0, with p-values less than 0.05, but our Variance Inflation Factor for these values is less than 4. Hence we selected these variables in our model as it will be loss of information if we don't select them.

10) Use your selected regression model to predict the three GDP data points for 2017 using a rolling window.

The next 3 data points of GDP (Q1, Q2,and Q3 2017) predicted by our model are:



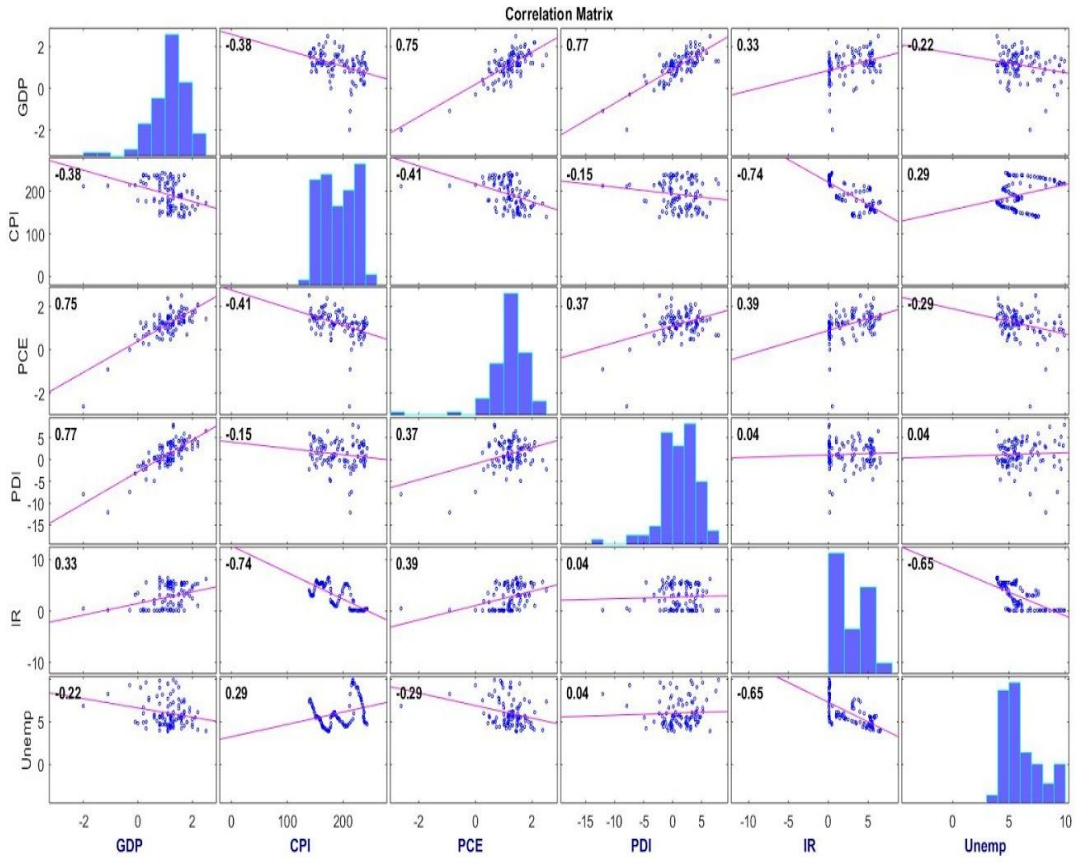
The actual data values

	A	B	C	D	E	F	G	H
1	Interest Rate	CPI	PCE	PDI	GDP	UE	Predicted	MAPE
2	0.7	243.752	1.03	0.082	0.805	4.7	0.956092	0.382692
3	0.95	243.79	0.875	1.58	1.01	4.4	1.044582	0.02424
4	1.15	246.373	1	1.98	1.28	4.3	1.17501	0.362023
5								0.768955

b) How does your regression perform?

We obtained a prediction accuracy of 77% on our prediction of 2017 GDP values, calculated using Mean Absolute Percentage Error (MAPE).

c) Provide graphs or plots to illustrate this.



The correlation plot shows linear relationship GDP and PCI, GDP and PDI and shows that Unemployment and CPI are negatively correlated to GDP, which seems reasonable.

When unemployment is going down, GDP is high as a result of high employment.

d) What conclusions can you make?

In conclusions, we can say that our model with 4 predictors is the best model as it includes all information required in predicting the GDP.

Reference:

[1] https://www.investopedia.com/terms/g/gdp.asp

[2] https://www.investopedia.com/terms/e/expenditure-method.asp

[3] https://www.investopedia.com/terms/i/income-approach.asp

[4] Ruppert, David. (2004). Statistics and Finance: An Introduction. Ithaca, NY: Springer

[5] https://www.thebalance.com/2008-financial-crisis-timeline-3305540