MAT 3672 Project 1 Regression Immersion

Le' Sean Roberts

AN IDEA OF REGRESSION

Regression is identified as a statistical method applied in various fields or industries to identify strength and/or character between a dependent variable and a collection of other variables being independent variables. The dependent variable is studied under the idea or belief of some law or rule or model based on independent variables. Idealistically, the independent variables aren't believed to be to dependent on each other or any other variable in the scope of the experiment in question. The dependent variable (DV) whose variation is observed w.r.t. to altering the inputs for the independent variables (IVs) - regressors being statistics jargon. Exploratory Data Analysis (EDA) is applied to confirm decent relations between the DV and IVs to determine whether choices for the IVs are practical; the (Pearson) correlation measure is a primitive method of variable selection. For the assumed condition that the IVs are not considerably dependent on each other, applying EDA, or applying the (Pearson) correlation measure in particular for confirmation. EDA can be very crucial for the 6-point process in statistical analysis. The (Pearson) correlation measure to be accompanied by multiple hypothesis tests data results for evaluation of the chosen models. Hypothesis Testing can possibly be applied to determine the relevance or significance of the IVs upon the DV, and as well serve as indicators of good or poor model fits. This project concerns applying both EDA and hypothesis tests to determine significance of IVs and wellness of model fits.

Two Unemployment Rate models are considered, namely, investigation of a bivariate regression model and a multivariate regression model:

1. The identification and analysis of a predictor or IV for Unemployment Rate, namely, Economic Growth in the bivariate case. Economic growth is the (positive) change in the Gross Domestic Product (GDP); recognizing GDP as the measure of the total monetary or market value of all finished goods and services produced within a country's borders in a specific time period. An increase in the output of goods and services can often be identified with a strong labor market. Identifying two hypotheses –

$$H_0 : \ Economic\, Growth\, is\, a\, decent\, predictor\, for\, Unemploment\, Rate$$

$$H_1: \ Economic\, Growth\, is\, not\, a\, decent\, predictor\ for\, Unemploment\, Rate$$

2. The identification and analysis of predictors or IVs for Unemployment Rate, namely, Economic Growth, Inflation and the Fed Funds Rate.
Economic growth is the (positive) change in the Gross Domestic Product (GDP); recognizing GDP as the measure of the total monetary or market value of all finished goods and services produced within a country's borders in a specific time period. An increase in the output of goods and services can often be identified with a strong labor market.
Inflation is the rate at which prices for goods and services rise, which can be translated as the decline in purchasing power. High inflation can lead to a slow down in the money supply (via credit, liquidity, consumer spending, firms' projects/investments, etc.). Thus, employment can be influenced.
The Fed Funds Rate is the target interest rate set by the Federal Open Market Committee of the Federal Reserve; as part of monetary policy, a target rate at which commercial banks borrow and lend their excessive reserves to each other overnight. The feral funds rate can influence short-term rates on consumer loans/mortgages and credit cards, and the stock market as well. In return influence on the money supply has effects on capital, liquidity and credit/debt of firms which in turn influences employment. The service industry in the United States of America has considerable weight in the generation of goods and services.
Identifying two hypotheses –

$$H_0: \ Indentified\, IVs\, are\, decent\, predictors\ for\, Unemploment\, Rate$$

$$H_0: \ Indentified\, IVs\, are\, poor\, predictors\ for\, Unemploment\, Rate$$

Data Wrangling & Exploratory Data Analysis

Restating the idea that EDA can be applied to confirm decent relations between the DV and IVs to determine whether choices for the IVs are practical; the correlation method is a primitive means of variable selection.

BIVARIATE CONTEXT - Starting with context, namely, stating a null hypothesis and a alternative hypothesis. Beginning with a bivariate model for Economic Growth. The null hypothesis to be that Economic Growth is highly influenced by Employment Rare; the alternative hypothesis to be that Economic Growth isn't highly influenced by the Employment Rate.

BIVARIATE OBSERVATIONS - Data sets for Growth and the Employment Rate to be assimilated for observation and analysis.

```
library(tidyverse)
library(tidymodels)
library(FactoMineR)
library(stats)
```

```
# Incorporating USA GDP
library(readxl)
USA_Growth <- read_excel("C:/Users/verlene/OneDrive/Desktop/USA GDP Growth 1961-2021.xlsx"
head(USA_Growth)
```

```
# A tibble: 6 x 4
   Year GDP          `GDP per Capita`  Growth
  <dbl> <chr>                   <dbl>   <dbl>
1  2021 $22,996.10B             69288  0.0567
2  2020 $20,893.74B             63028 -0.034
3  2019 $21,372.57B             65095  0.0229
4  2018 $20,527.16B             62805  0.0292
5  2017 $19,479.62B             59915  0.0226
6  2016 $18,695.11B             57867  0.0167
```

```
USA_Growth
```

```
# A tibble: 61 x 4
    Year GDP          `GDP per Capita`  Growth
   <dbl> <chr>                   <dbl>   <dbl>
 1  2021 $22,996.10B             69288  0.0567
 2  2020 $20,893.74B             63028 -0.034
 3  2019 $21,372.57B             65095  0.0229
 4  2018 $20,527.16B             62805  0.0292
 5  2017 $19,479.62B             59915  0.0226
 6  2016 $18,695.11B             57867  0.0167
 7  2015 $18,206.02B             56763  0.0271
 8  2014 $17,550.68B             55124  0.0229
```

```
 9  2013 $16,843.19B           53291  0.0184
10  2012 $16,253.97B           51784  0.0228
# i 51 more rows
```

```r
# Incorporating the dependent variable
library(readr)
index <- read_csv("C:/Users/verlene/OneDrive/Desktop/index.csv")
head(index)
```

```
# A tibble: 6 x 10
   Year Month Day   `Federal Funds Target Rate` `Federal Funds Upper Target`
  <dbl> <chr> <chr>                       <dbl>                        <dbl>
1  1954 07    01                             NA                           NA
2  1954 08    01                             NA                           NA
3  1954 09    01                             NA                           NA
4  1954 10    01                             NA                           NA
5  1954 11    01                             NA                           NA
6  1954 12    01                             NA                           NA
# i 5 more variables: `Federal Funds Lower Target` <dbl>,
#   `Effective Federal Funds Rate` <dbl>, `Real GDP (Percent Change)` <dbl>,
#   `Unemployment Rate` <dbl>, `Inflation Rate` <dbl>
```

```r
index
```

```
# A tibble: 904 x 10
    Year Month Day   `Federal Funds Target Rate` `Federal Funds Upper Target`
   <dbl> <chr> <chr>                       <dbl>                        <dbl>
 1  1954 07    01                             NA                           NA
 2  1954 08    01                             NA                           NA
 3  1954 09    01                             NA                           NA
 4  1954 10    01                             NA                           NA
 5  1954 11    01                             NA                           NA
 6  1954 12    01                             NA                           NA
 7  1955 01    01                             NA                           NA
 8  1955 02    01                             NA                           NA
 9  1955 03    01                             NA                           NA
10  1955 04    01                             NA                           NA
# i 894 more rows
# i 5 more variables: `Federal Funds Lower Target` <dbl>,
#   `Effective Federal Funds Rate` <dbl>, `Real GDP (Percent Change)` <dbl>,
#   `Unemployment Rate` <dbl>, `Inflation Rate` <dbl>
```

```r
index_ascend <- index[order(-index$Year), ]
index_ascend
```

```
# A tibble: 904 x 10
    Year Month Day   `Federal Funds Target Rate` `Federal Funds Upper Target`
   <dbl> <chr> <chr>                       <dbl>                        <dbl>
 1  2017 01    01                             NA                         0.75
 2  2017 02    01                             NA                         0.75
 3  2017 03    01                             NA                         0.75
 4  2017 03    16                             NA                         1
 5  2016 01    01                             NA                         0.5
 6  2016 02    01                             NA                         0.5
 7  2016 03    01                             NA                         0.5
 8  2016 04    01                             NA                         0.5
 9  2016 05    01                             NA                         0.5
10  2016 06    01                             NA                         0.5
# i 894 more rows
# i 5 more variables: `Federal Funds Lower Target` <dbl>,
#   `Effective Federal Funds Rate` <dbl>, `Real GDP (Percent Change)` <dbl>,
#   `Unemployment Rate` <dbl>, `Inflation Rate` <dbl>
```

```r
USA_Growth_new <- USA_Growth |>
  filter(Year >= 1962, Year <= 2016)
USA_Growth_new
```

```
# A tibble: 55 x 4
    Year GDP         `GDP per Capita`  Growth
   <dbl> <chr>                  <dbl>   <dbl>
 1  2016 $18,695.11B            57867  0.0167
 2  2015 $18,206.02B            56763  0.0271
 3  2014 $17,550.68B            55124  0.0229
 4  2013 $16,843.19B            53291  0.0184
 5  2012 $16,253.97B            51784  0.0228
 6  2011 $15,599.73B            50066  0.0155
 7  2010 $15,048.96B            48651  0.0271
 8  2009 $14,478.06B            47195 -0.026
 9  2008 $14,769.86B            48570  0.0012
10  2007 $14,474.23B            48050  0.0201
# i 45 more rows
```

```r
index_ascend_new <- index_ascend  |>
  select(Year, Month, `Unemployment Rate`) |>
  filter(Month == 12 , Year >= 1962) |>
  distinct() |>
  na.omit()
head(index_ascend_new)
```

```
# A tibble: 6 x 3
   Year Month `Unemployment Rate`
  <dbl> <chr>               <dbl>
1  2016 12                    4.7
2  2015 12                    5
3  2014 12                    5.6
4  2013 12                    6.7
5  2012 12                    7.9
6  2011 12                    8.5
```

```r
index_ascend_new
```

```
# A tibble: 55 x 3
    Year Month `Unemployment Rate`
   <dbl> <chr>               <dbl>
 1  2016 12                    4.7
 2  2015 12                    5
 3  2014 12                    5.6
 4  2013 12                    6.7
 5  2012 12                    7.9
 6  2011 12                    8.5
 7  2010 12                    9.3
 8  2009 12                    9.9
 9  2008 12                    7.3
10  2007 12                    5
# i 45 more rows
```

```r
# Merging Data
Merged_data <- merge(USA_Growth_new, index_ascend_new)
Merged_data
```

```
   Year          GDP GDP per Capita  Growth Month Unemployment Rate
```

1	1962	$605.10B	3244	0.0610	12	5.5
2	1963	$638.60B	3375	0.0440	12	5.5
3	1964	$685.80B	3574	0.0580	12	5.0
4	1965	$743.70B	3828	0.0640	12	4.0
5	1966	$815.00B	4146	0.0650	12	3.8
6	1967	$861.70B	4336	0.0250	12	3.8
7	1968	$942.50B	4696	0.0480	12	3.4
8	1969	$1,019.90B	5032	0.0310	12	3.5
9	1970	$1,073.30B	5234	-0.0028	12	6.1
10	1971	$1,164.85B	5609	0.0329	12	6.0
11	1972	$1,279.11B	6094	0.0526	12	5.2
12	1973	$1,425.38B	6726	0.0565	12	4.9
13	1974	$1,545.24B	7226	-0.0054	12	7.2
14	1975	$1,684.90B	7801	-0.0021	12	8.2
15	1976	$1,873.41B	8592	0.0539	12	7.8
16	1977	$2,081.83B	9453	0.0462	12	6.4
17	1978	$2,351.60B	10565	0.0554	12	6.0
18	1979	$2,627.33B	11674	0.0317	12	6.0
19	1980	$2,857.31B	12575	-0.0026	12	7.2
20	1981	$3,207.04B	13976	0.0254	12	8.5
21	1982	$3,343.79B	14434	-0.0180	12	10.8
22	1983	$3,634.04B	15544	0.0458	12	8.3
23	1984	$4,037.61B	17121	0.0724	12	7.3
24	1985	$4,338.98B	18237	0.0417	12	7.0
25	1986	$4,579.63B	19071	0.0346	12	6.6
26	1987	$4,855.22B	20039	0.0346	12	5.7
27	1988	$5,236.44B	21417	0.0418	12	5.3
28	1989	$5,641.58B	22857	0.0367	12	5.4
29	1990	$5,963.14B	23889	0.0189	12	6.3
30	1991	$6,158.13B	24342	-0.0011	12	7.3
31	1992	$6,520.33B	25419	0.0352	12	7.4
32	1993	$6,858.56B	26387	0.0275	12	6.5
33	1994	$7,287.24B	27695	0.0403	12	5.5
34	1995	$7,639.75B	28691	0.0268	12	5.6
35	1996	$8,073.12B	29968	0.0377	12	5.4
36	1997	$8,577.55B	31459	0.0445	12	4.7
37	1998	$9,062.82B	32854	0.0448	12	4.4
38	1999	$9,631.17B	34515	0.0479	12	4.0
39	2000	$10,250.95B	36330	0.0408	12	3.9
40	2001	$10,581.93B	37134	0.0095	12	5.7
41	2002	$10,929.11B	37998	0.0170	12	6.0
42	2003	$11,456.44B	39490	0.0280	12	5.7
43	2004	$12,217.19B	41725	0.0385	12	5.4

```
44 2005 $13,039.20B          44123  0.0348     12                 4.9
45 2006 $13,815.59B          46302  0.0278     12                 4.4
46 2007 $14,474.23B          48050  0.0201     12                 5.0
47 2008 $14,769.86B          48570  0.0012     12                 7.3
48 2009 $14,478.06B          47195 -0.0260     12                 9.9
49 2010 $15,048.96B          48651  0.0271     12                 9.3
50 2011 $15,599.73B          50066  0.0155     12                 8.5
51 2012 $16,253.97B          51784  0.0228     12                 7.9
52 2013 $16,843.19B          53291  0.0184     12                 6.7
53 2014 $17,550.68B          55124  0.0229     12                 5.6
54 2015 $18,206.02B          56763  0.0271     12                 5.0
55 2016 $18,695.11B          57867  0.0167     12                 4.7
```

```r
Merger_okay <- Merged_data |>
  select(Growth, `Unemployment Rate`)
Merger_okay
```

```
    Growth Unemployment Rate
1   0.0610               5.5
2   0.0440               5.5
3   0.0580               5.0
4   0.0640               4.0
5   0.0650               3.8
6   0.0250               3.8
7   0.0480               3.4
8   0.0310               3.5
9  -0.0028               6.1
10  0.0329               6.0
11  0.0526               5.2
12  0.0565               4.9
13 -0.0054               7.2
14 -0.0021               8.2
15  0.0539               7.8
16  0.0462               6.4
17  0.0554               6.0
18  0.0317               6.0
19 -0.0026               7.2
20  0.0254               8.5
21 -0.0180              10.8
22  0.0458               8.3
23  0.0724               7.3
```

```
24  0.0417              7.0
25  0.0346              6.6
26  0.0346              5.7
27  0.0418              5.3
28  0.0367              5.4
29  0.0189              6.3
30 -0.0011              7.3
31  0.0352              7.4
32  0.0275              6.5
33  0.0403              5.5
34  0.0268              5.6
35  0.0377              5.4
36  0.0445              4.7
37  0.0448              4.4
38  0.0479              4.0
39  0.0408              3.9
40  0.0095              5.7
41  0.0170              6.0
42  0.0280              5.7
43  0.0385              5.4
44  0.0348              4.9
45  0.0278              4.4
46  0.0201              5.0
47  0.0012              7.3
48 -0.0260              9.9
49  0.0271              9.3
50  0.0155              8.5
51  0.0228              7.9
52  0.0184              6.7
53  0.0229              5.6
54  0.0271              5.0
55  0.0167              4.7
```

```r
# Now, beginning the EDA
summary(Merger_okay)
```

```
     Growth          Unemployment Rate
 Min.   :-0.02600   Min.   : 3.400
 1st Qu.: 0.01950   1st Qu.: 5.000
 Median : 0.03290   Median : 5.700
 Mean   : 0.03076   Mean   : 6.062
 3rd Qu.: 0.04465   3rd Qu.: 7.200
```

```
 Max.   : 0.07240   Max.   :10.800
```

```r
# Standard Deviation-Variance finding
sd(Merger_okay$Growth)
```

```
[1] 0.02111011
```

```r
sd(Merger_okay$`Unemployment Rate`)
```

```
[1] 1.628734
```

```r
var(Merger_okay)
```

```
                        Growth Unemployment Rate
Growth             0.0004456368       -0.01809512
Unemployment Rate -0.0180951178        2.65277441
```

When investigating a relationship between two variables, commonly the first step is to exhibit how the data values are oriented graphically on a scatter diagram. On a scatter diagram, the closer the points lie to a straight line, the stronger the linear relationship between two variables. To quantify the strength of the relationship, we can calculate the correlation coefficient. Correlation is a statistic that measures the degree to which two variables move in relation to each other. In algebraic notation, if we have two variables x and y, and the data takes a the form of n pairs, say:

$$y[x_1, y_1], [x_2, y_2], [x_3, y_3], ...[x_n, y_n]$$

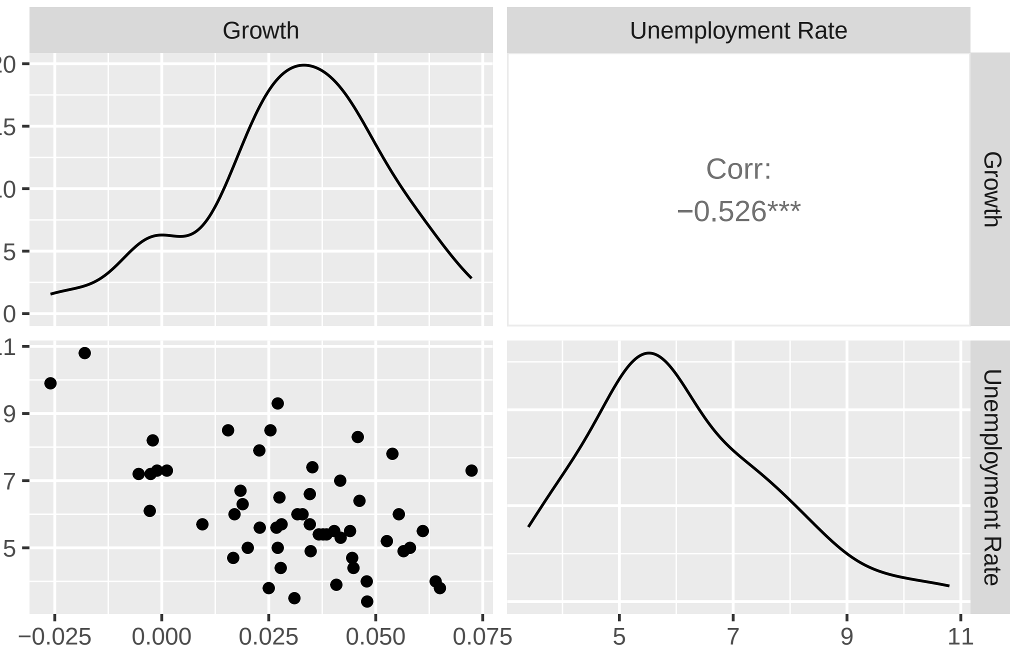
Then the (Pearson) correlation coefficient is given by the following equation:

$$r = N\sum XY - (\sum X \sum Y)/\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}$$

```r
library(GGally)
```

```
Registered S3 method overwritten by 'GGally':
  method from
  +.gg   ggplot2
```

```
ggpairs(Merger_okay)
```



Questions Conjured?

Why is a non-comoving relationship between Economic Growth and Employment apparent, when intuitively it seems to be the opposite?
Macroeconomics concerns multiple macro factors involving the dynamic of goods and services. Observed in prior display are non-normal distributions and a poor correlation of -0.525. Correlation measures association, but doesn't show if x causes y or vice versa—or if the association is caused by a third factor. The negative sign in the correlation value indicates a "non-cooperative" (linear) relationship w.r.t. the time range applied. However, a magnitude of 0.525 conveys weak correlation overall. Hence, weakly divergent movement for the time frame considered.
What distribution plays a role?
No ideal distribution can be readily called upon, namely, identified with apparent lack of symmetry in distribution (or formerly the skew), and having "tails" unrelated to any symmetry. The observed distributions are non-normal, say, no symmetry (has skew) and irregular tails (contrary kurtosis). The distributions are "realistic" rather than conforming to ideal probability densities (Gaussian, Gamma related, etc.).
Basic correlation measure is often a decent preliminary step towards variable selection. Yet, to dismiss the Employment Rate as non-influential upon Economic Growth may be

highly premature because not every period or environment has the same economic settings or conditions. However, transformations in the sectors/industries or transitions between sectors/industries can be the cause of lack of conformity to the intuitive thinking. A crude or primitive example, replacing plowing and planting workers with oxen and machinery; unemployment grows, yet, conventionally the production of goods and services should increase, with the possible savings in the long run.

Regression Model Development

REGRESSION STRUCTURE
Regression models involve the following components:

$$A, X_i, Y_i, e_i$$

Respectively, unknown parameters (scalar or vector), a vector of independent variables observed in the data, the dependent variable observed in the data, and the error terms not directly observed in the data.
Most regression models propose that

$$Y_i = f(X_i, A) + e_i$$

where the goal is to estimate such a function that quite closely fits the data. To carry out the regression analysis the function

$$f(X_i, A)$$

must be specified. Sometimes the form of such above function is based on prior knowledge about the relationship between the dependent variable and independent variable(s) without relying on the data; a luxury sometimes observed in the natural sciences and engineering. Else, for a target or dependent variable of interest, EDA is the alternative route to identify/choose independent variables (or features). Once a preference in independent variables is determined the estimation of the parameters are pursued. One of the most primitive methods of parameter estimation is the method of Ordinary Least Squares (OLS) which estimates the parameters that minimizes the sum of squared errors:

$$\sum_{i=1}^{n} (Y_i - f(X_i, A))^2$$

Note that data is always changing and/or growing, so the true value of the parameters are always changing.

OLS ASSUMPTIONS

1. The regression model possesses linearity in its coefficients and error terms.
2. The error terms' (ETs) population mean is zero. The ETs consider any variation in the DV that IVs fail to convey. For the idea case, stochastic chance determines the Ets' values. Else, it's daunting to develop unbiased values. Of consequence, a decent assumption is for the ETs' population mean to be zero. Positive average errors convey that the model under-predicts values, while negative average errors convey that the model over-predicts values.
3. There are no correlations between the IVs and ETs. If there are correlations, then it's possible to predict the ETs using the IVs; meaning the ETs represent predictable random error, which undermines the prior assumption.
4. Each of observation of the ETs is independent of each other.
5. The ETs variance is constant - Homoscedasticity. Variance remains constant across a single observation or a range of observations. Confirmation by plotting true values versus the residuals. If the spread of the residuals continues to increase in one direction, then the model fails to me the assumption of homoscedasticity.
6. The are no IVs that are perfect linear functions of other variables. Variables having 1 or -1 as coefficients among each other exhibits perfect correlation. Then there are some unneccesary variables applied. Perfect correlation is non-existent with OLS.
7. The ETs adheres to a normal by a normal distribution pattern. Such allows researchers to construct reliable prediction intervals, generate correct confidence intervals and conduct informative hypothesis testing.

STRENGTHS AND WEAKNESSES OF LINEAR REGRESSION

Strengths –

Linear Regression (LR) is a useful tool for EEDA and predictive analysis. Due to its "simplicity" and ease of implementation, where only basic algebra and calculus are required.

LR is highly interpretable and transparent, as each variable effect on the outcome can be observed and the model's fit to the data can be evaluated.

LR can also be applied for various types of data and various purposes.

LR can handle noise and outliers in the data which provides confidence intervals and hypothesis tests for coefficients.

Weaknesses –

LR is sensitive to multicollinearity, namely, some IVs may be highly correlated with each other, influencing the stability and precision of the coefficients.

LR can be prone to overfitting and underfitting, leading to poor generalization and prediction. Overfitting is an undesirable behavior that arises when the model in question gives accurate predictions for training data but not for new data. Underfitting, when a data model is unable to captiure the relationship between input and output variables with good accuracy, generating a high error rate both on training data and unseen data. The bias-variance trade-off/predicament must always be considered.; concerns a model's complexity, accuracy

of predictions, and how well it can makes predictions on unseen data not used to train the model. When increasing the number of "tunable" variables in a model there's more flexibility, and may better fit a training set; such generates lower error or bias. Yet, more complex models will tend to exhibit greater variance to the model fit each time a set of samples to create a new training set is taken.
Data dispersion may be too nonlinear or "patchy" for LR.

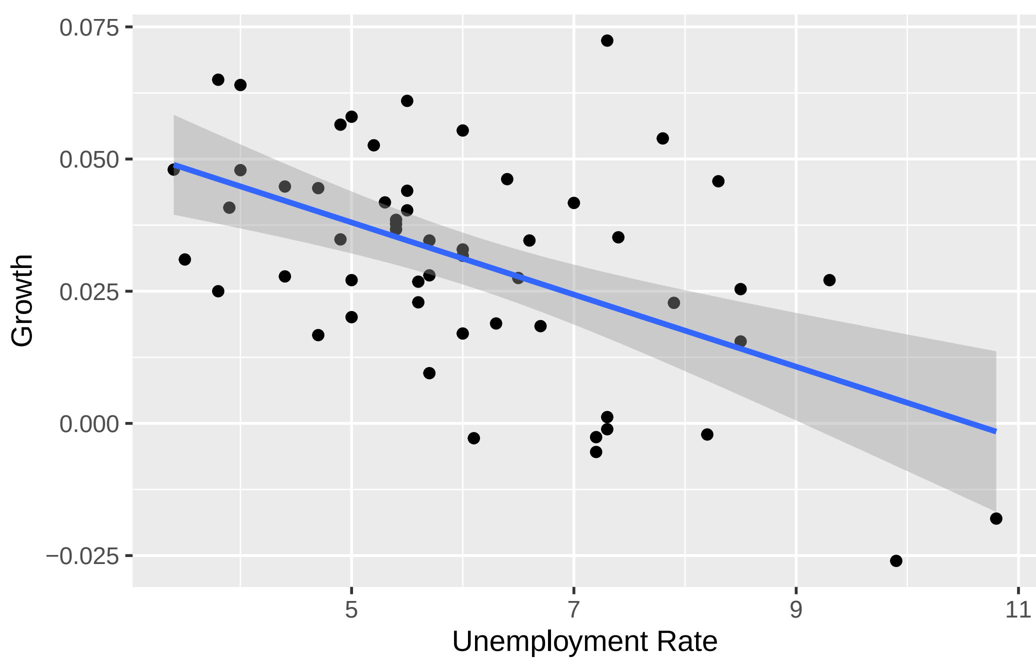
POSSIBLE EXTENSIONS OR IMPROVEMENTS

–Polynomial Regression: for nonlinear relationships to treat curved relations.
–Sinusoidal regression is also possible to treat cyclic/periodic relations.
–Multivariate Regression: when multiple independent variables are proven to be considerably relevant to the target (dependent variable) w.r.t. to applied data (via correlation or other methods). Multiple linear regression accounts for the combined effects of the independent variables.

BIVARIATE MODEL PLOTTING

```
bivariate_plot<-ggplot(Merger_okay, aes(x = `Unemployment Rate`, y = Growth)) + geom_point
bivariate_plot
```

```
`geom_smooth()` using formula = 'y ~ x'
```



SUMMARY STATISTICS PROVIDING MULTIPLE HYPOTHESIS TESTS FOR THE BIVARIATE MODEL

```r
model <- lm(Growth ~ `Unemployment Rate`, data = Merger_okay)
summary(model)
```

```

Call:
lm(formula = Growth ~ `Unemployment Rate`, data = Merger_okay)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.033303 -0.014242  0.001368  0.012416  0.050082 

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)          0.072113   0.009496   7.594 5.00e-10 ***
`Unemployment Rate` -0.006821   0.001514  -4.506 3.68e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01812 on 53 degrees of freedom
Multiple R-squared:  0.277,	Adjusted R-squared:  0.2633 
F-statistic:  20.3 on 1 and 53 DF,  p-value: 3.676e-05
```

SUMMARY STATISTICS INTERPRETATION

For the above linear model the summary statistics relates to a linear regression of the following form:

$$Growth = intercept + (UnemploymentRate) \times coefficient$$

We have the explicit model:

$$Growth = 0.72113 + (UnemploymentRate) \times (-0.006821)$$

to be used for predictions/forecasts.

In the summary statistics, a decent way to test the quality of the fit is to observe the residuals or differences between real values and predicted values. The idea is that the sum of the

residuals is approximately zero or as low as possible. In real life data orientation will not follow a straight line, so residuals are expected. The observes residuals are quite small or approximately zero.
Another measure to test if your linear model has a good fit is the *Coefficient of Determination* (*R-squared*), defined by the proportion of the total variability explained by the regression model:

$$R^2 = 1 - (\sum (y_i - f_i)^2 / \sum (y_i - \bar{y})^2)$$

The *R-squared* measure ranges from 0 to 1. Values in the lower half convey that the pursued model is poorly representative of the data. Values in the greater half convey that the model may represent well the data, but, when the value R-squared is very close to 1, such may convey falsified data. For THIS PARTICULAR data set, the R-squared value 0.277 conveying a poor model, namely, the model only explains of 28% of the data variability. An issue with R-squared is that it can't decrease as you add more independent variables to your model, rather, increase as the model is made more complex, even if the variables don't add anything to your prediction. Hence, the *Adjusted R-squared* measure may be a better alternative if adding more than one variable since it only increases if it reduces the overall error of the predictions. The Adjusted R-squared formula:

$$(R_a)^2 = 1 - [(n-1)/(n-k-1)](\sum (y_i - f_i)^2 / \sum (y_i - \bar{y})^2)$$

where n is the number of observations (sample size) and k being the number of independent variables or predictors in the model. The adjusted R-squared measure has the same measure range as R-squared, but it penalizes the addition of unnecessary predictors that don't significantly improve the model's explanatory power.
If *P(abs(t))* is sufficiently low one can reject a null hypothesis that the respective coefficient is 0. The p-value interpreted as the probability of seeing as much or more evidence for the alternative hypothesis than observed in our data, when the hypothesis is true. The p-value quantifies the the amount of evidence against the null hypothesis. The smaller it is, the more evidence against the null hypothesis, in favor of the alternative hypothesis. General tests size alpha to be 0.05; values much lower than 0.05 are observed for both coefficients. The power can't be applied because there is no ideal form for the probability density (Gaussian, Gamma related, etc.) involving the applied bivariate data with the model.
F-test and F-Statistic. In addition to observing whether predictors have a considerable effect, there's also concern for whether at least one predictor has a significant effect. Such translates to hypothesis testing of the following form:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_1 : \exists\, i : 1 \leq \, i \leq \, p-1 \, : \beta_i \neq 0$$

Under the null hypothesis the F-statistic will be F-distributed with (p-1, n-p) degrees of freedom. The probability of the observed data under the null hypothesis is then the p-value, where a p-value less than 0.05 indicates tha the model is likely significant; a value of 3.676-05 is observed.

Overall, observed are hypothesis tests which recognize significance for both the coefficients and the variables. However, there's poor model fit based on R-squared.

NOTE: BONFERRONI CORRECTION - SETTING SUCH ASIDE FOR FURTHER STUDIES IN ONE'S ACADEMIC FUTURE CONCERNING USE OF T-TESTS (IMPLYING NORAMILTY PRESENCE).
NOTE: such prior explained summary statistics measures will be applied in further development, namely, for the multivariate model development.

CASE OF ERRONEOUS BIVARIATE MODELING DUE TO BEING THE MOST PRIMITIVE REGRESSION MODEL

```
GDP <- read_csv("C:/Users/verlene/Downloads/GDP.csv")
UNRATE <- read_csv("C:/Users/verlene/Downloads/UNRATE.csv")
```

```
head(UNRATE)
```

```
# A tibble: 6 x 2
  DATE       UNRATE
  <date>      <dbl>
1 1948-01-01    3.4
2 1948-02-01    3.8
3 1948-03-01    4
4 1948-04-01    3.9
5 1948-05-01    3.5
6 1948-06-01    3.6
```

```
head(GDP)
```

```
# A tibble: 6 x 2
  DATE          GDP
  <date>      <dbl>
1 1947-01-01  243.
2 1947-04-01  246.
3 1947-07-01  250.
4 1947-10-01  260.
5 1948-01-01  266.
6 1948-04-01  273.
```

```r
data_merger <- merge(UNRATE, GDP)
head(data_merger)
```

```
        DATE UNRATE     GDP
1 1948-01-01    3.4 265.742
2 1948-04-01    3.9 272.567
3 1948-07-01    3.6 279.196
4 1948-10-01    3.7 280.366
5 1949-01-01    4.3 275.034
6 1949-04-01    5.3 271.351
```

```r
summary(data_merger)
```

```
      DATE                UNRATE            GDP
 Min.   :1948-01-01   Min.   : 2.600   Min.   :  265.7
 1st Qu.:1966-10-24   1st Qu.: 4.425   1st Qu.:  836.0
 Median :1985-08-16   Median : 5.500   Median : 4415.4
 Mean   :1985-08-16   Mean   : 5.733   Mean   : 7114.3
 3rd Qu.:2004-06-08   3rd Qu.: 6.775   3rd Qu.:12257.2
 Max.   :2023-04-01   Max.   :14.700   Max.   :27063.0
```

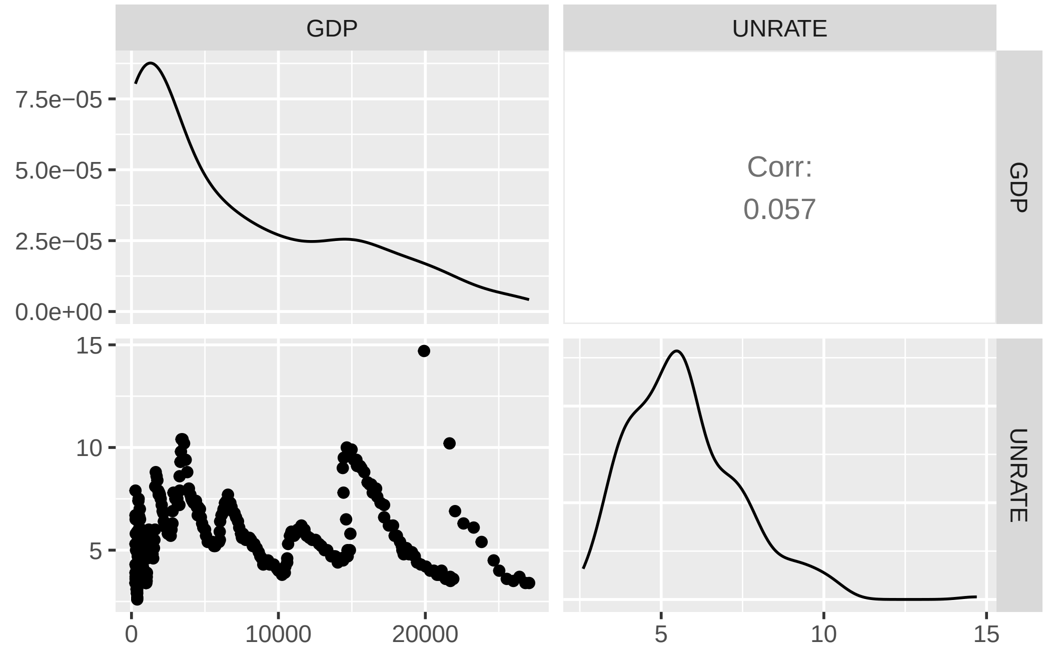
```r
sd(data_merger$UNRATE)
```

```
[1] 1.741773
```

```r
sd(data_merger$GDP)
```

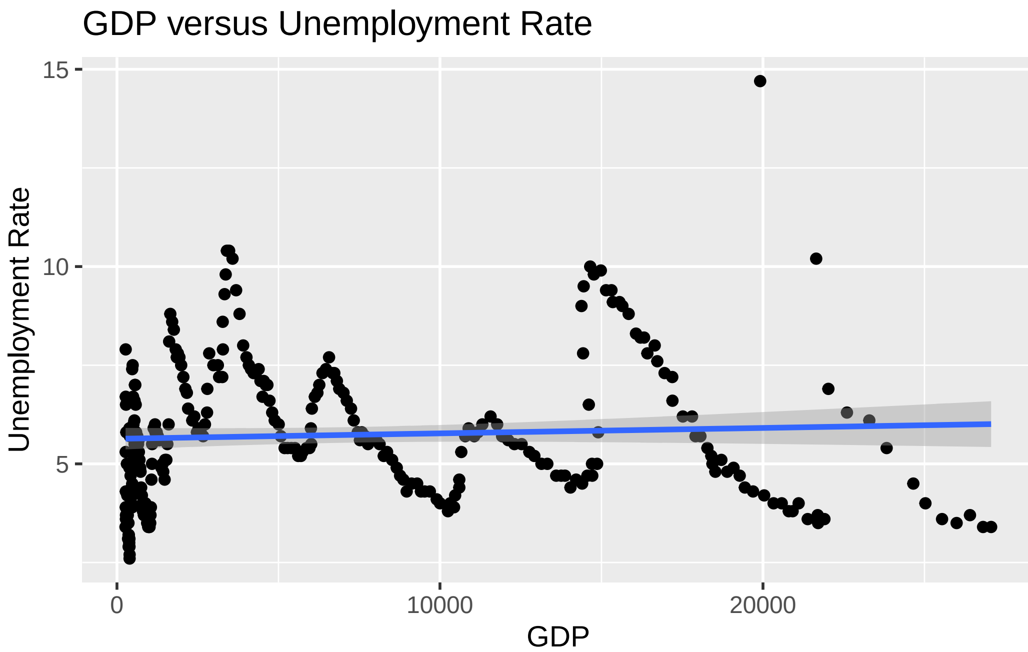
```
[1] 7230.814
```

```r
data_merger_clean <- data_merger |>
  select(GDP, UNRATE)
library(GGally)
ggpairs(data_merger_clean)
```




```r
regression_plot <- ggplot(data_merger_clean, aes(x = GDP, y = UNRATE)) + geom_point() + ge
regression_plot
```

```
`geom_smooth()` using formula = 'y ~ x'
```

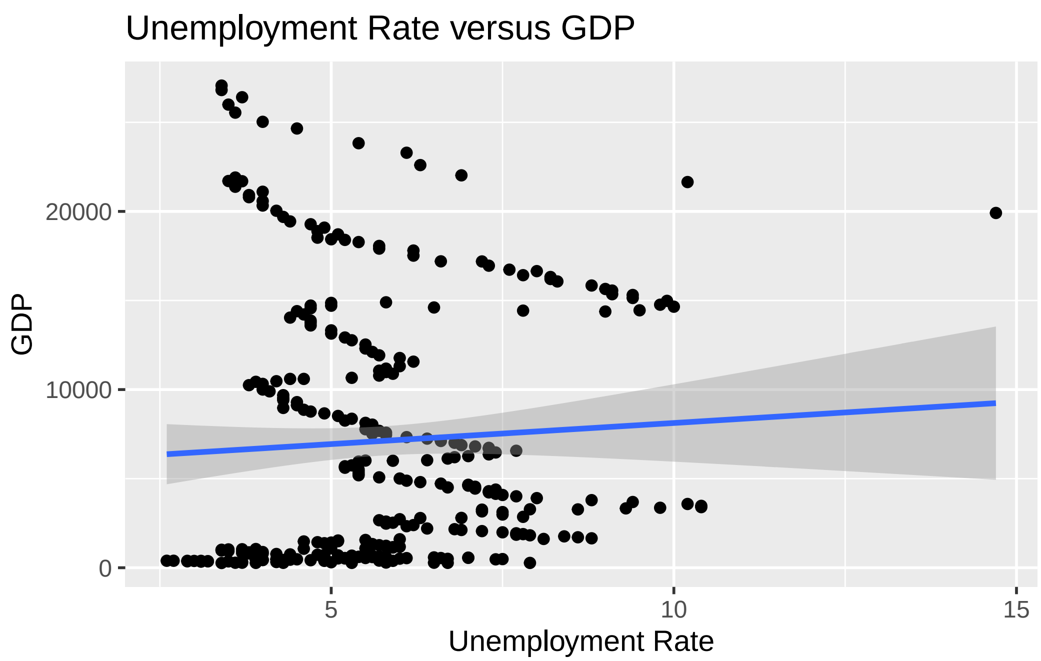



```r
new_model <- lm(GDP ~ UNRATE, data = data_merger_clean)
summary(new_model)
```

```
Call:
lm(formula = GDP ~ UNRATE, data = data_merger_clean)

Residuals:
   Min     1Q Median     3Q    Max 
 -7357  -5994  -3058   5192  20501 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   5757.4     1433.5   4.016 7.48e-05 ***
UNRATE         236.7      239.3   0.989    0.323    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7231 on 300 degrees of freedom
Multiple R-squared:  0.003251,	Adjusted R-squared:  -7.166e-05 
F-statistic: 0.9784 on 1 and 300 DF,  p-value: 0.3234
```

```r
regression_plot_inverted <- ggplot(data_merger_clean, aes(x = UNRATE, y = GDP)) + geom_poi
regression_plot_inverted
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```r
new_model_2 <- lm(UNRATE ~ GDP, data = data_merger_clean)
summary(new_model_2)
```

```
Call:
lm(formula = UNRATE ~ GDP, data = data_merger_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0404 -1.3153 -0.1688  1.0604  8.7914

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.635e+00  1.407e-01  40.043   <2e-16 ***
GDP         1.373e-05  1.388e-05   0.989    0.323
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.742 on 300 degrees of freedom
Multiple R-squared:  0.003251,  Adjusted R-squared:  -7.166e-05
F-statistic: 0.9784 on 1 and 300 DF,  p-value: 0.3234
```

Observing priors with economic data of GDP and Unemployment Rate, are scatter plot behaviors where bivariate regression models aren't befitting of the data in both possible Cartesian orientations.

CAUSALITY BLUNDERS WITH BIVARIATE REGRESSION MODEL CASE EXAMPLE
The trivial or ideal case of pairing the number of (houses/premises) fires data with data for the number of firefighter units to treat fire disasters. Empirically (or my common knowledge), there's high correlation, however, there's also the possible rudimentary assumption that "more firefighters lead to more fire disasters". A classical case of underfitting. The causes of fires, categories classifying the seriousness of fires, distance from nearest fire emergency support, civilian fire treatment/management, etc. should be considered. As well, recalling that correlation measures association, but doesn't show if x causes y or vice versa—or if the association is caused by a third factor.

Multivariate Data Wrangling & EDA

The claimed IVs for the regression model are Employment Rate, Inflation and the Fed Funds Rates. The null hypothesis to be that Economic Growth is highly influenced by Employment Rate, Inflation and the Fed Funds Rate; the alternative hypothesis to be that Economic Growth isn't highly influenced by the Employment Rate, Inflation and the Fed Funds Rate.

```r
MV_index_ascend_new <- index_ascend  |>
  select(Year, Month, `Unemployment Rate`, `Inflation Rate`, `Federal Funds Target Rate`)
  filter(Month == 12 , Year >= 1962) |>
  distinct() |>
  na.omit()
head(MV_index_ascend_new)
```

```
# A tibble: 6 x 5
   Year Month `Unemployment Rate` `Inflation Rate` `Federal Funds Target Rate`
  <dbl> <chr>               <dbl>            <dbl>                       <dbl>
```

```
1  2008 12                     7.3             1.8                        1
2  2007 12                     5               2.4                        4.5
3  2006 12                     4.4             2.6                        5.25
4  2005 12                     4.9             2.2                        4
5  2004 12                     5.4             2.2                        2
6  2003 12                     5.7             1.1                        1
```

```
# Pursuing development of dataframe for for DV with IVS of interest
MV_merger <- merge(MV_index_ascend_new, USA_Growth_new)
MV_merger
```

```
   Year Month Unemployment Rate Inflation Rate Federal Funds Target Rate
1  1982    12              10.8            4.5                    9.0000
2  1983    12               8.3            4.8                    9.3750
3  1984    12               7.3            4.7                    9.0000
4  1985    12               7.0            4.3                    8.0000
5  1986    12               6.6            3.8                    5.8750
6  1987    12               5.7            4.2                    6.8125
7  1988    12               5.3            4.7                    8.3750
8  1989    12               5.4            4.4                    8.5000
9  1990    12               6.3            5.2                    7.5000
10 1991    12               7.3            4.4                    4.7500
11 1992    12               7.4            3.3                    3.0000
12 1993    12               6.5            3.2                    3.0000
13 1994    12               5.5            2.6                    5.5000
14 1995    12               5.6            3.0                    5.7500
15 1996    12               5.4            2.6                    5.2500
16 1997    12               4.7            2.2                    5.5000
17 1998    12               4.4            2.4                    4.7500
18 1999    12               4.0            1.9                    5.5000
19 2000    12               3.9            2.6                    6.5000
20 2001    12               5.7            2.7                    2.0000
21 2002    12               6.0            1.9                    1.2500
22 2003    12               5.7            1.1                    1.0000
23 2004    12               5.4            2.2                    2.0000
24 2005    12               4.9            2.2                    4.0000
25 2006    12               4.4            2.6                    5.2500
26 2007    12               5.0            2.4                    4.5000
27 2008    12               7.3            1.8                    1.0000
           GDP GDP per Capita  Growth
1   $3,343.79B          14434 -0.0180
2   $3,634.04B          15544  0.0458
```

```
3   $4,037.61B          17121  0.0724
4   $4,338.98B          18237  0.0417
5   $4,579.63B          19071  0.0346
6   $4,855.22B          20039  0.0346
7   $5,236.44B          21417  0.0418
8   $5,641.58B          22857  0.0367
9   $5,963.14B          23889  0.0189
10  $6,158.13B          24342 -0.0011
11  $6,520.33B          25419  0.0352
12  $6,858.56B          26387  0.0275
13  $7,287.24B          27695  0.0403
14  $7,639.75B          28691  0.0268
15  $8,073.12B          29968  0.0377
16  $8,577.55B          31459  0.0445
17  $9,062.82B          32854  0.0448
18  $9,631.17B          34515  0.0479
19 $10,250.95B          36330  0.0408
20 $10,581.93B          37134  0.0095
21 $10,929.11B          37998  0.0170
22 $11,456.44B          39490  0.0280
23 $12,217.19B          41725  0.0385
24 $13,039.20B          44123  0.0348
25 $13,815.59B          46302  0.0278
26 $14,474.23B          48050  0.0201
27 $14,769.86B          48570  0.0012
```

```r
# Variables of interest
MV_merger_okay <- MV_merger |>
  select(Growth, `Unemployment Rate`, `Inflation Rate`,
         `Federal Funds Target Rate`)
MV_merger_okay
```

```
   Growth Unemployment Rate Inflation Rate Federal Funds Target Rate
1  -0.0180              10.8            4.5                    9.0000
2   0.0458               8.3            4.8                    9.3750
3   0.0724               7.3            4.7                    9.0000
4   0.0417               7.0            4.3                    8.0000
5   0.0346               6.6            3.8                    5.8750
6   0.0346               5.7            4.2                    6.8125
7   0.0418               5.3            4.7                    8.3750
8   0.0367               5.4            4.4                    8.5000
```

```
9   0.0189             6.3           5.2                7.5000
10 -0.0011             7.3           4.4                4.7500
11  0.0352             7.4           3.3                3.0000
12  0.0275             6.5           3.2                3.0000
13  0.0403             5.5           2.6                5.5000
14  0.0268             5.6           3.0                5.7500
15  0.0377             5.4           2.6                5.2500
16  0.0445             4.7           2.2                5.5000
17  0.0448             4.4           2.4                4.7500
18  0.0479             4.0           1.9                5.5000
19  0.0408             3.9           2.6                6.5000
20  0.0095             5.7           2.7                2.0000
21  0.0170             6.0           1.9                1.2500
22  0.0280             5.7           1.1                1.0000
23  0.0385             5.4           2.2                2.0000
24  0.0348             4.9           2.2                4.0000
25  0.0278             4.4           2.6                5.2500
26  0.0201             5.0           2.4                4.5000
27  0.0012             7.3           1.8                1.0000
```

```
# Now beginning the MV EDA
summary(MV_merger_okay)
```

```
     Growth          Unemployment Rate Inflation Rate  Federal Funds Target Rate
 Min.   :-0.01800   Min.   : 3.900    Min.   :1.100   Min.   :1.000
 1st Qu.: 0.02345   1st Qu.: 5.150    1st Qu.:2.300   1st Qu.:3.500
 Median : 0.03480   Median : 5.700    Median :2.700   Median :5.500
 Mean   : 0.03073   Mean   : 5.993    Mean   :3.174   Mean   :5.294
 3rd Qu.: 0.04125   3rd Qu.: 6.800    3rd Qu.:4.350   3rd Qu.:7.156
 Max.   : 0.07240   Max.   :10.800    Max.   :5.200   Max.   :9.375
```

```
# Standard-Deviation-Variance finding
sd(MV_merger_okay$`Unemployment Rate`)
```

```
[1] 1.477244
```

```
sd(MV_merger_okay$`Inflation Rate`)
```

```
[1] 1.142734
```

```r
sd(MV_merger_okay$`Federal Funds Target Rate`)
```

```
[1] 2.548556
```

```r
ggpairs(MV_merger_okay)
```

SUMMARY STATISTICS PROVIDING MULTIPLE HYPOTHESIS TESTS FOR THE MULTIVARIATE MODEL

```r
MV_model <- lm(`Unemployment Rate` ~ Growth + `Inflation Rate` + `Federal Funds Target Rat
summary(MV_model)
```

```
Call:
lm(formula = `Unemployment Rate` ~ Growth + `Inflation Rate` +
    `Federal Funds Target Rate`, data = MV_merger_okay)

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-1.62105 -0.87371  0.00904  0.49553  2.25778

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  4.84017    0.76952   6.290 2.04e-06 ***
Growth                     -35.30131   13.62968  -2.590   0.0164 *  
`Inflation Rate`             0.82245    0.33604   2.447   0.0224 *  
`Federal Funds Target Rate`  -0.07049    0.15824  -0.445   0.6601    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.111 on 23 degrees of freedom
Multiple R-squared:  0.4999,    Adjusted R-squared:  0.4347 
F-statistic: 7.664 on 3 and 23 DF,  p-value: 0.001004
```

SUMMARY STATISTICS INTERPRETATION FOR THE MULTIVARIATE MODEL

Observed is a multivariate linear model of the following form:

$$UR = \beta_0 + (Growth) \times \beta_1 + (IR) \times \beta_2 + (FFR) \times \beta_3$$

Observed is a explicit multivariate linear model of the following form:

$$UR = 4.84017 + (Growth) \times (-35.30131) + (IR) \times (0.82245) + (FFR) \times (-0.07049)$$

to be used for predictions/forecasts.

The observed residual values are not really close to 0, so there's early indication of poor model fit.
If P(abs(t)) is sufficiently low one can reject a null hypothesis that the respective coefficient is 0.
Considering Adjusted R-squared instead of basic R-squared because we have a multivariate model with the "independent variables". Observation of value 0.4347, hence an apparent poor model fit.
For the *F-test and F-Statistic*, concern for whether at least one predictor has a significant effect; observing a p-value of 0.001004 being much less than 0.05. Hence, the model my be significant.

Overall, hypothesis tests recognize the significance of coefficients and variables, except for the federal funds rate. An adjusted R-squared of 0.4347 exhibits a weak model fit.

Conclusion

This project concerns applied both EDA and hypothesis tests to determine significance of IVs and wellness of model fits. With the R environment exhibited was fast development of primitive data analysis and regression modeling. Various measures and statistics identified as data analysis and hypothesis tests were implemented. The data applied with regression models did not provide strong results in terms of model fits regardless of much encourage with the significance of coefficients and IVs. The data range applied can be categorized as long term behaviour where phenomenon like mean reversion may.....

References

1. FRED St.Louis. (2023, September 28). *Gross Domestic Product.* FRED. https://fred.stlouisfed.org/series/GDP
2. FRED St. Louis. (2023, October 6). *Unemployment Rate.* FRED. https://fred.stlouisfed.org/series/UNRATE
3. Reserve, F. (2017, March 16). *Federal Reserve Interest Rates, 1954-Present.* Kaggle. https://www.kaggle.com/datasets/federalreserve/interest-rates
4. Vyas, M. (2022, December 1). *USA GDP Growth Dataset 1961-2021.* Kaggle. https://www.kaggle.com/datasets/malayvyas/usa-gdp-dataset-19612021