STAT-UB 103 Final Project

Daphne Chan
Ishan Pranav
Ellen Ryoo
Claire Shi

May 8, 2023

Abstract

Working with real data is the best way to learn statistics. This project is an opportunity to do that.

1 Introduction

1.1 Data sources

While exploring potential variables to examine, we discovered three data sets available from the Federal Reserve Economic Data website: the Federal Funds Effective Rate ("interest rates"), the unemployment rate among all persons in the United States between 15 and 64 years of age ("unemployment"), and the gross domestic product of the United States ("GDP"). We grouped the data points by quarter and narrowed our focus to the 40 observations between the quarter beginning January 1, 2013, and the quarter beginning October 1, 2022—a 10-year period.

- Interest rates: `https://fred.stlouisfed.org/series/FEDFUNDS`
- Unemployment: `https://fred.stlouisfed.org/series/LRUN64TTUSQ156S`
- GDP: `https://fred.stlouisfed.org/series/GDP`

1.2 Motivation

As business students, we are interested in the relationship between macroeconomic measurements because they are the lifeblood of business and are directly linked to the topics we study in our other courses. We hope to discover empirical relationships between the variables that confirm our understanding of economics or reveal insights that our hypotheses overlook.

1.3 Hypotheses

Before exploring the data set, we hypothesize a positive relationship between the interest rate and the GDP. Recalling that during periods of low economic output (low GDP), the Federal Reserve generally implements an expansionary monetary policy, we predict that the interest rate decreases alongside GDP. On the other hand, we expect a negative relationship between the interest rate and unemployment. This is because high unemployment is associated with low economic output, so the Federal Reserve might decrease interest rates to encourage growth.

2 Scatterplots

2.1 Interest rates *vs.* unemployment

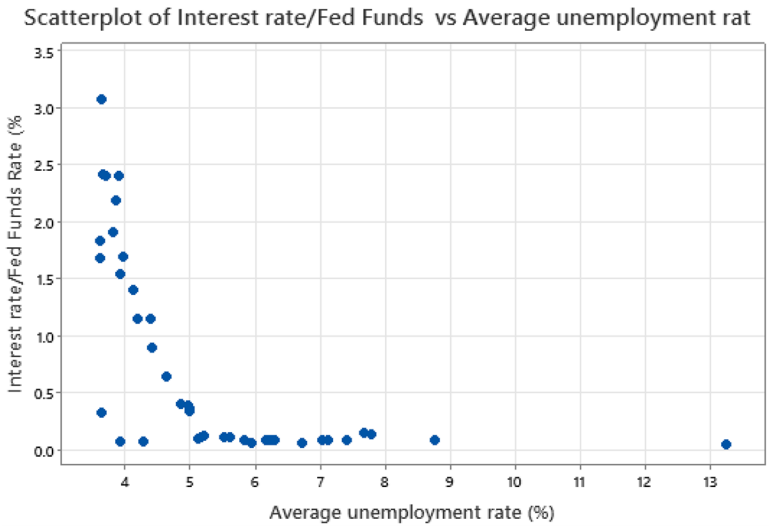

Figure 1: Scatterplot of unemployment (horizontal axis) and interest rates (vertical axis).

As expected, there is a negative relationship between unemployment and interest rates. The trendline is flat for the portion with high unemployment rates because it is very unusual for interest rates to fall below zero. On the very right, there is one outlier.

There are also three data points significantly below the major trendline. Those observations were collected during the COVID-19 pandemic, which is a potential "cause" for the unusual data. Even when the unemployment rate was relatively low, the Federal Reserve was not confident enough in the general economic situation to raise interest rates.

2.2 Interest rates *vs.* GDP

The pattern in Figure 2 indicates a strong positive correlation up to just under USD$22,000 billion. The positive correlation falls out of expectations beginning in the last quarter of 2019, then resumes in the second half of 2022.

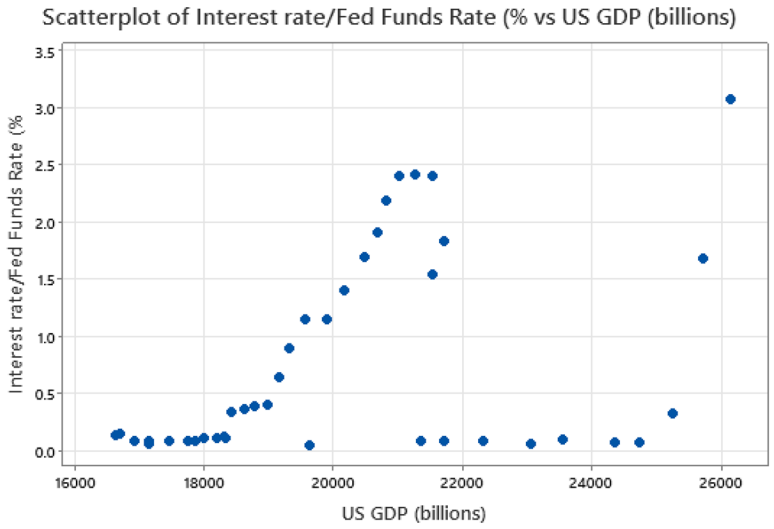

Figure 2: Scatterplot of GDP (horizontal axis) and interest rates (vertical axis).

One potential "cause" for the abnormalities is the COVID-19 pandemic, which began at the end of 2019. During the pandemic, GDP dropped drastically from USD$21,538 billion in the first quarter of 2020 to USD$19,636 billion in the second quarter of 2020. Although GDP increased from roughly $20 trillion to $25 trillion, the interest rate remained near zero because the Federal Reserve was not confident in the nation's economic performance. This is unusual since periods of high output are typically associated with more robust economies. Once the economy stabilized and GDP began to increase steadily, the Federal Reserve increased interest rates to curb inflation.

3 Other variables

This project considers the impact of GDP and unemployment on interest rates. However, other variables may be useful predictors of the interest rate. For example, inflation, consumer spending, investment, government spending, and profits from imports and exports all contribute to this metric. Since the Federal Funds Rate is determined by the Federal Reserve, it is a product of great deliberation and is influenced by many economic indicators

that reflect the overall health of the market.

4 Graphical summaries

Compared to the scatterplots (Figure 1 and Figure 2), the graphical summaries (Figure 3, Figure 4, and Figure 5) provide insufficient information about outliers. The graphs and boxplots in the graphical summaries only consider one variable at a time. In the case of unemployment, only one outlier meets the "one-and-a-half-times-the-interquartile-range" standard for determining an outlier. Meanwhile, the irregularities—like low-interest rates during the high-GDP COVID-19 era—go unnoticed when variables are examined one at a time. For this reason, the scatterplot can provide more insight into whether a specific combination of two variables is unusual.

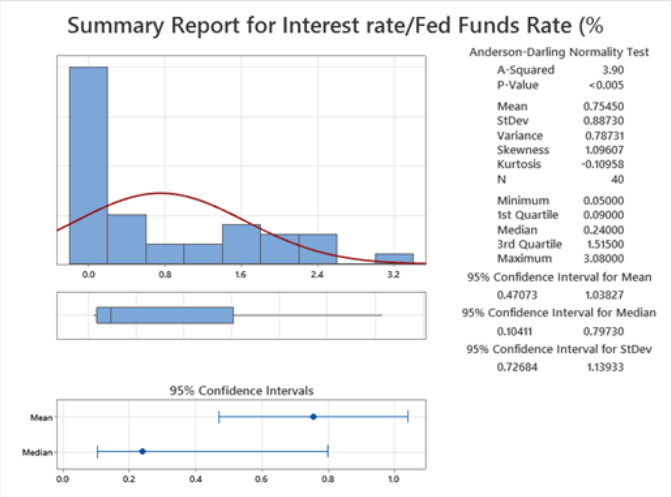

Figure 3: Summary report for interest rates. There are no apparent outliers.

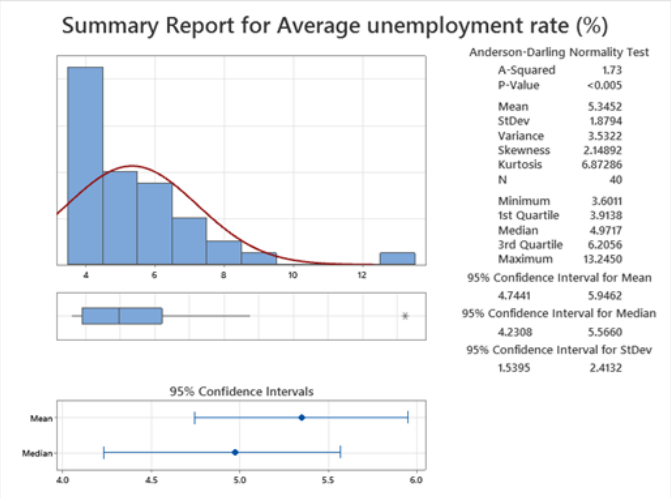

Figure 4: Summary report for unemployment. The maximum (13.2 percent) was an outlier in the second quarter of 2020 (represented by a star on the boxplot).

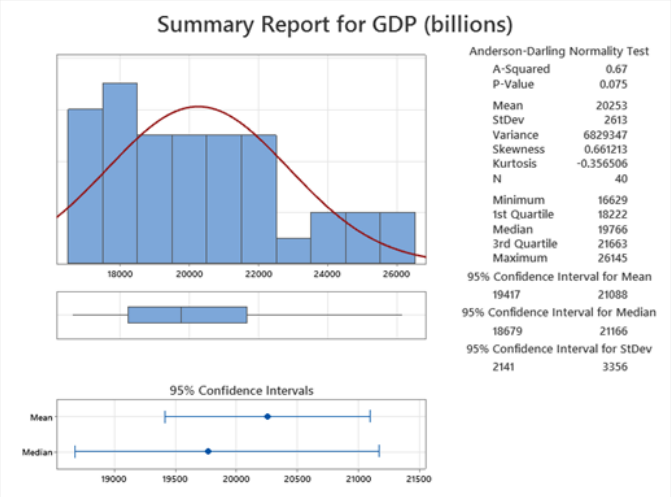

Figure 5: Summary report for GDP. There are no apparent outliers.

5 Size-dependent variability

GDP does not suffer from significant size-dependent variability. The histogram is close to symmetrical; the mean is not significantly larger than the median, relative to the number of data points; the median line on the boxplot is approximately in the middle of the box; and the line on the right of the box is just slightly longer than the line on the left of the box. There are no outliers. Applying a *log*-transformation on the GDP variable does not provide value in statistical analysis.

Meanwhile, the opposite is true for interest rates and unemployment. The histograms for these variables are right-skewed and the boxplots depict the median line on the low (left) side. While there are no outliers for the interest rate variable, the single outlier for the unemployment variable is on the high side (13.2 percent). The interest rate and unemployment variables suffer from size-dependent variability, so applying a *log*-transformation can bring the high outliers in line with the rest of the data and magnify the lower observations, allowing clearer analysis of the data.

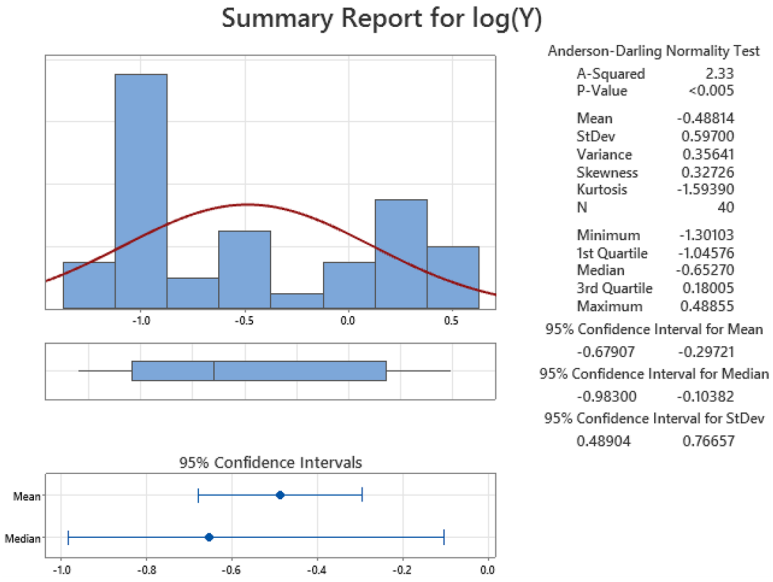

Figure 6: Summary report for interest rates after applying a *log*-transformation. There are no apparent outliers.

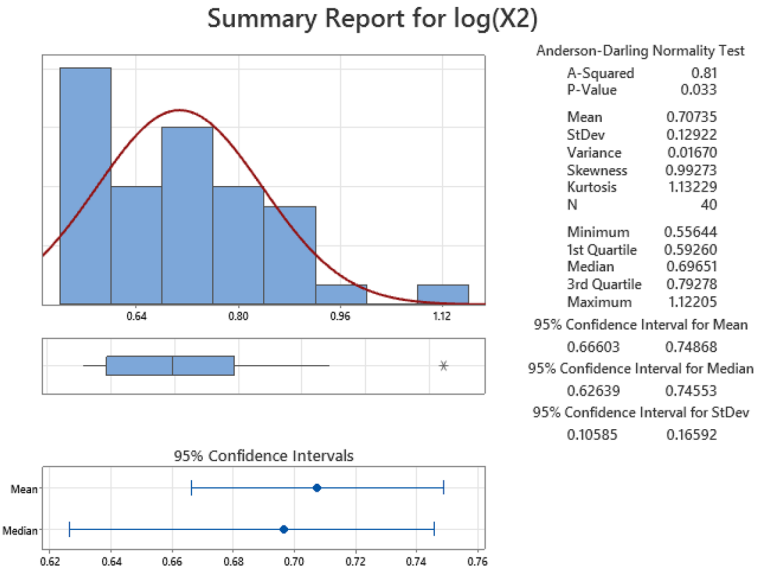

Figure 7: Summary report for unemployment after applying a *log*-transformation. The outlier of 13.2 percent remains apparent.

6 Logarithmic-scale scatterplots

6.1 Results of *log*-transformation and conclusion

As depicted in Figure 8 and Figure 9 below, taking the natural logarithm of the interest rate and unemployment observations yields a clearer negative relationship between unemployment and interest rates and a clearer positive relationship between GDP and interest rates. The correlation is quite strong in both cases.

Having analyzed the data we confirmed our hypothesis: there is a strong positive relationship between GDP and interest rates and a strong negative relationship between unemployment and interest rates.

We conclude that there exists a positive correlation between economic output (as measured by GDP) and the logarithm of the interest rate and a negative correlation between the logarithm of the unemployment rate and the logarithm of the interest rate.

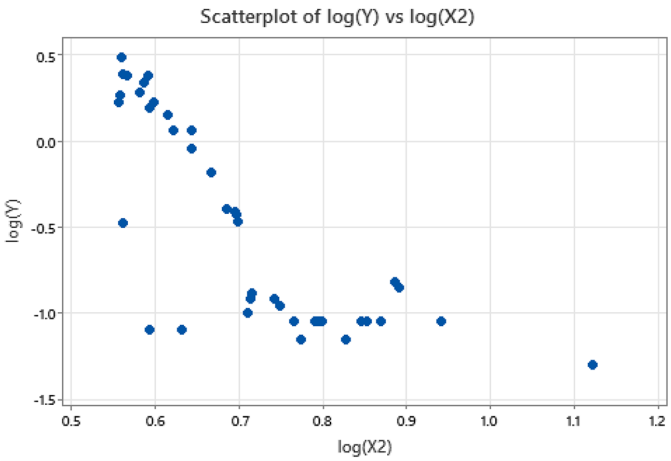

Figure 8: Scatterplot of unemployment (horizontal axis, logarithmic) and interest rates (vertical axis, logarithmic).

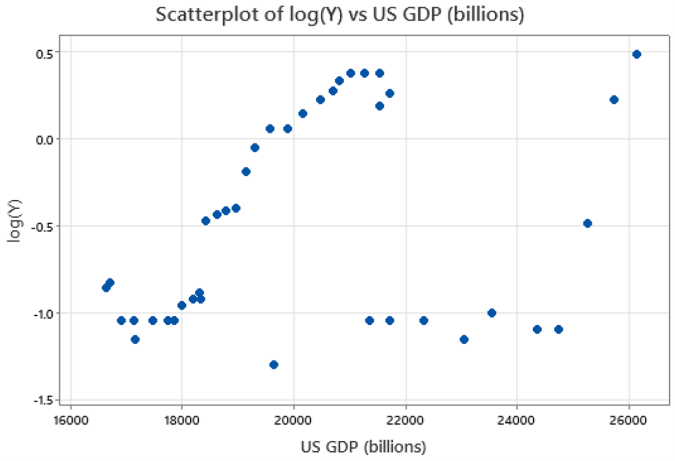

Figure 9: Scatterplot of GDP (horizontal axis) and interest rates (vertical axis, logarithmic).

7 Anderson–Darling test

7.1 Gaussian probability plot

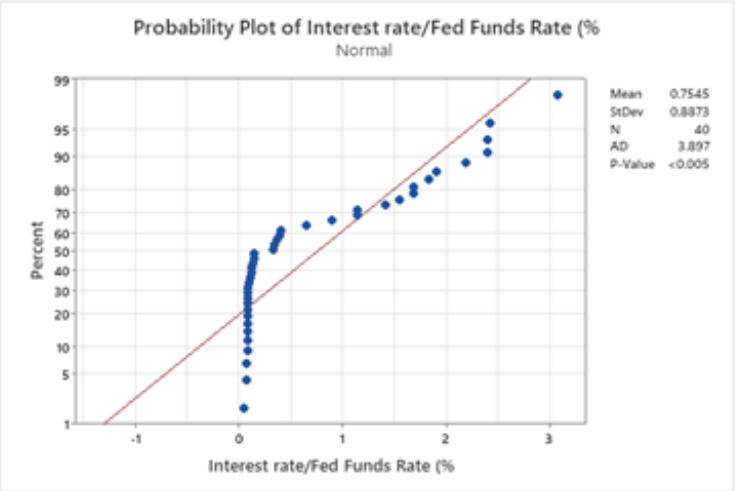

Figure 10: Normal probability plot for interest rates.

The pattern in the plot of interest rates (Figure 10) seems to indicate non-normality, as the data do not fall in a straight line. In particular, the points on the left end bend vertically below the line which indicates that the left tail is longer than that of a normal distribution. The points on the right end stray away from the line before bending back up.

7.2 *P*-value

The P-value resulting from the Anderson–Darling test seems to indicate non-normality since the value is less than 0.005. In the Anderson–Darling hypothesis test, the null hypothesis is that the data follows a Gaussian (normal) distribution, and the alternative hypothesis is that the data do not follow a normal distribution. A low P-value indicates that the test statistic obtained is very unlikely given that the null hypothesis. Using a significance level of five percent (α=0.05), there is convincing evidence to reject the null

hypothesis: The P-value is far smaller than 5 percent. Therefore, we conclude that interest rates are not normally distributed.

7.3 Comparison with summary report

The normal curve depicted on the summary report for the interest rate (Figure 3) is heavily skewed to the right. This finding aligns with the descriptive statistics: We see a large portion of the data accumulated between 0.0 and 0.5. As the interest rate increases, there are fewer and fewer points included within the range, providing a skewed distribution.

8 Anderson–Darling test after logarithmic scaling

8.1 Normal probability plot

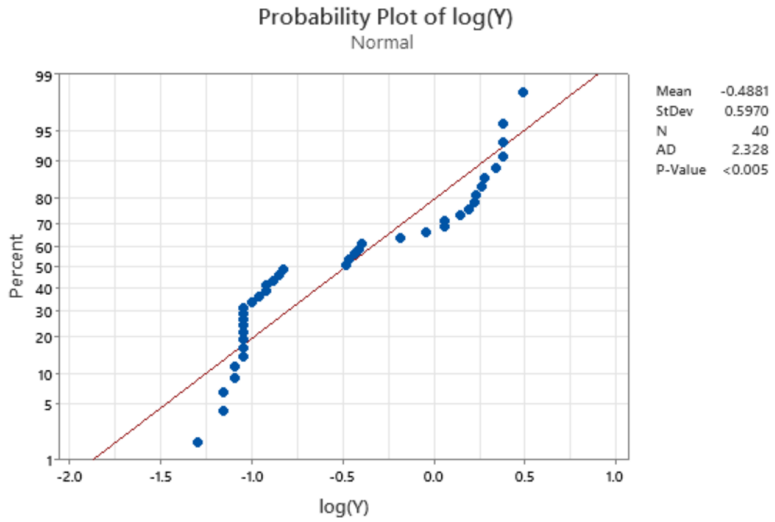

Figure 11: Gaussian probability plot for the logarithm of the interest rate.

The pattern in the plot of *log*-scaled interest rates (Figure 11) seems to indicate non-normality because the points do not form a straight line. Instead, they bend downwards on the left side of the plot, which means that the lower tail is longer.

8.2 *P*-value

The P-value resulting from the Anderson–Darling test also indicates non-normality because it is very close to zero. There is convincing evidence at the $\alpha = 0.05$ significance level to reject the null hypothesis. Therefore, we conclude that the logarithm of the interest rate is not normally distributed.

8.3 Comparison with summary report

Once again, the findings based on the plot and the Anderson–Darling test align with the summary report for the logarithm of the interest rate in Figure 6. After applying a *log*-transformation, the distribution now fits the regression line slightly better. According to the descriptive statistics, there is still a large portion of the data accumulated near -1.0 and 0.3, but applying the *log*-transformation has resulted in less skew.

9 Simple linear regression

In the first module of this project, we decided to take the logarithm of the interest rate (response variable) and unemployment (predictor) while leaving GDP (predictor) alone.

9.1 Interest rate *vs.* unemployment

Let $\log(y)$ represent the interest rate on $\log(x)$ represent unemployment, both on *log* scales. Let β_1 represent the unemployment coefficient, β_0 represent the intercept, and ϵ represent the residual.

$$\log(y) = \beta_1 \log(x) + \beta_0 + \epsilon.$$

$$b_1 = -3.574.$$

The sample unemployment coefficient (b_1) is negative, indicating that the interest rate decreases as unemployment increases. Let H_0 denote the null hypothesis and H_1 denote its alternative. Let α represent the significance level.

Regression Equation

log(Y) = 2.040 - 3.574 log(X2)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.040	0.341	5.97	0.000	
log(X2)	-3.574	0.475	-7.52	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.383348	59.83%	58.77%	53.26%

Figure 12: Linear regression analysis for interest rate (response, logarithmic) and unemployment (predictor, logarithmic).

$$H_0 : \beta_1 = 0.$$
$$H_1 : \beta_1 < 0.$$
$$P(t \leq -7.52 \,|\, H_0) \approx \frac{0.000}{2} \approx 0.000.$$
$$\alpha = 0.05.$$

Given that there is no relationship between unemployment and the interest rate, the probability of obtaining a Student t-statistic as extreme as (or more extreme than) -7.52 is approximately zero. There is convincing evidence to reject the null hypothesis at the 5 percent significance level. We can conclude beyond a reasonable doubt that a negative linear relationship exists between unemployment and the interest rate.

The negative relationship is also reflected in the scatterplot of unemployment and the logarithm of the interest rate in Figure 8.

Regression Equation

log(Y) = -1.911 + 0.000070 US GDP (billions)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-1.911	0.720	-2.65	0.012	
US GDP (billions)	0.000070	0.000035	1.99	0.053	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.575482	9.46%	7.08%	0.00%

Figure 13: Linear regression analysis for interest rate (response, logarithmic) and GDP (predictor).

9.2 Interest rate *vs.* GDP

Let $\log(y)$ represent the interest rate on a *log* scale and x represent GDP. Let β_1 and the GDP coefficient, β_0 represent the intercept, and ϵ represent the residual.

$$\log(y) = \beta_1 x + \beta_0 + \epsilon.$$

$$b_1 = 0.000070.$$

The slope coefficient for GDP is positive, indicating that the interest rate increases as GDP increases.

$$H_0 : \beta_1 = 0.$$

$$H_1 : \beta_1 > 0.$$

$$P(t \geq 1.99 \,|\, H_0) \approx \frac{0.053}{2} \approx 0.0265.$$

$$\alpha = 0.05.$$

Given that there is no relationship between GDP and the interest rate, the probability of obtaining a Student t-statistic as extreme as (or more extreme than) 1.99 is approximately 2.65 percent. There is sufficient evidence to reject the null hypothesis at the 5 percent significance level. We can conclude that a positive linear relationship exists between GDP and the interest rate.

The negative relationship is also reflected in the scatterplot of the logarithm of the interest rate and GDP in Figure 9.

10 Multiple linear regression

Regression Equation

log(Y) = 2.828 - 3.871 log(X2) - 0.000029 US GDP (billions)

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	2.828	0.830	3.41	0.002	
log(X2)	-3.871	0.554	-6.99	0.000	1.36
US GDP (billions)	-0.000029	0.000027	-1.04	0.305	1.36

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.382924	60.97%	58.86%	48.13%

Figure 14: Linear regression analysis for interest rate (response, logarithmic) and unemployment (predictor, logarithmic) and GDP (predictor).

$$\log(y) = \beta_2 \log(x_2) + \beta_1 x_1 + \beta_0 + \epsilon.$$

Let $\log(y)$ represent the interest rate (*log*-scaled), $\log(x_2)$ represent unemployment (*log*-scaled), and x_1 represent GDP. Let β_2 and β_1 represent their respective coefficients, β_0

represent the intercept, and ϵ represent the residual.

$$\alpha = 0.05.$$

$$H_0 : \beta_2 = 0.$$
$$H_1 : \beta_2 \neq 0.$$
$$P(t \leq -6.99 \,|\, H_0) \approx 0.000.$$

The P-value suggests that unemployment is a significant predictor of the interest rate.

$$H_0 : \beta_1 = 0.$$
$$H_1 : \beta_2 \neq 0.$$
$$P(t \leq -0.000029 \,|\, H_0) \approx 0.305.$$

Based on the P-value, GDP may not be a significant predictor of the interest rate.

$$H_0 : \beta_2 = \beta_1 = 0.$$
$$H_1 : \beta_2 \neq 0 \text{ or } \beta_1 \neq 0.$$
$$P(F \geq 28.90) \approx 0.000.$$

There is convincing evidence to reject the null hypothesis at the five-percent significance level. We can conclude beyond a reasonable doubt that at least one of the two variables—unemployment and GDP—is a significant predictor of the interest rate.

$$r^2 = 60.97\%.$$

Approximately 60.97 percent of the variation in the interest rate can be explained by variations in unemployment and GDP using a least-squares regression line. This suggests that the linear relationship is moderately strong. No, the r^2 is not appreciably higher than the r^2 from the simple regression that only includes unemployment ($r^2 = 59.83\%$), but it is much higher than the regression that only includes GDP ($r^2 = 9.46\%$).

11 Comparison of regression methods

There is an inconsistency between the coefficient for GDP in the simple linear regression compared to the multiple linear regression. In the simple regression, the coefficient is positive, whereas it is negative in the multiple regression. Both coefficients have small magnitudes.

12 Cook's distance

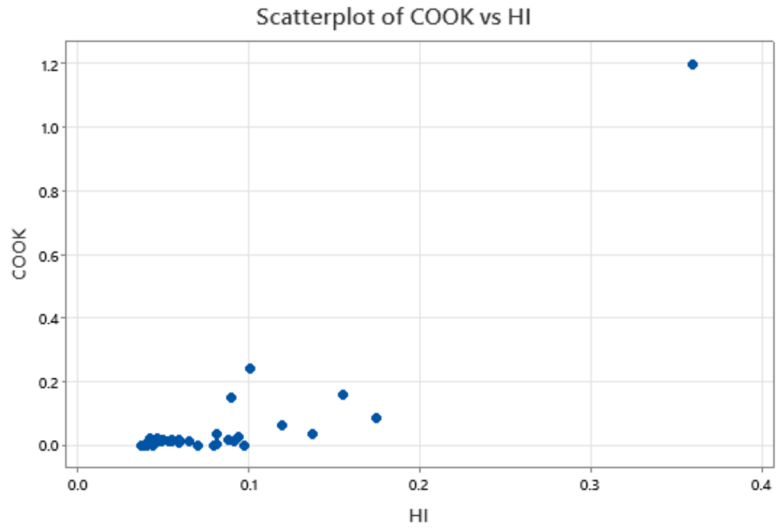

Figure 15: Scatterplot of leverage (horizontal axis) and Cook's distance (vertical axis).

$$\frac{2(k+1)}{n} = \frac{2(2+1)}{40} = 0.15.$$

$$\frac{3(k+1)}{n} = \frac{3(2+1)}{40} = 0.225.$$

Using $2(k+1)/n$ as the standard for determining leverage points, there are three points with leverage greater than 0.15. Using $3(k+1)/n$, there is only one point with high

leverage. The largest leverage is 0.359638, the second-largest is 0.174313, and the third-largest is 0.155365. The Cook's distances are 1.19731, 0.08490, 0.16138, respectively. The results suggest that $3(k+1)/n$ is the appropriate standard, and that the point with leverage of 0.359638 is necessarily a bad leverage point (since its Cook's distance is greater than 1). At that point, the interest rate is -1.30103, the logarithm of the unemployment rate is 1.12205, and the GDP is \$19636.731 billion. Based on the scatterplots of the *log*-scaled interest rate and the *log*-scaled unemployment (Figure 8) and GDP (Figure 9), this point appears to be an outlier.

Based on the r^2, the significance of the coefficients, and the Cook's distances, we believe that the model fits decently, but is not the best choice. Since the GDP coefficient is not statistically significant, it may be beneficial to re-assess whether the model without GDP is more appropriate. A nested F-test may be required to determine if a reduced model is better than the complete one.

13 Residual analysis

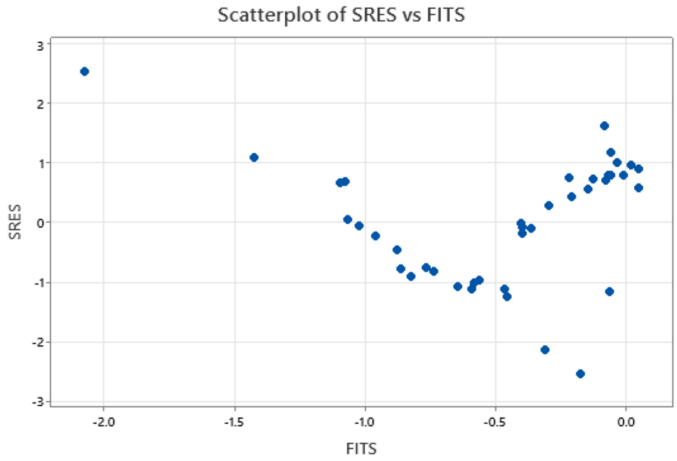

Figure 16: Scatterplot of fitted values (horizontal axis) and standardized residuals (vertical axis).

Based on the scatterplot of the standardized residuals *vs.* the fitted values in Figure 16,

there is evidence of non-constant variance. There is a strong pattern to the points in the scatterplot. Since that assumption is violated, we believe that the regression model may be spurious. It may be beneficial to consider additional or alternative variables in the model.

14 Automatic methods for selecting predictors

14.1 Akaike information criterion, corrected (AICc)

Let n represent the sample size, k represent the number of predictors, and SSE represent the residual sum of squares. The corrected Akaike information criterion is given by the formula:

$$\text{AICc} = \ln(\text{SSE}) + \frac{2(k+2)}{n-k-3}.$$

For the simple linear regression of interest rate (logarithmic) *vs.* unemployment (logarithmic), SSE is approximately 12.585 and AICc is approximately 1.1992. For the simple linear regression of interest rate (logarithmic) *vs.* GDP, SSE is 5.584 and AICc is 0.3865. For the multiple linear regression, SSE is 5.4253 and AICc is 0.35774. Since 0.35774 is the minimum AICc, we conclude that, based on the Aikaike criterion, the multiple regression is the best model.

14.2 Coefficient of determination (r-squared)

The multiple regression also has the highest coefficient of determination ($r^2 = 60.97\%$) compared to the regression that only includes unemployment ($r^2 = 59.83\%$) and the regression that only includes GDP ($r^2 = 9.46\%$). We arrive at the same conclusion: Based on the coefficient of determination, the multiple regression remains the best model.

15 Conclusion

The best model appears to be the multiple regression model. This model predicts the logarithm of the Federal Funds Effective Rate (the interest rate) using the logarithm of the unemployment rate among all US persons between 15 and 64 years of age and the US GDP.

The equation for the regression model is given by:

$$\log(y) = b_2 \log(x_2) + b_1 x_1 + b_0 + e.$$

$$\widehat{\log(y)} = -3.871 \log(x_2) - 0.000029 x_1 + 2.828.$$

The sample *log*-unemployment coefficient ($b_2 = -3.871$) is negative, indicating that as unemployment increases, interest rates decrease. The sample GDP coefficient ($b_1 = -0.000029$) is also negative, suggesting that as GDP increases, interest rates decrease slightly. While the findings confirm our hypothesis about the relationship between unemployment and interest rates, the role of GDP in predicting interest rates remains somewhat questionable. This specific model contradicts our preconception of a positive relationship between interest rates and GDP.