
U-3 Labor Underutilization Rate Forecasting

December 11, 2024

Kurt Fischer
Samuel Martinez Koss
Samuel Park
Minh Phan



Agenda

1. Problem Definition
2. Data Overview
3. Model Development
4. Result Interpretation
5. Closing Argument



Agenda

- 1. Problem Definition**
- 2. Data Overview**
- 3. Model Development**
- 4. Result Interpretation**
- 5. Closing Argument**



Problem Definition

The U-3 Labor Underutilization Rate is a key economic indicator that directly influences policy decisions, labor market strategies, and business planning. Predicting this rate with accuracy is challenging due to the dynamic nature of the labor market, its dependence on macroeconomic factors (e.g., GDP trends), and changing social trends. Currently, stakeholders lack a timely, data-driven forecasting model capable of integrating multiple datasets and providing actionable insights for the upcoming Employment Situation Report in November 2024.

Main Objectives

- **Predict the U-3 Labor Underutilization Rate:** Create an accurate, data-driven model to predict the November 2024 U-3 Labor Underutilization Rate, addressing the need for timely and reliable labor market forecasts.
- **Integrate Diverse Datasets:** Combine and preprocess data from multiple sources, including macroeconomic indicators as well as social indicators to capture the complexities of labor market dynamics.
- **Identify Key Predictive Factors:** Analyze and prioritize the most impactful variables influencing the U-3 rate to improve model performance and interpretability.

Analysis Scope

- **Datasets**
 - Integrate data from trusted sources, such as the U.S. Department of Labor Statistics, the Federal Reserve and U.S. Census Bureau.
- **Timeframes**
 - Analyze data spanning 2004–2024, spanning 20 years.
 - From July 2004 up to October 2024.
- **Key Variables and Features**
 - Macroeconomic indicators: Monthly unemployment rate, government spending, etc.
 - Social indicators: recessions, pandemic, etc..
- **Methodology**
 - Prepare and clean the data to address inconsistencies and fill any missing values.
 - Conduct preliminary exploratory analysis to understand the datasets.
 - Use feature selection to identify the most significant factors influencing labor underutilization.
 - Develop predictive models leveraging both statistical methods.
 - Validate the models using test datasets and evaluation metrics to ensure accuracy and reliability.
 - Evaluate model performance using goodness-of-fit indicators.
- **Boundaries and Exclusions**
 - The analysis is centered on the U-3 Labor Unemployment Rate as the sole outcome variable.
 - Data is limited to the United States.

Agenda

1. Problem Definition
2. Data Overview
3. Model Development
4. Result Interpretation
5. Closing Argument



Data Overview

Index	Data Source	Timeframes	Description	Key Features
1	<u>Federal Reserve Economic Data</u>	Monthly (1948/01 to 2024/11)	Monthly unemployment rate in percentage	Unemployment Rate
2	<u>U.S. Census Bureau</u>	Monthly (1992/01 to 2024/12)	Advance monthly sales data for retail and food services in the U.S.	Advance Monthly Sales for Retail and Food Services
3	<u>U.S. Census Bureau</u>	Monthly (1992/01 to 2024/12)	Advance monthly retail inventories in the U.S.	Advance Retail Inventories
4	<u>U.S. Census Bureau</u>	Monthly (1992/01 to 2024/12)	Advance wholesale inventories in the U.S.	Advance Wholesale Inventories
5	<u>U.S. Census Bureau</u>	Monthly (2004/01 to 2024/12)	High-frequency data on new business formations in the United States.	Business Formation Statistics

Data Overview

Index	Data Source	Timeframes	Description	Key Features
6	<u>U.S. Census Bureau</u>	Monthly (1963/01 to 2024/12)	National and regional data on new single-family homes sold and for sale	New Home Sales
7	<u>Federal Reserve Economic Data</u>	Monthly (1992/01 to 2024/10)	Monthly U.S. GDP history in billions of dollars	GDP
8	<u>Federal Reserve Economic Data</u>	Monthly (1854/12 to 2024/11)	Binary indicator for U.S. recession periods based on NBER's peak-to-trough analysis	Recession Indicators (1 for recession, 0 for No recession)
9	<u>Our World in Data</u>	Monthly (2020/01 to 2024-08)	Comprehensive analysis of pandemic trends	Confirmed COVID-19 Cases per Million People
10	<u>U.S. Census Bureau</u> (Several other data)	Monthly data	Relevant other data	Construction Spending, International Trade, Manufacturers, New Residential Construction

Training Data Set

Continuous

Categorical

Target

Date	Unemployment Rate	Advance Monthly Sales for Retail and Food Services	Advance Retail Inventories	Advance Wholesale Inventories	Business Formation Statistics	Construction Spending	International Trade in Goods and Services	Manufacturers Shipments, Inventories, and Orders	New Home Sales	New Residential Construction	US Monthly GDP History	NBER Based Recession Indicators for the US	Confirmed COVID 19 Cases per Million People
2021-09-01	4.7	620529.0	616586.0	736967.0	428542.0	1669575.0	-77095.0	519478.0	731.0	1625.0	24078.505046	No Recession	12047.380
2021-10-01	4.5	630562.0	622996.0	753664.0	428180.0	1685471.0	-68942.0	530938.0	683.0	1719.0	24574.834426	No Recession	12253.685
2021-11-01	4.1	637775.0	635873.0	765248.0	434368.0	1728158.0	-80153.0	537547.0	787.0	1766.0	24708.440563	No Recession	7789.745
2021-12-01	3.9	632515.0	663693.0	785340.0	428051.0	1757320.0	-80792.0	543199.0	834.0	1913.0	25047.839010	No Recession	7158.870
2022-01-01	4.0	644034.0	682107.0	796786.0	436130.0	1810368.0	-85565.0	550914.0	798.0	1915.0	25028.434631	No Recession	15956.250
2022-02-01	3.8	650522.0	690948.0	819127.0	424993.0	1852805.0	-87070.0	556300.0	781.0	1860.0	25167.058326	No Recession	59861.560
2022-03-01	3.6	664167.0	707318.0	839309.0	413451.0	1878681.0	-101914.0	571865.0	713.0	1879.0	25450.980043	No Recession	12826.660
2022-04-01	3.7	673245.0	710620.0	859103.0	433922.0	1918254.0	-85376.0	576007.0	636.0	1835.0	25593.071580	No Recession	2992.840
2022-05-01	3.6	671040.0	720176.0	878251.0	428999.0	1930664.0	-84255.0	583765.0	648.0	1712.0	25820.539872	No Recession	3406.520
2022-06-01	3.6	675702.0	732487.0	892510.0	405805.0	1918140.0	-81215.0	590870.0	543.0	1745.0	26003.761547	No Recession	7899.660
2022-07-01	3.5	671067.0	739908.0	897756.0	423821.0	1925909.0	-71040.0	581578.0	519.0	1719.0	26069.275734	No Recession	8978.780
2022-08-01	3.6	675107.0	746521.0	909447.0	419403.0	1915377.0	-68522.0	580871.0	644.0	1542.0	26398.241174	No Recession	11486.220
2022-09-01	3.5	673312.0	743570.0	912319.0	423488.0	1914299.0	-71217.0	581689.0	556.0	1613.0	26348.516093	No Recession	8430.860
2022-10-01	3.6	681748.0	742353.0	917839.0	429760.0	1907841.0	-75266.0	587576.0	577.0	1560.0	26667.759068	No Recession	5441.220

Continuous Variables	Categorical Variable
Advance Monthly Sales for Retail and Food Services, Advance Retail Inventories, Advance Wholesale Inventories, Business Formation Statistics, Construction Spending, International Trade in Goods and Services, Manufacturers Shipments, Inventories, and Orders, New Home Sales, New Residential Construction, US Monthly GDP History	NBER Based Recession Indicators for the US Confirmed COVID 19 Cases per Million People

Training Data Set

Purpose: Understanding macroeconomic trends and their relationship to employment levels. The dataset combines economic, business, housing, and health metrics to analyze factors influencing the unemployment rate, the target variable.

Size & Temporal Coverage: The data spans 246 months, from July 2004 to December 2024, offering a monthly granularity that captures both short-term fluctuations and long-term trends.

Missing Data & Handling: COVID-19 Cases Data from before 2020 was missing because it was unrecorded. Therefore, manual fillings were required.

	Unemployment Rate	Advance Monthly Sales for Retail and Food Services	Advance Retail Inventories	Advance Wholesale Inventories	Business Formation Statistics	Construction Spending	International Trade in Goods and Services
count	245.000000	244.000000	244.000000	244.000000	244.000000	2.440000e+02	244.000000
mean	5.814286	462378.704918	568945.672131	577591.368852	276674.672131	1.255360e+06	-52175.717213
std	2.116446	115661.782735	104021.343753	167697.294320	91502.780781	3.799926e+05	13468.005671
min	3.400000	318549.000000	423384.000000	326753.000000	159034.000000	7.583760e+05	-101914.000000
25%	4.200000	367905.000000	480906.000000	428701.500000	212348.250000	9.873002e+05	-62308.750000
50%	5.000000	435787.500000	557213.000000	575452.000000	228894.500000	1.158060e+06	-48752.500000
75%	7.300000	514834.000000	628942.750000	666829.250000	296191.500000	1.439492e+06	-41455.500000
max	14.800000	718867.000000	824656.000000	923660.000000	546415.000000	2.173968e+06	-25840.000000

Agenda

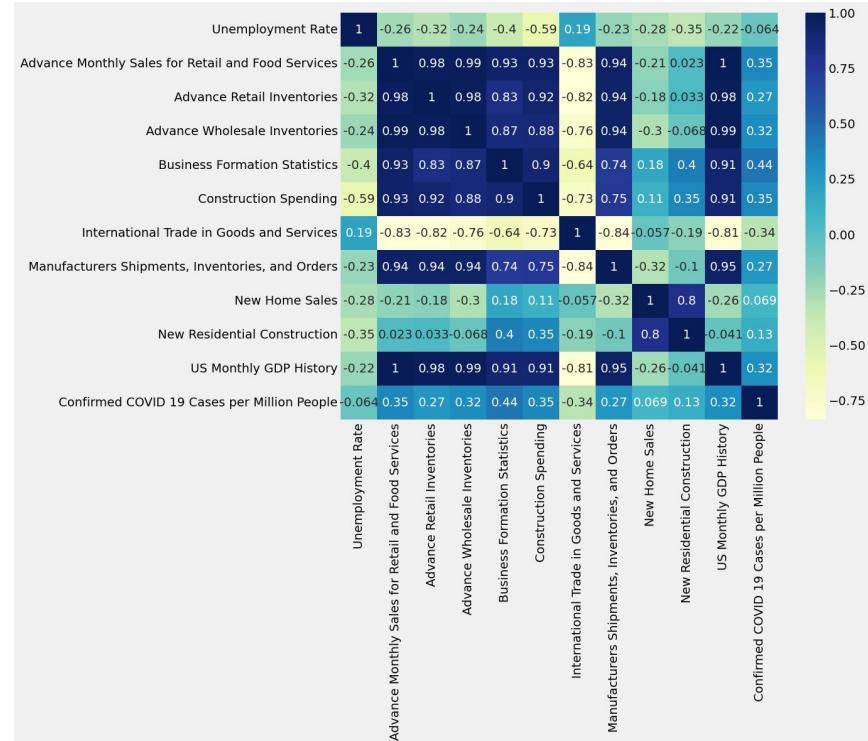
1. Problem Definition
2. Data Overview
3. Model Development
4. Result Interpretation
5. Closing Argument



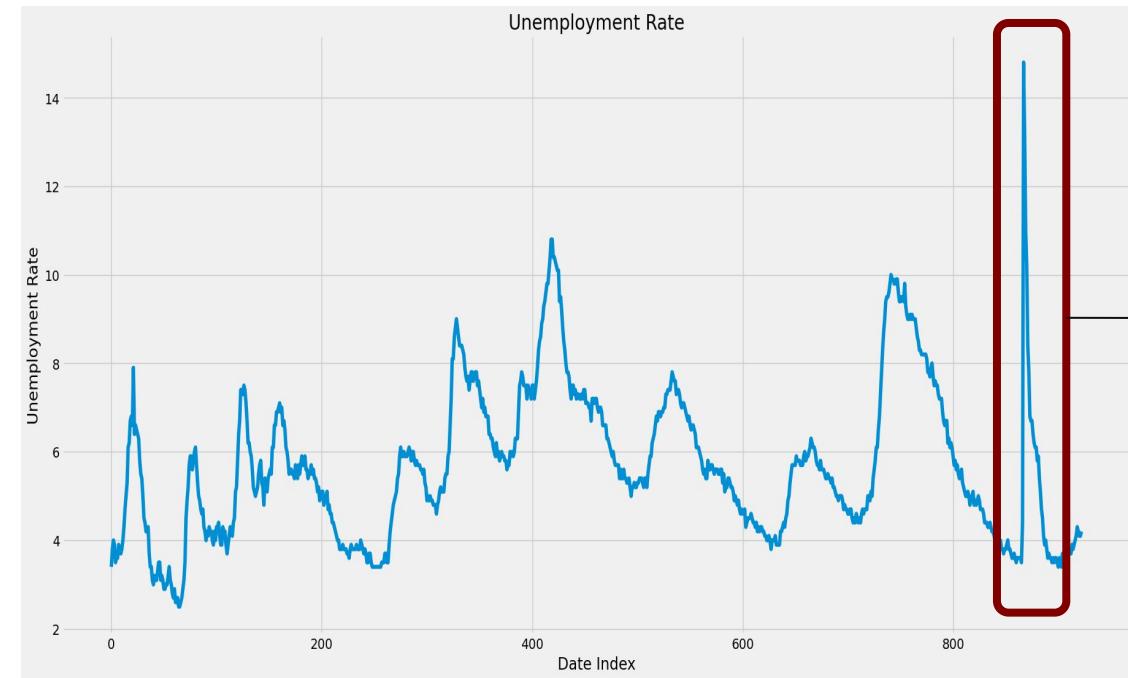
Model Development: Correlation Analysis

Exploratory analysis of numerical predictors motivates future exclusion of redundant series

- Predictors associated with construction have high correlation with unemployment rate
- The majority of series with high overall cross-correlation come from the US Census Bureau's economic indicators

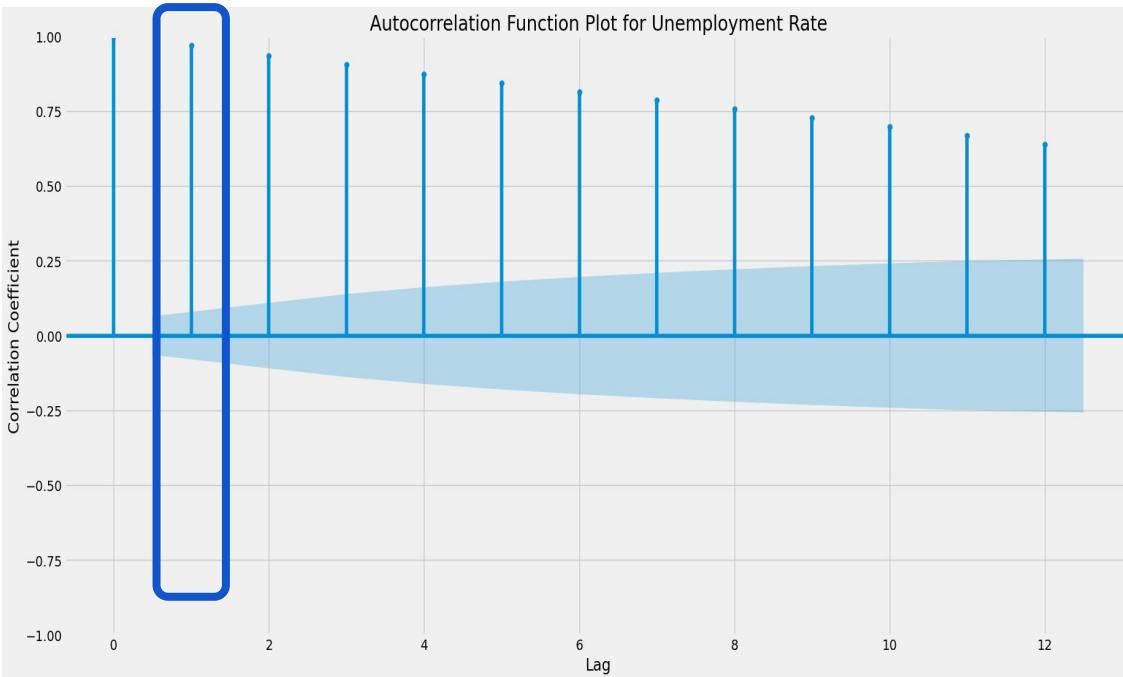


Model Development: Unemployment Time Series



Shock highlights the effects
of the COVID-19 pandemic on
the United States
unemployment rate

Model Development: Unemployment Time Series

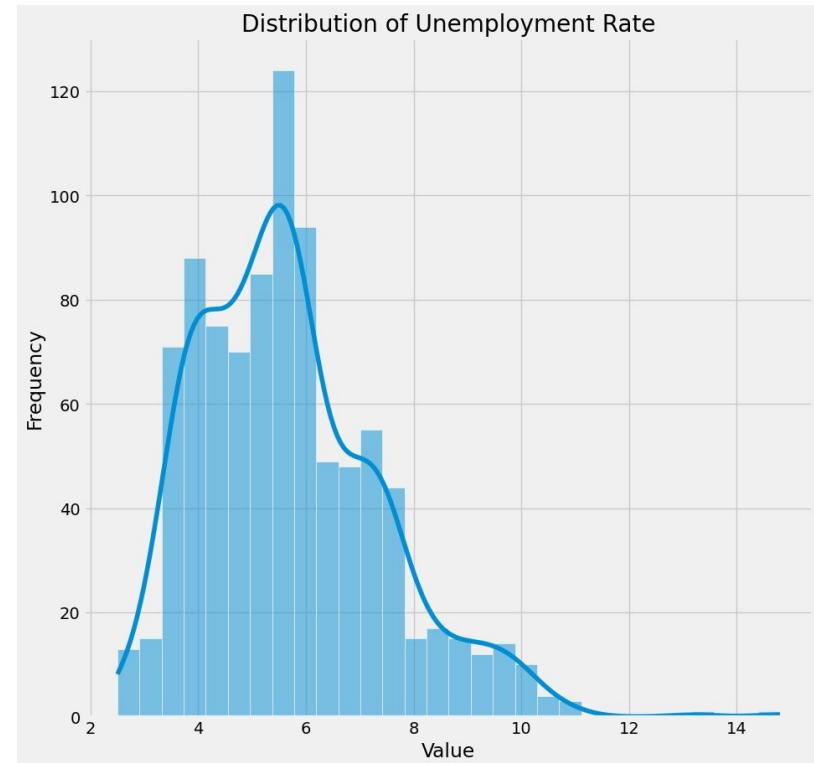
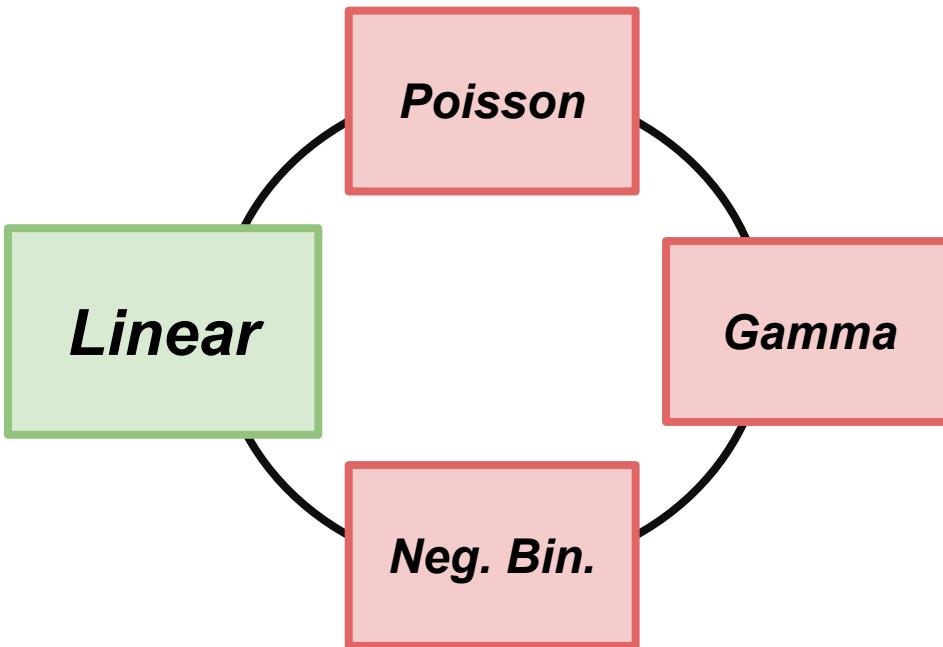


High autocorrelation shows how unemployment could be modeled as an autoregressive function

As a result, we generated a lag predictor:

```
[5] # Include Lag 1 Variable  
lag_merged_df = merged_df.copy()  
lag_merged_df['Lag 1'] = lag_merged_df['Unemployment Rate'].shift(1)
```

Model Development: GLM Selection



Model Development: Forward Selection

```
[134] # The FSig is the sixth element in each row of the Ftest
def takeFSig(s):
    return s[6]

enter_threshold = 0.1
q_show_diary = False
step_diary = []

var_in_model = ['Intercept']

# Step 0: Enter Intercept
X0 = train_data[var_in_model]
y0 = train_data['Underutilization', 1.0)
result_list = Regression.linearRegression(X0, y0)
m0 = len(result_list[5])

# residual_variance = result_list[2] and residual_df = result_list[3]
SSe0 = result_list[2] * result_list[3]

step_diary.append([0, 'Intercept', SSe0, m0] + 4 * [numpy.nan])

# Forward Selection Steps
for istep in range(candidate_count):
    FTest = []
    for pred in candidate_name:
        X = train_data[[pred]]
        if (pred == 'cat'):
            u = X[pred].astype('category')
            ufreq = u.value_counts(ascending = True)
            X[pred] = u.cat.reorder_categories(list(ufreq.index)).copy()
            X = pandas.get_dummies(X.astype('category'), dtype = float)
            X = X.join(X)

        result_list = Regression.linearRegression(X, y)
        m1 = len(result_list[5])
        SSe1 = result_list[2] * result_list[3]

        df_numerator = m1 - m0
        df_denom = n_sample - m1
        if (df_numerator > 0 and df_denom > 0):
            Fstat = ((SSe0 - SSe1) / df_numerator) / (SSe1 / df_denom)
            FSig = F.fif(Fstat, df_numerator, df_denom)
            FTest.append([pred, SSe1, m1, Fstat, df_numerator, df_denom, FSig])

    # Show F Test results for the current step
    if (q_show_diary):
        print("----- | Test Results for the Current Forward Step -----")
        print("Step Number: ", istep)
        print("Step Diary:")
        print("Variable Candidate | Residual Sum of Squares | N Non-Aliased Parameters | F Stat | F DF1 | F DF2 | F Sig")
        for row in FTest:
            print(row)

    FTest.sort(key = takeFSig, reverse = False)
    FSig = takeFSig(FTest[0])
    if (FSig <= enter_threshold):
        enter_var = FTest[0][0]
        SSe0 = FTest[0][1]
        m0 = FTest[0][2]
        step_diary.append([istep+1] + FTest[0])
        X = train_data[enter_var]
        if (enter_var in cat_name):
            X = pandas.get_dummies(X.astype('category'), dtype = float)
            X = X.join(X)
        var_in_model.append(enter_var)
        candidate_name.remove(enter_var)
    else:
        break

forward_summary = pandas.DataFrame(step_diary, columns = ['step', 'Variable Entered', 'Residual Sum of Squares', 'N Non-Aliased Parameters', 'F Stat', 'F DF1', 'F DF2', 'F Sig'])
```

By using forward selection, we:

- Reduce cost of model training
- Reduce overfitting
- Increase generalizability
- Increase interpretability
- Address multicollinearity concerns

Model Development: Forward Selection

Step		Variable Entered	Residual Sum of Squares	N	Non-Aliased Parameters	F Stat	F DF1	F DF2	F Sig
0	0	Intercept	880.6572		1	NaN	NaN	NaN	NaN
1	1	Lag 1	6.5687		2	30,605.6867	1.0000	230.0000	0.0000
2	2	NBER Based Recession Indicators for the US	5.1238		3	64.5771	1.0000	229.0000	0.0000
3	3	Confirmed COVID 19 Cases per Million People	4.9721		4	6.9567	1.0000	228.0000	0.0089
4	4	New Residential Construction	4.8902		5	3.8013	1.0000	227.0000	0.0524
5	5	New Home Sales	4.7590		6	6.2309	1.0000	226.0000	0.0133

Predicted Unemployment Rate (%) = 0.4966

$$\begin{aligned}
 & + 0.9736 \cdot \text{Lag 1} \\
 & - 0.2339 \cdot \text{No Recession Indicator} \\
 & + 0.0000 \cdot \text{Recession Indicator} \\
 & - 0.0000002 \cdot \text{COVID-19 Cases} \\
 & \boxed{- 0.0003 \cdot \text{New Residential Construction}} \\
 & + 0.0003 \cdot \text{New Home Sales}
 \end{aligned}$$

Sanity Check:

Model Development: Review of Parameter Table

Step	Variable Entered	Residual Sum of Squares		N Non-Aliased Parameters	F Stat	F DF1	F DF2	F Sig
		Estimate	Standard Error					
	Intercept	0.4965846	0.0824746	6.0210583	0.0000000	0.3340670	0.6591022	
	Lag 1	0.9735618	0.0080234	121.3403378	0.0000000	0.9577515	0.9893720	
	NBER Based Recession Indicators for the US_No Recession	-0.2339326	0.0403614	-5.7959439	0.0000000	-0.3134654	-0.1543997	
	NBER Based Recession Indicators for the US_Recession	0.0000000	0.0000000	NaN	NaN	0.0000000	0.0000000	
	Confirmed COVID 19 Cases per Million People	-0.0000023	0.0000022	-1.0113405	0.3129350	-0.0000067	0.0000021	
	New Residential Construction	-0.0002551	0.0000831	-3.0692530	0.0024084	-0.0004189	-0.0000913	
	New Home Sales	0.0002845	0.0001140	2.4961842	0.0132679	0.0000599	0.0005090	

- **Recession Indicators** cannot be used for forecasting purposes as recessions are determined after they happen
- The confidence interval for **COVID 19 Cases** includes 0 – inclusion is likely due to data processing / code errors
- Coefficient for **New Home Sales** does not make intuitive sense
 - Consider: Why would the unemployment rate increase when homes are sold?

Model Development: Improved Forward Selection

Step		Variable Entered	Residual Sum of Squares	N	Non-Aliased Parameters	F Stat	F DF1	F DF2	F Sig
0	0	Intercept	880.6572		1	NaN	NaN	NaN	NaN
1	1	Lag 1	6.5687		2	30,605.6867	1.0000	230.0000	0.0000
2	2	New Residential Construction	6.0338		3	20.3020	1.0000	229.0000	0.0000
3	3	New Home Sales	5.4667		4	23.6523	1.0000	228.0000	0.0000
4	4	Manufacturers Shipments, Inventories, and Orders	5.3866		5	3.3732	1.0000	227.0000	0.0676
5	5	Construction Spending	5.0335		6	15.8567	1.0000	226.0000	0.0001
6	6	Advance Monthly Sales for Retail and Food Serv...	4.5922		7	21.6226	1.0000	225.0000	0.0000

Predicted Unemployment Rate (%)=0.7750

$$\begin{aligned}
 & + 0.7750 \cdot \text{Lag 1} \\
 & - 0.0002 \cdot \text{New Residential Construction} \\
 & - 0.0001 \cdot \text{New Home Sales} \\
 & - 0.0002 \cdot \text{Manufacturers Shipments, Inventories, and Orders} \\
 & - 0.0003 \cdot \text{Construction Spending} \\
 & + 0.0003 \cdot \text{Advance Monthly Sales for Retail and Food Services}
 \end{aligned}$$

Agenda

1. Problem Definition
2. Data Overview
3. Model Development
- 4. Result Interpretation**
5. Closing Argument



Result Interpretation: Parameters Table

	Estimate	Standard Error	t Significance	Lower 95 CI	Upper 95 CI	
Intercept	0.7750108	0.1806383	4.2904013	0.0000265	0.4190516	1.1309701
Lag 1	0.9783997	0.0092169	106.1526445	0.0000000	0.9602372	0.9965622
New Residential Construction	-0.0002142	0.0000890	-2.4071526	0.0168843	-0.0003895	-0.0000388
New Home Sales	-0.0001318	0.0001611	-0.8182025	0.4141072	-0.0004492	0.0001856
Manufacturers Shipments, Inventories, and Orders	-0.0000005	0.0000004	-1.2611888	0.2085469	-0.0000014	0.0000003
Construction Spending	0.0000007	0.0000001	6.0585890	0.0000000	0.0000005	0.0000009
Advance Monthly Sales for Retail and Food Services	-0.0000021	0.0000004	-4.6500070	0.0000057	-0.0000029	-0.0000012

- Estimate interpretation:** For a unit increase in the variable, unemployment rate changes by X holding all other variables constant.
- Example:** For each new home sold, the predicted unemployment rate decreases by 0.0002142% holding all other variables constant.

Result Interpretation: Prediction of November 2024 Unemployment Rate

	0
0	5.5255441
1	5.4292050
2	5.3161163
3	5.2858854
4	5.4060875
...	...
227	4.0480713
228	4.1308056
229	4.3324526
230	4.2287927
231	4.1504778

- This DataFrame contains all of the model's predicted unemployment rates for each month in the last 20 years, from July 2004 up to October 2024.
- The predicted value for November 2024 is **4.1419%**, which is decently close to the actual value of 4.2%.

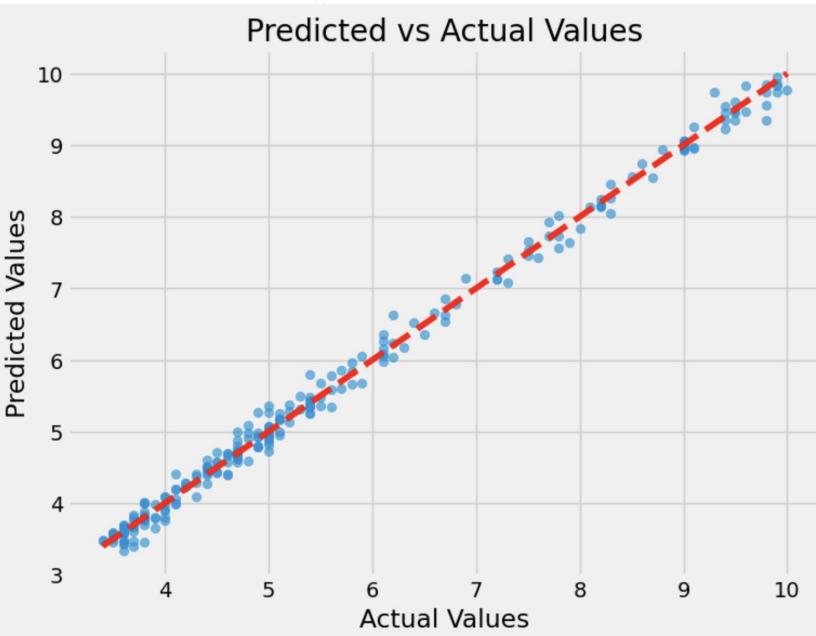
Prediction for month of: 2024-11-01 **4.1419 %**

Model Evaluation: Goodness of Fit

R-Squared: 0.9948

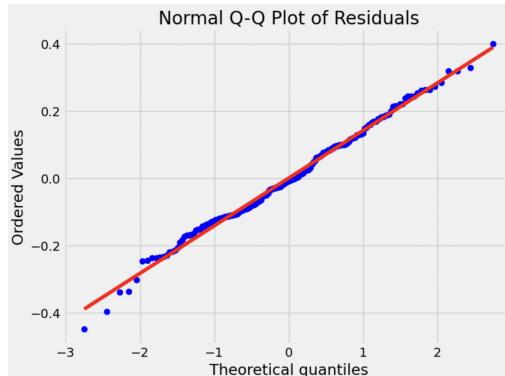
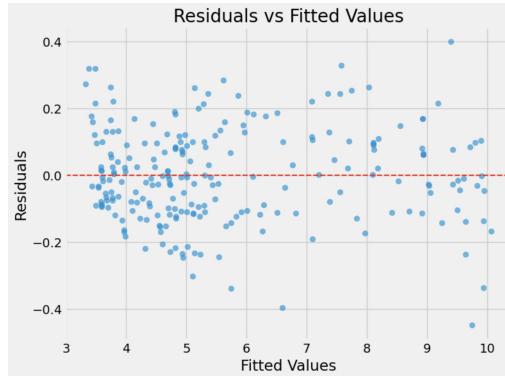
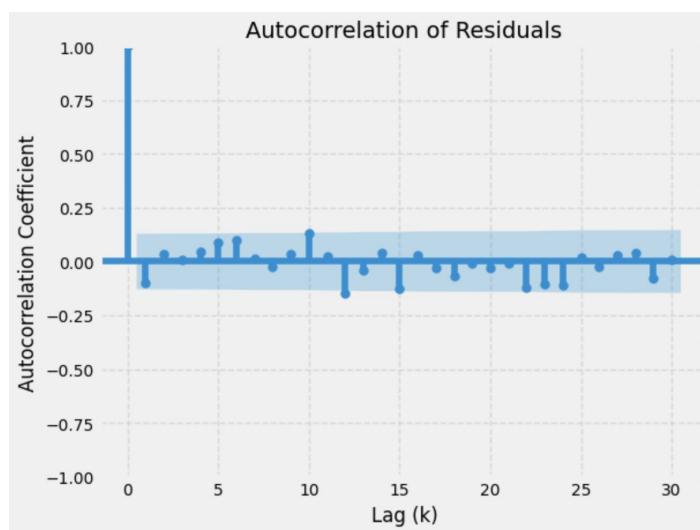
Mean Squared Error: 0.0198

Root Mean Squared Error: 0.1407



- **R-Squared:** 99.48% of the variability in unemployment rate can be explained by the variability in the selected features of the model, which is an extremely good value.
- **MSE/RMSE:** Typical deviation of predicted unemployment rate from the actual value is 0.0198%, which is very low.
- We can see that the points roughly fall on the $y = x$ line in the **Predicted vs. Actual Values** graph, indicating how good the model is at predicting unemployment rate.

Model Evaluation: Are Assumptions for Linear Regression Valid?



- **Autocorrelation of Residuals:** Autocorrelation coefficients should be contained within the 95% confidence bands (independence)
- **Residuals vs. Fitted Values:** Residuals should be randomly scattered around zero with no clear pattern (linearity and homoscedasticity)
- **Normal Q-Q Plot of Residuals:** Points should fall roughly on the red line, with deviations more acceptable near the ends (normality)

Agenda

1. Problem Definition
2. Data Overview
3. Model Development
4. Result Interpretation
5. Closing Argument



Closing Argument

Unemployment Model Prediction for November 2024: **4.1419%**
FED Reported U3 Underutilization Rate, November 2024: **4.2%**

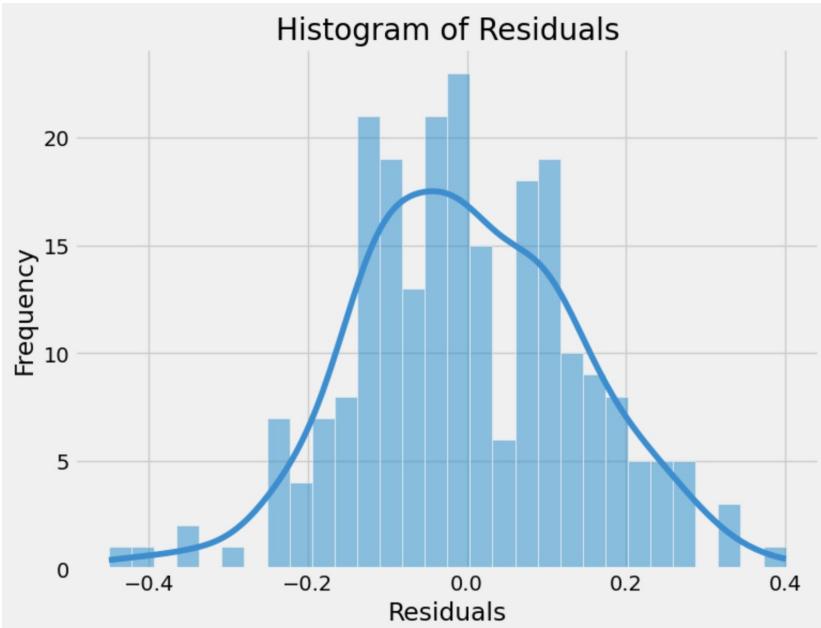
Our team:

- Contextualized the civilian unemployment situation in the United States with trustworthy data that demonstrates an understanding of the dimensionality of the problem. For example, by considering the effects of the COVID-19 pandemic
- Analyzed complex data using statistical methodologies and models covered in this course, like GLM training and residual analysis
- Trained a model with strong predictive power, forecasting the unemployment rate with both precision and accuracy

Thank You

Appendix

Histogram of Residuals



Shapiro-Wilk Test Statistic: 0.9941, p-value: 4.9907e-01

Residuals are normally distributed (fail to reject H₀).

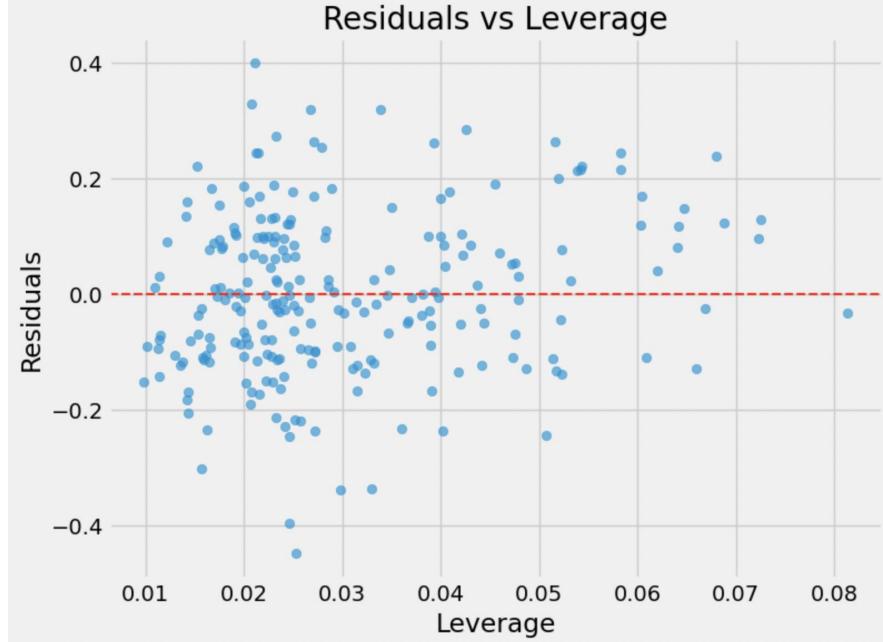
Anderson-Darling Test Statistic: 0.49164081469740495

Critical Values: [0.566 0.645 0.774 0.903 1.074]

Significance Levels: [15. 10. 5. 2.5 1.]

Residuals are normally distributed (fail to reject H₀).

Residuals vs Leverage



	Feature	Shapley Value	Percentage Contribution
1	Lag 1	2263.7941	45.05%
2	New Residential Construction	766.2259	15.25%
3	Construction Spending	727.1658	14.47%
4	New Home Sales	498.1334	9.91%
5	Advance Monthly Sales for Retail and Food Services	428.9974	8.54%
6	Manufacturers Shipments, Inventories, and Orders	340.7347	6.78%