UNIWERSYTET
WARSZAWSKI

Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

Real GDP Estimator with Keynesian Regressors

Ege Güney Kıymaç
Index No. 447987
e.kiymac@student.uw.edu.pl

Sung Kwan Chiu
Index No. 444585
s.chiu@student.uw.edu.pl

November 27, 2025

Contents

1 Introduction **3**
1.1 Overview . . . 3
1.2 Hypothesis . . . 3
1.3 Literature Review . . . 3

2 Data **4**
2.1 Series Characteristics . . . 4
2.1.1 Real GDP . . . 4
2.1.2 CPI . . . 4
2.1.3 Unemployment Rate . . . 4

3 Methodology **5**
3.1 Data Preparation and Examination . . . 5
3.1.1 Dataset Structure Validation . . . 5
3.1.2 Design Matrix Formation . . . 5
3.1.3 Rescaling . . . 6
3.2 Model Estimation . . . 6
3.3 BLUE Evaluation . . . 6

4 Results **7**
4.1 Design Matrix . . . 7
4.1.1 Autoregressive Order . . . 7
4.1.2 Exogenous Order . . . 8
4.2 Model Fit . . . 9
4.3 Residual Analysis . . . 9
4.4 Gauss–Markov Conditions . . . 9
4.5 Residual Distribution . . . 9

1 Introduction

1.1 Overview

Real GDP is a prominent measure of output due to its simplicity, interpretability, and effectiveness in communicating the overall activity of an economy. Alongside academic settings, Real GDP has retained its relevance in politics as one of the go-to measures in public communications regarding economic growth and prosperity. As per its wide-spread adoption, the measure has accumulated a plethora of literature referencing it from a multitude of perspectives.

This paper aims to show that empirical observations of Real GDP in the U.S. economy can be modelled optimally with a **Gauss–Markov** compliant $ARX(p, q)$ specification where the exogenous contributors are theoretical descriptors of output.

1.2 Hypothesis

In the history of economic equilibria, multiple models have proposed mathematically defined relationships between Real GDP and many indicators. From the Keynesian approach, two of said indicators are:

- Price levels (CPI) through the **Aggregate Demand Curve**
- Unemployment rates through **Okun's Law**

The formal hypothesis of this paper is that an $ARX(p, q)$ representation of Real GDP, incorporating CPI and unemployment:

$$ARX_y(p,\ q) = C + \sum_{i=1}^{p} \beta_i y_{t-i} + \sum_{j=1}^{q} \left[\phi_j CPI_{t-j} + \gamma_j u_{t-j}\right] + \varepsilon_t \tag{1}$$

yields OLS parameter estimates that satisfy the **Gauss–**Markov conditions for time-series estimators.

1.3 Literature Review

...

2 Data

The proposed ARX model relies on two exogenous variables alongside the autoregressive target. All three variables are sourced from the Federal Reserve Bank of St. Louis database "FRED". The implementation uses quarterly data from 1990Q1 to 2025Q1 and the following FRED series:

- Real GDP (y): "GDPC1"
- CPI (CPI): "CPIAUCSL"
- Unemployment Rate (u): "UNRATE"

NOTE: All series are seasonally adjusted by FRED.

2.1 Series Characteristics

2.1.1 Real GDP

FRED's Real GDP series is measured in chained 2017 dollars (Bn). The series is updated quarterly, which matches the frequency of the model natively.

2.1.2 CPI

The CPI series of choice for the model records the US-wide city average CPI for all urban consumers. The series is expressed as a chain index with base period 1982-1984=100. Entries are recorded monthly, and quarter-end observations are used to align with the Real GDP series.

2.1.3 Unemployment Rate

Unemployment rate in FRED is recorded as percentages with a monthly frequency. Similar to CPI, quarter-end observations are used to match the frequencies among the series.

3 Methodology

The proposed methodology of this paper can be split into three main components:

- Data Preparation and Examination
- Model Estimation
- BLUE Evaluation

Each component is detailed in the following subsections.

3.1 Data Preparation and Examination

Data preparation does not involve any steps that are specific to a given column of the dataset. Therefore, $X_{i,j}$ notation will be used to denote any given column/observation and β_i will be used to denote any given coefficient. In addition to formatting the X matrix to match the model structure, the preparation steps emphasize transforming the raw data for interpretability while preserving its underlying distributions.

3.1.1 Dataset Structure Validation

More often than not, the performance of a non-robust, linear estimator is governed by the quality of the input variables. Ideally, we look for a dataset that's **free of multicollinearity, extreme outliers, and missing values** while being **correlated to the target**. This general construction achieves a model where all variables have uniquely identifiable contributions to the output, often leading to lower variances and better predictive power.

Autoregressive models present a challenge in this regard since variables selected to be correlated to the target (y) are also correlated to the lagged representation of it within X. Acknowledging this issue, the endogenous portion of the model will only be checked to ensure there are no singularities by perfect correlation inside X. The presence/absence of multicollinearity will be displayed through a correlation heatmap.

Furthermore, data continuity will be prioritized over outlier removal since the model itself is time-dependent. Outliers, if any, will be addressed if they are found to be damaging to the model's performance after the initial estimation.

Finally, any missing values will be handled through linear interpolation of the observations prior and after the occurrence.

3.1.2 Design Matrix Formation

X is formed by converting the model equation specified in Equation 1 into matrix form. The inner product operation specified as $ARX(p,\ q)$ can be written as a matrix multiplication of observations and coefficients:

$$\begin{bmatrix} 1 & X_{1,1} & \cdots & X_{1,m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,m} \end{bmatrix} \times \begin{bmatrix} C \\ \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} \tag{2}$$

where n is the number of observations and m is the number of features (including the intercept) dictated by the $\{p,\ q\}$ orders of the model.

As shown in the above equation, a closed form matrix representation of the model requires p and q to be determined. In this paper, the selection of these hyperparameters will follow the simple and idealistic approach of selecting the lowest possible orders that lead to an efficient estimator.

3.1.3 Rescaling

Rescaling the input matrix (X) allows better comparison of coefficients and ensures that $\sum X^T \varepsilon = 0$ must hold within very tight tolerances[1]. A common approach to achieve this stability is standardization, which transforms each feature to have a mean of 0 and a standard deviation of 1. Standardization is performed column-wise on the input using the formula:

$$Z = \frac{X_{i,j} - \bar{X}_i}{S_i}$$

where $\bar{X}_i$ is the mean of the i-th column, and S_i is the [unbiased] standard deviation of the i-th column.

This transformation has no effect on the error profile of the fitted model as the distribution shapes and inter-variable interactions are perfectly preserved.[2] However, the transformation brings all variables to the same scale and units. As a result, the **Kolmogorov–Smirnov (KS) test** which is sensitive to such differences now becomes available.

3.2 Model Estimation

...

3.3 BLUE Evaluation

...

[1] Differences in magnitude between X columns can lead to numerical ambiguity in orthogonality checks. It is generally preferrable to eliminate any differences in the order of magniutes in X.

[2] The statement holds for any unpenalized estimator of the form $\hat{\beta} = \arg\min_\beta L(y, X\beta)$, since rescaling X can be absorbed into a reparameterization of β.

4 Results

4.1 Design Matrix

First step of the design matrix creation is the identification of optimal $\{p,\ q\}$ parameters for the model. There are two distinct methods to select the autoregressive and exogenous lag order of the model:

- Autoregressive Order: **Autocorrelation Function (ACF)** and **Partial Autocorrelation Function (PACF)** plots of the dependent variable are examined to identify significant lags. The ACF plot helps to identify the overall correlation structure, while the PACF plot isolates the direct effect of each lag.
- Exogenous Order: **Lagged Correlation Matrices** and **Cross-Correlation Function (CCF)** plots between the dependent variable and each exogenous variable are analyzed to determine significant lags. The CCF plot reveals the correlation between the dependent variable and lagged values of the exogenous variables.

4.1.1 Autoregressive Order

Application of ACF yielded a very common pattern of monotonically decreasing correlations, not outlining any specific lag order as superior. Looking for more conclusive insights, PACF was applied.

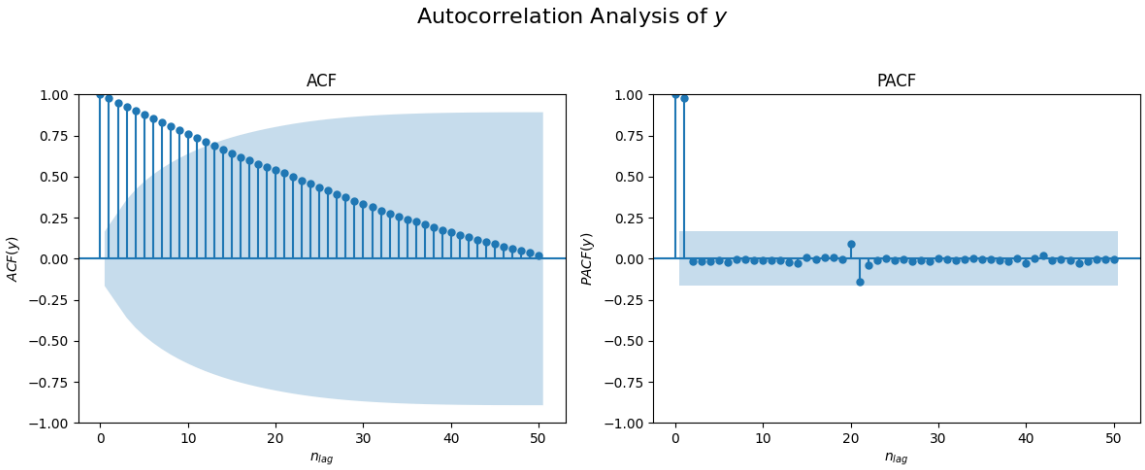

Figure 1: ACF and PACF plots of the dependent (target) variable.

Figure 1 clearly shows the aforementioned decay in ACF. Conversely, PACF shows lag-1 is the only entry with actual partial contribution into the descriptive structure of y.[3] Interestingly, the small spikes around lag-20 of the PACF plot (corresponding to 20 quarters or 5 years) indicate that the COVID-19 shock still has a minor effect on the 2025 Real GDP.

[3]The first spike in both ACF and PACF refers to lag-0 (the variable itself) which is equal to 1 by definition.

In any case, the visualization of the autocorrelation structure presents strong evidence towards the $p = 1$ selection, which is used for the model generation in this paper.

4.1.2 Exogenous Order

The exogenous order selection process started with the visualization of the simple correlation vector y and the lags of exogenous variables.

$\mathbf{i_{lag}}$	$\mathbf{corr(y,\ CPI_{t-i})}$	$\mathbf{corr(y,\ u_{t-i})}$
0	0.989	-0.221
1	0.988	-0.255
2	0.988	-0.221
3	0.988	-0.190
4	0.988	-0.236
5	0.988	-0.205

Table 1: Pearson correlation coefficients between the dependent variable and lags of exogenous variables.

Inspecting the table of coefficients, we that CPI consistently stays at a coefficient of 0.988. This indicates that the autoregressive process of CPI is likely very strong, and related to the autoregressive process explaining y. On the other hand, u shows a more reasonable set of coefficients. Although coefficients of u are also relatively stable, they reside in the reasonable range of $[-0.255,\ -0.190]$. However a lack of decay or noticeable regime changes in the matrices leaves this simple test inconclusive.

To remedy the existence of multicollinearly dependent AR processes, a whitening procedure will be applied to variables before proceeding to the application of CCF. In this case, an $AR(1)$ model will be fitted to each exogenous variable. As the variables are re-scaled to $\bar{X}_i = 0$, no intercept term will be included in the whitening models. After obtaining β_{CPI} and β_u, the following structure will be applied:

$$\begin{aligned} \hat{X}_i &= X_i \beta_i \\ \hat{y}_i &= y \beta_i \\ \rho_{y,i} &= \text{corr}(\hat{y}_i, X_i - \hat{X}_i) \end{aligned} \tag{3}$$

The transformation removes the autoregressive structure from X_i by converting it to errors. To preserve scale equality, y is transformed into $\hat{y}_i$ using the β_i discovered. The resulting correlation coefficients do not absorb processes already explained by the AR models. Applying said transformation, the CCF plots tell a much clear story.

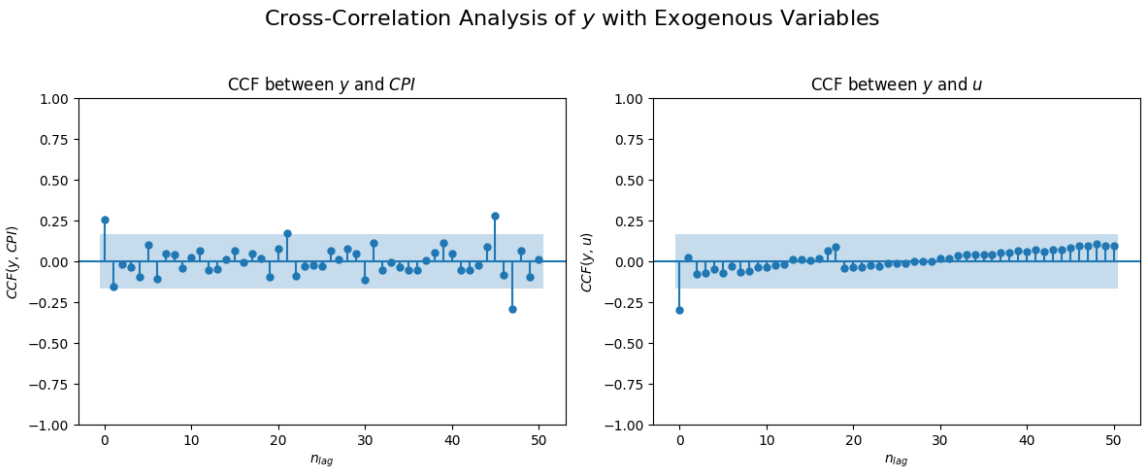

Figure 2: CCF plots of the whitened exogenous variables and the dependent variable.

4.2 Model Fit

...

4.3 Residual Analysis

...

4.4 Gauss–Markov Conditions

...

4.5 Residual Distribution

...