UNIWERSYTET WARSZAWSKI

Uniwersytet Warszawski
Wydział Nauk Ekonomicznych

ARX Model of U.S. Real GDP with Keynesian Exogenous 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

December 1, 2025

Prepared under the supervision of Dr. A.W. Kolasa.

Abstract

This paper presents an autoregressive model of Real GDP with exogenous regressors that are connected to output in the Keynesian framework. A relatively standard/conventional statistical test suite is augmented with newly discovered modifications to the Kolmogorov–Smirnov test to facilitate time-series data and the interdependence brought by it. Building on the original validation methods that are rigorously tested over the decades, the paper manages to present moderate evidence that the estimator specified is efficient, well conditioned, and interpretable.

Contents

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 for 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.

Statistic	Value
$\bar{X}$	16,213.22
S_X	3,837.33
Skew(X)	0.063
Kurt(X)	-0.945

Quantile	Value
0.00	9,951.91
0.25	13,191.67
0.50	16,485.35
0.75	19,062.71
1.00	23,586.54

Figure 1: Descriptive statistics and quantiles of Real GDP.

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.

Statistic	Value
$\bar{X}$	207.32
S_X	49.10
Skew(X)	0.343
Kurt(X)	-0.708

Quantile	Value
0.00	127.50
0.25	163.90
0.50	207.60
0.75	238.99
1.00	319.09

Figure 2: Descriptive statistics and quantiles of CPI.

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.

Statistic	Value
$\bar{X}$	5.71
S_X	1.75
Skew(X)	1.249
Kurt(X)	2.068

Quantile	Value
0.00	3.40
0.25	4.40
0.50	5.40
0.75	6.60
1.00	14.80

Figure 3: Descriptive statistics and quantiles of Unemployment Rate.

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 Model Order Selection

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.

While ACF and PACF can directly be applied to the dependent variable, CCF results can be heavily distorted by nonstationary and the presence of AR structures in the variables being compared. Therefore, the **Box–Jenkins Prewhitening** procedure will be applied to expose a clearer signal of the underlying cross-correlations. The procedure transforms the exogenous variables into residuals of their AR representation, then uses the β coefficient to re-scale the dependent variable. (Hackhard and Romer, 2025) For our purposes, the process will be applied to normalized data, making the inclusion of an intercept redundant. Formally, the $AR(1)$ variant of the process can be shown as:

$$\begin{aligned}\hat{X}_i &= (L^1 X_i)\beta_i \\ \hat{y}_i &= y\beta_i \\ \rho_{y,i} &= \text{corr}(\hat{y}_i, X_i - \hat{X}_i)\end{aligned} \tag{3}$$

Where L^1 is the lag-1 operator.

Following this selection procedure, the candidate $\{p,\ q\}$ are used to form a design matrix for the final confirmation of the **Variance Inflation Factor** (VIF) between exogenous variables. VIF is computed as:

$$VIF_{X_i} = \frac{1}{1 - R^2_{X_i}} \tag{4}$$

Where $R^2_{X_i}$ is the coefficient of determination of the regression of X_i on all other columns of X. In our case, a lag of the target variable is known to be present in X. As exogenous variables are selected to explain the variance in y, it is expected that they will attain a high R^2 as regressors of y_{t-i}. Referring back to Equation 4, we observe that this phenomenon will introduce artificially inflated VIF calculations as R^2 acts as a deflator of the denominator. Therefore, VIF calculations will only be used to calculate the inflation factor between exogenous variable (via a reduced design matrix $X^{(exog)}$), confirming the two variables are not explaining the same portion of variance in y.[1]

3.1.4 Rescaling

Rescaling the input matrix (X) allows better comparison of coefficients and ensures that $\sum X^\top \varepsilon = 0$ must hold within very tight tolerances[2]. 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:

$$Z = \frac{X_{i,j} - \bar{X}_i}{S_i} \tag{5}$$

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.[3] However, the transformation brings

[1] See Section 6.1 for further discussion on the VIF application in this context.

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

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

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.

3.1.5 Analysis of Normalized Input Distributions

With each regressor standardized to zero mean and unit variance, a one unit change corresponds to a 1σ shift in the original scales. However, this unit equality does not make the "likelihood" (or "extremeness") of different variables comparable. The "Z-scores" are comparable in likelihood only if the variables subject to comparison are identically distributed. This part of the methodology works towards identifying such equalities in X.

Concretely, the aforementioned KS test can be used to check for equality of distributions. However, the KS' assumption of **i.i.d. samples** is virtually impossible to satisfy in time-series data. To address this, a modification of the **Bootstrap KS Test** (Præstgaard, 1995) can be applied.

The specific test statistic is computed by calculating

$$D_{i,j} = \sup_x |F_{X_i}(x) - F_{X_j}(x)|$$

Where F_{X_i} and F_{X_j} are the Empirical CDFs of the respective regressors. Afterward, a comparison of the bootstrapped distribution of $D^*_{i,j}$ to the observed $D_{i,j}$ is performed.

The empirical p-value is estimated as the proportion defined by:

$$p_{i,j} = \frac{1}{B+1}\left(\sum_{b=1}^{B}(\mathbb{1}_{\{D^{*(b)}_{i,j} \geq D_{i,j}\}}) + 1\right) \tag{6}$$

Where B is the number of bootstrap samples. To preserve the in-block serial dependence of observations, we employ a **Block Bootstrap** approach instead of the standard bootstrap resampling.[4]

3.2 Model Estimation

The model estimation is conducted using the standard **Ordinary Least Squares** (OLS) approach. Using the derived design matrix X we minimize:

$$\begin{aligned} J(\beta) &= ||y - X\beta||_2^2 \\ &= (y - X\beta)^T(y - X\beta) \end{aligned} \tag{7}$$

[4]See Section 6.2 for further discussion on the time-series application of the KS test.

Which transforms into the first-order condition:

$$\begin{aligned} X^\top \hat{\beta} &= X^\top y \\ \hat{\beta} &= (X^\top X)^{-1} X^\top y \end{aligned} \tag{8}$$

Where $\hat{\beta}$ is the vector of estimated coefficients. This gives us a closed-form, single-step solution to the estimation problem.

3.3 BLUE Evaluation

After estimating the model coefficients, we evaluate the Gauss–Markov conditions to confirm whether the resulting fit is BLUE. Previous stages of the methodology involve checks for the pre-fit Gauss–Markov conditions:

- **Linearity of Parameters:** by definition the model employed is a linear combination of regressors and coefficients. Therefore, this condition is satisfied at the stage of model selection.
- **Full Rank:** $\text{rank}(X) = k$ must hold for $X^\top X$ to be invertible. Therefore, this condition must hold for the model to be estimable via OLS.

The remaining Gauss–Markov conditions concern ε and are verified post-estimation:

- **Spherical Errors:** This condition holds if $\text{Var}(\varepsilon) = \sigma^2 I_n$; implies that the residuals are homoskedastic and uncorrelated. The two criteria can be tested separately using the **Breusch–Pagan Test** for homoskedasticity and the **Breusch–Godfrey Test** for autocorrelation.
- **Strict Exogeneity:** In textbook terms strict exogeneity requires $\mathbb{E}[\varepsilon_i \mid X] = 0$. However, this condition is not directly testable. It's often assumed to hold given weak exogeneity holds, and the design matrix X shows no direct signs of endogenous interactions. In our paper, we will rely on the Breusch–Godfrey test, VIF, and ACF/PACF analysis to check for evidence against weak exogeneity. Given the application of RESET is inappropriate for level based time-series data, as a final check, we will apploy the **Harvey–Collier Test** for to assess model specification. Strong exogeneity will then be assumed, and therefore any finding of BLUE will be referred to as "approximate" under the modelling assumptions.

Noticeably, all conditions except spherical errors can be verified through deterministic calculations using the information attained at-or-before estimation. Fittingly, the tests involved in verifying spherical errors are relatively simple and lead to intuitive test statistics.

The Breusch–Pagan test checks for heteroskedasticity by first deriving the residual based dependent variable g:

$$\begin{aligned} \varepsilon^2 &= (y - X\hat{\beta})^2 \\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n} \\ g &= \frac{\varepsilon^2}{\hat{\sigma}^2} \end{aligned} \tag{9}$$

Where $\hat{\sigma}^2$ is the mean of squared residuals. Therefore, g is a mean-scaled vector of squared errors. Afterward, an auxiliary regression of g on X is performed to attain the R^2 of the fit. The test statistic is then computed:

$$nR^2(g,\ \hat{g}) \sim \chi^2_{k-1} \tag{10}$$

Where k is the number of regressors (including the intercept).

The Breusch–Godfrey test creates a very similar statistic, again derived from a regression of residuals. The test is performed by estimating an $AR(p)$ model of the residuals. The statistic is again derived from R^2 with only a slight modification:

$$(n-p)R^2(\varepsilon,\ \hat{\varepsilon}) \sim \chi^2_p \tag{11}$$

Where p ($= k - 1$ from the BP test) is the number of lags used in the auxiliary regression.

Finally, the Harvey–Collier test is applied to check for model specification. The test is performed by fitting $n - k$ simple regressions of y using the design matrix X. Each fitted model is used to calculate a one-step ahead prediction error, giving us a vector of errors w. The test confirms $H_0 : \mathbb{E}[w_i] = 0\ \forall w$. This is achieved by computing:

$$\begin{aligned} \bar{w} &= \frac{1}{n-k} \sum_{i=1}^{n-k} w_i \\ s_w^2 &= \frac{1}{n-k-1} \sum_{i=1}^{n-k} (w_i - \bar{w})^2 \\ t_{HC} &= \left| \frac{\bar{w}}{s_w / \sqrt{n-k}} \right| \sim t_{n-k-1} \end{aligned} \tag{12}$$

4 Results

4.1 Design Matrix

First step of the design matrix creation is the identification of optimal $\{p,\ q\}$ parameters for the model.

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.

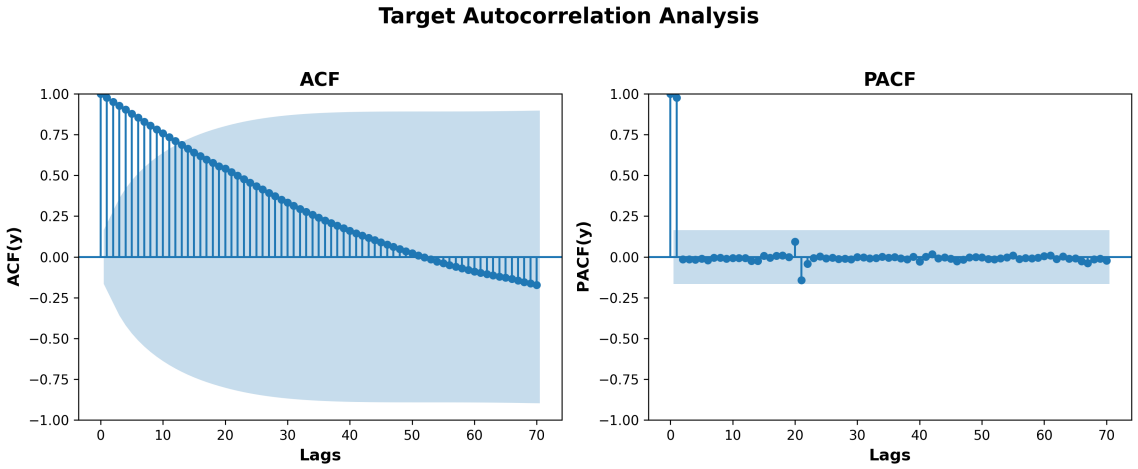

Figure 4: ACF and PACF plots of the dependent (target) variable.

Figure 4 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.[5] 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.

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.

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

$\mathbf{i_{lag}}$	$\mathbf{corr(y,\ CPI_{t-i})}$	$\mathbf{corr(y,\ u_{t-i})}$
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 see 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. This inconclusiveness was expected, and the CCF plots of whitened variables were examined for clearer insights.

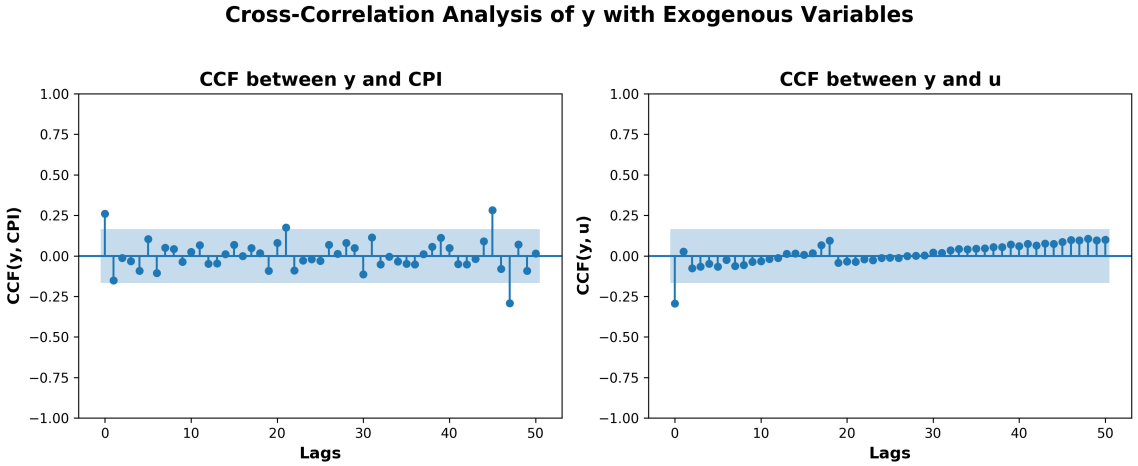

Figure 5: CCF plots of the whitened exogenous variables and the dependent variable.

Although not as clear as the PACF results, Figure 4.1.2 carries much stronger insights into the cross-correlation structure compared to raw correlation coefficients. CPI shows several spikes above the significance bounds with notable lags being $\{0, 45, 47\}$. For u, we observe that lag-0 is the only significant entry. Based on the observations, lag-0 is the only selection that ensures significant lags of both exogenous variables are included. Therefore, the cross-correlation analysis yields $q = 0$ to be

the optimal selection.[6]

Using $q = 0$ introduces the problem of **data leakage**, as current quarter values of exogenous variables would not be available at the time of estimation. This results in a design choice of either using the closest available selection in time (i.e. lag-1) or in significance (selection of lag-45 or lag-47 for CPI). Since introducing 45 or 47 more coefficients per variable would essentially guarantee overfitting, the final decision was in favor of using $q = 1$.[7]

4.1.3 ARX Equation

With the order specification finalized, we were left with an $ARX(1,\ 1)$ model. The objective equation is as follows:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 CPI_{t-1} + \beta_3 u_{t-1} + \varepsilon_t \tag{13}$$

At this stage, all model variables were known, and rescaling was performed to standardize the data. With a concrete design matrix, a final check of VIF between the two exogenous variables was performed to confirm the model parameters did not yield multicollinear exogenous components. $VIF_{CPI,u}$ was found to be 1.023, indicating near-perfect orthogonality between exogenous regressors. Moreover, after standardization, we computed $\text{cond}(X) = 14.75$. The VIF and condition number both indicate a relatively well-conditioned design matrix.

4.1.4 Variable Distribution Equality

The standardized design matrix was used to conduct KS tests to confirm the equality of distributions. For demonstration purposes, both a simple KS test, and a bootstrapped test with 1,000 iterations were performed.

[6] In Equation 1, $q = 0$ would mean not including exogenous variables. But as a lag order selection, it should be interpreted as using the current observations as regressors.

[7] Further regularization techniques, or using sparse lagged estimators are indeed possible but are deemed outside the scope of this paper.

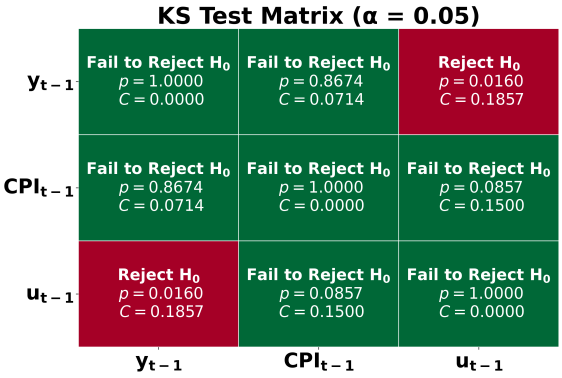

Figure 6: KS Test Matrix

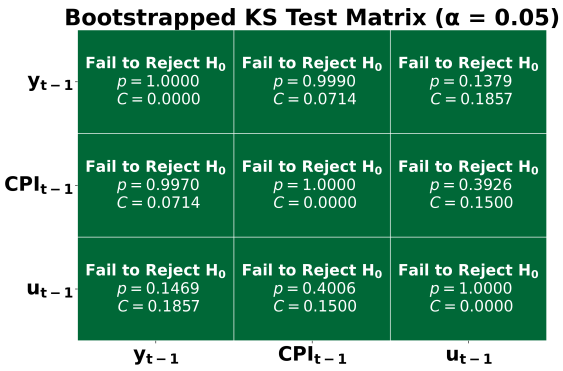

Figure 7: Bootstrapped KS Test Matrix

We observe that the only difference in results between the two tests is in the $\{y_{t-1},\ u_{t-1}\}$ pair. Both tests use H_0: $F_x = F_y$ therefore failing to reject means there was not enough evidence to refute the equality of distributions. However, it is important to note that these tests are not proof of equality, but rather an indication that the distributions are similar enough to not be statistically different. In this paper, we proceed respect the results of the bootstrapped test (Figure 7) and assume approximate equality of distributions for all variable pairs.

NOTE: The p-values annotated in the Standard KS matrix are approximations derived by a 101-term Taylor expansion. (Expansion is ill-defined at $D = 0$ for an even number of terms) The test results are evaluated by a comparison of statistics and are not supposed to yield concrete p-values.

4.2 Model Fit

The fit was performed using Ordinary Least Squares (OLS) on the matrix expansion of Equation 13. The estimation yielded the following coefficients:

Variable	Coefficient
C (β_0)	16,257.27
y_{t-1} (β_1)	3,795.25
CPI_{t-1} (β_2)	14.83
u_{t-1} (β_3)	65.54

Table 2: Estimated coefficients of the ARX(1, 1) model.

With a normalized design matrix, the intercept column expectedly has to take on the re-scaling of inputs. Also, the strong influence of y_{t-1} is inline with the autocorrelation analysis in Section 4.1.1. The only notable observation is the positive coefficient of u_{t-1}, which both by theory and correlation

analysis should have been a negative value. This anomaly is likely attributable to the normalization step introducing negative values into an otherwise positive-only variable.

The overall model fit statistics are as follows:

Statistic	Value
R^2	0.9972
Adjusted R^2	0.9971
RMSE	201.30
MAPE	0.62%
F-statistic	12,002.50
p-value (F)	≈ 0
$X^\top \hat{\varepsilon}$	$-6.25 \times 10^{-9} \approx 0$

Figure 8: Overall model fit statistics.

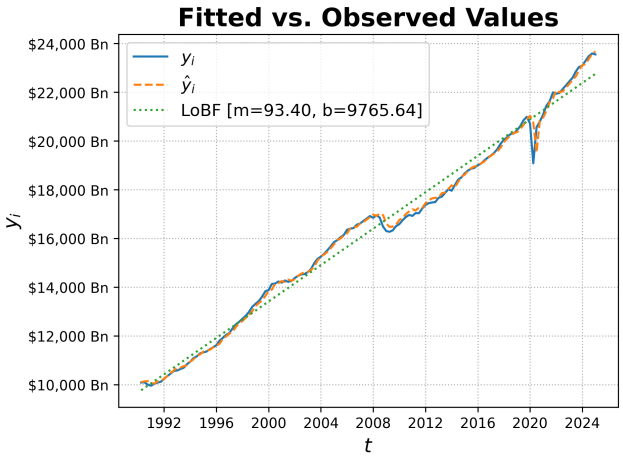

Figure 9: Fitted vs. Observed Values

All metrics indicate a good fit to in-sample data. However, the extremely high R^2 and F-statistic are clear signs of overfitting. Overall, the fit behavior indicates a model where coefficients are interpretable, but predictions should be handled with care. Figure 9, displays 2 expected locations where residuals noticeably deviate from 0:

- The 2008 Financial Crisis
- The 2020 COVID-19 Shock

As expected from a generalized model, the shocks cause an increase in error. However, it's arguable that the model still responds to the shocks more than what would be expected from a well-generalized estimator. With the highly autocorrelated y and the presence of y_{t-1} in the design matrix, the unregularized ARX model expectedly "follows" the shocks. In fact, despite the reasonable fit statistics, the use of **Ridge** or **Lasso** regularization would yield better generalization for the tradeoff of a slightly reduced R^2.

To test the predictive performance of the model, a quick test with the only available out-of-sample data (Q2 2025) was conducted. A prediction of $\$23,632.14$ Bn was achieved against a true value of $\$23,770.98$ Bn. This results in a percentage error of -0.58% which is within the expected range by MAPE.

4.3 Residual Analysis

Given the fit information above, we can safely expect well-behaved residuals. However, a thorough analysis of residuals is of course necessary to confirm the validity of OLS assumptions. Plotting residuals over time presents the opportunity to inspect the exact magnitudes of the errors in the shock periods discussed in Section 4.2.

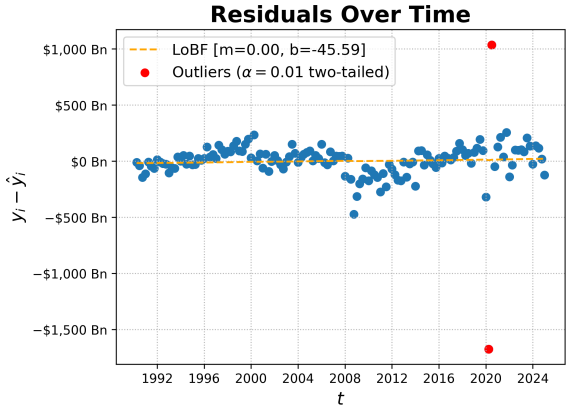

Figure 10: Residuals over time.

Figure 10 outlines two abnormalities in residuals using a 0.99 confidence interval to define outliers. Both extremes occur in the COVID-19 period, but looking at the 2008 period, we clearly see an overestimation of y. Inspecting the line of best fit shows us that the residuals are centered around 0, with a slope of 0. The residuals are heavily concentrated around 0, indicating a tighter empirical distribution compared to a Normal distribution. Fitting a Normal and t-student distribution to the residuals confirms this observation.

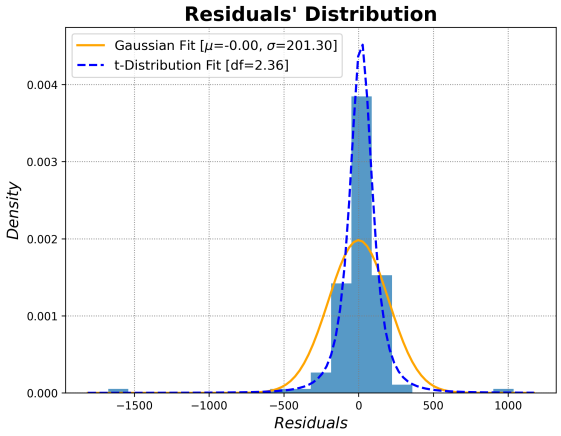

Figure 11: Residual density with fitted distributions.

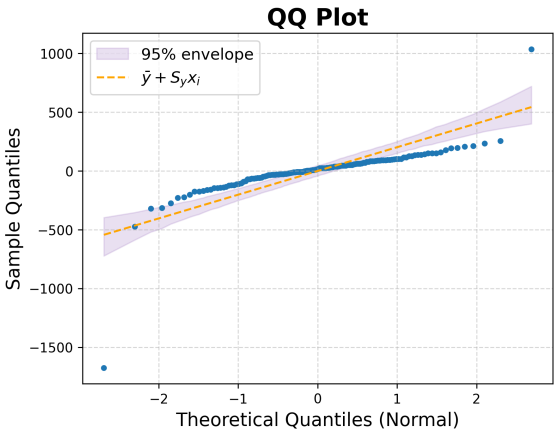

Figure 12: QQ Plot of Residuals vs. $N(0,1)$ quantiles.

The QQ plot of residuals against the Normal distribution (Figure 12) shows that most observations lie on a line with a noticeably smaller slope than the reference line $\bar{y} + S_y x_i$, indicating that the bulk of the residuals is less dispersed than a Normal distribution with the same mean and standard deviation. However, there are extreme points in both tails that lie well outside the 95% simulation envelope, reflecting occasional large shocks. Together with the fitted Student-t distribution with $\widehat{\text{df}} \approx 2.36$, this suggests a very concentrated core with heavier-than-Normal tails rather than Gaussian residuals. As an added confirmation, the Shapiro–Wilk test rejects normality with $W = 0.62$ and $p = 1.81 \times 10^{-17} \approxeq 0$.

As a final descriptor of residual behavior, we inspect the stationarity using the ADF test. Knowing the residuals are zero-centered, we fit an ADF regression with no deterministic terms. With $ADF = -2.25$ we obtain a p-value of 0.02 and reject the null hypothesis of a unit root at the 5% significance level. This confirms the stationarity of residuals and presents further evidence towards homoskedasticity.

4.4 Gauss–Markov Conditions

Although we present strong evidence towards all Gauss-Markov conditions being satisfied, we further confirm our findings using the tests outlined in Section 3.3.

- **Linearity:** The model is linear in parameters by design (Equation 13 & Equation 1).
- **Full Rank:** We computed a lack of perfect multicollinearity with $VIF_{CPI,u} = 1.023 < 5$ and $\text{cond}(X) = 14.75 < 30$

- **Spherical Errors:** Results of the BP test yielded $BP = 3.88$ and $p = 0.73$ failing to reject the null hypothesis of homoskedasticity. Additionally, BG tests for lags $[1, 8]$ (2 years) all failed to reject the null hypothesis of no autocorrelation with the most significant result being $BG(7) = 6.14$ and $p = 0.48$; while the least significant was $BG(3) = 5.19$ and $p = 0.61$.
- **Strict Exogeneity:** While strict exogeneity is not testable, the BG test and correlation analysis of X support the weak exogeneity assumption. Therefore, we elect to proceed with the assumption of exogeneity given the results of all prior diagnostics. The discussed final check of Harvey–Collier test results in $HC = 0.83$ and $p = 0.41$ failing to reject the null hypothesis of $\mathbb{E}[w_i] = 0 \ \forall w$.

5 Conclusion

With the help of normalization of variables, we were able to construct a well-behaved $ARX(1,1)$ model that, against expectations, showed no evidence against BLUE. Although the model showed signs of overfitting in-sample, the statistical tests and equality of variable distributions support the conclusion of an extremely interpretable estimator where coefficients may show connections to empirical elasticities of the used Keynesian regressors.

An unexpected finding was made with the unemployment coefficient, but the model as a whole shows enough stability and in-sample goodness-of-fit to warrant further comparison of empirical coefficients to economic theory. Moreover, we were able to display the use new and developing statistical tests and adapt to the unique challenges brought by the time-series context.

All in all, we believe that out model provides a decent starting point for further research and analysis. While the limited out-of-sample testing did not yield results that immediately display predictive instability, we strongly believe that as more data becomes available, the model will exhibit the well-known symptoms of overfitting. Therefore, we want to stress that the idea behind the creation of this model is to explore where theory and empirical evidence meet, and we encourage future researchers to build upon our findings.

6 Remarks

6.1 Remark 1: VIF Application on a Reduced Design Matrix

Although VIF just a measure and has no constraints on its application, the conventional use is to check for multicollinearity among the complete X matrix. In models with AR components, this application of VIF leads to two issues (assuming the dependent variable is actually autocorrelated):

- The AR components inside X will create processes equivalent to the an AR estimator of y. To demonstrate, let's assume a relationship $y_t \sim y_{t-1}$ exists in an $AR(2)$ model. As shifting the variable in time does not change the assumed relationship, we also have $y_{t-1} \sim y_{t-2}$. As a consequence, the design matrix (which is composed of y_{t-1} and y_{t-2}) will exactly replicate an $AR(1)$ estimator of y. With lags selected to maximize R_y^2, we end up also maximizing the R^2 in the VIF calculation.
- The point above extends to exogenous variables as well. Again, assuming $y_t \sim x_{t-1}$ exists, we also have $y_{t-1} \sim x_{t-2}$. Although $y_{t-1} \sim x_{t-2}$ never directly occurs, the design matrix in this case produces $y_{t-1} \sim x_{t-1}$. This is not an exact replication of the designed relationship, but by autocorrelation of y, it is a very safe assumption that $y_{t-1} \sim x_{t-1}$ will produce an R^2 comparable to the original estimator. As a result, we will again be maximizing the R^2 in the VIF calculation if we're constructing the model to explain the dependent variable as best as it can.

The points discussed show the well-known fact that VIF is not designed for these scenarios. Therefore, many studies, alongside this paper, elect to use VIF on a reduced X matrix stripped of AR components. (Niu and Li, 2022 & Karlström et al., 2023)

6.2 Remark 2: Time-Series Application of KS Test

The i.i.d. assumption of the classic KS test was addressed by Præstgaard with a bootstrapping approach. Yet, the examples, lemmas, and theorems presented in the paper were derived from row-exchangeable samples.

With row-exchangeability, Præstgaard was able to apply standard bootstrap resampling while preserving the continuity of the empirical distribution. However, time-series data inherently requires the row-structure to be preserved; hence rendering the standard bootstrap inappropriate. To preserve the time-dependent dynamics, our methodology elected to use a **Block Bootstrap** approach. This specific adaptation is not documented nor theoretically justified in Præstgaard's work; consequently, while we regard it as a reasonable and practically reliable extension for the purposes of this study, it remains an empirical choice that invites further theoretical investigation.

A block bootstrap 1-sample KS test was derived in a recent preprint paper (Chandy et al., 2025) and applications of bootstrap techniques in KS style tests were demonstrated in multiple studies. (Psaradakis, 2003 & Wyłupek, 2023) These works place bootstrap KS tests on a more solid theoretical footing, even though our specific use case is undocumented.

7 Bibliography

References in the Paper

Chandy, M., Schifano, E., Yan, J., & Zhang, X. (2025, November). Nonparametric Block Bootstrap Kolmogorov-Smirnov Goodness-of-Fit Test [arXiv:2511.05733 [stat]]. https://doi.org/10.48550/arXiv.2511.05733

Block-Bootstrap KS Test for 1-Sample.

Hackhard, B., & Romer, M. (2025, July). 9 Prewhitening; Intervention Analysis – STAT 510 — Applied Time Series Analysis. Retrieved November 28, 2025, from https://online.stat.psu.edu/stat510/Lesson09

Whitening for CCF.

Harvey, A. C., & Collier, P. (1977). Testing for functional misspecification in regression analysis. *Journal of Econometrics*, *6*(1), 103–119. https://doi.org/10.1016/0304-4076(77)90057-4

Karlström, A., Hill, J., & Johansson, L. (2023). Data-Driven Soft Sensors in Refining Processes – Pulp Property Estimation Using ARX - Models. *BioResources*, *18*(4), 8163–8186. Retrieved November 29, 2025, from https://ojs.bioresources.com/index.php/BRJ/article/view/22443

Niu, M., & Li, G. (2022). The Impact of Climate Change Risks on Residential Consumption in China: Evidence from ARMAX Modeling and Granger Causality Analysis [Publisher: Multidisciplinary Digital Publishing Institute]. *International Journal of Environmental Research and Public Health*, *19*(19), 12088. https://doi.org/10.3390/ijerph191912088

Præstgaard, J. (1995). Permutation and Bootstrap Kolmogorov-Smirnov Tests for the Equality of Two Distributions. *Scandinavian Journal of Statistics*, *22*(3), 305–322. Retrieved November 28, 2025, from https://www.jstor.org/stable/4616362

Bootstrap KS Test Paper provides Lemma 1 about the test validity in row-exchangeable data (provided a traditional bootstrap is used instead of a block) In our paper a block-bootstrap is used to preserve the time continuity any arbitrary block generated by the bootstrapping function; which deviates from the exact proof presented in this reference.

Psaradakis, Z. (2003). A Bootstrap Test for Symmetry of Dependent Data Based on a Kolmogorov–Smirnov Type Statistic [Publisher: Taylor & Francis _eprint: https://doi.org/10.1081/SAC-120013116]. *Communications in Statistics - Simulation and Computation*, *32*(1), 113–126. https://doi.org/10.1081/SAC-120013116

Wyłupek, G. (2023). A nonparametric test for paired data [Publisher: Elsevier]. *Journal of Multivariate Analysis*, *198*(C). https://doi.org/10.1016/j.jmva.2023.105229

References in the Source Code

Breusch, T. S. (1978). Testing for Autocorrelation in Dynamic Linear Models* [_eprint: https://onlinelibrary.wiley.com/doi 8454.1978.tb00635.x]. *Australian Economic Papers*, *17*(31), 334–355. https://doi.org/10.1111/j.1467-8454.1978.tb00635.x

David, D., & Wayne, F. (1979). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association*, *74*(366), 427–431. https://doi.org/10.2307/2286348
Original ADF Test.

Durbin, J., & Watson, G. S. (1950). Testing for Serial Correlation in Least Squares Regression: I. *Biometrica*, *37*(3/4), 409–428. https://doi.org/10.2307/2332391

Godfrey, L. G. (1978). Testing for Higher Order Serial Correlation in Regression Equations when the Regressors Include Lagged Dependent Variables. *Econometrica*, *46*(6), 1303–1310. https://doi.org/10.2307/1913830

MacKinnon, J. (1996). Numerical Distribution Functions for Unit Root and Cointegration Tests. *Journal of Applied Econometrics*, *11*(6), 601–618. Retrieved November 21, 2025, from https://www.jstor.org/stable/2285154
Improved ADF Test.

Præstgaard, J. (1995). Permutation and Bootstrap Kolmogorov-Smirnov Tests for the Equality of Two Distributions. *Scandinavian Journal of Statistics*, *22*(3), 305–322. Retrieved November 28, 2025, from https://www.jstor.org/stable/4616362
Bootstrap KS Test Paper provides Lemma 1 about the test validity in row-exchangeable data (provided a traditional bootstrap is used instead of a block) In our paper a block-bootstrap is used to preserve the time continuity any arbitrary block generated by the bootstrapping function; which deviates from the exact proof presented in this reference.

Royston, J. P. (1982). An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Journal of the Royal Statistical Society*, *31*(2), 115–124. https://doi.org/10.2307/2347973
Updated SW Coefficients.

Royston, J. P. (1995). Remark AS R94: A Remark on Algorithm AS 181: The W-test for Normality. *Journal of the Royal Statistical Society*, *44*(4), 547–551. https://doi.org/10.2307/2986146
Updated ADF Coefficients.

Shapiro, S. S., & Wilk, M. B. (1965). An Analysis of Variance Test for Normality (Complete Samples). *Biometrica*, *52*(3/4), 591–611. https://doi.org/10.2307/2333709
Original SW Test.