State-Dependent Fiscal Policy Effects: ARIMA-Based Bootstrap Meets Local Projections

May 12, 2025

Abstract

Using an original panel assembled by Jordà and Taylor—17 OECD countries observed from 1978 to 2019—I replicate their instrumented local-projection estimates of how real GDP per capita reacts to a one-percentage-point change in the cyclically adjusted primary balance. I then reassess the strength of those estimates with simulation-based bootstrap: a time-series (ARIMA) model is fit to the regression errors and used to generate 10 000 synthetic data sets, allowing the stacked Wald statistic for each subsample to be evaluated in a way that accounts for the short, serially correlated nature of annual macro data. The bootstrap reveals that the asymptotic tests used in the paper overstate precision. Once finite-sample uncertainty is taken into account, fiscal tightening only remains significant when the economy is above trend, but the evidence becomes much less certain during downturns and in the full panel. These findings suggest that the output cost of consolidation is robust in booms yet harder to pin down in slumps, and they highlight the value of simulation-based inference for policy analysis based on short time-series panels.

Contents

1 Introduction

Fiscal policy interventions are believed to exert a substantial influence on national output, yet convincing evidence of a *causal* impact remains elusive because governments typically loosen budgets when growth falters and tighten them in booms. Jordà and Taylor (2024) confront this endogeneity with an instrumented local-projection (IV–LP) design and—using a panel of 17 advanced OECD economies—report large, state-dependent output responses to discretionary changes in the cyclically adjusted primary balance (CAPB). To evaluate the statistical significance of these estimated responses, they rely on stacked Wald (χ^2) statistics benchmarked against asymptotic critical values.

A growing literature cautions that annual macro panels are short and serially correlated, so the *effective* number of observations is far smaller than the raw sample size; standard asymptotics may therefore under-state uncertainty (e.g. Plagborg-Møller and Wolf, 2021a). Motivated by this concern, the present thesis first *replicates* Jordà & Taylor's IV–LP estimates exactly—maintaining the same data, instruments, and specifications—and then subjects their joint test statistic to a finite-sample assessment. The assessment employs an ARIMA-based residual bootstrap that mirrors the dependence structure of the data, generating synthetic samples under the null hypothesis of no fiscal effect; this allows the construction of a reference distribution for the $\chi^2_{\mathrm{df}=H+1}$ statistic against which the observed value can be compared.

By comparing fiscal-policy-induced changes in output (and their associated χ^2 statistics) with the null distribution estimated from these simulated samples, I evaluate whether the statistically significant results reported in Jordà and Taylor (2024) can be distinguished from outcomes that might plausibly arise under the null hypothesis. Put differently, we examine whether the observed evidence rises above baseline noise sufficiently to warrant a causal interpretation. This core exercise motivates the sections that follow: we outline the local-projection framework, detail the instrumental-variable strategy that addresses endogeneity, and describe the ARIMA-based bootstrap that grounds our inference procedure. Together, these methodological steps allow us not only to test whether expansions or recessions respond differently to fiscal shocks, but also to assess whether the estimated causal relationships remain robust—rather than reflecting random fluctuations. The analysis thus carries both theoretical significance for understanding macroeconomic dynamics and practical implications for the design of counter-cyclical policy.

2 Literature Review

Local projections (LPs), introduced by Jordà (2005), have become a standard tool for estimating impulse responses in macroeconomic research due to their flexibility and transparency. They are especially attractive in settings with externally identified shocks and short time series. For example, Nakamura and Steinsson (2018) use LPs with high-frequency monetary shocks to trace the effects of Federal Reserve announcements on interest rates and expectations, while Ramey and Zubairy (2014) apply LPs to narrative fiscal shocks to assess whether govern-

ment spending multipliers vary across business cycle states. These and other applications have contributed to the growing reliance on LPs across empirical macroeconomics—but also raise questions about the reliability of inference in the small samples and persistent environments where LPs are often applied.

Recent theoretical and simulation-based research has challenged the robustness of standard inference procedures used with LPs. Plagborg-Møller and Wolf (2021b) show that LPs and VARs estimate the same impulse responses in population, implying that differences in empirical results must reflect finite-sample behavior rather than structural robustness. Herbst and Johannsen (2024) further demonstrate that LP coefficients and their standard errors are biased in small samples—particularly when using Newey–West corrections with serially correlated data. Stock and Watson (2018) similarly find that LP-based confidence intervals may severely under-cover true effects, and recommend bootstrap or simulation-based methods as more reliable alternatives. Related concerns appear in Kilian and Kim (2011) and Brugnolini (2018), who show that design choices like truncation horizons or smoothing can distort both the estimates and their uncertainty. Choi and Chudik (2019) reinforce these findings in panel settings, highlighting how oversmoothing and fixed-T inference problems can degrade coverage. Collectively, these contributions call for more careful treatment of standard errors in LP studies and motivate simulation-based inference methods that more accurately reflect the finite-sample and dynamic features of macroeconomic data.

More broadly, simulation-based methods such as bootstrapping and permutation testing have gained prominence in time series econometrics as tools to improve inference under autocorrelation and small-sample constraints. Wang and Van Keilegom (2007) develop a bootstrap-based test for parametric regression models with dependent errors, showing that resampling improves size control relative to asymptotic benchmarks. Romano and Tirlea (2021, 2024) extend permutation testing to time series by constructing studentized test statistics that retain validity under weak dependence, offering an exact or near-exact alternative to standard methods. Beyond these, Kugiumtzis (2002) reviews surrogate data techniques for testing nonlinearity in economic and financial time series. These approaches generate synthetic series that preserve key properties of the observed data while allowing hypothesis testing without strict parametric assumptions. Together, this body of work underscores that robust inference in time series increasingly depends on simulation-calibrated procedures that account for temporal dependence and model uncertainty.

These concerns extend beyond economics. In climate science, Mudelsee (2014) uses block and surrogate bootstraps to construct confidence intervals for trends in autocorrelated paleoclimate records, showing how resampling corrects underestimated uncertainty. In hydrology, Noguchi et al. (2011) apply sieve bootstraps to detect long-run shifts in ice phenology, finding that classical tests often overstate significance under serial dependence. In neuroscience, Saravanan et al. (2020) employ hierarchical bootstrapping to improve inference in nested time series data—structures that mirror macroeconomic panels with country and year-level clustering. These applications reinforce the case for ARIMA-based bootstrapping in this thesis:

to obtain valid joint inference in the presence of autocorrelation and short samples, where asymptotic methods may mislead.

In summary, while prior studies have extensively applied local projections to estimate fiscal and monetary policy effects, my thesis contributes by refining the inference methods used in these applications. Existing works, such as Ramey and Zubairy (2014) and Nakamura and Steinsson (2018), have successfully identified causal shocks and documented state-dependent responses, often using asymptotic standard errors or conventional resampling techniques. Building on recent research by Plagborg-Møller and Wolf (2021b), Herbst and Johannsen (2024), and others, this thesis extends simulation-based inference to better reflect the challenges posed by short, serially correlated macro panels. Specifically, I implement an ARIMA-based bootstrap procedure that captures the dynamic structure of regression residuals, allowing for more realistic finite-sample inference. This application to IV–LP fiscal policy analysis complements existing approaches by providing a simulation-calibrated benchmark for statistical significance, while preserving the flexibility and transparency that make local projections attractive.

3 Technical Report on Jordà and Taylor (2024)

Jordà and Taylor (2024) examine the dynamic effects of fiscal policy using a panel of 17 advanced OECD economies observed annually from 1978 to 2019. The outcome variable is log real GDP per capita, and the policy variable of interest is the annual change in the cyclically adjusted primary balance (dCAPB), sourced from the narrative fiscal consolidations database of Guajardo et al. (2014), updated through 2019. To reduce confounding from automatic stabilizers, the analysis restricts attention to announced, discretionary fiscal changes.

Additional control variables include lagged output growth, lagged fiscal shocks, changes in the debt-to-GDP ratio, and the lag of the HP-filtered cyclical component of log output. The HP filter is applied with smoothing parameter $\lambda = 100$, following standard practice for annual macroeconomic data (Hodrick and Prescott, 1997).

3.1 Causal Identification

Although local projections offer a flexible method for tracing dynamic responses, they do not on their own identify causal effects. In practice, fiscal policy often responds endogenously to prevailing macroeconomic conditions. This creates a risk that ordinary least squares (OLS) estimates may attribute movements in output to fiscal actions that were themselves partly driven by the economic cycle.

To address this concern, Jordà and Taylor (2024) isolate fiscal shocks that are plausibly exogenous to current conditions by using an instrumental variable strategy. Conceptually, they seek to estimate the difference in expected output under two counterfactual paths—one with a discretionary fiscal intervention and one without:

$$R_{\mathrm{dCAPB},y}(h,\delta) \;=\; \mathbb{E}\big[y_{i,t+h} \;\big|\; \mathrm{dCAPB}_{i,t} = \mathrm{dCAPB}_{i,0}{+}\delta;\, \mathbf{x}_{i,t}\big] - \mathbb{E}\big[y_{i,t+h} \;\big|\; \mathrm{dCAPB}_{i,t} = \mathrm{dCAPB}_{i,0};\, \mathbf{x}_{i,t}\big].$$

By instrumenting $\text{dCAPB}_{i,t}$ with a variable uncorrelated with short-term output fluctuations, they estimate the causal impact of discretionary policy shifts, not mere correlations.

3.2 Two-Stage Estimation Strategy

The empirical approach uses a two-stage least squares (2SLS) estimation embedded within a local projection framework.

Stage 1: Instrumenting Discretionary Fiscal Policy. The first stage regresses the fiscal policy variable on an instrument that captures the magnitude of announced fiscal measures, normalized relative to each country's historical baseline:

$$\text{dCAPB}_{i,t} = \alpha^{(1)} + \mathbf{X}'_{i,t}\,\boldsymbol{\gamma} + \lambda\,\texttt{size\_transformed}_{i,t} + u_{i,t}, \tag{1}$$

where $\texttt{size\_transformed}_{i,t} = \texttt{size}_{i,t} - \overline{\texttt{size}}_i$ represents the deviation of fiscal intervention size from its country-specific mean. The fitted values $\widehat{\text{dCAPB}}_{i,t}$ from this regression constitute the exogenous fiscal shocks used in the second stage.

Stage 2: Local Projection Estimation. The second stage estimates the dynamic output response to a fiscal shock over horizons $h = 0, \ldots, H$ using local projections:

$$\Delta_h y_{i,t} = \alpha_h^{(2)} + \beta_h\,\widehat{\text{dCAPB}}_{i,t} + \mathbf{X}'_{i,t}\,\delta_h + \alpha_i + \gamma_t + \varepsilon_{i,t}^{(h)}, \tag{2}$$

where $\Delta_h y_{i,t} = y_{i,t+h} - y_{i,t-1}$ denotes the h-year cumulative change in log real GDP per capita. Country fixed effects α_i and year fixed effects γ_t account for unobserved heterogeneity and global shocks, respectively. The coefficient β_h traces the impulse response of output to a one-unit exogenous change in dCAPB.

3.3 State-Dependent Responses and Replication

The original analysis further explores heterogeneity in fiscal effects across business cycle phases. Observations are classified into booms or slumps based on the sign of the lagged cyclical component $c_{i,t-1}$ from the HP-filtered decomposition:

$$y_{i,t} = \tau_{i,t} + c_{i,t}.$$

A period is defined as a boom if $c_{i,t-1} > 0$ and a slump otherwise. Local projections are estimated separately for each regime.

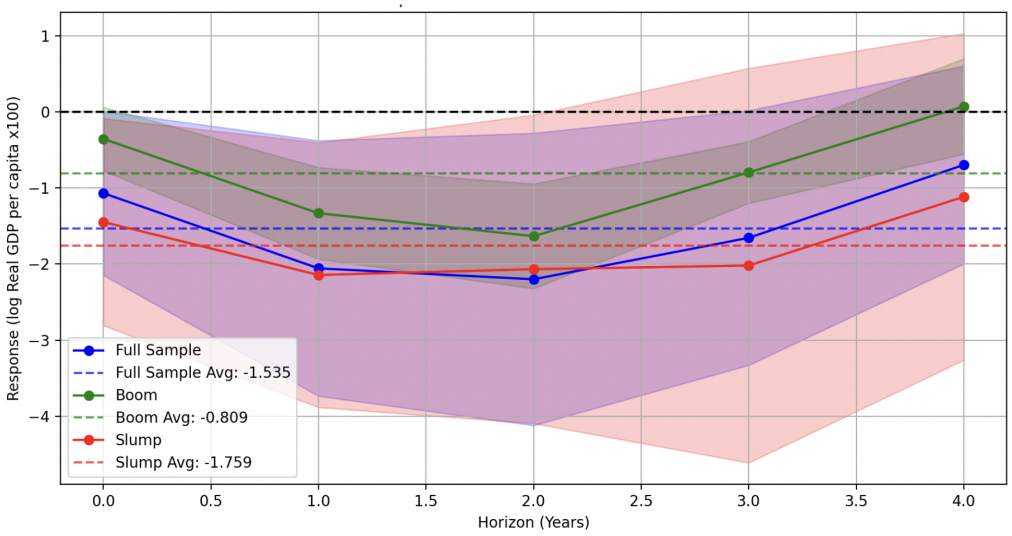

Figure 1: Replicated cumulative impulse responses of log real GDP per capita to a one-percentage-point tightening in CAPB, estimated via IV–LP with 95% bootstrap confidence bands. Output losses are largest and most precisely estimated during booms.

Impulse responses show more pronounced and statistically precise output contractions during booms, while estimates in slumps are weaker and less robust. Figure 1 presents a replication of these state-dependent results, produced using the authors' official dataset and replication package, reimplemented in Python for consistency.

3.4 Joint Inference and Asymptotic Testing

To formally assess whether the dynamic fiscal effects are jointly significant, Jordà and Taylor (2024) compute a Wald-type statistic that tests the null hypothesis that all impulse responses are zero:

$$\chi^2_{\text{joint}} = \hat{\boldsymbol{\beta}}' \, \mathbf{V}^{-1} \, \hat{\boldsymbol{\beta}}, \tag{3}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_H)'$ and $\mathbf{V}$ is the estimated covariance matrix. Under the null, $\chi^2_{\text{joint}} \sim \chi^2_{H+1}$ asymptotically.

Table 1 summarizes the χ^2 test values and p-values as reported in their figures. All three samples reject the null at conventional levels using asymptotic critical values, with the strongest significance observed during booms.

Table 1: Joint Significance of Cumulative Fiscal Impulse Responses in Jordà and Taylor (2024)

Subsample	**Observed χ^2**	***p*-value**
Full sample	19.4	0.002
Booms ($c_{t-1} > 0$)	58.3	0.000
Slumps ($c_{t-1} \leq 0$)	17.2	0.004

Note: Joint Wald statistics test $R(h) = 0$ across horizons $h = 0, \ldots, 4$, with 5 degrees of freedom. p-values correspond to asymptotic χ^2_5 distribution.

4 ARIMA-Based Bootstrap Inference

While Jordà and Taylor (2024) rely on asymptotic χ^2 critical values to assess the joint significance of impulse responses, this approach assumes a large effective sample size and weak serial dependence in the residuals. These assumptions are problematic for short annual macro panels, where persistent dynamics can distort standard errors and lead to overconfident inference (Plagborg-Møller and Wolf, 2021a).

To address this, I implement a residual-based bootstrap procedure that simulates counterfactual samples under a data-generating process calibrated to the observed residual dynamics. My objective is to assess whether the dynamic response coefficients from the IV–LP regressions are statistically distinguishable from zero once finite-sample variation is accounted for.

The goal is to construct a finite-sample reference distribution for the joint Wald statistic by simulating counterfactual data consistent with the null hypothesis:

$$H_0 : \beta_0 = \beta_1 = \cdots = \beta_H = 0.$$

Under this null, the observed impulse response path arises solely from stochastic noise governed by the estimated ARIMA process. This allows us to obtain p-values that better reflect small-sample uncertainty than those implied by asymptotic χ^2 distributions.

4.1 ARIMA Model Specification and Selection

To capture the temporal dependence structure present in the residuals of the IV–LP regressions, we fit an autoregressive integrated moving average (ARIMA) model of order (p, d, q). The ARIMA framework provides a flexible, parametric means of modeling autocorrelation through both autoregressive and moving-average components, allowing the resampled series to retain features of the original data's dynamic behavior.

Because the residual process exhibits stationarity (as established in Appendix A.2), we restrict attention to the class of ARIMA$(p, 0, q)$ models—stationary ARMA processes with no differencing. The general form is:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2), \tag{4}$$

where y_t denotes the residual at time t, c is a deterministic constant, and the innovations ε_t are assumed to be serially uncorrelated, homoskedastic, and Gaussian. The autoregressive (AR) component captures persistence in the series via its dependence on past values of y_t, while the moving average (MA) component accounts for serial correlation through lagged shocks.

This specification can be compactly rewritten using the backshift operator B as:

$$\Phi(B)\, y_t = c + \Theta(B)\, \varepsilon_t,$$

where $\Phi(B) = 1 - \phi_1 B - \cdots - \phi_p B^p$ and $\Theta(B) = 1 + \theta_1 B + \cdots + \theta_q B^q$ are lag polynomials applied to the series and the shocks, respectively.

Model selection is conducted by fitting multiple (p, q) combinations and minimizing the Akaike Information Criterion (AIC), which penalizes complexity to mitigate overfitting. The AIC is defined as:

$$\text{AIC} = 2k - 2\ln(\hat{L}),$$

where k is the number of free parameters and $\ln(\hat{L})$ is the maximized log-likelihood under the assumed Gaussian error distribution. The latter takes the form:

$$\ln(\hat{L}) = -\tfrac{n}{2}\ln(2\pi) - \tfrac{n}{2}\ln(\hat{\sigma}^2) - \tfrac{1}{2\hat{\sigma}^2}\sum_{t=1}^{n}\hat{\varepsilon}_t^2,$$

where $\hat{\sigma}^2$ is the estimated innovation variance and $\hat{\varepsilon}_t$ are the one-step-ahead prediction errors.

In our application, the AIC-minimizing specification is ARIMA(3,0,2), which incorporates three lags of the residual series and two lags of the innovations. The resulting model is:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2}, \tag{5}$$

with parameter set

$$\{\phi_1, \phi_2, \phi_3, \theta_1, \theta_2, c, \sigma^2\},$$

corresponding to $k = 7$ in the AIC formula.

The inclusion of both AR and MA terms allows the model to simultaneously account for medium-run persistence and short-run innovations in the residual process. This dual capability is essential for generating realistic counterfactual paths in the bootstrap procedure, as it ensures that the simulated series replicate both the amplitude and autocorrelation structure observed in the empirical data.

Stationarity Verification. Since ARIMA-based resampling assumes a stable process, we first confirm that the residuals are stationary. This is verified via both the Augmented Dickey-Fuller (ADF) test and a root condition test based on the AR polynomial. The ADF test rejects the null of a unit root at conventional significance levels across all subsamples (Full, Boom, Slump), and the estimated characteristic roots all lie strictly outside the unit circle.

Full details, test statistics, and visualizations (including the eigenvalue plot) are reported in Appendix A.2.

Interpretation. The AR components capture medium-run persistence in the residuals, while the MA components model short-run innovations. This flexible structure ensures that simulated paths reproduce the residual dynamics observed in the actual data—crucial for generating credible counterfactuals in our bootstrap procedure.

4.2 Bootstrap Algorithm

We simulate counterfactual economies that are consistent with the null hypothesis H_0 by assuming that changes in dCAPB exert no effect on output. Let $B = 10{,}000$ denote the number of bootstrap replications and H the forecast horizon. The steps below describe how we construct and use these synthetic data to approximate the finite-sample distribution of the joint Wald statistic:

1. **Estimate restricted model under the null.** Begin by estimating a fixed effects regression of log real GDP on year and country dummies, excluding dCAPB and its instruments:

$$y_{i,t} = \alpha_i + \gamma_t + \varepsilon_{i,t},$$

where $y_{i,t} = \log(\text{GDP}_{i,t})$, and the residuals $\hat{\varepsilon}_{i,t}$ capture the portion of output fluctuations unexplained by structural fiscal policy. These residuals serve as the basis for modeling the null distribution.

2. **Fit an ARIMA(3,0,2) model to the residuals.** Let $\hat{\varepsilon}_t$ denote the pooled residual series. Estimate the stationary ARIMA model:

$$\hat{\varepsilon}_t = c + \phi_1\hat{\varepsilon}_{t-1} + \phi_2\hat{\varepsilon}_{t-2} + \phi_3\hat{\varepsilon}_{t-3} + \eta_t + \theta_1\eta_{t-1} + \theta_2\eta_{t-2},$$

where $\eta_t \sim \mathcal{N}(0, \hat{\sigma}^2)$ is white noise.

3. **Simulate B synthetic residual paths.** For each bootstrap replication $b = 1, \ldots, B$, generate a time series $\{\varepsilon_t^{*(b)}\}$ by simulating from the fitted ARIMA(3,0,2) model. Each draw preserves the estimated autocorrelation and variance structure under the null.

4. **Construct synthetic output series.** For each b, reconstruct a synthetic log-GDP series:

$$y_{i,t}^{*(b)} = \widehat{y}_{i,t}^{\text{FE}} + \varepsilon_{i,t}^{*(b)},$$

where $\widehat{y}_{i,t}^{\text{FE}} = \hat{\alpha}_i + \hat{\gamma}_t$ are the deterministic fixed effects from the restricted model. This step ensures that the synthetic series retain the observed mean structure but reflect stochastic variation only through bootstrapped residuals.

5. **Reclassify business cycle states.** Apply the HP filter to the synthetic series $y_{i,t}^{*(b)}$ to obtain:

$$y_{i,t}^{*(b)} = \tau_{i,t}^{*(b)} + c_{i,t}^{*(b)},$$

where $\tau_{i,t}^{*(b)}$ minimizes the HP objective:

$$\sum_t (y_{i,t}^{*(b)} - \tau_{i,t}^{*(b)})^2 + \lambda \sum_t (\Delta^2 \tau_{i,t}^{*(b)})^2, \quad \lambda = 100.$$

A period is classified as a boom if $c_{i,t-1}^{*(b)} > 0$, and a slump otherwise.

6. **Re-estimate the IV–LP model.** For each synthetic dataset b, repeat the two-stage IV–LP procedure across all horizons $h = 0, \ldots, H$, recovering a new sequence of impulse response estimates:
$$\hat{\boldsymbol{\beta}}^{*(b)} = (\hat{\beta}_0^{*(b)}, \ldots, \hat{\beta}_H^{*(b)})'.$$

7. **Compute the joint Wald statistic.** For each b, form the Wald statistic under the null:
$$\chi^{2*(b)}_{\text{joint}} = (\hat{\boldsymbol{\beta}}^{*(b)})' \, \mathbf{V}^{-1} \, \hat{\boldsymbol{\beta}}^{*(b)},$$
where $\mathbf{V}$ is the covariance matrix estimated from the original data. This yields a simulated sampling distribution for the test statistic under H_0.

8. **Compute the empirical bootstrap p-value.** Let ${\chi^2_{\text{joint}}}^{\text{obs}}$ denote the test statistic from the observed data. The bootstrap p-value is:
$$\hat{p}_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} \mathbf{1} \left\{ \chi^{2*(b)}_{\text{joint}} \geq {\chi^2_{\text{joint}}}^{\text{obs}} \right\}.$$
This quantity reflects the probability of observing a statistic as extreme as the one in the actual data, under the null.

This algorithm ensures that finite-sample inference reflects the dependence structure, heteroskedasticity, and short time series that characterize the empirical macroeconomic environment. By simulating economies where fiscal shocks exert no causal influence, we obtain a credible benchmark for evaluating whether the observed dynamic responses are statistically distinguishable from noise.

4.3 Cyclical Classification in Simulated Economies

As described in Section 3.3, the original IV–LP analysis by Jordà and Taylor (2024) distinguishes between booms and slumps using the cyclical component of log real GDP. To maintain consistency with this state-dependent structure, we apply the same classification to each bootstrap dataset.

For each simulated output series $y_{i,t}^{*(b)}$, we extract the cyclical component $c_{i,t}^{*(b)}$ via the Hodrick–Prescott (HP) filter, defined as the solution to the minimization problem:

$$\min_{\{\tau_{i,t}^{*(b)}\}} \sum_{t=1}^{T} \left(y_{i,t}^{*(b)} - \tau_{i,t}^{*(b)} \right)^2 + \lambda \sum_{t=2}^{T-1} \left[\left(\tau_{i,t+1}^{*(b)} - \tau_{i,t}^{*(b)} \right) - \left(\tau_{i,t}^{*(b)} - \tau_{i,t-1}^{*(b)} \right) \right]^2, \tag{6}$$

where $\lambda = 100$ is the smoothing parameter appropriate for annual macroeconomic data (Hodrick and Prescott, 1997).

The smoothed trend $\tau_{i,t}^{*(b)}$ yields the cyclical component as:

$$c_{i,t}^{*(b)} = y_{i,t}^{*(b)} - \tau_{i,t}^{*(b)}.$$

A period t is classified as a boom if $c_{i,t-1}^{*(b)} > 0$ and a slump otherwise.

Figure 2 illustrates the distribution of the joint Wald statistics across $B = 10{,}000$ bootstrap replications, stratified by boom/slump classification and projection horizon. The histograms visualize the empirical null distribution under the fitted ARIMA(3,0,2) process, with rejection regions marked relative to conventional significance thresholds.

4.4 Interpretation of Inference

If the observed test statistic ${\chi^2_{\text{joint}}}^{\text{obs}}$ lies in the upper tail of the bootstrap distribution, we reject the null hypothesis of no dynamic policy effect. Conversely, if it falls within the range commonly generated under simulated noise, the evidence for a systematic fiscal impact weakens.

This simulation-based approach strengthens the robustness of our conclusions. It supplements traditional inference with a test that remains valid even under complex residual structures and finite-sample distortions—ultimately providing a more reliable basis for evaluating the statistical significance of estimated policy responses.

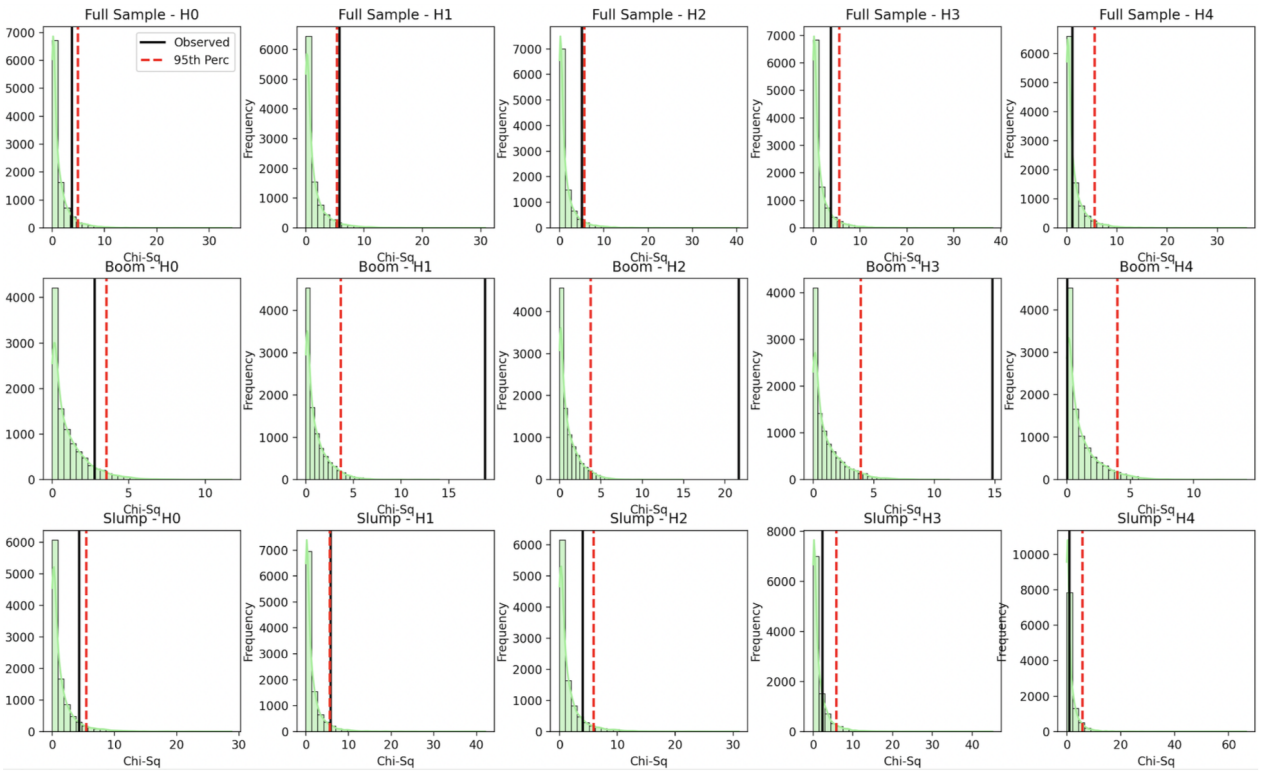

Figure 2: Histogram-style bootstrap distributions of the χ^2 statistic for each horizon $h = 0, \ldots, 4$ (columns) and subsample (rows). Green bars show the empirical distribution from 10,000 bootstrap replications; the red dashed line marks the 95th percentile, and the black solid line indicates the observed value in the original data. Values to the right of the red line correspond to rejections of the null hypothesis at the 5% level under finite-sample inference.

5 Results

Table 2: Joint Significance of Fiscal Policy Responses and 95% Critical Values (df=5 for Parametric)

	Observed χ^2	95% Parametric	95% Bootstrap
Full sample	19.454	11.07	23.080
	(0.002***)		(0.069*)
Slump	17.552	11.07	25.071
	(0.000***)		(0.099*)
Boom	58.085	11.07	15.524
	(0.004***)		(0.000***)

Note: p-values in parentheses * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

Our primary results, presented in Table 2, emphasize the importance of rigorous bootstrap inference methods in validating the dynamic effects of fiscal policy. Table 2 highlights notable differences between parametric critical values and those obtained from our ARIMA-based bootstrap approach, with bootstrap critical values consistently higher, as serial correlation and finite-sample uncertainty are incorporated, setting a much stricter benchmark for statistical significance.

Table 2 shows that, while parametric inference might suggest strong evidence of fiscal policy effectiveness, these findings become less clear when benchmarked against our bootstrap-based distributions. For the Full Sample and Slump subsamples, the joint Wald statistics (19.454 and 17.552, respectively) are highly significant using standard parametric methods but only marginally significant when assessed through bootstrapping. In contrast, the Boom subsample exhibits a robustly significant observed Wald statistic (58.085), maintaining strong statistical significance even with bootstrap inference.

These results highlight that, while fiscal policy effectiveness indeed appears state-dependent, conventional parametric tests may overstate significance. Crucially, our robust ARIMA-based bootstrap procedure reveals that even under stringent testing that accounts explicitly for random variability and finite-sample artifacts, the strong impact of fiscal policy during economic expansions remains convincingly replicable.

6 Discussion

The thesis set out to replicate the findings of Jordà and Taylor (2024) and to reassess their statistical strength under a finite-sample lens. Replication was successful: the observed stacked Wald statistics in Table 2 match the originals almost exactly. Yet once the benchmark for inference shifts from asymptotic to bootstrap probabilities, **only one of the three samples—the boom state—retains unambiguous significance.** Parametric p-values of 0.002 (full sam-

ple) and < 0.001 (slumps) swell to 0.069 and 0.099, respectively, when the ARIMA-based bootstrap is applied, whereas the boom p-value tightens from 0.004 to < 0.001. Our findings clearly demonstrate the critical need for robust econometric validation. Parametric methods alone, commonly used in existing literature, risk identifying spurious or overstated policy effects that may, in reality, arise from random fluctuations or finite-sample peculiarities. By contrast, our ARIMA-based bootstrap analysis rigorously simulates plausible random paths of the economic data, effectively distinguishing genuine policy-induced impacts from statistical noise.

The evidence therefore supports a sharp asymmetry: discretionary fiscal *tightening* is robustly contractionary when economies are above trend, but its impact becomes statistically uncertain—or at least far less certain—in slumps. This complements earlier studies that find larger *expansionary* multipliers in recessions (Auerbach and Gorodnichenko, 2012), suggesting that the direction of the shock matters as much as the state of the cycle. If policymakers tighten budgets during booms, they should anticipate clear output losses; if they tighten in slumps, the payoff is highly uncertain once finite-sample variability is acknowledged. Because the bootstrap admits a non-negligible 8–10 % chance that the slump-state statistic could arise under the null, cost–benefit analyses should weigh downside risks more heavily than point estimates would imply.

7 Conclusion

This thesis set out to revisit the findings of Jordà and Taylor (2024), focusing on whether the dynamic effects of fiscal policy they estimate remain robust when viewed through the lens of finite-sample uncertainty. Using an ARIMA-based bootstrap approach, I reassessed their results with a method that accounts for the short and serially correlated nature of macroeconomic data.

The main takeaway is one of refinement rather than contradiction. While Jordà and Taylor's estimates suggest significant output contractions following fiscal tightening, my analysis shows that this effect is particularly strong and reliable during economic booms. In contrast, the evidence becomes less clear in slumps and in the overall sample once we account for the randomness that naturally arises in small samples. Rather than undermining their results, this perspective helps clarify when and where the estimated effects are most compelling.

The broader message is that careful attention to inference methods—especially in contexts with limited data and persistent dynamics—can enrich our understanding of policy impacts. Simulation-based tools like the bootstrap provide a useful complement to standard approaches, offering additional insight into the robustness of empirical findings.

Looking ahead, there are many opportunities to build on this work. Beyond the specific case of fiscal policy, similar robustness checks could be applied to other areas of macroeconomic research—such as monetary policy, labor market dynamics, or financial shocks—where short panels and persistent dynamics pose similar challenges for inference. Methodologically,

exploring alternative bootstrap designs, like block or wild bootstraps, or integrating Bayesian approaches that formally incorporate prior information, are natural next steps. More broadly, there is scope for developing practical tools that bridge the gap between simulation-based and asymptotic inference, making robust inference more accessible to applied researchers. I intend to pursue these questions further in graduate study, with the goal of contributing to both the methodological and empirical literatures in macroeconomics as part of a PhD research program.

References

Auerbach, A. J. and Gorodnichenko, Y. (2012). Measuring the output responses to fiscal policy. American Economic Journal: Economic Policy, 4(2):1–27.

Brugnolini, L. (2018). About local projection impulse response function reliability. Manuscript, University of Rome "Tor Vergata".

Choi, C.-Y. and Chudik, A. (2019). Estimating impulse response functions when the shock series is observed. Economics Letters, 180:71–75.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74:427–431.

Guajardo, J., Leigh, D., and Pescatori, A. (2014). Expansionary austerity? international evidence. Journal of the European Economic Association, 12(4):949–968.

Herbst, E. P. and Johannsen, B. K. (2024). Bias in local projections. Journal of Econometrics, 240:105655.

Hodrick, R. J. and Prescott, E. C. (1997). Postwar u.s. business cycles: An empirical investigation. Journal of Money, Credit and Banking, 29(1):1–16.

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. American Economic Review, 95(1):161–182.

Jordà, Ò. and Taylor, A. M. (2024). Evaluating the effects of macroeconomic shocks using simulation-based inference. Forthcoming, Journal of Monetary Economics.

Kilian, L. and Kim, Y. (2011). How reliable are local projection estimators of impulse responses? Review of Economics and Statistics, 93(4):1460–1466.

Kugiumtzis, D. (2002). Surrogate data test on time series. In Modelling and Forecasting Financial Data, pages 267–282. Springer US.

Mudelsee, M. (2014). Climate Time Series Analysis: Classical Statistical and Bootstrap Methods, volume 51 of Atmospheric and Oceanographic Sciences Library. Springer, 2nd edition.

Nakamura, E. and Steinsson, J. (2018). High-frequency identification of monetary non-neutrality: The information effect. Quarterly Journal of Economics, 133(3):1283–1330.

Noguchi, K., Gel, Y. R., and Duguay, C. R. (2011). Bootstrap-based tests for trends in hydrological time series, with application to ice phenology data. Journal of Hydrology, 410(3-4):150–161.

Plagborg-Møller, M. and Wolf, C. K. (2021a). Local projections and VARs. Econometrica, 89(2):611–646.

Plagborg-Møller, M. and Wolf, C. K. (2021b). Local projections and vars estimate the same impulse responses. Econometrica, 89(2):955–980.

Ramey, V. A. and Zubairy, S. (2014). Government spending multipliers in good times and in bad: Evidence from u.s. historical data. NBER Working Paper No. 20719.

Romano, J. P. and Tirlea, M. (2021). Permutation testing for dependence in time series. Journal of Time Series Analysis, 42(6):793–811.

Romano, J. P. and Tirlea, M. (2024). Least squares-based permutation tests for time series regressions. arXiv preprint arXiv:2404.06238.

Saravanan, V., Berman, G. J., and Sober, S. J. (2020). Application of the hierarchical bootstrap to multi-level data in neuroscience. Neuron Behavior Data Analysis and Theory, 3(5).

Stock, J. H. and Watson, M. W. (2018). Identification and estimation of dynamic causal effects in macroeconomics using external instruments. Economic Journal, 128(610):917–948.

Wang, L. and Van Keilegom, I. (2007). Nonparametric test for the form of parametric regression with time series errors. Statistica Sinica, 17(1):369–386.

A Appendix

A.1 VAR vs. LP: Model Formulations

Although the main text outlines conceptual differences between Vector Autoregressions (VARs) and Local Projections (LPs), here we provide a more explicit illustration of both approaches using equations.

Vector Autoregression (VAR). Suppose we have a K-dimensional vector of endogenous variables, $\mathbf{X}_t = (X_{1t}, X_{2t}, \ldots, X_{Kt})'$, observed over $t = 1, 2, \ldots, T$. A p-lag VAR is written as:

$$\mathbf{X}_t = \mathbf{c} + \mathbf{A}_1\mathbf{X}_{t-1} + \mathbf{A}_2\mathbf{X}_{t-2} + \cdots + \mathbf{A}_p\mathbf{X}_{t-p} + \mathbf{u}_t, \quad \mathbf{u}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \tag{7}$$

where $\mathbf{c}$ is a K-vector of intercepts, $\mathbf{A}_i$ are $K \times K$ coefficient matrices, and $\mathbf{u}_t$ is a K-vector of error terms with covariance matrix $\mathbf{\Sigma}$. The impulse responses are typically obtained by inverting this system or by applying a structural identification (e.g., using Cholesky decomposition, sign restrictions, or other methods) to isolate the causal effect of one variable's innovation on the others.

Local Projections (LP). By contrast, Local Projections compute impulse responses more directly, one horizon at a time. For a single dependent variable y_t, we define an h-period-ahead outcome y_{t+h} and regress it on the shock of interest at time t, plus other controls and fixed effects. A simplified version of the LP regression takes the form:

$$y_{t+h} = \alpha_h + \beta_h \cdot \text{Shock}_t + \gamma_h'\mathbf{X}_t + \varepsilon_{t+h}, \tag{8}$$

where $\mathbf{X}_t$ might include lags of y_t, lags of the shock, or additional exogenous variables. Repeating this regression across horizons $h = 0, 1, 2, \ldots, H$ produces a sequence of estimates $\hat{\beta}_h$ that trace out the impulse response function. Notably, each horizon h is estimated independently, freeing the analysis from the requirement to specify an entire system of equations as in VAR models.

The principal benefit of LPs lies in their ability to flexibly accommodate non-linearities or state-dependent effects (e.g., booms vs. slumps) without recalibrating a large system of equations. Additionally, identification can be incorporated more simply by instrumenting the shock term (Shock_t) with external or narrative-based instruments. At the same time, VAR-based approaches may provide more coherent, system-wide restrictions and can exploit cross-equation information at the cost of stricter assumptions on system structure and identification.

A.2 Stationarity

A key requirement for reliable time-series estimation is that the data-generating process remain stationary, meaning its mean, variance, and autocovariances do not change over time. To verify

stationarity in our data, we perform two complementary tests: the Augmented Dickey-Fuller (ADF) test and an eigenvalue root test based on the AR polynomial.

ADF Test. We first conduct Augmented Dickey-Fuller (ADF) tests (Dickey and Fuller, 1979) to statistically assess stationarity. The ADF test checks whether a time series possesses a unit root, implying non-stationarity. Formally, the ADF test evaluates the null hypothesis (H_0) that a unit root exists against the alternative hypothesis (H_A) that the series is stationary (or trend-stationary). Specifically, the ADF regression is expressed as:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^{k} \delta_i \Delta y_{t-i} + \varepsilon_t,$$

where $\Delta y_t = y_t - y_{t-1}$, t represents a time trend, and k is the number of terms of lagged difference included to account for the autocorrelation in the residuals.

In this setup, the null hypothesis and the alternative hypothesis are:

$$H_0 : \gamma = 0 \quad \text{(unit root present, implying non-stationarity)}$$

$$H_A : \gamma < 0 \quad \text{(no unit root, indicating stationarity).}$$

Applying the ADF test to the residual series from each subsample (Full, Boom, and Slump), we find strong evidence against the presence of a unit root in all cases. Specifically, the Full and Boom samples yield highly significant ADF statistics of -25.93 and -18.91, respectively, requiring no additional lagged terms to correct for autocorrelation. The Slump sample selects 14 lags and yields an ADF statistic of -3.85, which remains statistically significant at conventional significance levels. These results confirm the suitability of our subsequent ARIMA modeling and bootstrap procedures, which rely fundamentally on stationarity.

Eigenvalue Root Test. As a secondary check, we require that all roots (eigenvalues) of the AR polynomial lie inside the unit circle. For an AR(p) model with characteristic polynomial

$$\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p,$$

is stationary if all polynomial roots (solutions to $\Phi(z) = 0$) lie strictly outside the unit circle ($|z| > 1$). Equivalently, all eigenvalues of the associated companion matrix—defined as the reciprocals of these roots—must lie strictly inside the unit circle ($|1/z| < 1$). This equivalence ensures that shocks diminish over time, a crucial property of stationary series. Our fitted ARIMA(3,0,2) for Full and Slump and ARIMA(1,0,1) for Boom meet this condition.

To visually verify stationarity, Figure 3 plots the eigenvalues (roots) of the AR polynomial from the fitted ARIMA(3,0,2) model. All points (roots) lying strictly inside the unit circle confirm stationarity of the underlying residual series.

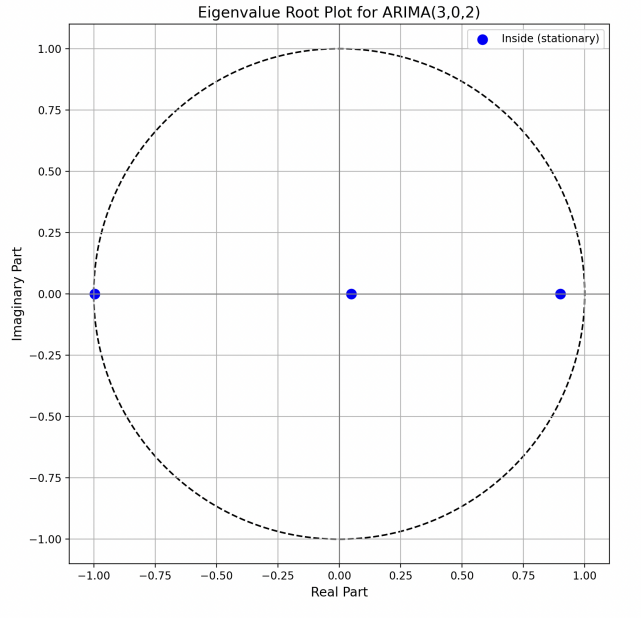

Figure 3: Eigenvalue (Root) Plot for ARIMA(3,0,2). All eigenvalues lie inside the unit circle, confirming stationarity.