SAPIENZA
UNIVERSITÀ DI ROMA

On the Reliability of OLS in High Dimensions

Facoltà di Economia
Master's Degree in Economics

Alberto Leorati
ID number 2051643

Advisor
Prof. Giuseppe Ragusa

Academic Year 2023/2024

Thesis not yet defended

On the Reliability of OLS in High Dimensions
Master Thesis. Sapienza University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: leoratialberto@gmail.com

"Sic transit gloria mundi"

Abstract

This thesis aims to investigate the reliability of the Ordinary Least Squares (OLS) estimator in high-dimensional settings, where the number of regressors K is large relative to the sample size N. The OLS requires the regressors to be a negligible fraction of the sample size to ensure its properties. However, the common econometric models tend to use several control variables, such that the previous assumptions may still not hold. While the i.i.d. framework has been well-studied, the time series needs further enhancements. This thesis studies the OLS behavior in a time series context, focusing on the inference performance and the coverage errors. Using both theoretical derivation and Monte Carlo simulations, the instability of the OLS is illustrated as the dimensionality increases. This work analyses two possible solutions: the lag augmentation regression and the endogenous instrumental variable. Their inference performances are tested and compared in case of correct and incorrect model specifications.

Acknowledgment

I would like to express my profound gratitude to my advisor, Professor Giuseppe Ragusa, for his invaluable guidance and insightful feedback throughout the course of this thesis. His ability to stimulate my intellectual curiosity has allowed me to grow academically and acquire knowledge. Most importantly, I am deeply thankful to him for nurturing my passion for research, particularly in econometrics. His mentorship has been instrumental in shaping the direction of my academic journey.

I am also grateful to my colleagues and friends for their support throughout this process, and I would especially like to mention Simone. Our discussions will remain an indelible memory of my time in Rome.

Finally, I would like to express my deepest thanks to my family for their immense efforts and sacrifices, which have been a cornerstone of my academic success.

Contents

Chapter 1

Introduction

The Ordinary Least Squares (OLS) is one of the most common estimations used in empirical analysis due to its simplicity and versatility. For instance, the OLS estimates policies and interventions' structural, causal, or treatment effects in econometric applications. The main advantage of this technique lies in the linear dependence assumption between the variables, which ensures an intuitive and naive model interpretation.

The OLS estimator has the following form: $\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$ and computing this estimation, in practice, requires only three technical assumptions. The first one is needed to ensure that the design matrix $\mathbf{X}'\mathbf{X}$ is invertible, which relies on the full rank conditions. Second, the fourth moment of the random variables must be finite to avoid outliers issues or distributions that don't approximate the Gaussian one: $E[x_i^4] < \infty$ and $E[u_i^4] < \infty$. Finally, the last assumption states that the variables must be independently and identically distributed.

Under these few conditions, considering a linear model of the form:

$$y_i = \mathbf{X}'\beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2),$$

the OLS estimator $\hat{\beta}_{OLS}$ coincides with the linear projection coefficient. In other words, it captures the correlation between the dependent and the independent variable. Introducing the stronger assumption $E[\epsilon_i|\mathbf{X}] = 0$, which means the error term is mean-independent of the regressors, $\hat{\beta}_{OLS}$ assumes a causal interpretation

since it can be seen as the marginal effect of x_i on y_i, expressed as:

$$\beta_j = \frac{\partial E[y|x]}{\partial x_j}.$$

The reliability of the OLS estimator is based on two fundamental statistical results: the Law of Law of Large Numbers (LLN) and the Central Limit Theorem (CLT). Respectively, the first one ensures that the $\hat{\beta}_{OLS}$ is a consistent estimator for the true β parameter, meaning that:

$$\lim_{n\to\infty} P(|\hat{\beta}_{OLS} - \beta| > \epsilon) = 0 \quad \text{for any } \epsilon > 0 \tag{1.1}$$

Looking at the notation in 1.1, the consistency is guaranteed when n approximates infinity. In real applications, this means the sample size must be large enough.

The CLT, instead, is crucial to ensure the asymptotic normality of the estimator:

$$\sqrt{N}(\hat{\beta}_{OLS} - \beta) \xrightarrow{d} \mathcal{N}(0, \mathbf{V}_{robust}), \tag{1.2}$$

where

$$\mathbf{V}_{robust} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1},$$

and $\xrightarrow{d}$ denotes convergence in distribution.

Let's focus on 1.2. It states that the more the sample size N increases and the more the $\hat{\beta}_{OLS}$ distribution converges to a Gaussian with zero mean and $\mathbf{V}_{robust}$ variance.

If these laws hold, then for large N, the OLS estimator can be easily computed using the sample averages:

$$\hat{\beta}_{\text{OLS}} = \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} x_i y_i\right)$$

In empirical works are usually employed multivariate regressions, where the number of regressors K is greater than one. It follows that the OLS has to estimate K parameters. In such contexts, the crucial factor that ensures the asymptotic properties is the ratio between the number of regressors and the number of observations.

Keeping fixed K, N has to go to infinity, thus:

$$\frac{K}{N} \to 0 \tag{1.3}$$

1.3 to hold, is essential that N is much larger than K. The more variables the researcher wants to use, the larger the required sample size.

These theoretical foundations are robust asymptotically. This means they ensure their properties only if the sample size tends to infinity. However, in practical econometric applications, the sample size N is always finite. Thus, a general researcher's concern regards the necessary sample size to approximate the asymptotic results as closely as possible. Furthermore, since getting data is usually costly, the researcher has to face an evident trade-off. He is willing to enlarge the sample size, only if the cost is overcome by better estimation performance. Typically, with well-behaved, i.i.d. variables and standard assumptions, a sample size of approximately 30 is sufficient. However, as the number of regressors K grows, maintaining the ratio K/N approaching zero becomes unrealistic.

In recent empirical research, due to the availability of large datasets, it is common to include several control variables in the regression model to rule out endogeneity issues and estimate the causal effect. This practice can lead to a situation where K/N does not approach zero. Consequently, the question arises whether OLS remains consistent and reliable.

The goal of this thesis is to explore the reliability of OLS in high-dimensional contexts. Chapter 2 analyzes the application of OLS within the independent and identically distributed (i.i.d.) framework, addressing the impact of a high-dimensional setting on the estimator's properties with a particular focus on heteroskedasticity. Chapter 3 extends this discussion to time series data, examining autocorrelation and weak exogeneity issues. In this context, allowing for high dimensions is more problematic due to the particular structure of time series datasets. If the ratio K over T doesn't approach zero the OLS leads to wrong inference. Chapter 4 presents two methods to improve the reliability of inference in high-dimensional models within time series data. The first is the Endogenous Instrumental Variable by Mikusheva and Sølvsten [2023], while I propose the Lag Augmentation regression. This chapter

shows using both theoretical derivations and Monte Carlo simulations the improved inference performance. Finally, Chapter 5 gives practical guidelines simulating an empirical application where the researcher doesn't know the data-generating process. While the Endogenous IV solves effectively the problem if the model is correctly specified, Lag Augmentation is more flexible and easier to implement.

This thesis first provides an overview of the key statistical inference challenges associated with using OLS in high-dimensional settings. The primary contribution, however, lies in offering theoretical guidelines for addressing the unique issues that an empirical researcher encounters in time series analysis.

Chapter 2

High Dimension OLS within i.i.d. context

2.1 Independent and Identically Distributed definition

A sequence of random variables $\{X_i\}_{i=1}^n$ is said to be independent and identically distributed (i.i.d.) if it satisfies the following properties[1]:

1. **Independence:** Each random variable X_i is independent of every other random variable X_j for $i \neq j$. Mathematically, this means that the joint probability distribution of any subset of the variables equals the product of their marginal distributions:

$$P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) = \prod_{i=1}^{n} P(X_i \leq x_i).$$

2. **Identical Distribution:** All random variables X_i have the same probability distribution. That is, they share the same cumulative distribution function (CDF), $F(x)$:

$$P(X_i \leq x) = F(x) \quad \text{for all } i.$$

As said in the previous Chapter 1 , the i.i.d. assumption is crucial in econometrics

[1] Following Casella and Berger [2002]

for two reasons[2]:

1. **Law of Large Numbers (LLN):** The sample average converges to the expected value as the sample size increases. Formally, if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ, then:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{p} \mu \quad \text{as} \quad n \to \infty.$$

2. **Central Limit Theorem (CLT):** The distribution of the sample mean approaches a normal distribution as the sample size grows, regardless of the original distribution of the variables. Formally, if $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with mean μ and variance σ^2, then:

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2) \quad \text{as} \quad n \to \infty.$$

2.2 Introducing High Dimension

Consider the standard linear model, which can be expressed as follows:

$$y_i = \beta x_{i,n} + \gamma w_{i,n} + \epsilon_i \tag{2.1}$$

In this model, y_i represents the dependent variable for the i-th observation. The term $x_{i,n}$ denotes the regressor of interest for the i-th observation, which is the primary focus of our investigation. The vector $w_{i,n}$ includes the control variables for the i-th observation, which can potentially be large. The coefficient β is associated with the regressor of interest and reflects its linear impact on the dependent variable. The vector γ consists of coefficients for the control variables, representing their respective impacts on the dependent variable. The inclusion of control variables $w_{i,n}$ aims to provide a more accurate estimation of the effect of $x_{i,n}$ on y_i by mitigating potential omitted variable bias.

In econometrics, the importance of high-dimensional data arises from the need to include a large number of control variables to ensure exogeneity. Ensuring that

[2]Based on Wooldridge [2020]

the regressor of interest is exogenous is crucial for getting unbiased and consistent estimates of β. By incorporating a comprehensive set of control variables, researchers can account for confounding factors. Referring to Equation 2.1, where the set of controls is represented by $w_{i,n}$, the number of regressors K can be potentially large enough to violate the $\frac{K}{N} \to 0$ assumption. The relevant question is if the OLS properties still hold within this setup.

Huber [1973] was one of the first to address this issue, demonstrating that when the number of regressors K increases at the same rate as the sample size N, the fitted regression values do not follow an asymptotically normal distribution. This result implies that the standard inferential procedures, which rely on the normality of the estimators, may not be valid in high-dimensional settings.

Another important result was derived by Mammen [1993]. He showed that asymptotic normality holds for arbitrary contrasts of OLS estimators, provided that the dimension of the covariates is at most a vanishing fraction of the sample size. Mammen's results highlight the importance of the relative growth rates of K and N in ensuring the normality of the estimators, which is crucial for making valid statistical inferences.

More recent work by El Karoui [2013] extends the understanding of asymptotic normality in high-dimensional settings. They showed that under the assumption of a Gaussian distribution for the regressors and homoskedasticity, certain estimated coefficients and contrasts in linear models are asymptotically normal even when K grows as fast as N.

Cattaneo et al. [2018] provided an important contribution to the literature. By using the Partialling-out Theorem Frisch and Waugh [1933], it is possible to focus only on a small subset of the β coefficients, thereby eliminating the need to consistently estimate all the coefficients. Thus, the OLS estimator $\hat{\beta}_n$ for the parameter β in a high-dimensional linear model is asymptotically normal.

Formally, following the latter setting, the OLS estimator is given by:

$$\hat{\beta}_n = \left(\sum_{i=1}^{n} \hat{v}_{i,n} \hat{v}'_{i,n} \right)^{-1} \left(\sum_{i=1}^{n} \hat{v}_{i,n} y_{i,n} \right). \tag{2.2}$$

Here, $\hat{v}_{i,n}$ are transformed regressors defined as:

$$\hat{v}_{i,n} = \sum_{j=1}^{n} M_{ij,n} x_{j,n}, \tag{2.3}$$

where the transformation matrix $M_{ij,n}$ is equal to:

$$M_{ij,n} = 1(i = j) - w'_{i,n} \left(\sum_{k=1}^{n} w_{k,n} w'_{k,n} \right)^{-1} w_{j,n}. \tag{2.4}$$

Let's explain these previous mathematical steps. The matrix $M_{ij,n}$, defined in 2.4, projects the regressors of interests $x_{j,n}$ onto the space orthogonal to the set of controls $w_{k,n}$. Indeed, the transformed regressors $\hat{v}_{i,n}$ are free from the influence of the set of controls. At this point, the goal is to establish that the scaled and centered OLS estimator $\hat{\beta}_n$ converges in distribution to a normal distribution as the sample size n grows:

$$\hat{\Omega}_n^{-1/2} \sqrt{n} (\hat{\beta}_n - \beta) \xrightarrow{d} N(0, I).$$

Here, $\hat{\Omega}_n$ is defined as:

$$\hat{\Omega}_n = \hat{\Gamma}_n^{-1} \hat{\Sigma}_n \hat{\Gamma}_n^{-1},$$

which corrects for the scaling and correlation structure of the estimators, ensuring that the distribution of the scaled estimator converges to a standard normal distribution. This requires finding an estimator $\hat{\Sigma}_n$ for the variance of $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{v}_{i,n} u_{i,n}$, where $u_{i,n}$ are the error terms.

The matrix $\hat{\Gamma}_n$ is defined as:

$$\hat{\Gamma}_n = \frac{1}{n} \sum_{i=1}^{n} \hat{v}_{i,n} \hat{v}'_{i,n},$$

which is essentially the sample covariance matrix of the transformed regressors $\hat{v}_{i,n}$.

Defining $\hat{u}_{i,n}$ as:

$$\hat{u}_{i,n} = \sum_{j=1}^{n} M_{ij,n} (y_{j,n} - \hat{\beta}'_n x_{j,n}),$$

they represent the estimated residuals from the fitted model, which are calculated by first projecting the observed responses $y_{j,n}$ onto the space orthogonal to the control variables $w_{i,n}$.

The consistent estimator $\hat{\sigma}_n^2$ for the error variance is given by:

$$\hat{\sigma}_n^2 = \frac{1}{n - d - K_n} \sum_{i=1}^{n} \hat{u}_{i,n}^2.$$

By incorporating these residuals, the variance-covariance estimator $\hat{\Sigma}_n^{HO} = \hat{\sigma}_n^2 \hat{\Gamma}_n$ is shown to be consistent under homoskedasticity, even when the ratio K/N converges to a constant α, where $\alpha > 0$.

The table 2.1 below presents the results of a Monte Carlo simulation conducted in R, focusing on the performance of the OLS estimator. The data-generating process follows a standard linear model with i.i.d. random variables and a homoskedastic error term. The number of observations N is fixed at 200, while the number of regressors K increases in increments of 10. The simulation runs 1000 iterations for each value of K. The primary metrics of interest include the mean absolute bias, standard deviation of the OLS estimates, and the coverage ratio of the 95% confidence intervals for the first coefficient β_1. This analysis highlights the stability and reliability of the OLS estimator under varying model dimensions. All the coverage ratios shown in the table remain close to the nominal level.

In contrast, estimating the variance-covariance matrix $\hat{\Sigma}_n^{HE}$ consistently within heteroskedasticity, in general, requires that $\frac{K}{N} \to 0$.

2.3 Allowing for Heteroskedasticity

White [1980] made a significant contribution to the econometrics field and statistics by introducing a method that enables robust inference in the presence of heteroskedasticity, without requiring knowledge of its specific functional form. Under certain conditions, he demonstrated the following result:

$$\hat{V}_n \equiv \frac{1}{n} \sum_{i=1}^{n} \hat{u}_i^2 X_i' X_i \xrightarrow{\text{a.s.}} \frac{1}{n} \sum_{i=1}^{n} E(u_i^2 X_i' X_i),$$

Table 2.1. Summary Statistics

K Values	Mean Abs Biases	Std Devs	Coverage Ratios
10	0.05624613	0.07066531	0.947
20	0.05577348	0.07022365	0.958
30	0.05567538	0.07009441	0.949
40	0.05910524	0.07306609	0.942
50	0.05635722	0.06987367	0.955
60	0.05478210	0.06943393	0.943
70	0.05854709	0.07221670	0.951
80	0.05663447	0.07171054	0.940
90	0.05436853	0.06829198	0.954
100	0.05617699	0.06994393	0.946
110	0.05532986	0.07021099	0.952
120	0.05562322	0.07132876	0.933
130	0.05682328	0.07182415	0.939
140	0.05643585	0.07148004	0.941
150	0.05532055	0.06924462	0.936
160	0.05696586	0.07089184	0.945
170	0.05628453	0.07127992	0.936
180	0.05649078	0.07075241	0.934

where $\hat{u}_i \equiv y_i - X_i\hat{\beta}$ represents the i-th OLS residual, and a.s. denotes *"almost sure convergence"*. This result implies that the finite-sample covariance matrix estimator is given by:

$$(X'X)^{-1}\left(\sum_{i=1}^{n}\hat{u}_i^2 X_i'X_i\right)(X'X)^{-1}$$

usually known as the Sandwich Covariance Estimator, expressed as:

$$(X'X)^{-1}X'\Omega X(X'X)^{-1},$$

where Ω is a diagonal matrix with $\hat{u}_i^2$ on the diagonal:

$$\Omega = \text{diag}(\hat{u}_1^2, \hat{u}_2^2, \ldots, \hat{u}_n^2).$$

It is crucial to acknowledge the trade-off associated with the White covariance estimator. As a non-parametric estimator, it does not necessitate specifying the form of heteroskedasticity. However, this advantage is offset by a slower rate of asymptotic

convergence, which implies that the White estimator may exhibit finite-sample bias. The main problems of the original White covariance estimator are two. First, it requires no fat tail condition, i.e. the heteroskedasticity must be bounded and not extreme. Second, it is needed low partial leverage, that excludes possible outliers. However, even in a balanced design, the essential condition to do inference is $\frac{K}{N} \to 0$. Therefore, in high-dimensional designs where K constitutes a significant portion of the sample size, the Eicker-Huber-White (EHW) estimator is prone to underperformance. This is still true even if the $\hat{\beta}$ estimator is asymptotic normal.

Within a standard framework, Long and Ervin [2000] shown EHW estimator inconsistency if $N < 250$. MacKinnon and White [1985] raised concerns about the small sample properties of the EHW estimator and proposed three alternatives known as HC1, HC2, and HC3. Asymptotically they are equal, however, they present different small sample properties. MacKinnon [2013] presented an exhaustive work on this topic.

Cattaneo et al. [2018] demonstrated that when the number of regressors K increases proportionally to the number of observations N, traditional HKc methods become inconsistent. To address this issue, Cattaneo proposed an innovative method based on the minimum norm quadratic unbiasedness, which remains consistent as long as $K/N \leq 1/2$.

Further advancing the field, Jochmans [2022] introduced a consistent estimator that performs reliably even when $K/N = 1$. This development ensures that consistent estimation is achievable in high-dimensional settings, even with heteroskdastic errors, where the number of regressors is equal to the number of observations.

Two distinct simulation procedures were undertaken to evaluate the performance of various heteroskedasticity-consistent HC covariance estimators within high-dimensional contexts. The experimental framework mirrors the previous setup with i.i.d. variables. However, instead of assuming homoskedasticity, the first simulation introduces a conservative form of heteroskedasticity:

$$\epsilon_t = Z_t \cdot \sqrt{|x_{1t}|}$$

where:

- ϵ_t represents the heteroskedastic error term at time t.
- Z_t is a random variable following a standard normal distribution, i.e., $Z_t \sim \mathcal{N}(0,1)$.
- $|x_{1t}|$ denotes the absolute value of the variable x_1 at time t.

This configuration models a scenario where the variance of the error term is proportional to the square root of the absolute value of the first regressor. Such a setup induces heteroskedasticity that is bounded and not extreme, aligning with the assumption of non-fat-tailed variance. The results are presented in Table 2.2. While the normal standard error is inconsistent across all values of K, the HC0 estimator performs adequately only when K is a negligible fraction of the sample size, fixed at 200. However, its performance deteriorates as K increases. Among the estimators evaluated, HC3 demonstrates the best performance. The coverage ratio for HC3 is close to the nominal level of 0.95, though it tends to be conservative as the coverage ratio consistently exceeds 0.95.

K	Normal	HC0	HC1	HC2	HC3
10	0.8376	0.948	0.926	0.939	0.952
20	0.8359	0.929	0.941	0.926	0.954
30	0.8447	0.911	0.931	0.945	0.965
40	0.8496	0.895	0.927	0.929	0.963
50	0.8592	0.891	0.913	0.934	0.962
60	0.8695	0.867	0.934	0.918	0.972
70	0.8725	0.847	0.923	0.929	0.978
80	0.8776	0.855	0.904	0.934	0.984
90	0.8810	0.805	0.914	0.927	0.981
100	0.8824	0.771	0.935	0.930	0.987
110	0.8925	0.754	0.913	0.921	0.996
120	0.8937	0.749	0.912	0.920	0.994
130	0.9003	0.690	0.894	0.911	0.994
140	0.9096	0.655	0.925	0.922	0.998
150	0.9127	0.632	0.918	0.931	0.999
160	0.9185	0.588	0.904	0.912	0.999
170	0.9221	0.511	0.919	0.915	1.000
180	0.9189	0.441	0.927	0.918	1.000

Table 2.2. Coverage Ratios with bounded Heteroskedasticity

In the second simulation, an extreme form of heteroskedasticity was introduced:

$$\epsilon_t = Z_t \cdot \exp\left(|x_{1t}|\right)$$

where:

- ϵ_t represents the heteroskedastic error term at time t.
- Z_t is a random variable following a standard normal distribution, i.e., $Z_t \sim \mathcal{N}(0, 1)$.
- $|x_{1t}|$ denotes the absolute value of the variable x_1 at time t.
- $\exp\left(|x_{1t}|\right)$ represents the exponential function applied to the absolute value of x_{1t}.

This configuration models a scenario where the variance of the error term grows exponentially with the absolute value of the first regressor. This setup induces a more extreme form of heteroskedasticity, thereby relaxing the low partial leverage condition, which is a crucial assumption for the reliability of HCk estimators. The results are presented in Table 2.3.

K	Normal	HC0	HC1	HC2	HC3
10	0.645	0.918	0.943	0.922	0.952
20	0.659	0.908	0.917	0.943	0.954
30	0.650	0.895	0.933	0.940	0.951
40	0.685	0.880	0.922	0.926	0.958
50	0.690	0.857	0.920	0.926	0.960
60	0.687	0.834	0.910	0.915	0.955
70	0.723	0.812	0.901	0.909	0.965
80	0.718	0.775	0.907	0.906	0.972
90	0.778	0.768	0.901	0.889	0.967
100	0.770	0.739	0.901	0.895	0.976
110	0.767	0.690	0.882	0.888	0.981
120	0.790	0.660	0.879	0.881	0.992
130	0.820	0.622	0.875	0.874	0.991
140	0.830	0.583	0.884	0.887	0.996
150	0.837	0.558	0.886	0.868	1.000
160	0.865	0.496	0.873	0.878	1.000
170	0.878	0.467	0.900	0.885	0.999
180	0.877	0.397	0.892	0.906	1.000

Table 2.3. Coverage Ratios with extreme Heteroskedasticity

Chapter 3

High Dimension OLS within Time Series

3.1 Time Series and the serial correlation issue

This Chapter starts with a brief revision of the time series framework and its particular challenges. The main sources are Hamilton [1994] and Greene [2020].

A time series is a sequence of data points measured at equally spaced points in time. It is an ordered set of observations $\{y_t\}$, where t represents the time index and y_t is the value of the observed variable at time t.

Employing a frequentist approach with time series data may present certain challenges. Unlike the iid context, a time series observation represents a single realization from a population. It is a single occurrence of a random event. In other words, the event can not be repeated and the counterpart can not be observed. Consequently, two consecutive observations, for instance, at $t = 1$ and $t = 2$, may be generated from different populations. To address this issue, it is essential for the time series to exhibit covariance stationarity. Covariance stationarity, or weak stationarity, requires that the mean, variance, and autocovariance remain constant over time. This implies that the statistical properties of the series are stable and do not change with time, making it possible to model and predict the series more accurately.

Covariance stationarity works by ensuring the following conditions:

1. **Constant Mean**: The expected value $E(y_t)$ is the same for all t.
2. **Constant Variance**: The variance $\text{Var}(y_t)$ is constant and does not depend on t.
3. **Constant Autocovariance**: The covariance between y_t and y_{t+k} depends only on the lag k and not on the specific time t.

However, the weak stationarity assumption alone is insufficient to guarantee the asymptotic properties of the OLS estimator. Inference within time series datasets requires additional assumptions compared with the i.i.d. context. The extra necessary conditions to ensure consistency of the OLS in time series are weak stationarity and mixing conditions, also known as weak dependence. The latter is crucial to ensure that serial dependence is weak enough such that observations distant in time are approximately independent.

In fact, the second problem that arises in time series data is the serial correlation, i.e. observations that are temporally close are generally serially correlated. Thus, the data exhibit a lack of independence, a fundamental requirement for applying the LLN and the CLT.

Mathematically, this can be expressed as:

$$\text{Cov}(y_t, y_{t+k}) \neq 0 \quad \text{for} \quad k \neq 0$$

where $\text{Cov}(y_t, y_{t+k})$ represents the covariance between observations at time t and $t + k$. Running the OLS may raise problems.

Consider the linear model:

$$y_t = \beta_0 + \beta_1 x_{1t} + \epsilon_t$$

where ϵ_t represents the error term at time t. In the context of time series data, the error term ϵ_t is often autocorrelated. This autocorrelation can be expressed as:

$$\epsilon_t = \rho \epsilon_{t-1} + u_t$$

where ρ is the autocorrelation coefficient and u_t is a white noise error term with

mean zero and constant variance. The presence of autocorrelation implies that the error term at time t is correlated with the error term at time $t-1$.

Combining these, the standard time series linear model with autocorrelated errors is:

$$y_t = \beta_0 + \beta_1 x_{1t} + \rho\epsilon_{t-1} + u_t$$

It follows:

$$\mathbb{E}[\epsilon\epsilon' \mid X] = \sigma^2\Omega$$

where σ^2 represents the variance of ϵ_t conditioning on X, and Ω is a function of the lag that represents the autocorrelation structure. In general, to establish the consistency of the OLS estimator, it is required that the matrix $Q_t = \frac{1}{T}X'X$ converges to a positive definite matrix Q due to the Law of Large Numbers[1]. However, in this case, we have

$$Q_t = \frac{1}{T}X'\Omega X,$$

which is equal to

$$\frac{1}{T}\sum_t \sum_s \rho_{ts} x_t' x_s.$$

To ensure that Q_t converges to Q, weak dependence is required. Specifically, this condition can be stated as follows:

$$\sup_{t\geq 1} \sum_{s=1}^{\infty} |\rho_{ts}| < \infty,$$

where ρ_{ts} denotes the autocorrelation coefficient between observations at times t and s. This condition implies that the autocorrelation coefficients ρ_{ts} must decay sufficiently fast as $|t-s|$ increases, ensuring that the dependence between observations decreases over time. To ensure weak dependence, a common requirement is that the mixing coefficients $\alpha(n)$, defined as

[1]See Greene [2020] for further details.

$$\alpha(n) = \sup_{A \in \mathcal{F}_{-\infty}^{t}, B \in \mathcal{F}_{t+n}^{\infty}} |P(A \cap B) - P(A)P(B)|,$$

must satisfy the condition:

$$\sum_{n=1}^{\infty} \alpha(n) < \infty.$$

If the previous conditions are satisfied, then the LLN holds and the OLS is consistent with time series data. However, the inference using the $\hat{\beta}_{OLS}$ is correct only if the CLT holds. The latter requires an additional requirement compared to LLN: the error term must be either a linear process or a Martingale Difference Sequence.

The linear process means that the error term u_t can be modeled as a sum of i.i.d. random variables called innovations. Mathematically, u_t can be expressed as:

$$u_t = \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j},$$

where ϵ_t are i.i.d. errors, and ψ_j are constants. This allows for serial dependence in the errors but ensures that the dependence is controlled and decays over time, depending on the values of ψ_j.

Instead, in a Martingale Difference Sequence, the error term u_t has the property that its conditional expectation given past information is zero. Formally:

$$E(u_t \mid \mathcal{F}_{t-1}) = 0,$$

where $\mathcal{F}_{t-1}$ represents the information set up to time $t-1$. This means that given all past information, the expected value of u_t is zero, implying that u_t contains no predictable pattern based on past observations. In a martingale difference sequence, there can still be some dependence between the errors at different periods, but this dependence is unpredictable.

Even if the OLS estimators for β are unbiased, making inferences using the standard errors calculated under the assumption of i.i.d. errors can lead to incorrect conclusions. This is because these standard errors tend to underestimate the true

standard errors when the errors are autocorrelated. As a result, the confidence intervals may be too narrow. Newey and West [1987] developed a method to account for autocorrelation in the error terms, adjusting the standard deviation of the estimator.

Table 3.1 shows the coverage ratios of three types of standard errors: Normal, White (heteroscedasticity-consistent), and Newey-West (heteroscedasticity and autocorrelation consistent). A simple linear regression model generates the data, where the dependent variable y_t is modeled as a function of the independent variable x_t plus an error term ϵ_t. Both x_t and ϵ_t are generated following an AR(1) process with a coefficient of 0,7. The number of simulations is 10,000 for every sample sample size $T = 100, 250, 500$, and 1000. The coverage ratio is calculated as the proportion of times the true parameter lies within the 0.95 confidence interval.

	Coverage Ratio			
Type of SE	T = 100	T = 250	T = 500	T = 1000
Normal	0.7414	0.7445	0.7597	0.7482
White	0.7265	0.7371	0.7554	0.7462
Newey-West	0.8352	0.8902	0.9179	0.9297

Table 3.1. Coverage Ratios for Different Sample Sizes

The results indicate that the coverage ratios for both the Normal and White standard errors are significantly below the nominal 0.95 level. This suggests that these estimators tend to underestimate the true standard errors, leading to inconsistency even as the sample size increases. Conversely, the Newey-West standard errors exhibit better performance in terms of coverage ratio. However, the convergence rate to the asymptotic value of 0.95 is slow. Even with a sample size of 1000 observations and only one regressor ($K = 1$), the coverage ratio remains below the theoretical level. This finding suggests that in high dimensions potential issues may arise.

Table 3.2 illustrates the coverage ratio performance across different degrees of autocorrelation within the model. The same setup as before is used. However, two distinct models are employed. Model 1 exhibits autocorrelation only in the error term with a coefficient of 0.7, instead, Model 2 incorporates autocorrelation in both the independent variable and the error term with coefficients set to 0.9.

Sample Size	Model 1			Model 2		
	Normal	White	Newey-West	Normal	White	Newey-West
T = 100	0.95	0.943	0.901	0.498	0.471	0.738
T = 250	0.95	0.945	0.914	0.475	0.461	0.826
T = 500	0.95	0.949	0.935	0.479	0.470	0.872
T = 1000	0.95	0.948	0.935	0.476	0.472	0.910

Table 3.2. Coverage Ratios across different degrees of autocorrelation

A trade-off becomes evident from the results. In Model 1, characterized by a lower degree of autocorrelation, both the Normal and White standard errors exhibit better performance compared to the Newey-West standard error across all sample sizes. This is due to the slower convergence of the Newey-West standard error to the asymptotic coverage ratio of 0.95. Conversely, in Model 2, which presents a more complex level of autocorrelation, the outcomes are different. The Normal and White standard errors perform poorly, while the Newey-West standard error demonstrates significantly better performance. However, it requires a larger sample size to achieve the nominal coverage level of 0.95.

This may serve as further evidence that, in high-dimensional settings, the reliability of OLS in time series analysis could be compromised.

3.2 Weak Exogeneity

The exogeneity assumption is fundamental in OLS, ensuring that the $\hat{\beta}_{OLS}$ is consistent. Furthermore, it allows to interpret the estimated parameter as a causal effect between the variables. Mathematically, the OLS estimator is given by:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y.$$

The classical linear model:

$$Y = X\beta + \epsilon,$$

substituting Y into the OLS estimator:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'(X\beta + \epsilon),$$

expanding the multiplication:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon.$$

The equation reduces to:

$$\hat{\beta}_{OLS} = \beta + (X'X)^{-1}X'\epsilon$$

The strong exogeneity assumption states that the error terms ϵ are uncorrelated with the independent variables X:

$$E[\epsilon|X] = 0.$$

Taking the expectation of $\hat{\beta}_{OLS}$:

$$E[\hat{\beta}_{OLS}] = E\left[\beta + (X'X)^{-1}X'\epsilon\right]$$

Since β is constant, its expectation is itself, and by the strong exogeneity assumption, the expected value of $(X'X)^{-1}X'\epsilon$ is zero:

$$E[\hat{\beta}_{OLS}] = \beta + (X'X)^{-1}X'E[\epsilon|X] = \beta$$

Thus, the expected value of the OLS estimator $\hat{\beta}_{OLS}$ is equal to the true parameter vector β.

In the time series framework, the strong exogeneity assumption may be unrealistic. Mathematically, the strong exogeneity condition in a time series context can be stated as:

$$E[\epsilon_t|\ldots, x_{t-1}, x_t, x_{t+1}, \ldots] = 0$$

This assumption implies that the error term at time t is uncorrelated with the past, present, and future values of the independent variables. Stock and Watson

[2019] provided a clear and detailed explanation of the topic. The main point is that simultaneous causality can exist in time series and economic contexts, where x and y may affect each other. Specifically, given the available set of information today, economic agents will consider it in their decision-making. For instance, if the x variable is possible to forecast, today's value of y_t will also incorporate future information. To better understand it, let's consider an example.

Suppose the researcher is interested in the stock price of an oil company and he wants to study the influence of geopolitical shocks on the price. It turns out that the regression could be:

$$y_t = \alpha_0 + \beta_1 x_t + \gamma w_t + \varepsilon_t \tag{3.1}$$

where y_t is the stock price of the oil firm, x_t is a variable that captures the geopolitical shock and w_t is a set of controls. It is reasonable to assume that the geopolitical shocks are independent of the oil prices at the current time t, such that geopolitical shocks can be seen as random. However, the inverse relation may not. Suppose tensions that can affect the availability of oil in the future arise. The market, in general, will respond by increasing the oil demand. Thus, the price will rise even if the geopolitical shock has not occurred. In economics, the expectations about the future are relevant because the agents tend to anticipate the events. As a consequence, strong exogeneity is violated. Thus, the weak exogeneity assumption may be more realistic. One of the first to formalize this problem was Engle et al. [1983]. Mathematically, weak exogeneity can be stated as:

$$E[\varepsilon_t | x_t, x_{t-1}, x_{t-2}, \ldots] = 0$$

It turns out that, given all the other assumptions of OLS, relaxing strong exogeneity and allowing for weak exogeneity is not problematic. This is because the Q_t matrix still asymptotically converges to a positive definite matrix Q [2]. Indeed, in order to ensure consistency, it is required contemporaneous exogeneity, which is less restrictive than strong and weak exogeneity. Formally:

[2] See Hamilton [1994] for further details.

$$E[\varepsilon_t|x_t] = 0$$

To investigate the impact of relaxing the strong exogeneity assumption to weak exogeneity, a Monte Carlo simulation was employed. Data are generated according to a classical linear model $y_t = \beta x_t + u_t$, where u_t is a Gaussian error and x_t^* follows an AR(1) process with a coefficient of 0.9. To introduce weak exogeneity, a feedback effect à la Granger [1969] is implemented: $x_t = x_t^* + \alpha u_{t-1}$, where α is fixed at 1.5. For each sample size, 10000 simulations are run and the coverage ratios are computed.

	T = 30	**T = 50**	**T = 100**
Coverage Ratio	0.946	0.9483	0.948

Table 3.3. Coverage Ratio of the Normal Standard Error for Different Values of T

The simulation results in Table 3.3 demonstrate that relaxing the strong exogeneity assumption to weak exogeneity is not problematic in this setup. The coverage ratios for the normal standard errors converge quickly to the nominal 0.95 confidence level.

3.3 High Dimension

As seen in the previous sections, in the context of time series analysis, managing data becomes more complex compared to the independent and identically distributed setting, and several challenges emerge. These challenges include autocorrelation, non-stationarity, and weak exogeneity, which complicate the estimation and inference processes.

However, ensuring that $\frac{K}{T} \to 0$ produces reliable inference. In practice, as the sample size T increases, maintaining a relatively smaller number of parameters K becomes crucial. This balance allows the underlying model to capture the essential features of the time series without being overwhelmed by noise or spurious patterns.

A common contemporary approach to estimating causal effects involves incorporating multiple controls in regression models. This can lead to a situation where

$\frac{K}{T} \to \alpha$, with $\alpha > 0$. The pertinent question is whether OLS inference remains reliable under these conditions. According to Mikusheva and Sølvsten [2023], when accounting for weak exogeneity, autocorrelation, and high-dimensional settings, OLS estimators lose their consistency, thus rendering traditional OLS inference unreliable in such contexts.

3.3.1 *OLS estimator Bias*

Considering the following linear model:

$$y_t = \beta x_{1t} + \gamma x_{jt} + \epsilon_t \quad \text{with} \quad j \geq 2$$

where each x_{it} is generated as an AR(1) process:

$$x_{it} = \rho x_{i,t-1} + u_{it}.$$

The error terms are independent, and the following holds:

$$\epsilon_t \sim N(0, \sigma_\epsilon^2), \quad u_{it} \sim N(0, \sigma_u^2)$$

$$\text{Cov}(\epsilon_t, u_{it}) = 0 \quad \forall (i,t), \quad \text{Cov}(u_{it}, u_{jt}) = 0 \quad \forall i \neq j,\, t$$

The expected value of ϵ_t conditional on x_{jt} is zero for every $j \geq 2$, meaning strong exogeneoity:

$$\mathbb{E}[\epsilon_t \mid x_{jt}] = 0 \quad \text{for all} \quad j \geq 2.$$

The regressor of interest x_{1t}, however, is weakly exogenous because it presents a feedback effect:

$$x_{1t} = x_{1t}^* + \alpha \epsilon_{t-1}$$

where x_{1t}^* is generated like all other regressors as an AR(1) process. Thus, the expected value of ϵ_t conditional on x_{1t} is different from zero:

$$\mathbb{E}[\epsilon_t \mid x_{1t}] \neq 0.$$

Since x_{1t} is the regressor of interest, and all the other regressors are exogenous, we can utilize the Frisch-Waugh-Lovell Theorem [Frisch and Waugh, 1933] to project onto the space orthogonal to the matrix of exogenous regressors, denoted as $\tilde{X}_{-1}$, using the Partitioning-out operator:

$$M_{-1} = I - \tilde{X}_{-1} \left(\tilde{X}'_{-1} \tilde{X}_{-1} \right)^{-1} \tilde{X}'_{-1}.$$

Let X_1 be the $(T \times 1)$ vector that contains all the x_{1t} observations. It follows that the OLS estimator is:

$$\hat{\beta}_{\text{OLS}} = \left(X'_1 M_{-1} X_1 \right)^{-1} X'_1 M_{-1} y.$$

To establish the consistency of OLS, we can rewrite the estimator in terms of the error term ϵ:

$$\hat{\beta}_{\text{OLS}} - \beta = \left(X'_1 M_{-1} X_1 \right)^{-1} X'_1 M_{-1} \epsilon.$$

Since we condition on $\tilde{X}_{-1}$, the first term does not pose problems. However, endogeneity arises if there is a correlation between the error term and the first regressor x_{1t}. As shown by Mikusheva and Sølvsten [2023], the expectation is:

$$\mathbb{E}[X'_1 M_{-1} \epsilon \mid \tilde{X}_{-1}] = \alpha \sigma^2 \sum_t M^*_{tt-1}. \tag{3.2}$$

The origin of the bias can be understood by examining the decomposition of the regressor X_1. We decompose x_{1t} into two components: $x_{1t} = x^*_{1t} + \alpha \epsilon_{t-1}$, where x^*_{1t} is exogenous, so $\mathbb{E}[\tilde{X}'_1 M_{-1} \epsilon \mid \tilde{X}_{-1}] = 0$. However, the second component, $\alpha \epsilon_{t-1}$, correlates with the error term through the projection onto the space orthogonal to the exogenous regressors. This leads to the summation term $\alpha \mathbb{E}\left[\sum_{s,t} M^*_{st} \epsilon_{s-1} \epsilon_t \mid \tilde{X}_{-1} \right]$, where $\sum_t M^*_{tt-1}$ corresponds to the lower diagonal of the projection matrix. The visualization below shows how this lower diagonal appears in the transformation matrix:

$$\begin{pmatrix} \epsilon_1\epsilon_1 & \epsilon_1\epsilon_2 & \epsilon_1\epsilon_3 & \cdots & \epsilon_1\epsilon_T \\ \epsilon_2\epsilon_1 & \epsilon_2\epsilon_2 & \epsilon_2\epsilon_3 & \cdots & \epsilon_2\epsilon_T \\ \epsilon_3\epsilon_1 & \epsilon_3\epsilon_2 & \epsilon_3\epsilon_3 & \cdots & \epsilon_3\epsilon_T \\ \vdots & \epsilon_4\epsilon_2 & \epsilon_4\epsilon_3 & \ddots & \vdots \\ \epsilon_T\epsilon_1 & \vdots & \epsilon_T\epsilon_3 & \epsilon_T\epsilon_{T-1} & \epsilon_T\epsilon_T \end{pmatrix}$$

The main diagonal of this matrix represents $\mathbb{E}[\epsilon_t\epsilon_t] = \sigma^2$, while the off-diagonal elements, $\mathbb{E}[\epsilon_{t-1}\epsilon_t] = \rho_i$, correspond to the autocorrelation between consecutive error terms. The matrix structure is:

$$\begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ \rho_2 & \sigma^2 & 0 & \cdots & 0 \\ 0 & \rho_3 & \sigma^2 & \cdots & 0 \\ \vdots & 0 & \rho_4 & \ddots & \vdots \\ 0 & \vdots & 0 & \rho_K & \sigma^2 \end{pmatrix}$$

The magnitude of the bias depends on the average autocorrelation coefficient $\bar{\rho}$, which represents the average of the individual autocorrelation coefficients ρ_i for the regressors. Since there are K regressors, the bias is approximately:

$$\mathbb{E}[X_1' M_{-1}\epsilon \mid \tilde{X}_{-1}] = \alpha\sigma^2 \frac{K}{T}\bar{\rho}. \tag{3.3}$$

Thus, the three main factors affecting the bias are the magnitude of α, the ratio $\frac{K}{T}$, and the average autocorrelation $\bar{\rho}$ across regressors. In settings with high-dimensional data where $\frac{K}{T}$ does not approach zero, OLS becomes inconsistent.

3.3.2 *OLS estimator Variance Bias*

This issue extends to the standard deviation of the OLS estimator. Under weak exogeneity, the OLS estimate $\hat{\beta}_{\text{OLS}}$ is biased as a result of the correlation between $\mathbf{X}$ and ϵ_t. The inconsistency in the variance estimate $\hat{\sigma}^2$ emerges because the conventional OLS variance estimator is:

$$\hat{\sigma}^2 = \frac{\mathbf{y}'\mathbf{M}\mathbf{y}}{T-K}$$

where $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the projection matrix.

The correct variance should account for the bias induced by the weak exogeneity, leading to the following expression:

$$\frac{\hat{\sigma}^2}{\sigma^2} = 1 - \frac{\sigma^2 \alpha' \bar{S}^{-1} \alpha}{T - K} \left(\sum_{t=2}^{T} \tilde{M}_{tt-1} \right)^2 + o_p(1) \tag{3.4}$$

where:

- σ^2 represents the true variance of the error term ϵ_t.
- α is the vector that captures the feedback mechanism in the regressors.
- $\bar{S}^{-1}$ is the inverse of the scaled design matrix. The $\bar{S}^{-1}$ term scales the bias according to the variability and correlation structure of the regressors.
- $\sum_{t=2}^{T} \tilde{M}_{tt-1}$ captures the lower diagonal trace of the projection matrix $\mathbf{M}$, which reflects the correlation between successive periods' regressors. This term highlights how the autocorrelation in the regressors contributes to the bias. See the previous section for a comprehensive explanation.
- $\frac{1}{T-K}$ represents the degrees of freedom adjustment in the variance estimation. T is the total number of observations, and K is the number of regressors.
- $o_p(1)$ is a term that goes to zero as the sample size T increases, under the probabilistic order $o_p(1)$. It indicates that additional small terms in the bias expression become negligible in large samples.

As derived before from the 3.2 to the 3.3, the $\sum_{t=2}^{T} \tilde{M}_{tt-1}$ term is equal to $\frac{K}{T}\bar{\rho}$, then:

$$\frac{\hat{\sigma}^2}{\sigma^2} = 1 - \frac{\sigma^2 \alpha' \bar{S}^{-1} \alpha}{T - K} \left(\frac{K}{T} \bar{\rho} \right)^2 + o_p(1). \tag{3.5}$$

Since the numerator in equation 3.5 does not significantly impact the analysis, we define it as $\Phi = \sigma^2 \alpha' \bar{S}^{-1} \alpha$. Consequently, we have:

$$\frac{\hat{\sigma}^2}{\sigma^2} = 1 - \frac{\Phi}{T - K} \left(\frac{K}{T} \bar{\rho} \right)^2 + o_p(1). \tag{3.6}$$

Equation 3.6 illustrates that the OLS variance estimate $\hat{\sigma}^2$ is biased downward, potentially leading to an underestimation of the standard deviation and, consequently, to optimistic confidence intervals.

To see this, consider that when the number of observations T is held constant, the derivative of expression 3.3, which captures the bias in the OLS estimator, with respect to K is:

$$\frac{\partial}{\partial K}\left(\alpha\sigma^2\frac{K}{T}\bar{\rho}\right) > 0.$$

Similarly, the derivative of expression 3.6, which reflects the bias in the variance of the OLS estimator, with respect to K is:

$$\frac{\partial}{\partial K}\left(-\frac{\Phi}{T-K}\left(\frac{K}{T}\bar{\rho}\right)^2\right) < 0.$$

In conclusion, as K increases, the bias in the OLS estimator increases, while the variance of the estimator decreases. This derivation also makes it clear that when the autocorrelation is zero, these issues do not arise.

3.3.3 *Confidence Interval*

The confidence interval for an OLS estimator $\hat{\beta}$ is typically given by:

$$\hat{\beta}_{OLS} \pm z_{\alpha/2} \cdot \mathrm{SE}(\hat{\beta}_{OLS})$$

where $\mathrm{SE}(\hat{\beta})$ is the standard error of the estimator, and $z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to the desired confidence level.

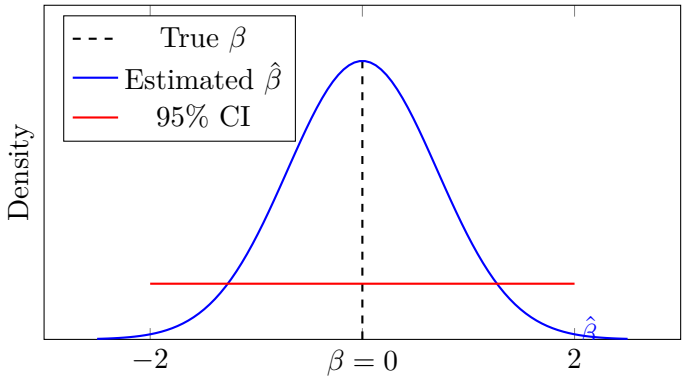
As the number of regressors K increases, the bias in the OLS estimator $\hat{\beta}_{\mathrm{OLS}}$ increases due to weak exogeneity. This bias causes $\hat{\beta}_{\mathrm{OLS}}$ to systematically deviate from the true parameter value β.

Simultaneously, the variance of the OLS estimator decreases because the estimated standard deviation $\hat{\sigma}$ is lower than the true standard deviation σ. This reduction in the estimated standard deviation results from the bias in the variance estimator under weak exogeneity, leading to a smaller standard error.

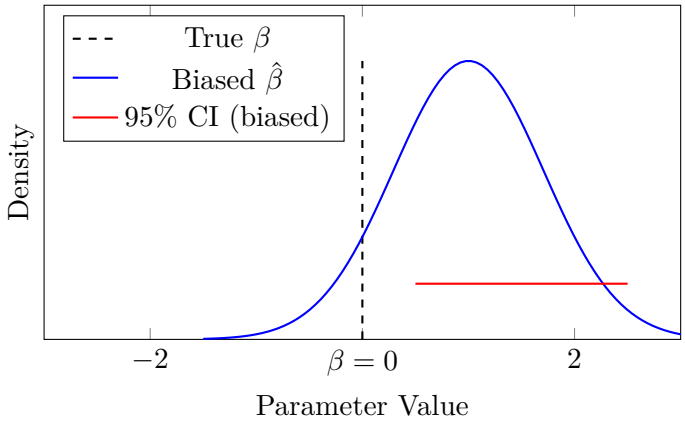
Due to these factors, the resulting confidence interval is narrower than the

theoretical confidence interval would be if the bias were absent and the standard deviation were correctly estimated. This narrowing effect causes the actual coverage probability of the confidence interval to fall below the nominal 95% level. In other words, the confidence interval fails to contain the true parameter value β as frequently as expected, leading to a coverage ratio that is less than 95%. The coverage ratio deteriorates as the number of regressors K increases. Specifically, the larger the value of K, the more pronounced the decline in the coverage ratio performance.

Small Number of Regressors K



Large Number of Regressors K



The two graphs 3.3.3 illustrate the effect of increasing the number of regressors K on the coverage probability of the confidence interval for the OLS estimator $\hat{\beta}$. In the first graph, we consider a scenario with a small number of regressors K. Here, the true parameter value β is represented by the black dashed line at $\beta = 0$. The

blue curve represents the distribution of the OLS estimator $\hat{\beta}$. Because the number of regressors K is small, the OLS estimator is unbiased, meaning that $\hat{\beta}$ is centered around the true β. The distribution is symmetric around the true parameter value. The red line indicates the 95% confidence interval. In this scenario, the confidence interval is sufficiently wide to include the true parameter value β in 95% of cases, as expected from a properly specified model. This means that the coverage probability of the confidence interval is close to the nominal level of 95%.

The second graph represents the scenario where the number of regressors K is large. Again, the true parameter value β is shown by the black dashed line at $\beta = 0$. However, with a large K, the OLS estimator $\hat{\beta}$ becomes biased. This is illustrated by the blue curve, which is now shifted to the right of the true β, indicating that $\hat{\beta}$ systematically deviates from the true parameter value. The bias causes the estimator to no longer be centered around β. Furthermore, the red line shows that the confidence interval is now narrower, due to the underestimated standard error. This narrower confidence interval fails to cover the true β as frequently as it should. As a result, the actual coverage probability of the confidence interval falls below the nominal 95% level. This degradation in coverage illustrates the adverse impact of having a large number of regressors, leading to a decline in the reliability of the confidence interval.

Table 3.4 below presents the outcomes of a Monte Carlo simulation designed to examine the properties of OLS estimators under conditions where all previously identified issues are present. For each value of K, a Monte Carlo simulation with $T = 200$ and $\rho = 0.9$ was conducted. Normally distributed error terms (ϵ_t) were first generated, and $T \times K$ matrices V were initialized. The V_t matrix was simulated as a VAR(1) process using the autoregression coefficient ρ. The Cholesky decomposition of $(V'V/T)$ was then computed to obtain $X_{\tilde{t}}$ and ensure that the regressors are independent and exogenous. The variable of interest, x_{1t}, was generated using the first exogenous regressor combined with a feedback effect $\alpha = 1.5$ and the lagged error term (ϵ_{t-1}). The true coefficient vector β was defined as a zero vector. The outcome variable y was computed by multiplying X with β and adding ϵ_t, where X is a new matrix composed of the first weak exogenous regressor and all the other

exogenous regressors. Finally, a linear model was fitted, and the first coefficient estimate $\hat{\beta}_1$ was stored.

Table 3.4. Simulation Results

K_values	Mean_abs_biases	Std_devs	Coverage_ratios
10	0.03188510	0.03867485	0.946
20	0.03862833	0.04202530	0.894
30	0.04342498	0.04168722	0.868
40	0.05289169	0.04248582	0.797
50	0.05895662	0.04380035	0.759
60	0.06436209	0.04442067	0.706
70	0.07592833	0.04847724	0.621
80	0.07725356	0.04646007	0.635
90	0.08273144	0.04881821	0.598
100	0.08613708	0.04944025	0.594
110	0.08976419	0.05319210	0.561
120	0.09174973	0.05480303	0.562
130	0.08702714	0.05776020	0.628
140	0.08921463	0.05957892	0.647
150	0.08329007	0.06053504	0.715

The coverage ratios illustrate the deteriorating performance of the OLS estimator as the ratio K/T increases. Notably, even when the number of regressors is about 10% the number of observations, the coverage ratio is significantly below the nominal level of 0.95.

Graph 3.1 below shows the dynamic of the bias and the standard deviation across different K values of the previous simulation.

Graph 2: Mean Absolute Bias and Standard Deviation Across K Values

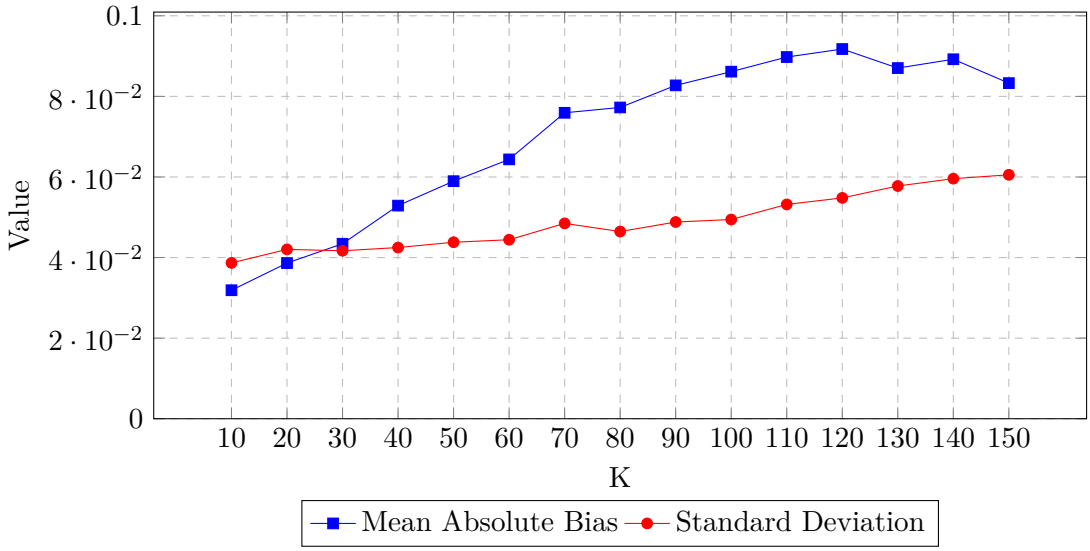


Figure 3.1. Mean Absolute Bias and Standard Deviation Across K Values

Chapter 4

Methods to improve the inference reliability

4.1 The Endogeneous/Invalid instrument

The issue of weak exogeneity is a critical challenge in high-dimensional settings, particularly in the context of time series analysis where feedback effects are prevalent. Weak exogeneity implies that the error term ϵ_t has a zero conditional expectation given the present and past values of the regressors but allows for correlation with future regressor values. This can lead to significant biases in the OLS estimators, especially as the number of regressors K increases with the sample size T.

To address this problem, Mikusheva and Sølvsten [2023] propose a novel estimator that corrects for the bias induced by weak exogeneity. The key idea is to construct a bias-corrected estimator that mimics an instrumental variables (IV) approach. Specifically, they use a linear combination of the regressors and their future values to form an intentionally endogenous instrument. This instrument induces an endogeneity bias along the feedback direction, which offsets the bias originating from weak exogeneity in the OLS estimator.

Consider the linear model:

$$y_t = \mathbf{X}_t'\boldsymbol{\beta} + \epsilon_t, \quad t = 1, \ldots, T,$$

where $\mathbf{X}_t \in \mathbb{R}^K$ are the regressors, which are weakly exogenous:

$$E[\epsilon_t | \mathbf{X}_t, \mathbf{X}_{t-1}, \ldots] = 0.$$

The bias in the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$ arises because the design matrix $\frac{1}{T}\mathbf{X}'\mathbf{X}$ remains asymptotically random and correlates with $\mathbf{X}'\mathbf{y}$ under weak exogeneity.

To correct this bias, they propose the following estimator:

$$\hat{\boldsymbol{\beta}}_{BC} = (\mathbf{X}'\mathbf{M}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}\mathbf{y}, \tag{4.1}$$

where $\mathbf{M}$ is a projection matrix that orthogonalizes $\mathbf{X}$ with respect to its future values.

This approach ensures that the bias correction is based solely on the regressors, making it applicable regardless of the outcome variable. The bias-corrected estimator $\hat{\boldsymbol{\beta}}_{BC}$ is shown to be consistent and asymptotically Gaussian under the assumption of weak exogeneity, even when there are multiple periods of feedback effects.

Since this method relies on an instrumental variables (IV) approach, the instrument must be relevant. The relevance condition implies that there must be a non-zero correlation between the instrument and the endogenous regressor. In this context, the relevance is derived from the feedback effect. Mathematically, the instrument's relevance can be assessed by the correlation:

$$\rho = \mathrm{Corr}(\mathbf{z}_t, \mathbf{X}_t),$$

where $\mathbf{z}_t$ is the instrument formed by future values of $\mathbf{X}_t$ and $\mathbf{X}_t$ is the endogenous regressor.

If this correlation is close to zero, the instrument fails to capture the necessary variation in the endogenous regressor, leading to weak instrument issues. This is particularly problematic because the bias of the IV estimator increases as the instrument's relevance decreases.

Another critical aspect is the accurate specification of the instrument to account for the feedback effect under weak exogeneity, especially when feedback occurs over

multiple periods. The derivation process involves forming the instrument by using future values of the regressors in a way that offsets the bias induced by feedback. Mathematically, the instrument is constructed as a linear combination of current and future values of the regressors, similar to:

$$\mathbf{z}_t = \mathbf{x}_t - \sum_{\ell=1}^{L} \gamma_\ell \mathbf{x}_{t+\ell},$$

where L denotes the number of feedback periods considered, and γ_ℓ are parameters chosen to ensure that the resulting instrument $\mathbf{z}_t$ induces an endogeneity bias that offsets the bias originating from weak exogeneity. The parameters γ_ℓ are selected to satisfy a system of equations that minimizes the bias in the feedback direction.

However, if the number of feedback periods L is misspecified, the method may fail to fully correct the bias. The number of feedback periods plays a crucial role in capturing the dynamic structure of the feedback. Incorrect specification of L can reduce the instrument's relevance and lead to weak instrument problems. For instance, if the true feedback effect occurs at lag k but the instrument is constructed with a different lag length, the correlation between the instrument and the endogenous regressor may weaken, introducing further bias and inconsistency in the estimator.

In the paper, the parameter γ_ℓ that reduces the bias is determined by solving a non-linear equation that minimizes the feedback effect in the OLS bias. The equation to solve for each feedback period ℓ is given by:

$$\operatorname{tr}\left[D'\left(I - \sum_{m=1}^{L} \gamma_m D^m\right) M_\gamma\right] = 0, \quad \forall\, \ell = 1, \ldots, L,$$

where D is the lead operator matrix (shifting indices forward by one period), and M_γ is the projection matrix orthogonal to the regressors, adjusted for the transformation induced by the feedback correction. The parameter γ_ℓ is chosen such that the bias induced by the weak exogeneity is offset by the instrument, and the equation ensures this cancellation occurs in the direction of the feedback.

Solving this equation is non-trivial, especially when multiple lags ($L > 1$) are considered. The following challenges arise:

1. **Uniqueness of the Solution**: According to the paper, the uniqueness of

the solution is only guaranteed if the number of regressors K is sufficiently small compared to the sample size T. Specifically, the solution is unique if $K < T/5$. When the number of regressors exceeds this threshold there is no guarantee that a unique solution exists. In such cases, numerical instability and convergence issues may arise.

2. **High Dimensionality**: When considering multiple feedback periods, the system's dimensionality increases, introducing additional complexity. Each additional lag period requires solving an extra equation. This leads to potential computational challenges, especially when the feedback direction is not well-identified or regressors are highly autocorrelated.

3. **Non-linear Nature of the System**: The system of equations to solve is non-linear, which makes it sensitive to the choice of initial values and the numerical method used for finding the solution.

These challenges underscore the difficulty of applying the bias correction method in practical settings, particularly when dealing with high dimensions, which is the aim of this research.

4.2 Addressing the Autocorrelation

The average autocorrelation coefficient of the regressors plays a critical role in determining the magnitude of the overall bias and impacts the reliability of the OLS estimator. This coefficient serves as the mechanism through which the feedback effect propagates into the dynamic model.

Table 4.1 shows its role. Keeping the previous simulation setup fixed, additional Monte Carlo simulations were conducted to evaluate the performance of OLS coverage ratios across different average ρ values of the regressors:

$$X_t = \rho X_{t-1} + \varepsilon_t.$$

In line with the theoretical findings, when there is no autocorrelation ($\rho = 0$), the coverage ratio is near the nominal value of 95%. Conversely, as ρ increases, the

coverage ratio performances worsen, since the feedback effect α is amplified.

Table 4.1. Coverage Ratios for different ρ values

K_values	$\rho = 0$	$\rho = 0.3$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 1$
10	0.942	0.947	0.958	0.929	0.952
20	0.953	0.940	0.940	0.913	0.914
30	0.953	0.930	0.899	0.888	0.884
40	0.951	0.925	0.877	0.820	0.781
50	0.952	0.903	0.849	0.766	0.738
60	0.950	0.907	0.815	0.738	0.668
70	0.955	0.888	0.797	0.708	0.633
80	0.932	0.879	0.767	0.633	0.596
90	0.954	0.851	0.736	0.613	0.566
100	0.954	0.855	0.704	0.607	0.577
110	0.948	0.867	0.717	0.609	0.547
120	0.935	0.850	0.741	0.578	0.572
130	0.938	0.835	0.716	0.631	0.622
140	0.950	0.858	0.746	0.676	0.653
150	0.938	0.867	0.797	0.710	0.695

In the presented model, the OLS estimation produces inaccurate inferences due to the bias in both the coefficient estimates and their standard deviations. Even under the assumption that the error terms (ϵ_t) in the data-generating process are i.i.d. and uncorrelated, OLS estimation can still lead to autocorrelation in the error term if the model is misspecified. Consequently, traditional standard deviation calculations exhibit a poor coverage ratio. The underlying issue is that x_t follows an autoregressive process of order 1 and the model is specified as $y_t = x_t + \epsilon_t$.

4.2.1 *Proof of Autocorrelation due to Misspecification*

Consider the true model specified as:

$$y_t = \beta_1 x_t + \epsilon_t, \tag{4.2}$$

where $\epsilon_t \sim$ i.i.d.$(0, \sigma^2)$ and is independent of x_t.

Suppose that the explanatory variable x_t presents a feedback effect from the error term ϵ_{t-1}:

$$x_t = x_t^* + \alpha\epsilon_{t-1}, \tag{4.3}$$

where:

- x_t^* is an exogenous AR(1) process:

$$x_t^* = \phi x_{t-1}^* + \nu_t, \tag{4.4}$$

  with $\nu_t \sim \text{i.i.d.}(0, \sigma_\nu^2)$ and independent of ϵ_t and ϵ_{t-1}.

- α is the coefficient capturing the feedback effect.

Suppose we estimate the model using OLS:

$$y_t = \hat{\beta}_1 x_t + u_t. \tag{4.5}$$

The residuals are:

$$\begin{aligned} u_t &= y_t - \hat{\beta}_1 x_t && (4.6)\\ &= (\beta_1 x_t + \epsilon_t) - \hat{\beta}_1 x_t && (4.7)\\ &= (\beta_1 - \hat{\beta}_1) x_t + \epsilon_t && (4.8)\\ &= \Delta\beta \cdot x_t + \epsilon_t, && (4.9) \end{aligned}$$

where $\Delta\beta = \beta_1 - \hat{\beta}_1$ represents the estimation bias.
Substitute $x_t = x_t^* + \alpha\epsilon_{t-1}$ into the expression for u_t:

$$\begin{aligned} u_t &= \Delta\beta \cdot (x_t^* + \alpha\epsilon_{t-1}) + \epsilon_t && (4.10)\\ &= \Delta\beta \cdot x_t^* + \Delta\beta\alpha\epsilon_{t-1} + \epsilon_t. && (4.11) \end{aligned}$$

Since x_t^* is exogenous and independent of ϵ_t and ϵ_{t-1}, the term $\Delta\beta \cdot x_t^*$ does not introduce autocorrelation related to ϵ_t. However, it may still contribute to autocorrelation in u_t if x_t^* is serially correlated (which it is, due to its AR(1) nature).

For simplicity, let's focus on the term involving ϵ_{t-1}, which directly introduces autocorrelation in u_t:

$$u_t = \Delta\beta\alpha\epsilon_{t-1} + \Delta\beta \cdot x_t^* + \epsilon_t. \tag{4.12}$$

To assess the autocorrelation in u_t, compute the covariance between u_t and u_{t-1}. First, compute the variance of u_t:

$$\begin{aligned}
\text{Var}(u_t) &= \text{Var}(\Delta\beta\alpha\epsilon_{t-1} + \Delta\beta \cdot x_t^* + \epsilon_t) && (4.13)\\
&= (\Delta\beta\alpha)^2\text{Var}(\epsilon_{t-1}) + (\Delta\beta)^2\text{Var}(x_t^*) + \text{Var}(\epsilon_t) && (4.14)\\
&\quad + 2\Delta\beta^2\alpha\text{Cov}(\epsilon_{t-1}, x_t^*) && (4.15)\\
&= (\Delta\beta\alpha)^2\sigma^2 + (\Delta\beta)^2\sigma_{x^*}^2 + \sigma^2, && (4.16)
\end{aligned}$$

since ϵ_t is independent of ϵ_{t-1} and x_t^*, and $\text{Cov}(\epsilon_{t-1}, x_t^*) = 0$.
Next, compute the covariance between u_t and u_{t-1}:

$$\begin{aligned}
\text{Cov}(u_t, u_{t-1}) &= \text{Cov}(\Delta\beta\alpha\epsilon_{t-1} + \Delta\beta x_t^* + \epsilon_t, \Delta\beta\alpha\epsilon_{t-2} + \Delta\beta x_{t-1}^* + \epsilon_{t-1}) && (4.17)\\
&= \Delta\beta^2\alpha^2\text{Cov}(\epsilon_{t-1}, \epsilon_{t-2}) + \Delta\beta^2\text{Cov}(x_t^*, x_{t-1}^*) && (4.18)\\
&\quad + \Delta\beta\alpha\text{Cov}(\epsilon_{t-1}, \epsilon_{t-1}) && (4.19)\\
&= \Delta\beta^2\text{Cov}(x_t^*, x_{t-1}^*) + \Delta\beta\alpha\sigma^2. && (4.20)
\end{aligned}$$

Here, we used:

- $\text{Cov}(\epsilon_{t-1}, \epsilon_{t-2}) = 0$ (since ϵ_t is i.i.d.).
- $\text{Cov}(\epsilon_{t-1}, \epsilon_{t-1}) = \sigma^2$.
- $\text{Cov}(x_t^*, x_{t-1}^*)$ is the autocovariance of x_t^* at lag 1.

Thus, the covariance simplifies to:

$$\text{Cov}(u_t, u_{t-1}) = \Delta\beta^2\gamma_{x^*}(1) + \Delta\beta\alpha\sigma^2, \tag{4.21}$$

where $\gamma_{x^*}(1) = \text{Cov}(x_t^*, x_{t-1}^*)$.

The autocorrelation coefficient at lag 1 is:

$$\rho_1 = \frac{\text{Cov}(u_t, u_{t-1})}{\text{Var}(u_t)}. \tag{4.22}$$

Substituting the expressions:

$$\rho_1 = \frac{\Delta\beta^2\gamma_{x^*}(1) + \Delta\beta\alpha\sigma^2}{(\Delta\beta\alpha)^2\sigma^2 + (\Delta\beta)^2\sigma_{x^*}^2 + \sigma^2}. \tag{4.23}$$

This expression shows that the autocorrelation coefficient ρ_1 depends on $\Delta\beta$, α, and the properties of x_t^*.

If $\Delta\beta = 0$, meaning $\hat{\beta}_1$ is an unbiased estimator of β_1, then:

$$u_t = \epsilon_t, \tag{4.24}$$

and since $\epsilon_t \sim \text{i.i.d.}(0, \sigma^2)$, there is no autocorrelation in u_t.

However, if $\Delta\beta \neq 0$, the residuals u_t depend on ϵ_{t-1} and x_t^*, introducing autocorrelation. Specifically, the presence of the term $\Delta\beta\alpha\epsilon_{t-1}$ means u_t follows an MA(1) process:

$$u_t = \Delta\beta\alpha\epsilon_{t-1} + e_t, \tag{4.25}$$

where $e_t = \Delta\beta x_t^* + \epsilon_t$. Setting $\Delta\beta\alpha = \theta$, we end up with the standard form of an MA(1):

$$u_t = \theta\epsilon_{t-1} + e_t, \tag{4.26}$$

with the standard properties:

$$\rho(k) = \begin{cases} 1, & \text{if } k = 0, \\ f(\theta), & \text{if } k = 1, \\ 0, & \text{if } k \geq 2. \end{cases} \tag{4.27}$$

where $f(\theta)$ means that the autocorrelation coefficient of order 1 $\rho(1)$ is a function of the parameter θ. By definition, the value of $f(\theta)$ is always bounded between 1

and 0.

In conclusion, when x_t follows an AR(1) process and the model is specified as $y_t = \hat{\beta}_1 x_t + u_t$, the residuals u_t will exhibit autocorrelation if $\Delta\beta$ is non-negligible.

The critical factor affecting the coverage ratio is the ratio K/T, where K is the number of regressors and T is the sample size. Under asymptotic conditions, where K/T tends to zero as T increases faster than K, the bias is given by:

$$\mathbb{E}[\Delta\beta] = \mathbb{E}[\beta_1 - \hat{\beta}_1] = 0. \tag{4.28}$$

This implies that the OLS estimator is unbiased in large samples with a relatively small number of regressors. However, in finite samples, as the ratio K/T increases, the reliability of the estimated β decreases, leading to a larger $\Delta\beta$. As $\Delta\beta$ increases, it introduces more pronounced autocorrelation in the residuals u_t, exacerbating the decline in the coverage ratio. This effect is observed in the poor coverage ratios in Table 4.1.

Given the results of the previous section, which specify the effect of the bias:

$$\mathbb{E}[X_1' M^{-1} \epsilon \mid \tilde{X}] = \alpha\sigma^2 \frac{K}{T}\bar{\rho}, \tag{4.29}$$

$$\frac{\hat{\sigma}^2}{\sigma^2} = 1 - \frac{\Phi}{T-K}\left(\frac{K}{T}\bar{\rho}\right)^2 + o_p(1) \tag{4.30}$$

increasing the ratio K/T has two negative effects. First, it directly amplifies the bias. Second, it increases the autocorrelation in the error term due to model misspecification.

As shown in the table 4.1, when autocorrelation is removed from the model, specifically when $\rho = 0$, the OLS estimator performs reliably even in high-dimensional settings, with the coverage ratio closely approximating the nominal value. To address the issue of autocorrelation and further enhance the properties of the OLS estimator in high-dimensional time series, I propose the use of lag augmentation methods as a systematic approach to mitigate the effects of autocorrelation.

4.3 Lag-augmentation Methods

A common and efficient method to address autocorrelation in the error term, as recommended by Olea and Plagborg-Møller [2021], is the lag-augmentation technique. The rationale behind this method is straightforward: autocorrelation in the error term indicates model misspecification. By incorporating lags of the variables, one can better capture the dynamics of the model and mitigate the autocorrelation in the residuals.

For instance, consider a simple autoregressive model where the dependent variable Y_t is

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \theta_0 X_t + \epsilon_t.$$

If the error term ϵ_t exhibits autocorrelation, one might augment the model by adding lags of both Y_t and X_t to capture the underlying dynamics more effectively:

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i Y_{t-i} + \sum_{j=0}^{q} \theta_j X_{t-j} + \epsilon_t.$$

In this augmented model, including lags of both the dependent variable Y_t and the independent variable X_t captures the temporal dynamics and addresses any potential autocorrelation in the residuals.

Even though the rationale for lag augmentation is conceptually simple, researchers often face practical challenges due to the unknown true data-generating process. This uncertainty leads to two significant issues: deciding which variables to include in the lags and the appropriate number of lags to include in the model for each variable.

In this thesis, I propose the lag-augmentation method to improve the reliability of OLS estimates. While lag augmentation is widely recognized for its efficacy in mitigating autocorrelation, its properties in high-dimensional settings have not been thoroughly explored. The primary rationale behind lag augmentation is that it cleanses the autocorrelation model. As demonstrated in the previous section 4.2, when $\rho = 0$, the coverage ratio is close to the nominal level, indicating that the inferences made using OLS are correct even when the ratio of K/T is high. Never-

theless, further investigation is needed to understand the behavior and properties of lag augmentation when applied in high-dimensional contexts.

Intuition on Autocorrelation Reduction

We begin with the original regression model:

$$y_t = \beta_1 x_{1t} + \beta_j x_{jt} + \epsilon_t \tag{4.31}$$

where:

- y_t is the dependent variable.
- x_{1t} is the primary explanatory variable of interest.
- x_{jt} represents other exogenous explanatory variables ($j \geq 2$).
- ϵ_t is the error term, assumed to be white noise: $\epsilon_t \sim \text{i.i.d.}(0, \sigma_\epsilon^2)$.

Assume that x_{1t} is influenced by the past error term ϵ_{t-1}, introducing endogeneity and autocorrelation:

$$x_{1t} = x_{1t}^* + \alpha \epsilon_{t-1}, \tag{4.32}$$

where:

- x_{1t}^* is an exogenous variable, following an AR(1) process:

$$x_{1t}^* = \phi x_{1,t-1}^* + \nu_t, \quad \nu_t \sim \text{i.i.d.}(0, \sigma_\nu^2), \quad \nu_t \perp \epsilon_s \ \forall s. \tag{4.33}$$

- α captures the feedback effect.

The OLS estimator for β_1 is given by:

$$\hat{\beta}_{\text{OLS}} = \left(X_1' X_1\right)^{-1} X_1' y, \tag{4.34}$$

where X_1 is the regressor matrix and y is the vector of dependent variables.

To assess the bias in the OLS estimator, we decompose it as follows:

$$\Delta\beta = \hat{\beta}_{\text{OLS}} - \beta_1 = \left(X_1' X_1\right)^{-1} X_1' \hat{u}. \tag{4.35}$$

As derived in the previous section, the residual $\hat{u}_t$ exhibits autocorrelation if the ratio K/T is sufficiently large. Specifically, it follows an MA(1) process:

$$\hat{u}_t = \theta_1 \epsilon_{t-1} + \epsilon_t, \tag{4.36}$$

where $\theta_1 = \Delta\beta\alpha$.

Substituting this into the OLS estimator expression, we get:

$$\hat{\beta}_{\text{OLS}} - \beta_1 = \left(X_1' X_1\right)^{-1} X_1' (\theta_1 \epsilon_{t-1} + \epsilon_t). \tag{4.37}$$

Since ϵ_{t-1} is correlated with u_t through θ_1, the expectation of the second term is non-zero, leading to bias:

$$\mathbb{E}\left[X_1' \epsilon\right] \neq 0. \tag{4.38}$$

More precisely, the bias depends on the ratio K/T.

$$\mathbb{E}[X_1' M^{-1} \epsilon \mid \tilde{X}] = \alpha\sigma^2 \frac{K}{T} \bar{\rho}, \tag{4.39}$$

where

$$\bar{\rho}(1) = \frac{\text{Cov}(u_t, u_{t-1})}{\text{Var}(u_t)} = \frac{\theta_1 \sigma^2}{\sigma^2(1+\theta_1^2)} = \frac{\theta_1}{1+\theta_1^2}. \tag{4.40}$$

Asymptotically, $\Delta\beta \to 0$, but in finite samples, where the ratio K/T is sufficiently large, $\Delta\beta$ remains non-zero. Consequently, the residuals $\hat{u}_t$ exhibits autocorrelation. The magnitude of this autocorrelation depends on θ_1, which is proportional to the bias in the estimated coefficient β_1.

A lag augmentation technique is employed to address this autocorrelation. Specifically, the lagged dependent variable y_{t-1} is introduced into the regression model:

$$y_t = \beta_1 x_{1t} + \beta_j x_{jt} + \gamma y_{t-1} + \epsilon_t. \tag{4.41}$$

The underlying intuition is that the feedback effect introduces autocorrelation at time $t-1$ due to endogeneity between ϵ_{t-1} and x_{1t}. By including y_{t-1}, which is influenced by ϵ_{t-1}, this technique helps mitigate the autocorrelation. Specifically:

$$y_{t-1} = \beta_1 x_{1,t-1} + \beta_j x_{j,t-1} + \epsilon_{t-1}. \tag{4.42}$$

Using the Partitioning-out Theorem, the new OLS estimator for β_1 becomes:

$$\hat{\beta}_1 = \left(X_1' M_{y_{t-1}} X_1\right)^{-1} X_1' M_{y_{t-1}} y, \tag{4.43}$$

where $M_{y_{t-1}} = I - P_{y_{t-1}}$ is the projection matrix that removes the effect of y_{t-1}, and $P_{y_{t-1}} = y_{t-1}\left(y_{t-1}' y_{t-1}\right)^{-1} y_{t-1}'$ is the projection onto the space spanned by y_{t-1}. Applying the $M_{y_{t-1}}$ matrix to X_1 means removing the effect of y_{t-1} on the regressors. We get transformed regressors that are orthogonal to y_{t-1}, and indeed orthogonal to ϵ_{t-1}.

Thus, the estimator becomes:

$$\hat{\beta}_1 = \left(\tilde{X}_1' X_1\right)^{-1} \tilde{X}_1' y, \tag{4.44}$$

where $\tilde{X}_1 = X_1' M_{y_{t-1}}$ represents the residuals from regressing X_1 on y_{t-1}.

The bias is:

$$\Delta\beta = \hat{\beta}_1 - \beta_1 = \left(\tilde{X}_1' X_1\right)^{-1} \tilde{X}_1' \hat{u}_t. \tag{4.45}$$

Substituting the MA(1) process for $\hat{u}_t$, we get:

$$\hat{u}_t = \epsilon_t + \theta_1 \epsilon_{t-1}. \tag{4.46}$$

$$\mathbb{E}\left[\tilde{X}_1 \hat{u}_t\right] = \mathbb{E}\left[\tilde{X}_1' (\epsilon_t + \theta_1 \epsilon_{t-1})\right]. \tag{4.47}$$

The key observation is that by including y_{t-1}, we effectively control for part of the autocorrelation. Since $M_{y_{t-1}}$ projects onto the space orthogonal to y_{t-1}, it removes the component of ϵ_{t-1}, since it is explained by y_{t-1}, thus reducing the bias.

$$\mathbb{E}\left[\tilde{X}_1'\epsilon_{t-1}\right] = 0. \tag{4.48}$$

It follows:

$$\mathbb{E}[\tilde{X}_1\hat{u}] = 0. \tag{4.49}$$

In order for the OLS estimator to be effective, it must correctly estimate the parameter γ associated with the lagged dependent variable y_{t-1}. Asymptotically, the estimated $\hat{\gamma}$ converges to the true γ. However, when the ratio K/T becomes sufficiently large, the estimates become less precise, and the difference $\Delta\gamma = \hat{\gamma} - \gamma$ increases.

As $\Delta\gamma$ increases, y_{t-1} fails to fully capture the feedback effect. Nevertheless, compared to the model without y_{t-1}, the influence of the feedback on the model is reduced. Consequently, the bias $\Delta\beta$ is also reduced, and the coverage ratio is improved.

If the specified model can fully capture the system dynamics, the LA is reliable, with a coverage ratio close to 95%, even in high dimensions. A simulation was run to show this in practice. Data are generated following Section 3.3.3. The y_t are generated from the $\mathbf{X_t}$ matrix, which contains K regressor x_t, all generated as an AR(1), with coefficient $\rho = 0.9$. The first regressor of interest x_{1t} presents a feedback effect. T is equal to 200. However, the regression controls, are the lagged $\mathbf{X_t}$ matrix and the lagged y_t. Thus, the LA regression has this form:

$$y_t = \beta_0 X_t + \sum_{i=1}^{p} \beta_i X_{t-i} + \sum_{j=1}^{q} \gamma_j y_{t-j} + \varepsilon_t \tag{4.50}$$

In particular, in the simulation, the number of regressors that generate the y_t is 20, the lags of $\mathbf{X_t}$ are 3 and the y_t lags are 4. The total number of variables in the regression is:

$$\textit{Total regressors} = 20 + 20 \cdot 3 + 4 = 84. \tag{4.51}$$

The summary statistics are the following:

The data are generated through an AR(1), however, the regression uses 3 and 4

Table 4.2. Simulation Results

K_values	mean_abs_biases	std_devs	coverage_ratios
84	0.13	0.13	0.941

lags. The model is overspecified, but in doing so, the residual correlation should be close to zero. As theoretically demonstrated before, in this context, the coverage ratio is correct.

In practice, fully capturing the dynamic of the system could be difficult. Thus, the next sections will explore what happens if the latter is left out of the model specification.

4.3.1 *First Model: Inclusion of* y_{t-1}

The first model includes the first lag of the dependent variable y_t alongside the current values of independent variables x_{1t} and x_{jt} with $j \geq 2$. This can be expressed as:

$$y_t = \beta x_{1t} + \gamma x_{jt} + \delta y_{t-1} + \epsilon_t \tag{4.52}$$

In this specification, the inclusion of y_{t-1} is crucial for controlling the feedback effect on the first regressor, specifically x_{1t}. It helps to account for the influence of past values of the dependent variable on its current value, represented by the parameter δ. By incorporating y_{t-1}, we aim to capture this dynamic feedback effect and mitigate the potential bias in the estimation of β.

However, while the inclusion of y_{t-1} helps address some of the autocorrelations, it does not eliminate the issue if the ratio K/T is large enough. Therefore, further model adjustments or the inclusion of additional lags might be necessary to address autocorrelation and endogeneity concerns fully.

This section analyzes the impact of adding a lagged dependent variable, y_{t-1}. Specifically, we consider the case where x_{1t} follows an AR(1) process and is weakly exogenous due to its correlation with ϵ_{t-1}. We demonstrate why the inclusion of y_{t-1} improves model performance and provide a simulation to illustrate this improvement.

First Model Simulation

The Lag Augmentation method aims to fully capture the dynamic of the system, such that the residuals are not autocorrelated. Since the data-generating process is not known, this may be not always possible. In particular, the performances depend on the researcher's ability to choose the correct controls.

The following simulation evaluates the performance of the LA when the model is not completely able to capture all the relevant dynamics, which should be the most common situation in practical applications.

The data are generated as in Section 3.3.3. The y_t are generated from the $\mathbf{X_t}$ matrix, which contains K regressor x_t, all generated as an AR(1), with coefficient $\rho = 0.9$. The first regressor of interest x_{1t} presents a feedback effect. T is equal to 200.

Running the following regression:

$$y_t = \beta_1 x_{1t} + \beta_j x_{jt} + \gamma y_{t-1} + \epsilon_t, \tag{4.53}$$

it means that we are not accounting for all the x_t autocorrelation, since they all follow an AR(1).

Table 4.3 presents the OLS coverage ratio when y_{t-1} is included as a control variable. The results indicate an enhancement in performance relative to the standard model that excludes this adjustment. However, as stated before, if the ratio K/T is large enough, there may be dynamics in the system that cause autocorrelation in the residuals. As a consequence, the coverage ratio is below the nominal one.

4.3.2 *Second Model: Addition of x_{1t-1} as a Lagged Variable*

Researchers generally include the same lag structure for both y_t and x_t. Therefore, the Ordinary Least Squares estimation is performed on the following equation:

$$y_t = \alpha + \beta x_{1t} + \beta_j x_{jt} + \gamma y_{t-1} + \delta x_{1t-1} + \epsilon_t \tag{4.54}$$

The results displayed in Table 4.4 indicate that incorporating the lagged regressor x_{1t-1} as a control variable leads to poorer model performance. This outcome can be

Table 4.3. Simulation Results

K_values	mean_abs_biases	std_devs	coverage_ratios
10	0.05633616	0.07117394	0.941
20	0.05788008	0.07260134	0.946
30	0.06053463	0.07593353	0.934
40	0.06062931	0.07494184	0.936
50	0.05824211	0.07445798	0.932
60	0.05913653	0.07453830	0.926
70	0.06389426	0.07934055	0.903
80	0.06339291	0.07929869	0.908
90	0.06399438	0.07960490	0.913
100	0.06457637	0.08195714	0.895
110	0.07002634	0.08759066	0.877
120	0.06881610	0.08566025	0.886
130	0.06808279	0.08461920	0.882
140	0.06816346	0.08618340	0.870
150	0.06827039	0.08606135	0.864

attributed to the feedback effect, which results in weak exogeneity. Consequently, the inclusion of x_{1t-1} introduces endogeneity and induces correlation with the error term ϵ_{t-2}.

Recall that the residuals of the first model without any lagged variables are:

$$\hat{u_t} = \Delta\beta\alpha\epsilon_{t-1} + \Delta\beta \cdot x_t^* + \epsilon_t. \tag{4.55}$$

If x_{1t-1} is included as a regressor with the parameter δ, it follows that the residuals of the models also depend on the $\Delta\delta = \hat{\delta} - \delta$. In particular:

$$\hat{u_t} = \Delta\delta \cdot \alpha \cdot \epsilon_{t-2} + \textit{(other terms)}. \tag{4.56}$$

Since x_{1t-1} is equal to:

$$x_{1t-1} = x_{1t-1}^* + \alpha\epsilon_{t-2}. \tag{4.57}$$

as a result:

$$\text{Cov}(x_{1t-1}, \hat{u_t}) \neq 0. \tag{4.58}$$

Therefore, while the inclusion of lagged independent variables such as x_{1t-1} aims

to capture dynamic relationships, it is crucial to carefully consider the potential endogeneity issues and their implications on the model's validity and reliability.

Table 4.4. Simulation Results

K_values	mean_abs_biases	std_devs	coverage_ratios
10	0.06112691	0.07705777	0.938
20	0.06101958	0.07713378	0.939
30	0.06598970	0.08114270	0.923
40	0.06619620	0.08059100	0.920
50	0.07220397	0.08545231	0.890
60	0.06958347	0.08336807	0.909
70	0.07586163	0.08993367	0.878
80	0.07542953	0.08798509	0.880
90	0.07767911	0.09155946	0.871
100	0.07854303	0.09114977	0.875
110	0.08551548	0.09895601	0.844
120	0.08250756	0.09552582	0.854
130	0.09027050	0.09933377	0.838
140	0.09297506	0.10520678	0.824
150	0.09700606	0.11030987	0.826

In conclusion, incorporating $x_{1,t-1}$ into the model may lead to inferior OLS performance due to the resulting endogeneity problem.

It is crucial to acknowledge that when the ratio K/T approaches zero, the coverage is satisfactory. However, issues arise if K increases at a faster rate than T, as this leads to autocorrelation, as demonstrated in the preceding section.

4.3.3 *Third Model: Addition of y_{t-2} to Eliminate the Feedback Effect*

The third model further includes the second lag of the dependent variable y_{t-2} to address the feedback effect and improve model performance. The equation for this model is:

$$y_t = \alpha + \beta x_{1t} + \beta_j x_{jt} + \gamma y_{t-1} + \delta x_{1t-1} \eta y_{t-2} + \epsilon_t$$

The findings indicate that the inclusion of y_{t-2} results in coverage ratios closely aligned with the nominal value of 0.95, signifying enhanced model accuracy. However,

this improvement comes at the cost of higher standard deviations, which can be attributed to the increased complexity introduced by the additional lagged variable. The inclusion of y_{t-2} helps to address feedback effects and mitigate endogeneity, but it also necessitates the estimation of additional parameters, thereby increasing the variability of the estimates.

Table 4.5. Summary Statistics for Different K Values

K_values	mean_abs_biases	std_devs	coverage_ratios
10	0.07239212	0.09324006	0.949
20	0.07741398	0.09882716	0.945
30	0.07431681	0.09504538	0.952
40	0.08068930	0.10272431	0.948
50	0.07893571	0.10112841	0.939
60	0.08162476	0.10595880	0.937
70	0.08445320	0.10974480	0.948
80	0.09209570	0.11976208	0.925
90	0.09091675	0.11865225	0.939
100	0.09480176	0.12315616	0.926
110	0.09901518	0.13093555	0.919
120	0.10544420	0.13741397	0.931
130	0.11594074	0.16429967	0.914
140	0.11873787	0.16179114	0.913
150	0.13290035	0.21688348	0.916

If you are concerned about weak exogeneity, your model should include more lags of the dependent variable y_t than the independent variable x_t.

4.3.4 *Fourth Model: Overspecification*

This section explores the situation in which the researcher overspecifies the model. In particular, in the simulation, the regression is:

$$y_t = \alpha + \beta_1 x_{1t} + \beta_j x_{jt} + \sum_{i=1}^{5} \gamma_i x_{1,t-i} + \sum_{j=1}^{6} \delta_j y_{t-j} + \epsilon_t \tag{4.59}$$

Intuitively, the lags of the independent variable of interest x_{1t} are 5, while the dependent variable y_t lags are 6.

The results are summarized in Table 4.6. Interestingly, the LA method does not exhibit a negative effect when additional lags are included as a control. The

econometric intuition behind this result is that adding an extra lag projects the system orthogonally to that variable. Consider the last control y_{t-6} in this specific case. Since the only term that could be correlated with y_{t-6} is the error term, and given that the error term is already free from autocorrelation, y_{t-6} becomes irrelevant to the regression. Consequently, the bias, standard deviation, and coverage ratio remain unchanged.

This significant result demonstrates that the output is asymmetric using a symmetric loss function, where the objective is to minimize the coverage error, defined as the difference between the nominal coverage ratio of 0.95 and the actual coverage ratio. The asymmetry arises from the fact that using fewer lagged controls incurs a higher cost, while adding additional controls proves inconsequential, leaving the model at a point of minimum error. In the LA approach, the misspecification cost is asymmetric. Instead, in the endogenous IV, the latter is symmetric since both the under and the over-specification are costly.

Table 4.6. Updated Summary Statistics for Different K Values

K_values	mean_abs_biases	std_devs	coverage_ratios
10	0.07736052	0.09829710	0.943
20	0.07431539	0.09351708	0.946
30	0.08249386	0.10502108	0.943
40	0.08090525	0.10635995	0.941
50	0.08523387	0.11968471	0.947
60	0.08517961	0.11015540	0.947
70	0.08382665	0.10702702	0.953
80	0.09347433	0.12003528	0.944
90	0.09548196	0.12494163	0.932
100	0.09941079	0.13030998	0.928
110	0.10501895	0.13561065	0.907
120	0.11468570	0.16110402	0.913
130	0.12283833	0.16303488	0.922
140	0.13359267	0.18109841	0.915
150	0.15079527	0.20631308	0.923

Chapter 5

Practical Implications

5.1 Lag-augmentation or Endogenous Instrument?

This chapter provides practical advice for a researcher tasked with conducting empirical analysis in a setting similar to that described earlier. One of the central challenges in such a framework is the correct specification of the model, particularly when the length of the lags and feedback are unknown. There are different setups: one where the model is correctly specified and others where it is misspecified. Furthermore, we will evaluate the performance of lag-augmented methods (LA) against that of endogenous instruments (IV), offering guidance on their use under different conditions.

According to theory, the effectiveness of an endogenous instrument is compromised when the ratio of variables K to time observations T exceeds 0.2 and in the presence of multiple feedbacks. It is important to note that, according to our simulations, the Matlab code provided by Mikusheva and Sølvsten [2023] can not find a finite solution in the presence of multiple feedbacks and a relatively small number of regressors (20/200). Conversely, lag-augmented methods can perform well even when the K/T ratio is close to 1, provided the residuals can be sufficiently cleaned from autocorrelation such that $\rho_\epsilon \approx 0$. While this may result in a higher estimator variance, the method remains valid under a broader range of conditions compared to the IV approach.

A key challenge for the researcher is that the data-generating process is unknown.

The IV method requires correct specification. There is a misspecification cost in terms of coverage error both for over and under specification. This can be defined as symmetric misspecification cost. In contrast, lag augmentation is more flexible, requiring only that the lag-specified number be greater than or equal to the true lag length (asymmetric misspecification cost). If the researcher overspecifies lags in the IV framework, the correlation between the instrument and bias decreases, undermining the instrument's effectiveness. In contrast, lag-augmented methods are less sensitive to over-specification, with the primary consequence being an increase in estimator variance. Simulations indicate that this increase is negligible unless the K/T ratio is large. However, in such cases, the IV approach is ineffective, leaving lag augmentation the only available method.

The competition between these two approaches becomes more evident in settings with relatively low dimensionality. Here, adding superfluous lags has minimal adverse impact on LA method performance. Given that the true model is unknown, lag-augmented methods offer a more practical and robust solution. It is important to note that even in cases where the K/T ratio is around 0.2, the performance of lag-augmentation remains strong. High feedback and near-unit-root correlation do not significantly detract from the asymptotic properties of the model. In this range of K values, the lag-augmented method performs with an accuracy close to 0.95. As the K/T ratio increases, however, the limitations of endogenous instruments become apparent, rendering lag augmentation the only feasible approach.

5.2 Empirical Example

Consider the usual setup where the data are generated with two feedback effects on the first regressor, such that:

$$x_{1t} = x_{1t}^{*} + a \cdot \epsilon_{t-1} + b \cdot \epsilon_{t-2} \tag{5.1}$$

In this study, the researcher works within a time-series framework, consisting of 200 observations and about 50 control variables. This setup is typical in time-series econometrics, particularly when dealing with models that involve multiple variables

and lags. Specifically, the model includes 4 control variables $z_{i,t}$, each controlled for with 10 lags. The y_t are generated from 10 x_t, where all are AR(1) and the first regressor is weak exogenous. The resulting model can be expressed as follows:

$$y_t = \alpha + \beta_1 x_{1t} + \delta x_{jt} + \sum_{i=1}^{4} \sum_{k=1}^{10} \gamma_{ik} z_{i,t-k} + \epsilon_t$$

As shown in the previous chapters, OLS results in incorrect inference when there is autocorrelation in the residuals. Since the model does not account for the AR(1) regressions structure, the feedback effect will propagate into the system, leading to the endogeneity issue. By setting the parameters $\rho = 0.8$ and the coefficients of the feedback effect $a = 1.5$ and $b = 1$, the OLS coverage ratio for the estimated β_1 when $K/T = 50/200$ is approximately 40%. Therefore, the inference performance of OLS in this context is poor.

The researcher now suspects the presence of a feedback effect and must specify the feedback length. Erroneously, he assumes that one lag is sufficient. He can proceed in two ways: either by using the lag-augmented OLS to account for one feedback effect, or by employing an endogenous IV approach, specifying one feedback effect.

Let's denote the lag-augmented regression as LA(p, q), where p is the number of equal lags for the independent variable x_{1t} and for the dependent variable y_t, while q stands for the number of extra lags of y_t. The p parameter accounts for the AR dynamic system. Instead, the parameter q accounts for the feedback effect. Thus, the LA(1,1) to account for one feedback is:

$$y_t = \alpha + \beta_1 x_{1t} + \delta_j x_{jt} + \beta_2 x_{1,t-1} + \sum_{i=1}^{4} \sum_{k=1}^{10} \gamma_{ik} z_{i,t-k} + \rho_1 y_{t-1} + \rho_2 y_{t-2} + \epsilon_t \quad (5.2)$$

The IV regressor is computed as:

$$Z = (I - \gamma D)X \quad (5.3)$$

where γ is the parameter that minimizes the impact of the feedback effect,

measured with the trace of the matrix[1]:

$$\operatorname{tr}\left(D'(I-\gamma D)M_\gamma\right)=0 \tag{5.4}$$

Thus, the estimated β_1 is:

$$\hat{\beta}_{IV}(\gamma)=\left(X'(I-\gamma D)X\right)^{-1}X'(I-\gamma D)y \tag{5.5}$$

Table 5.1 shows the results of the simulations.

Estimator	Bias	Standard Error	Coverage Ratio
OLS	0.083	0.034	0.40
LA(1,1)	0.085	0.087	0.837
IV	0.038	0.04	0.84

Table 5.1. Summary for OLS, LA, and IV with one specified feedback

The OLS coverage ratio is poor. Instead, LA and IV are better than OLS. Moreover, they are relatively close in coverage ratio performances. However, the IV has a lower bias and standard deviation.

Assume now that the researcher correctly specifies two feedback effects. The LA(1,2) regression has the following form:

$$y_t=\alpha+\beta_1 x_{1t}+\delta_j x_{jt}+\beta_2 x_{1,t-1}+\sum_{i=1}^{4}\sum_{k=1}^{10}\gamma_{ik}z_{i,t-k}+\sum_{k=1}^{3}\rho_k y_{t-k}+\epsilon_t \tag{5.6}$$

The corresponding results are presented in Table 5.2. The LA method demonstrates a coverage ratio that is close to the nominal value. In contrast, the IV method fails, as it is unable to find a solution with 50 regressors and 2 feedback effects.

Estimator	Bias	Standard Error	Coverage Ratio
OLS	0.083	0.034	0.40
LA(1,2)	0.119	0.128	0.947
IV	/	/	/

Table 5.2. Summary for OLS, LA, and IV with two specified feedbacks

[1]See Chapter 4 for the details.

Assume the researcher believes that the regressor of interest, x_{1t}, is subject to three feedback effects. Accordingly, the researcher specifies three feedbacks. Thus, the LA(1,3) regression is:

$$y_t = \alpha + \beta_1 x_{1t} + \delta_j x_{jt} + \beta_2 x_{1,t-1} + \sum_{i=1}^{4} \sum_{k=1}^{10} \gamma_{ik} z_{i,t-k} + \sum_{k=1}^{4} \rho_k y_{t-k} + \epsilon_t \tag{5.7}$$

and the results are presented in Table 5.3.

Estimator	**Bias**	**Standard Error**	**Coverage Ratio**
OLS	0.083	0.034	0.40
LA(1,3)	0.12	0.148	0.944
LA(1,5)	0.123	0.144	0.945
LA(3,3)	0.117	0.125	0.95
IV	/	/	/

Table 5.3. Summary for OLS, LA, and IV with feedback overspecification

Table 5.3 reports also the summary of the simulation conducted on LA(1,5) and the results are stable. While the IV mechanism fails to provide a solution for the non-linear system in case of multiple feedbacks, the LA method is robust to over-specification, as previously demonstrated in Section 4.3.4.

The LA(3,3) illustrates the results when the regression has 6 lags for y_t and 3 for the x_t. This demonstrates again that the over-specification does not have a relevant effect on the result.

Chapter 6

Conclusion

This thesis analyses the behavior of the OLS in high dimensions, particularly focusing on time series. While the OLS is widely used in time series empirical research, it is unreliable if the number of regressors is high enough compared to the sample size.

Chapter 3 clearly shows the OLS unreliability. The key issue is the autocorrelation. The regressor of interest is weak exogenous since it exhibits a feedback effect. The serial correlation structure of the model transports the feedback through time and the exogeneity assumption is violated. Thus, the estimator is biased and the standard errors are too optimistic. As a consequence, the confidence intervals are narrower and they fail to capture the true parameter as frequently as the theory states, leading to unreliable inference. Then, the coverage ratio is below the nominal value of 95%.

Mikusheva and Sølvsten [2023] proposed the Endogenous Instrumental Variable to obtain a bias-corrected OLS. The idea is to construct an IV correlated with the feedback effect that fully eliminates the bias. It is a linear combination of the weak exogenous regressor and its future values. The methodological procedure consists of solving a nonlinear equation for the parameter that allows to capture the feedback. Even if in theory works, in practice, solving the nonlinear equation is not trivial and the solution is not always guaranteed, especially if the system is complex (high dimension or multiple feedbacks).

I proposed an alternative system: the Lag Augmentation Method. This intuitive approach consists of adding to the original regression the lagged variables, both dependent and independent. It is commonly used in time series empirical applications,

but its theoretical properties in a high-dimensional context are still unexplored. Since the key problem of biased OLS inference is the serial correlation, the LA method reduces it. The simulations demonstrate that the coverage ratios are improved, with a value close to the nominal one.

Chapter 5 simulates an empirical application using both the endogenous IV and the LA method. If the model is correctly specified both methods show the correct coverage ratio, but the first one performs better in terms of variance and bias. Note that, during my simulations, in the case of multiple feedbacks, the IV method is not able to find a solution for the nonlinear equation. Instead, the LA method works. However, even if the coverage ratio is correct, this flexibility comes at the cost of higher variance and bias. Furthermore, the simulations show that the LA regression is robust to the over-specification. If the model captures almost all the autocorrelation, adding lags has no implications: the bias and the standard error remain constant and the coverage ratio is close to 95%.

Interestingly, usually applied researchers tend to use the same lag number for both the dependent and independent variables. However, if there is a suspect of feedback effect, it is recommended to include more dependent variable lags. As a rule of thumb, if the feedback effects are n and the independent variable lags are m, then the dependent variable lags should be $n + m$.

One promising direction for future research is to study in more detail the optimal number of lags to include in the LA method. Although the simulations in this thesis show that the LA method can improve inference, the choice of lag length remains an open question, especially if the data-generating process is unknown. Developing new methods or criteria to select the correct lag length in high-dimensional settings would be a valuable contribution to improving the robustness of this approach. Future research could also focus on testing and applying this approach to real data.

Bibliography

George Casella and Roger L. Berger. *Statistical Inference*. Duxbury, Pacific Grove, CA, 2nd edition, 2002.

Matias D. Cattaneo, Michael Jansson, and Whitney K. Newey. Inference in linear regression models with many covariates and heteroskedasticity. *arXiv preprint arXiv:1507.02493*, 2018. doi: 10.48550/arXiv.1507.02493.

Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013. doi: 10.48550/arXiv.1311.2445.

Robert F. Engle, David F. Hendry, and Jean-François Richard. Exogeneity. *Econometrica*, 51(2):277–304, 1983.

Ragnar Frisch and Frederick V. Waugh. Partial time regressions as compared with individual trends. *Econometrica*, 1(4):387–401, 1933. doi: 10.2307/1907330. URL `https://www.jstor.org/stable/1907330`.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3): 424–438, 1969.

William H. Greene. *Econometric Analysis*. Pearson, Upper Saddle River, NJ, 8th edition, 2020.

James Douglas Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

Peter J. Huber. Robust regression: Asymptotics, conjectures and monte carlo. *Annals of Statistics*, 1(5):799–821, 1973. doi: 10.1214/aos/1176342503.

Koen Jochmans. Heteroscedasticity-robust inference in linear regression models with many covariates. *Journal of the American Statistical Association*, 117(538): 887–896, 2022. doi: 10.1080/01621459.2020.1831924.

J. Scott Long and Laurie H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000. doi: 10.2307/2685594.

J. G. MacKinnon and H. White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29 (3):305–325, 1985. doi: 10.1016/0304-4076(85)90158-7.

James G. MacKinnon. Thirty years of heteroskedasticity-robust inference. *Proceedings of the Conference on Econometric Models*, 2013. URL `https://api.semanticscholar.org/CorpusID:118430779`.

Enno Mammen. Bootstrap and wild bootstrap for high dimensional linear models. *Annals of Statistics*, 21(1):255–285, 1993. doi: 10.1214/aos/1176349025.

Anna Mikusheva and Mikkel Sølvsten. Linear regression with weak exogeneity. *arXiv preprint arXiv:2308.08958*, 2023. URL `https://doi.org/10.48550/arXiv.2308.08958`.

Whitney K. Newey and Kenneth D. West. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987. doi: 10.2307/1913610.

José Luis Montiel Olea and Mikkel Plagborg-Møller. Local projection inference is simpler and more robust than you think. *Econometrica*, 89(4):1789–1823, July 2021.

James H. Stock and Mark W. Watson. *Introduction to Econometrics*. Pearson, fourth edition, 2019.

Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: Journal of the Econometric Society*, pages 817–838, 1980.

Jeffrey M. Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage Learning, Mason, OH, 7th edition, 2020.