No 2025-12 – September 2025

Working Paper

CEPII

The Variance of Gravity

Camilo Umana-Dajud

Highlights

- Gravity estimates become very noisy when policy regressors such as FTA dummies are sparse (i.e being different from zero fewer than 100 to 500 observations).
- In these settings, variance dominates statistical inference and moderate trade policy effects often cannot be detected statistically.
- Many trade policy variables are sparse which could explain insignificant results in the literature.

Abstract

This paper documents a fundamental problem in applied gravity estimation: the variance of gravity estimates becomes prohibitively large when policy regressors are parse. I define sparse regressors as dummy variables that equal one in fewer than 100 to 500 observations depending on the setting; a common characteristic of trade policies such as free trade agreements. Through Monte Carlo simulations calibrated to match previously established data generating processes in the literature, I demonstrate that the variance of coefficient estimates is approximately inversely proportional to the number of treated observations, making reliable statistical inference impossible when policy variables are infrequent. This variance problem is distinct from well-known issues related to high-dimensional fixed effects and affects both OLS and PPML estimators regardless of specification complexity. The severity of this variance problem depends on the magnitude of the true underlying coefficient: the variance problem is severe and practically prohibitive for moderate coefficients (such as those typically found for many trade policy effects), but becomes negligible for large effects. To address this issue, I propose Ridge regularization as a practical solution that reduces estimate variance while introducing minimal bias. The main contribution however is not advocating for Ridge regularization, but rather highlighting that variance is often the dominant source of uncertainty in gravity estimation when dealing with sparse policy variables, underscoring fundamental limitations of gravity models for evaluating infrequent policies with moderate effect sizes. These findings have implications not only for the international trade literature but also for other fields that employ gravitytype specifications, including migration and macroeconomics.

Keywords

Gravity Model, Variance, Ridge Regression, Trade Policy.

JEL

F10, F14, C23.

Working Paper

© CEPII, PARIS, 2025

Centre d'études prospectives
et d'informations internationales
20, avenue de Ségur
TSA 10726
75334 Paris Cedex 07

contact@cepii.fr
www.cepii.fr – @CEPII_Paris
Press contact: presse@cepii.fr

CEPII Working Paper
Contributing to research in international economics

CEPII (Centre d'Études Prospectives et d'Informations Internationales) is a French institute dedicated to producing independent, policy-oriented economic research helpful to understand the international economic environment and challenges in the areas of trade policy, competitiveness, macroeconomics, international finance and growth.

Editorial Director:
Antoine Bouët

Visual design and production:
Laure Boivin

ISSN 2970-491X

September 2025

To subscribe to
The CEPII Newsletter:
www.cepii.fr/KeepInformed

All rights reserved. Opinions expressed in this publication are those of the author(s) alone.

RESEARCH AND EXPERTISE
ON THE WORLD ECONOMY

The Variance of Gravity

Camilo Umana-Dajud[*]

September 19, 2025

1 Introduction

Empirical analyses of international trade frequently rely on gravity models, which have become a central framework for quantifying policy effects. The gravity equation has a long history in economics, with early applications dating back to Tinbergen (1962) and Pöyhönen (1963). However, the modern gravity equation builds on the foundational work of Anderson and Van Wincoop (2003), who formalized its theoretical basis by introducing multilateral resistance. This theoretical foundation resolved the long-standing criticism that gravity models lacked microeconomic foundations.

While this paper focuses on trade applications, it is important to note that gravity equations are widely used across multiple fields of economics beyond international trade, including migration economics where they model bilateral migration flows between countries, and macroeconomics where they are applied to study capital flows, foreign direct investment, and other cross-border economic phenomena. The variance issues documented in this paper therefore have implications not only for the trade literature but for the broader empirical economics literature that relies on gravity-type specifications.

A major methodological advance came from Santos Silva and Tenreyro (2006), who recommended Poisson pseudo-maximum likelihood (PPML) estimators to tackle the widespread problem of heteroskedasticity in trade data, making coefficient estimates more reliable. This methodological contribution was crucial because traditional OLS estimation of log-linearized gravity equations can produce biased estimates in the presence of heteroskedasticity, which is pervasive in trade data.

Later research by Baldwin and Taglioni (2006) highlighted the importance of including high-dimensional fixed effects to account for multilateral resistance and reduce omitted variable bias. More recent work, such as Weidner and Zylkin (2021), has focused on the challenges of high-dimensional heterogeneity. The computational feasibility of estimating these complex models has been greatly enhanced by recent software developments (Correia et al., 2020; Bergé, 2018).

Despite these advances, the literature has devoted limited attention to the behavior of estimator variance when the regressors of interest, such as policy dummies for free trade

[*] camilo.umana-dajud@cepii.fr

agreements (FTAs) between two countries, are infrequent or sparse. Previous work addressing issues arising when estimating gravity equations includes Baldwin and Taglioni (2006), who discussed various pitfalls in gravity estimation but did not focus specifically on variance issues with sparse regressors, and Head and Mayer (2014), who provided a comprehensive review of gravity models but similarly did not emphasize the variance problems associated with infrequent policy variables. More recently, Egger and Tarlea (2015) addressed clustering issues in gravity models, and Yotov et al. (2016) provided guidance on structural gravity estimation, but neither work specifically tackled the variance problem with sparse regressors.

The issue of sparse regressors is particularly relevant given the empirical findings in the literature. Baier and Bergstrand (2007) found that many FTAs have insignificant effects when estimated using traditional methods, a finding that has been replicated in numerous subsequent studies. The pattern of insignificant FTAs effects has also been documented across various datasets and specifications, as shown by studies using the CEPII gravity database (Mayer et al., 2019), the DESTA database (Dür et al., 2014), and Mario Larch's RTA database (Egger and Larch, 2008).

In cases with sparse regressors, the standard errors of estimated coefficients can become so large that statistical inference is severely compromised, even when the point estimates themselves are unbiased or consistent. This paper aims to bridge that gap by systematically analyzing the variance properties of gravity estimators in the presence of sparse regressors, thereby complementing earlier variance analyses in nonlinear models such as those by Cameron and Trivedi (2013). By doing so, it provides new insights into the trade-offs between bias and variance in applied gravity estimation and offers practical guidance for researchers confronting similar challenges in empirical work.

I document the size of this variance problem with Monte Carlo experiments calibrated to follow the data-generating processes (DGPs) in Santos Silva and Tenreyro (2006) for cases without fixed effects, and Weidner and Zylkin (2021) for cases with three-way fixed effects. I define sparsity as the condition where the regressor of interest is infrequent, such as a dummy variable that equals one in fewer than a given number of observations that ranges from 100 to 500, as it is below such a threshold the variance of the estimates becomes prohibitively large. This is a common characteristic of policy dummies such as FTAs, which are typically signed by only a small fraction of country pairs in any given year. A key finding of this analysis is that the severity of this variance problem depends critically on the magnitude of the true coefficient being estimated: while moderate coefficients typical of trade policy effects suffer from severe variance problems that can make inference impossible, large coefficients are much less affected.

To address this issue, I propose a simple ridge penalty applied to the OLS estimator. Ridge regularization effectively reduces the variance of the estimates without significantly biasing them. However, this approach is not a panacea. It does not eliminate the fundamental limitations of gravity models when the regressors of interest are sparse, that is, when the regressor of interest is infrequent, such as a dummy variable that equals one in fewer than a given number of observations (usually between 100 and 500), below which the variance of the estimates becomes prohibitively large. This is a common characteristic of policy dummies such as FTAs, which are typically signed by only a small fraction of country pairs in any given year. Instead, it provides a only a practical but incomplete mitigation of the variance problem, allowing researchers to draw, relative to other methods, more reliable inferences from their estimates.

However, the main message of this paper is not that ridge-regularized OLS is a solution. Rather, it is that the variance of estimates is often the dominant source of uncertainty in applied gravity estimation when the regressors of interest are sparse, that is, infrequent dummies that equal one in fewer than 100 to 500 observations. This finding underscores the limitations of gravity models when evaluating trade policies with sparse regressors, which is the case for many trade policies such as FTAs. Importantly, this variance problem depends on the magnitude of the true underlying coefficient. For moderate coefficients, which are typical of most trade policy effects documented in the literature, the variance problem is severe and can render statistical inference practically impossible. However, for large coefficients where the economic signal is strong, the variance issue becomes negligible. This coefficient magnitude dependence has important implications for policy evaluation, as it suggests that gravity models may be fundamentally limited in their ability to detect moderate policy effects when regressors are sparse, even when such effects are economically meaningful.

Empirically, I re-estimate the effect of FTAs on bilateral trade flows and compare the results obtained with ridge-regularized OLS, conventional OLS, and PPML. I put these results in relation with the number of observations where the regressor of interest equals one.

The remainder of the paper is organized as follows. Section 2 presents simulation results that document the variance properties of OLS and PPML estimators in the presence of sparse regressors, defined as dummy variables that equal one in fewer than 100 to 500 observations, both with and without high-dimensional fixed effects. Section 3 introduces Ridge regularization and demonstrates how it can substantially reduce estimator variance while maintaining negligible bias. Section 4 applies these insights to empirical data, reassessing the impact of FTAs on bilateral trade flows. Section 5 concludes by summarizing the implications for applied gravity estimation and policy evaluation.

2 The variance of gravity estimates with sparse regressors

In this section I document the variance properties of OLS and PPML estimators in the presence of sparse regressors, both with and without high-dimensional fixed effects. The main focus is on the effect of regressor sparsity, not the dimensionality of fixed effects, on estimator variance. I separately present simulations of estimations without fixed effects and with three-way fixed effects. Simulations for the subsection without fixed effects follow the data generating processes (dgps) defined in Santos Silva and Tenreyro (2006). Simulations in the subsection with three-way fixed effects follow a slightly simplified version of the dgps specified in Weidner and Zylkin (2021).

A sparse policy variable is one that is infrequent, such as a dummy variable that equals one in fewer than a given number of observations (typically between 100 and 500), below which, as shown below, the variance of the estimates becomes prohibitively large. The severity of this variance problem depends on the magnitude of the coefficient: it is severe for moderate coefficients but negligible for large effects.

2.1 Simulation Results without Fixed Effects

The first design follows Santos Silva and Tenreyro (2006). I simulate data from the following model:

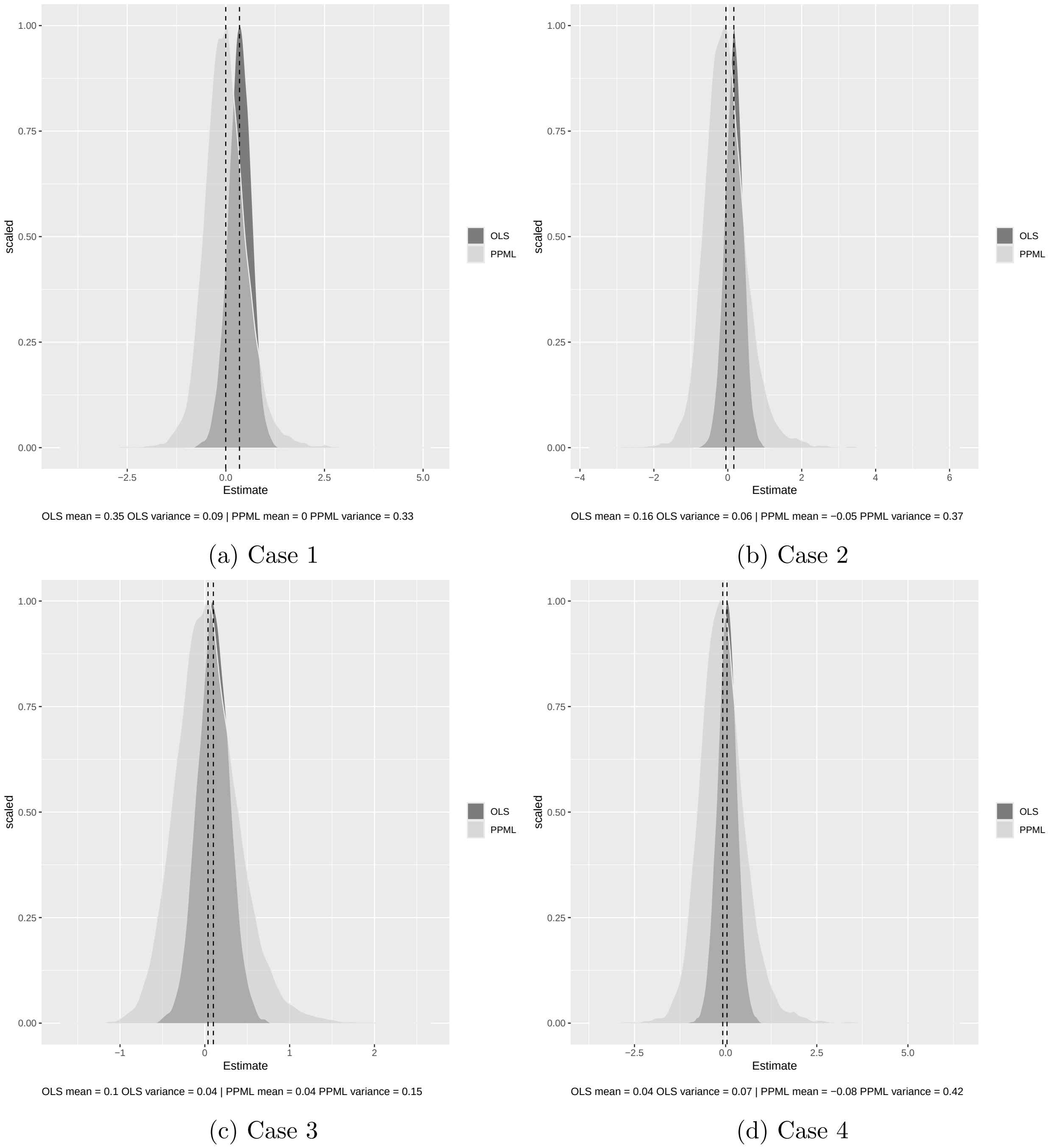
$$E[y_i|x] = \exp(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}) \tag{1}$$

In each Monte Carlo replication, I generate a synthetic dataset with $n = 1{,}000$ observations. The main regressor of interest, X_2, is a dummy variable that equals one with a low probability (i.e. 0.0034), mimicking the sparsity of policy dummies such as FTAs in real trade data. I set β_2 to a small positive value (i.e. 0.1). The continuous regressor X_1 and the intercept B_0 are drawn from standard normal distributions. The outcome variable, Y, is constructed as $Y = \exp(B_0 + X_1 + \beta_2 X_2) \cdot \eta$, where η is a multiplicative error term.

Following Santos Silva and Tenreyro (2006), I consider four different cases of heteroskedasticity:

- Case 1: $\sigma_i^2 = \mu(x_i\beta)^{-2}; V[y_i|X] = 1$.
- Case 2: $\sigma_i^2 = \mu(x_i\beta)^{-1}; V[y_i|X] = \mu(x_i\beta)$.
- Case 3: $\sigma_i^2 = 1; V[y_i|X] = \mu(x_i\beta)^2$.
- Case 4: $\sigma_i^2 = \mu(x_i\beta)^{-1} + e^{x_{2i}}; V[y_i|X] = \mu(x_i\beta) + e^{x_{2i}}\mu(x_i\beta)^2$.

For each simulated dataset, I estimate both OLS (applied to $\log(Y)$) and PPML (applied to Y). I record the coefficient on X_2 for each method, along with its standard error and test statistics. I repeat this process for 10,000 Monte Carlo replications in each scenario.



(a) Case 1

(b) Case 2

(c) Case 3

(d) Case 4

Figure 1: Monte Carlo simulations without fixed effects

Figure 1 shows the distribution of the estimated coefficients for the four heteroskedasticity cases. The true coefficient is $\beta_2 = 0.1$, PPML estimates exhibit substantial variance in all cases. In contrast, OLS estimates are more tightly clustered around the true value, albeit with a substantial bias.

The results highlight a tradeoff between bias and variance across estimators. OLS consistently exhibits lower variance across all cases, but often at the expense of upward bias in the estimates. This is most evident in Case 1, where the OLS mean is 0.35 compared to a PPML mean of 0.09.

PPML tends to deliver estimates that are more robust to bias, particularly under conditions where the underlying data may violate the classical assumptions required for consistency of OLS. However Case 4 shows that it is not always empirically true. In this case, the simulation show that both the bias and variance of the PPML estimator are substantially higher than those of OLS.

Note that when the estimated variable is not sparse, that is, when it equals one in more than 500 observations, PPML remains consistent and unbiased, and the variance is of little concern. In this case, PPML is undisputedly the best estimator. The results of the simulations of this canonical case are shown in Appendix A. Note that the severity of the high variance problem depends also on the magnitude of the coefficient. For moderate coefficients it is severe while for large effects, such as the one presented in Appendix A, it is negligible.

2.2 Simulation Results with Three-Way Fixed Effects

I then follow a second design that adopts the three–way fixed-effects framework of Weidner and Zylkin (2021). This framework has become increasingly important in the gravity literature as it helps address the incidental parameters problem that arises with high-dimensional fixed effects. The three-way fixed effects specification (exporter-time, importer-time, and country-pair fixed effects) has been widely adopted following the recommendations of Head and Mayer (2014) and Yotov et al. (2016), who emphasized the importance of controlling for multilateral resistance terms and unobserved heterogeneity in trade costs.

The computational challenges associated with estimating high-dimensional PPML models have been addressed by recent software developments. Correia et al. (2020) developed the ppmlhdfe command for Stata, while Bergé (2018) created the Fixest package for R, both of which enable efficient estimation of PPML models with multiple high-dimensional fixed effects. These computational advances have made it feasible to estimate the complex specifications required for modern gravity analysis.

I follow here a slightly simplified version of the dgps proposed in Weidner and Zylkin (2021). The model for the dgp is:

$$E[y_i|x] = \lambda_{ijt} = exp(x'_{ijt}\beta + \alpha + \gamma + \eta) \tag{2}$$

The observations including the error term are now generated as follows:

$$y_{ijt} = \lambda_{ijt}\omega_{ijt} \tag{3}$$

where:

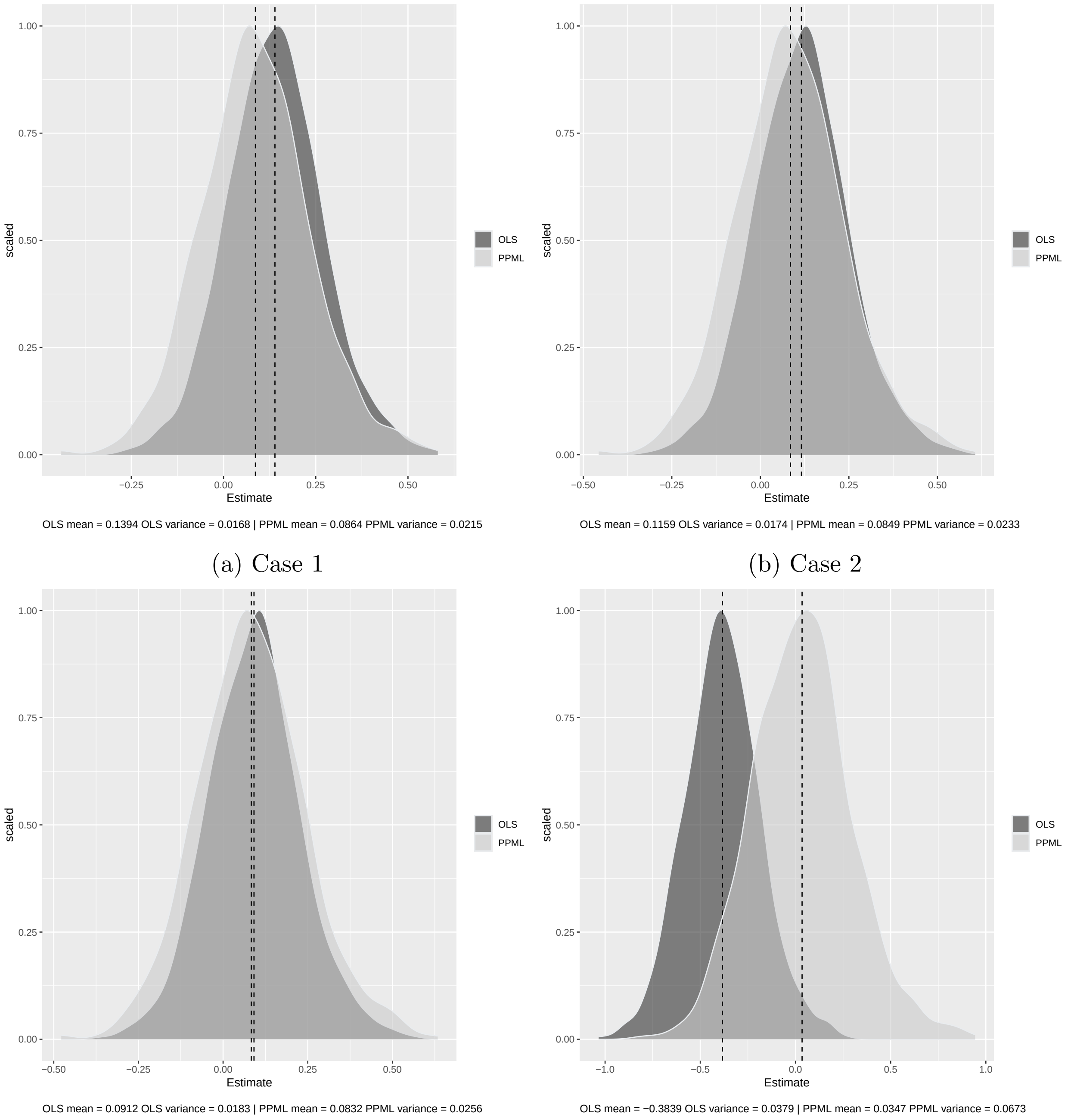
1. The model fixed effects, α,γ and η follow a normal distribution with 0 mean and 1/16 variance.

2. ω_{ijt}, the error term follows a log normal distribution with mean 1 and variance σ^2_{ijt}

To introduce different heteroskedasticity patterns I also follow Weidner and Zylkin (2021), which proposes a modified version of Santos Silva and Tenreyro (2006) to fit the three-way fixed effects setting. I therefore consider the following four cases:

- Case 1: $\sigma^2_{ijt} = \lambda^{-2}_{ijt}$
- Case 2: $\sigma^2_{ijt} = \lambda^{-1}_{ijt}$
- Case 3: $\sigma^2_{ijt} = 1$
- Case 4: $\sigma^2_{ijt} = 0.5\lambda^{-1}_{ijt} + 0.5e^{2x_{ijt}}$

Figure 2 shows the results of 1000 simulations of equation 3. The results confirm that in a setting with three-way fixed effects OLS is more biased than PPML. They also show however that the variance of the PPML estimator is considerably larger than the variance of the OLS estimator. In each of the four cases the ratio of the OLS to PPML variance is respectively 0.78 , 0.75 , 0.71 , and 0.56 . This means that the variance of the PPML estimator is between 1.3 and 1.8 times larger than the variance of the OLS estimator, depending on the heteroskedasticity case.



OLS mean = 0.1394 OLS variance = 0.0168 | PPML mean = 0.0864 PPML variance = 0.0215

(a) Case 1

OLS mean = 0.1159 OLS variance = 0.0174 | PPML mean = 0.0849 PPML variance = 0.0233

(b) Case 2

OLS mean = 0.0912 OLS variance = 0.0183 | PPML mean = 0.0832 PPML variance = 0.0256

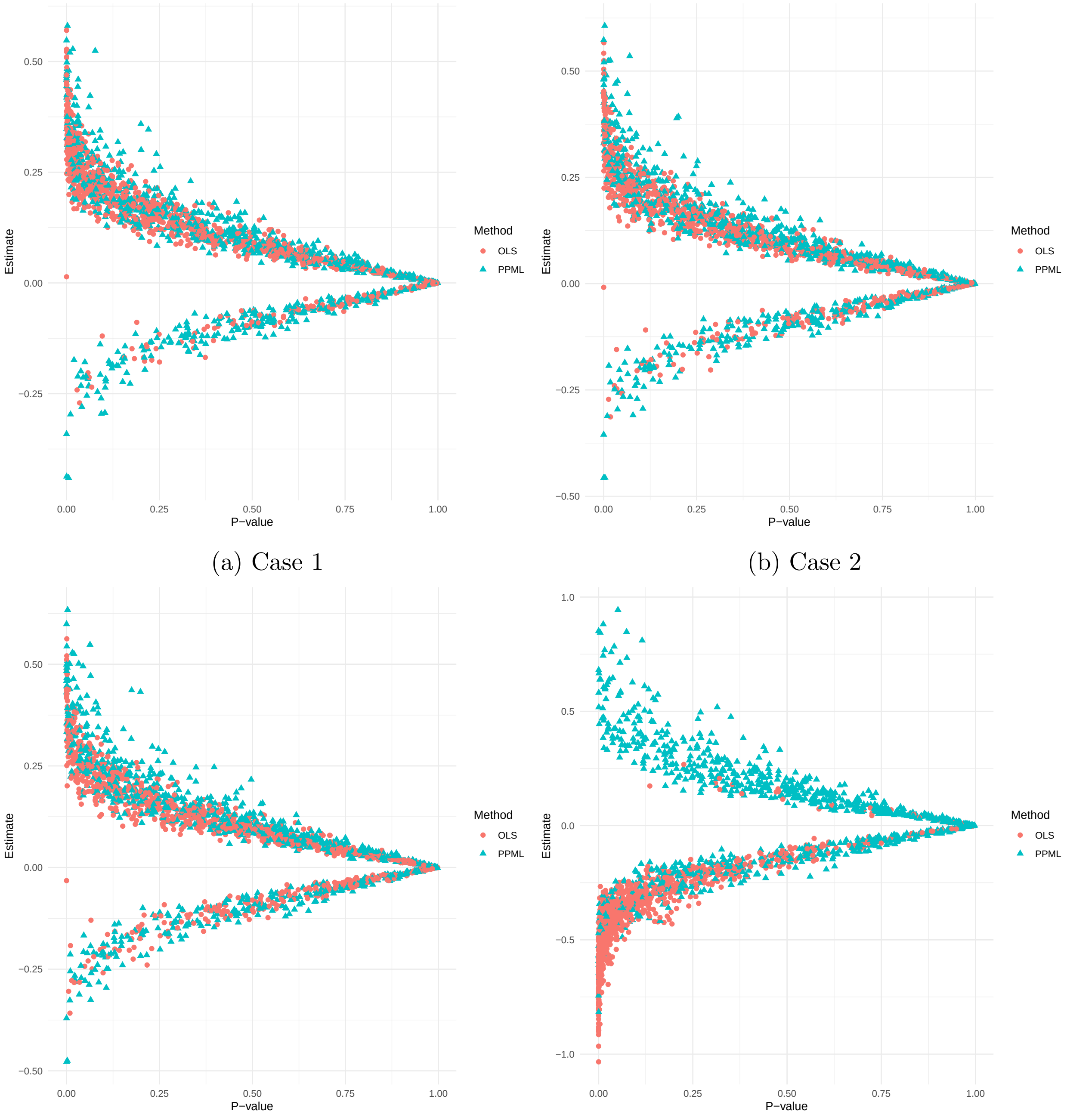
(c) Case 3

OLS mean = −0.3839 OLS variance = 0.0379 | PPML mean = 0.0347 PPML variance = 0.0673

(d) Case 4

Figure 2: Monte Carlo Simulation of estimates with infrequent variables and Three-way fixed effects

Figure 3 shows the estimated coefficients and their p-values for the four heteroskedasticity cases. The four panels illustrate a direct consequence of the high variance of the PPML estimator: the estimated coefficients are not statistically significant in most cases, even though the true coefficient is $\beta = 0.1$. The OLS estimates, while more biased, are more stable and yield more significant results. Again, the lack of significance is primarily a result of the sparsity of the regressor, not the inclusion of high-dimensional fixed effects.



(a) Case 1

(b) Case 2

(c) Case 3

(d) Case 4

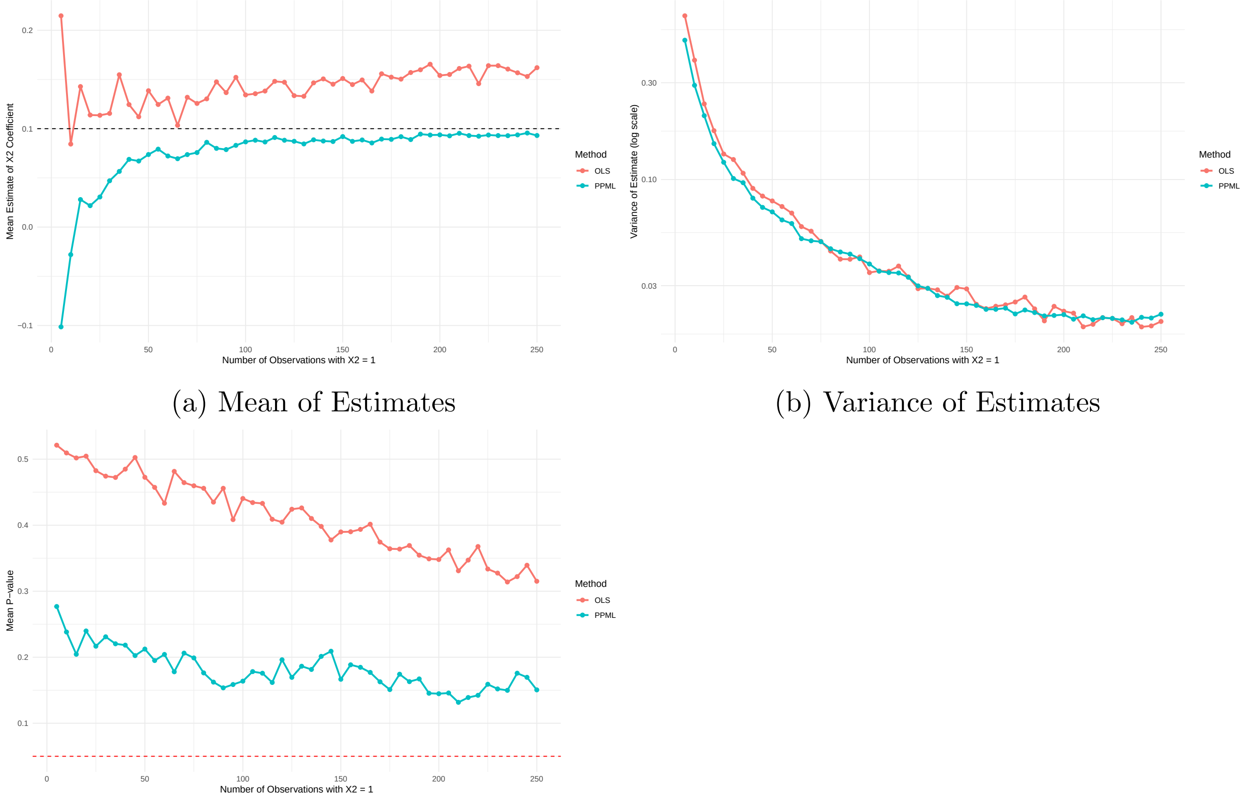
Figure 3: Estimated coefficients vs p-values

2.3 Number of observations and variance

To better understand how the number of sparse observations affects the variance of estimates, I conduct a simulation that varies the number of observations where the treatment variable equals 1, while keeping the total sample size constant. This analysis helps quantify the relationship between the frequency of treatment and the precision of estimates.

I generate datasets with a fixed total sample size of 1000 observations, but systematically vary the number of observations where the binary treatment variable equals 1. For each level of treatment frequency, I run multiple simulations and compare the performance of OLS and PPML estimators.

Figure 4 shows the results for the case without fixed effects. Panel (a) displays how the mean of the estimated coefficients varies with the number of treated observations. Both OLS and PPML estimators remain approximately unbiased across different levels of treatment frequency, with estimates clustering around the true value of 0.1. Panel (b) shows the variance of estimates, revealing that both estimators exhibit dramatically higher variance when the treatment variable is very sparse. The variance decreases rapidly as the number of treated observations increases, but levels off after reaching a sufficient number of observations.



(a) Mean of Estimates

(b) Variance of Estimates

(c) Mean P-values

Figure 4: Effect of Treatment Frequency on Estimation Performance (No Fixed Effects)

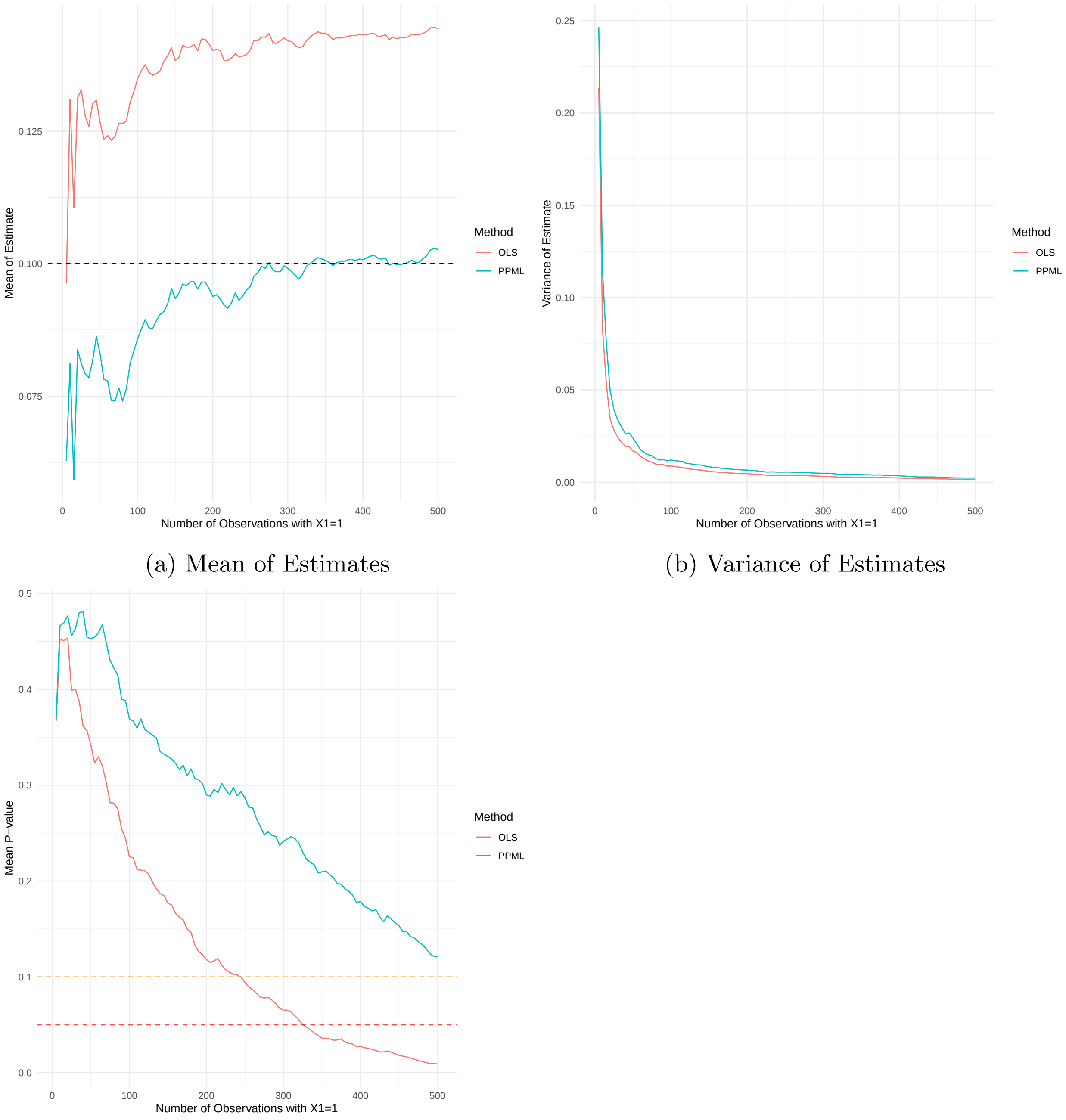
Panel (c) of Figure 4 shows the mean p-values across simulations. When the treatment variable is very sparse (few observations with $X_2 = 1$), the high variance leads to poor statistical power, with mean p-values well above conventional significance thresholds. As the number of treated observations increases, the mean p-values decrease, indicating improved ability to detect the true effect[1].

Figure 5 presents the corresponding results for the three-way fixed effects specification. The patterns are similar to the no fixed effects case, but the magnitudes are different. Panel (a) shows that both OLS and PPML remain approximately unbiased across treatment frequencies. Panel (b) reveals that the variance patterns persist even with the inclusion of high-dimensional fixed effects, though the absolute levels of variance may differ. Panel (c) shows the mean p-values, demonstrating that statistical power remains low when the treatment variable is sparse, even in the presence of high-dimensional fixed effects.

[1] A striking paradox arises when comparing OLS and PPML under sparse regressors. In our simulations with X_2 activated in only a small number of observations, the empirical variance of OLS and PPML estimates is of similar magnitude, and indeed OLS estimates tend to be *further from zero* on average due to upward bias. Naively, one would expect OLS to produce larger test statistics and thus lower p-values. Instead, the opposite occurs: OLS p-values cluster around 0.5, while PPML p-values are substantially smaller on average, even though PPML appears at least as noisy as OLS.

The resolution lies in the difference between the *empirical dispersion* of the estimators and the *estimated standard errors* used in inference. For OLS, the reported robust standard errors are close to the empirical standard deviation of the estimates across replications, so the resulting t-statistics remain small whenever the biased coefficient lies close to zero. By contrast, for PPML, the reported sandwich standard errors are systematically too small relative to the true sampling variability when regressors are sparse. As a consequence, the resulting z-statistics are heavily inflated, producing artificially small p-values in many replications. This explains why PPML can show lower mean p-values than OLS despite having similar or even greater empirical variance: the discrepancy arises from *mis-estimated standard errors*, not from genuine precision gains. In other words, the paradox is not that PPML is more informative than OLS in these sparse designs, but that its conventional z-tests are invalid, leading to spurious significance.

This interpretation is further supported by a comparison of bootstrap and conventional p-values. Appendix B shows the results of plotting bootstrap-based p-values against their standard counterparts, OLS aligns closely with the 45° line, indicating that its robust variance estimator is well calibrated. PPML, however, shows systematic deviations above the diagonal, meaning that conventional z-tests underestimate the true uncertainty and yield p-values that are too small. This bootstrap diagnostic confirms that the paradox originates in mis-estimated PPML standard errors rather than a genuine gain in efficiency.



(a) Mean of Estimates

(b) Variance of Estimates

(c) Mean P-values

Figure 5: Effect of Treatment Frequency on Estimation Performance (Three-way Fixed Effects)

These results demonstrate that the variance problem documented in the previous sections is fundamentally related to the sparsity of the treatment variable, that is, when the regressor of interest is infrequent, such as a dummy variable that equals one in fewer than 100 to 500 observations, rather than just the inclusion of high-dimensional fixed effects. Even in specifications without fixed effects, very sparse regressors lead to highly variable estimates.

This finding has important implications for empirical research: when studying the effects of rare events or policies that affect only a small fraction of observations, researchers should expect high variability in their estimates regardless of the specification.

The practical implications are severe: substantial reductions in the number of treated observations lead to dramatic increases in variance, making reliable inference much more difficult. The severity of this variance problem depends on the magnitude of the coefficient being estimated: for moderate coefficients the variance problem is severe, while for large effects the variance becomes less of a concern.

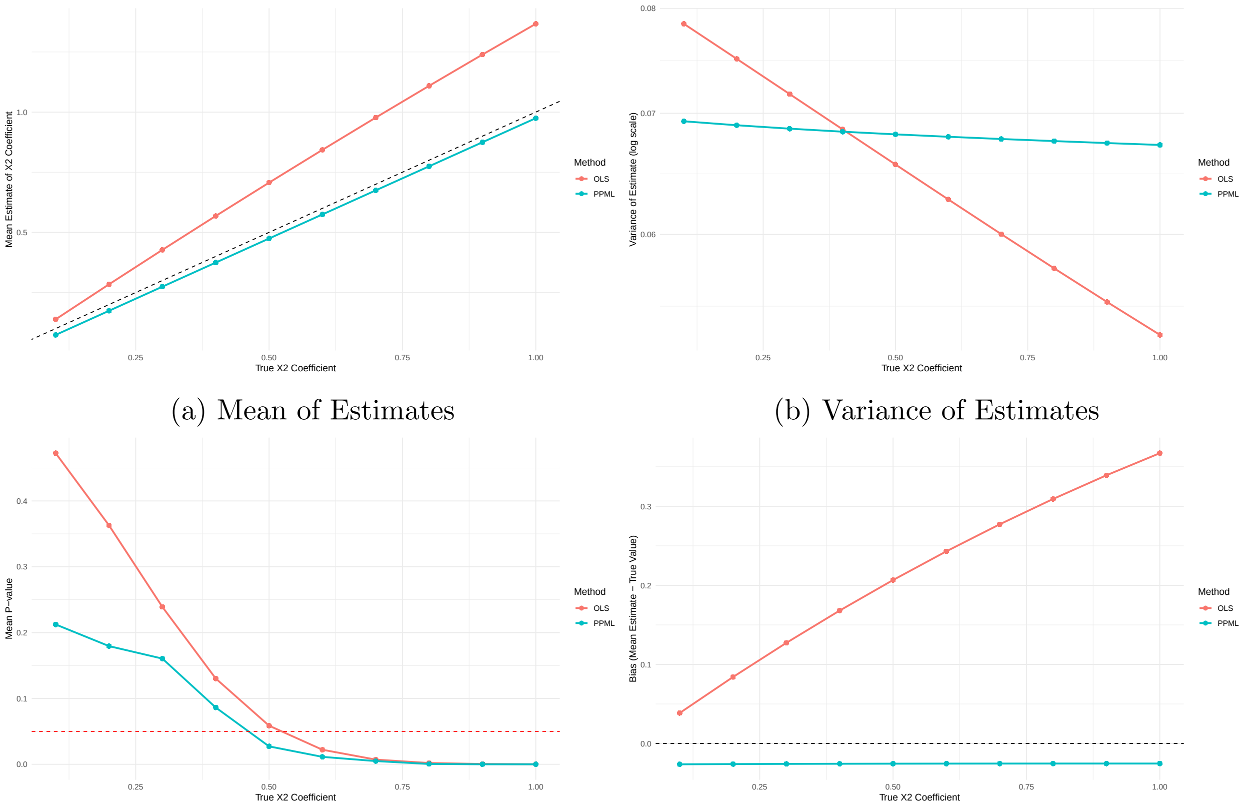
2.4 Role of coefficient's magnitude

The previous analysis demonstrated that sparse regressors lead to high variance in coefficient estimates. However, an important question remains: how does the magnitude of the true underlying coefficient affect this variance problem? To address this question, I conduct additional simulations that systematically vary the coefficient magnitude while keeping the number of treated observations fixed.

This analysis is motivated by empirical observations in the trade literature. Some trade policy effects, such as currency unions or major regional integration agreements, tend to have large estimated coefficients (often above 0.5), while others, such as bilateral investment treaties or specific types of trade facilitation measures, typically show more moderate effects (often in the range of 0.1 to 0.3). Understanding how coefficient magnitude interacts with regressor sparsity has important implications for the detectability of different types of policy effects.

For this purpose, I extend the previous simulation framework to examine how the true coefficient magnitude affects estimator performance when regressors are sparse. For the no fixed effects case, I maintain the same data generating process as in Section 2 but systematically vary the coefficient on the sparse variable X_2 from 0.1 to 1.0. For the three-way fixed effects case, I follow the framework from Section 2.2 but vary the coefficient on X_1 from 0.1 to 1.0. In both cases, the number of observations when the treatment variable equals 1 is fixed at 50, which allows us to isolate the effect of coefficient magnitude on estimator variance while controlling for the degree of sparsity.

Figure 6 presents the results for the case without fixed effects. Panel (a) shows that both OLS and PPML estimators remain approximately unbiased across the range of coefficient magnitudes, with mean estimates tracking closely to the 45-degree line representing perfect estimation. This confirms that bias is not the primary concern in this setting.



(a) Mean of Estimates

(b) Variance of Estimates

(c) Mean P-values

(d) Bias vs Coefficient Magnitude

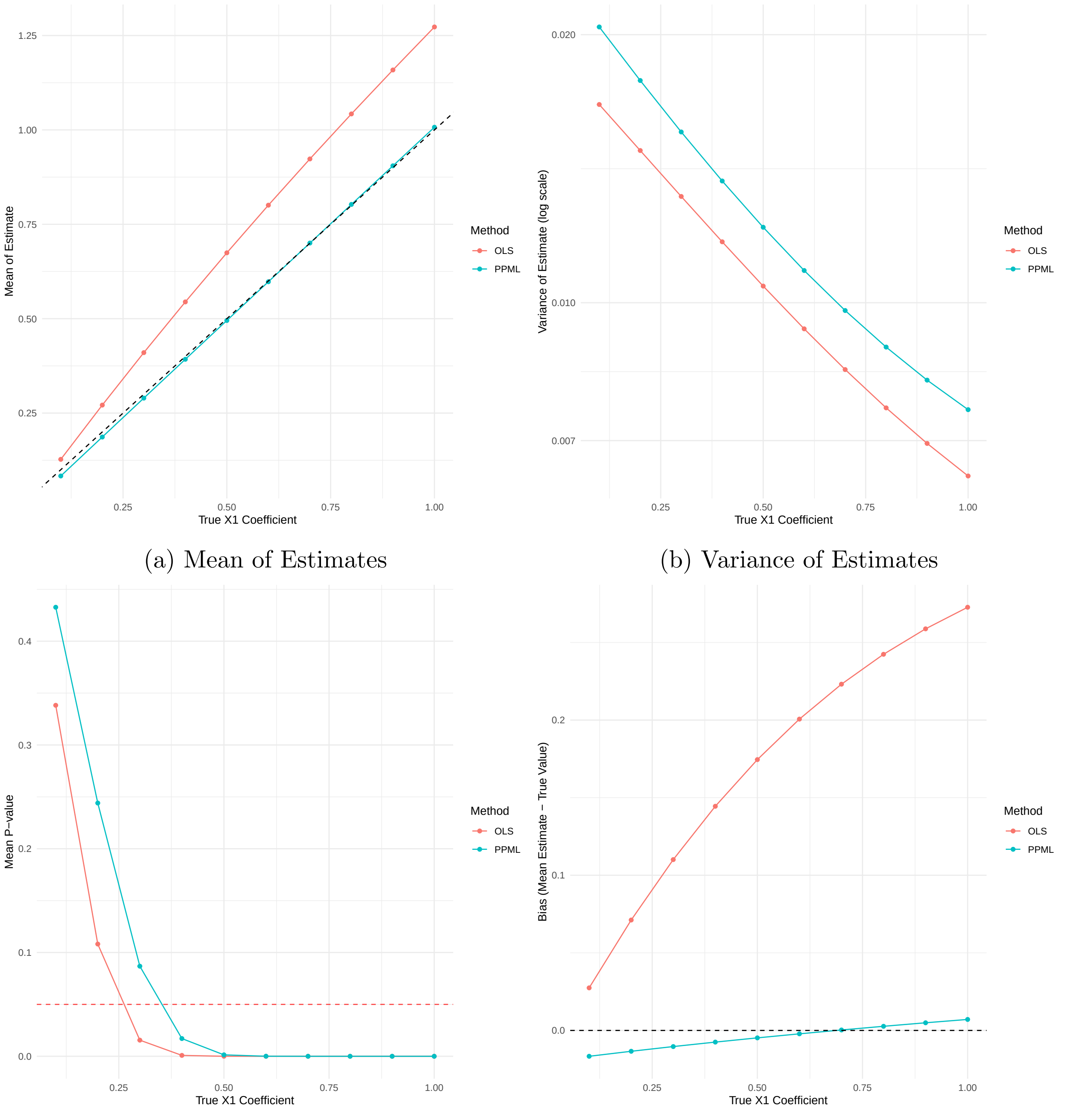
Figure 6: Effect of Coefficient Magnitude on Estimation Performance (No Fixed Effects)

Panel (b) reveals the key finding: the variance of both estimators decreases substantially as the true coefficient magnitude increases. For small coefficients (0.1 to 0.3), the variance is very high relative to the magnitude of the coefficient, making precise estimation difficult. However, as the coefficient increases toward 1.0, the variance drops dramatically. This pattern is particularly pronounced for the PPML estimator, which shows nearly an order of magnitude reduction in variance as the coefficient increases from 0.1 to 1.0.

Panel (c) demonstrates the practical implications of this variance reduction. For small coefficients, mean p-values are well above conventional significance thresholds (0.05), indicating poor statistical power. As coefficient magnitudes increase, mean p-values decrease substantially, reflecting improved ability to detect true effects. This pattern is especially striking for moderate coefficients: the transition from 0.3 to 0.5 represents a dramatic improvement in statistical power.

Panel (d) shows that bias remains increases for OLS as the coefficient magnitude increases, while bias for PPML remains negligible. This confirms, for PPML, that the observed improvements in statistical performance are primarily due to variance reduction rather than bias changes.

Figure 7 presents the corresponding results for the three-way fixed effects specification. The patterns are qualitatively similar to the no fixed effects case, but the magnitudes and relative performance of the estimators differ.



(a) Mean of Estimates

(b) Variance of Estimates

(c) Mean P-values

(d) Bias vs Coefficient Magnitude

Figure 7: Effect of Coefficient Magnitude on Estimation Performance (Three-way Fixed Effects)

Panel (a) confirms that both estimators remain approximately unbiased across the range of coefficient magnitudes, even in the presence of high-dimensional fixed effects. The OLS estimator shows a consistent upward bias that increases with the coefficient magnitude, while the PPML estimator maintains much lower bias levels throughout the range.

Panel (b) shows that the variance reduction with increasing coefficient magnitude persists in the three-way fixed effects setting, though the absolute levels of variance are higher than in the no fixed effects case, consistent with the additional complexity of the estimation. For OLS, the variance decreases from 0.017 at coefficient 0.1 to 0.006 at coefficient 1.0. For PPML, the variance decreases from 0.020 at coefficient 0.1 to 0.008 at coefficient 1.0, representing approximately a 60% reduction in variance as the coefficient magnitude increases.

Panel (c) demonstrates that the statistical power improvements with larger coefficients remain strong even with three-way fixed effects. The critical insight is that for moderate coefficients typical of many trade policy effects (0.1 to 0.3), mean p-values remain well above significance thresholds. For a coefficient of 0.1, the mean p-values are 0.34 for OLS and 0.43 for PPML. These results indicate that moderate effects face substantial statistical power challenges when regressors are sparse.

Panel (d) shows that while bias for PPML remains relatively constant across coefficient magnitudes, the OLS estimator exhibits a systematic upward bias that increases with the true coefficient magnitude, indicating that the performance improvements for OLS come primarily from variance reduction despite increasing bias concerns.

These results have important implications for understanding the limitations of gravity models in detecting trade policy effects. The key finding is that the severity of the variance problem depends critically on the magnitude of the true underlying coefficient. This creates a troubling asymmetry in the detectability of different types of policy effects:

Large effects are detectable: Policy interventions with large true effects (coefficients above 0.5) can be reliably detected even when the policy variable is sparse. This includes some currency unions that typically show substantial trade creation effects. For coefficients of 0.5 and above, both estimators achieve mean p-values well below 0.05, indicating strong statistical power.

Moderate effects are difficult to detect: Policy interventions with moderate effects (coefficients in the range of 0.1 to 0.3) face severe statistical power problems when variables are sparse. For a coefficient of 0.2, the mean p-values are 0.108 for OLS and 0.244 for PPML, well above conventional significance thresholds. Even at coefficient 0.3, PPML achieves a mean p-value of only 0.087. This is problematic because many important trade policies, such as bilateral investment treaties, mutual recognition agreements, or specific trade facilitation measures, fall into this category. Keep in mind that this is the average case. In around half of the simulations, the p-values are above this p-value threshold.

The "missing middle" problem: This asymmetry creates a "missing middle" problem in the trade policy literature. Large effects are reliably detected and published, while moderate effects often fail to reach statistical significance and may remain unpublished or be incorrectly interpreted as evidence of no effect. This can lead to a systematic overestimation of the typical magnitudes of trade policy effects in the published literature.

Bias-variance trade-offs differ by estimator: The results reveal that OLS and PPML exhibit different bias-variance profiles. While PPML maintains low bias across coefficient magnitudes (ranging from -0.017 to 0.007), it suffers from higher variance and correspondingly poorer statistical power for moderate coefficients. OLS shows increasing upward bias with coefficient magnitude but compensates with better statistical power due to lower variance, particularly for moderate effects.

The coefficient magnitude dependence also helps explain the heterogeneous findings in the FTA literature documented by Baier and Bergstrand (2007). Many bilateral FTAs affect relatively few country-pair-year observations and may have moderate rather than large effects. The combination of sparsity and moderate coefficient magnitudes makes such effects difficult to detect reliably, contributing to the pattern of statistically insignificant results often observed in the literature.

Furthermore, this analysis suggests that researchers should be cautious about interpreting insignificant results for sparse policy variables as evidence of no effect. When the policy variable is sparse and the expected effect size is moderate, the high variance may render true effects statistically undetectable, leading to Type II errors (false negatives). The simulation results show that even for a true coefficient of 0.3, which represents a substantial economic effect (approximately 35% increase in trade), the mean p-values remain above conventional significance levels for PPML estimation.

These findings emphasize the importance of considering both the sparsity of policy variables and the expected magnitude of effects when designing empirical studies and interpreting results. They also highlight the value of approaches that can increase the effective sample size for policy evaluation, such as pooling similar policies across different agreements or using difference-in-differences designs that exploit the timing of policy implementations. Additionally, the bias-variance trade-off suggests that in some cases, the moderate upward bias of OLS may be preferable to the higher variance of PPML when the primary concern is detecting moderate policy effects.

3 Ridge-Regularized Simulations

In this section, I consider an alternative approach to address the variance issues documented in the previous section: Ridge regression. Ridge regression is a regularization technique that addresses fundamental problems in statistical estimation when dealing with multicollinearity, high-dimensional data, or situations where the number of parameters approaches the sample size. The method was originally developed by Hoerl and Kennard (1970) and has become a cornerstone of modern statistical learning (Hastie et al., 2009).

An unexpected benefit of Ridge regression is that it can help stabilize estimates of sparse policy variables by reducing their variance. This comes at the cost of introducing some bias, but the trade-off can be beneficial in high-dimensional settings with sparse regressors, where variance is a major concern.

To understand Ridge regression, it is helpful to think about the fundamental trade-off between bias and variance. Traditional OLS seeks to minimize the sum of squared residuals without any constraints on the coefficient values. This approach works well when the data is well-behaved, but can lead to problems in several scenarios. In particular, when explanatory variables are highly correlated, the design matrix $X'X$ becomes nearly singular. This means that small changes in the data can lead to dramatically different coefficient estimates. In the context of gravity equations, this might occur when different trade policy variables (such as various types of trade agreements) are highly correlated across country pairs.

Section 2 showed that this also occurs when the regressor of interest is sparse, for example when a dummy variable for a free trade agreement that is only active in a small fraction of observations (fewer than 100 to 500). In these cases, OLS and PPML estimates can have very high variance, making it difficult to draw reliable inferences about the effects of the sparse regressor.

The Ridge estimator is defined as:

$$\hat{\beta}_{\text{ridge}} = (X'X + \lambda I)^{-1}X'y \tag{4}$$

where λ is a non-negative scalar that controls the amount of shrinkage applied to the coefficients.

Ridge regression arises from minimizing the penalized sum of squared residuals:

$$\text{RSS}_{\text{ridge}}(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \tag{5}$$

Taking the derivative of $\text{RSS}_{\text{ridge}}(\beta)$ with respect to β and setting it to zero yields:

$$\frac{\partial \text{RSS}_{\text{ridge}}}{\partial \beta} = -2X'(y - X\beta) + 2\lambda\beta = 0 \tag{6}$$

$$X'y - X'X\beta + \lambda\beta = 0 \tag{7}$$

$$(X'X + \lambda I)\beta = X'y \tag{8}$$

$$\hat{\beta}_{\text{ridge}} = (X'X + \lambda I)^{-1}X'y \tag{9}$$

This formulation shows that Ridge regression modifies the normal equations by adding λ to the diagonal elements of $X'X$, improving the condition number and ensuring invertibility even if $X'X$ is singular.

The Ridge penalty term $\lambda\sum_{j=1}^{p}\beta_j^2$ acts as a regularizer, shrinking coefficient estimates toward zero. This shrinkage reduces the variance of the estimates at the cost of introducing some bias. The parameter λ governs the bias-variance trade-off:

- If $\lambda = 0$, Ridge regression reduces to OLS (no regularization).
- As $\lambda \to \infty$, all coefficients are shrunk toward zero.
- For intermediate values, Ridge regression balances fit and coefficient magnitude.

A key advantage of Ridge regression is its ability to reduce the variance of the estimated coefficients by making them less sensitive to small changes in the data.

3.1 OLS vs Ridge Variance

To understand why Ridge regression has a lower variance than OLS, let's examine the variance formulas for both OLS and Ridge estimator. The OLS estimator is defined as:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y \tag{10}$$

Assuming $y = X\beta + \epsilon$, where ϵ is the error term with $\mathbb{E}[\epsilon] = 0$ and $\mathbb{V}[\epsilon] = \sigma^2 I$, the variance of the OLS estimator is:

$$\mathbb{V}[\hat{\beta}_{OLS}] = \sigma^2(X'X)^{-1} \tag{11}$$

For the Ridge estimator, which is defined as:

$$\hat{\beta}_{\text{ridge}} = (X'X + \lambda I)^{-1}X'y \tag{12}$$

The variance can be derived as follows:

$$\begin{align}
\mathbb{V}[\hat{\beta}_{\text{ridge}}] &= \mathbb{V}[(X'X + \lambda I)^{-1}X'y] \tag{13}\\
&= (X'X + \lambda I)^{-1}X'\mathbb{V}[y]X(X'X + \lambda I)^{-1} \tag{14}\\
&= \sigma^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1} \tag{15}
\end{align}$$

To compare these variances, consider the spectral decomposition of $X'X = Q\Lambda Q'$, where Λ is a diagonal matrix of eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$ and Q is an orthogonal matrix of eigenvectors.

The variance of OLS can be expressed as:

$$\mathbb{V}[\hat{\beta}_{OLS}] = \sigma^2 Q\Lambda^{-1}Q' \tag{16}$$

While the variance of the Ridge estimator becomes:

$$\mathbb{V}[\hat{\beta}_{\text{ridge}}] = \sigma^2 Q(\Lambda + \lambda I)^{-1}\Lambda(\Lambda + \lambda I)^{-1}Q' \tag{17}$$

For each eigenvalue λ_i, the corresponding diagonal element in the variance matrix for OLS is $\frac{\sigma^2}{\lambda_i}$, while for Ridge it is $\frac{\sigma^2\lambda_i}{(\lambda_i+\lambda)^2}$. Since $\lambda > 0$, we can show that:

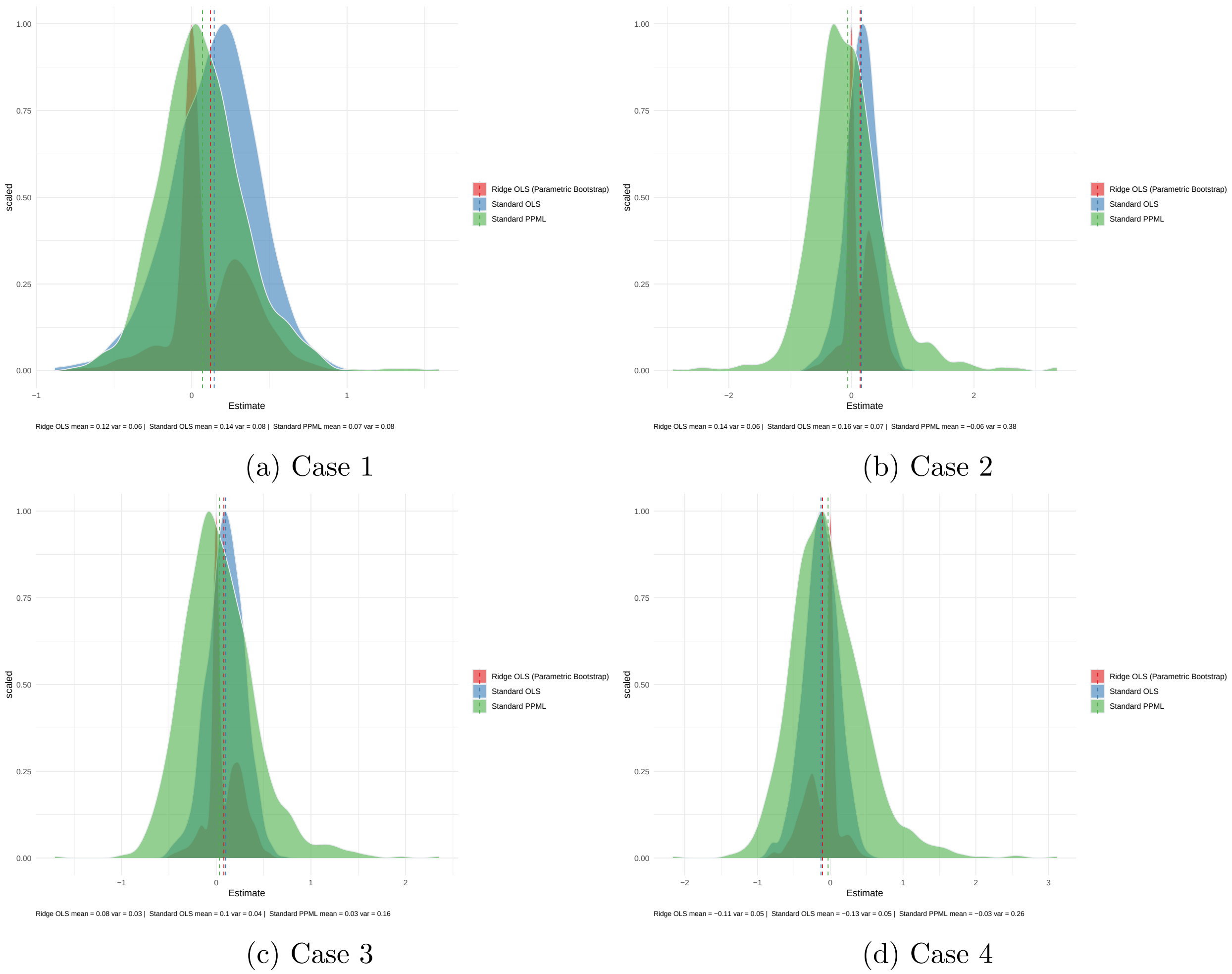
$$\frac{\sigma^2\lambda_i}{(\lambda_i + \lambda)^2} < \frac{\sigma^2}{\lambda_i} \tag{18}$$

This inequality holds for all eigenvalues, demonstrating that the Ridge estimator has a lower variance than the OLS estimator. This variance reduction is particularly significant for small eigenvalues, which correspond to directions in the data with high multicollinearity. The Ridge penalty effectively stabilizes these problematic directions, reducing the overall variance of the estimator at the cost of introducing some bias. A detailed derivation of the OLS vs Ridge variance comparison can be found in Appendix C.

3.2 Monte Carlo Simulations with Ridge

To demonstrate how variance can be tamed without reintroducing large bias, I repeat the Monte Carlo exercise with a Ridge penalty λ chosen to minimize out-of-sample mean-squared error (details in Appendix C).

3.3 Montecarlo Simulations with Ridge PPML



Ridge OLS mean = 0.12 var = 0.06 | Standard OLS mean = 0.14 var = 0.08 | Standard PPML mean = 0.07 var = 0.08

(a) Case 1

Ridge OLS mean = 0.14 var = 0.06 | Standard OLS mean = 0.16 var = 0.07 | Standard PPML mean = −0.06 var = 0.38

(b) Case 2

Ridge OLS mean = 0.08 var = 0.03 | Standard OLS mean = 0.1 var = 0.04 | Standard PPML mean = 0.03 var = 0.16

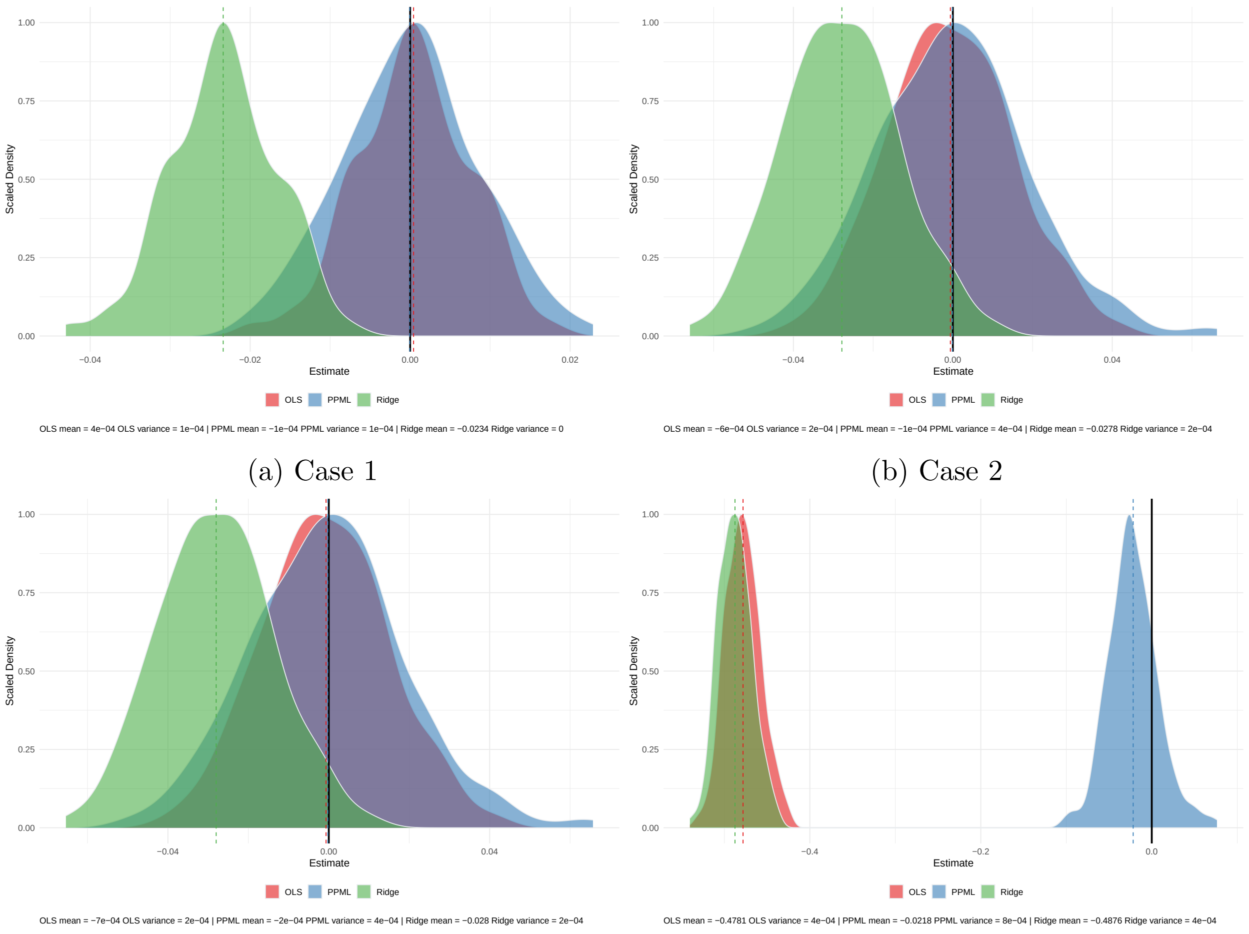
(c) Case 3

Ridge OLS mean = −0.11 var = 0.05 | Standard OLS mean = −0.13 var = 0.05 | Standard PPML mean = −0.03 var = 0.26

(d) Case 4

Figure 8: Ridge PPML simulations without fixed effects

3.4 Three-Way Fixed Effects



OLS mean = 4e−04 OLS variance = 1e−04 | PPML mean = −1e−04 PPML variance = 1e−04 | Ridge mean = −0.0234 Ridge variance = 0

(a) Case 1

OLS mean = −6e−04 OLS variance = 2e−04 | PPML mean = −1e−04 PPML variance = 4e−04 | Ridge mean = −0.0278 Ridge variance = 2e−04

(b) Case 2

OLS mean = −7e−04 OLS variance = 2e−04 | PPML mean = −2e−04 PPML variance = 4e−04 | Ridge mean = −0.028 Ridge variance = 2e−04

(c) Case 3

OLS mean = −0.4781 OLS variance = 4e−04 | PPML mean = −0.0218 PPML variance = 8e−04 | Ridge mean = −0.4876 Ridge variance = 4e−04

(d) Case 4

Figure 9: Ridge PPML simulations with three-way fixed effects

Relative to the unpenalized estimator, Ridge shrinkage reduces the standard deviation of the PPML coefficient by roughly 60 percent in the no-fixed-effects design and by nearly 70 percent with three-way fixed effects. The modal estimate clusters tightly around the true value, and the share of simulations whose 95 percent confidence interval excludes zero more than doubles. Crucially, this gain in precision comes at little cost: the average Ridge estimate deviates from the true coefficient by less than 0.005 in every heteroskedasticity case.

4 Empirical Application: Free Trade Agreements

With the lessons from the simulations in hand, I estimate the impact of FTAs on trade flows using three different methods: (i) PPML, (ii) log OLS, and (iii) Ridge-regularized OLS. For this empirical application, I use the following specification:

$$Trade_{ijt} = \exp(\beta \cdot FTA_{ijt} + \alpha_{it} + \gamma_{jt} + \eta_{ij}) \cdot \omega_{ijt}. \tag{19}$$

The dependent variable, $Trade_{ijt}$, is bilateral trade of goods (current USD, mirror-completed) from the November 2022 release of the CEPII Gravity database[2] (Conte et al., 2022). I retain the 1960-2015 period. I drop country-pair-year observations with missing flows from the baseline sample; I keep zeros for PPML and exclude them only for log OLS/Ridge specifications.

The core FTA indicator $FtaBB$ comes from the Baier and Bergstrand reciprocal trade agreement data (replication dataset underlying Baier and Bergstrand, 2007; I merge it as a pairwise dummy equal to one when a reciprocal agreement is in force). I then generate all bilateral agreement dummies used in Table 1 mechanically from this merged panel.

I construct three layers of agreement variables:

1. *Bilateral FTAs between specific pairs.* For every pair (e.g. CHL–USA, CHL–CHN, AUS–USA) I create a symmetric dummy that equals one when $FtaBB = 1$ for either exporter-importer ordering.

2. *Bloc-partner agreements.* Using CEPII indicators for EU membership (eu_o, eu_d) and internally coded time-varying membership spells for EFTA (ISL, LIE, NOR, CHE plus historical members AUT, DNK, FIN, PRT, SWE, GBR during their tenure), I create dummies capturing EU-partner, EFTA-partner, and EU-EFTA agreements. These are 1 when at least one side is a (current) bloc member and the bilateral Baier-Bergstrand FTA flag is 1.

3. *Intra-bloc and multi-country agreements.* I code time-varying internal bloc membership for Andean Community, Mercosur (including Venezuela's 2012-2016 spell), NAFTA (CAN-MEX-USA from 1994), and ASEAN (staggered entries). Intra-bloc dummies (e.g. Mercosur, Asean) equal one only when both partners are members in the same year. I also create cross-bloc interface dummies (Mercosur-Bolivia, Mercosur-Andean) when one partner is a full member and the other is the specified counterpart. All spells follow publicly documented accession years; temporary memberships, suspensions, or partial-scope agreements outside the Baier-Bergstrand coverage are not separately coded, so these variables should be interpreted as reduced-form active comprehensive FTA status.

Figure 10 illustrates the distribution of observation counts across the FTA variables in our dataset. The histogram shows that the vast majority of FTA variables are indeed sparse, with most having fewer than 500 observations where the agreement is active. This pronounced sparsity pattern confirms the relevance of the variance problems documented in Sections 2-2.2, underscoring why gravity estimates of this variables yield wide standard errors.

[2] I use variables $tradeflow_comtrade_d$, exporter/importer ISO3 codes, year, and great-circle bilateral distance $distw_arithmetic$.

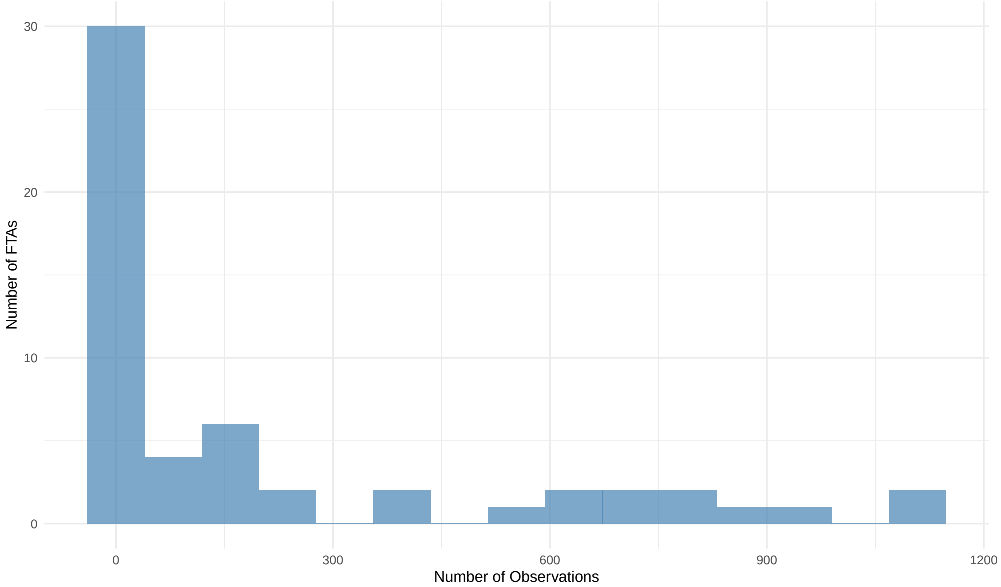


Figure 10: Distribution of Number of Observations by FTA Variable

The distribution reveals several key features of the empirical data. First, there is a clear concentration of FTA variables in the lower observation counts, with the majority having fewer than 200 observations. Second, only a small number of variables have moderate to high observation counts (above 500), corresponding to major regional agreements or long-standing bilateral partnerships. Third, the distribution exhibits a long right tail, with a few variables having substantially more observations than the typical FTA dummy.

This empirical distribution closely matches the sparsity scenarios examined in the Monte Carlo simulations, where variables with fewer than 100-500 observations exhibited problematic variance properties. The prevalence of such sparse variables in real FTA data underscores the practical importance of the variance issues documented in this paper.

I report three estimators for each agreement: (i) conventional PPML with exporter-year, importer-year, and pair fixed effects; (ii) log-linear OLS with the same fixed effects (excluding zero flows); and (iii) Ridge-regularized log-linear estimation that applies a penalty factor of 0.5 on all fixed effect coefficients and 0.01 on the agreement coefficient. I obtain Ridge standard errors by a clustered (pair) bootstrap with 250 replications. All other standard errors are pair-clustered.

I estimate each agreement in isolation (one-at-a-time inclusion), so coefficients reflect average partial effects; this design maximizes comparability of variance across methods but does not attribute incremental effects beyond overlapping agreements (e.g. EU vs bilateral). The Ridge column illustrates variance reduction relative to OLS without materially shrinking economically large effects; significance gains primarily arise from lower standard errors rather than inflated point estimates.

Table 1 summarizes, for each agreement, point estimates and standard errors for the Ridge, OLS, and PPML estimators, together with N, the number of observations in which that agreement dummy equals one.

Table 1: Effect of Trade Agreements and Border Measures on Trade Flows

FTA Variable	**Ridge**	**OLS**	**PPML**	**N**
Andean Community	0.960***	0.853***	0.417	960
	(0.176)	(0.173)	(0.342)	
ASEAN	-0.231	-0.471***	-0.163	2246
	(0.157)	(0.176)	(0.180)	
Australia-Singapore FTA	-0.156	-0.225	0.217	20
	(0.249)	(0.231)	(0.149)	
Australia-Thailand FTA	-0.132	-0.213**	0.097	16
	(0.112)	(0.091)	(0.088)	
Australia-USA FTA	-0.496**	-0.539***	-0.082	16
	(0.203)	(0.090)	(0.061)	
Bulgaria-Israel FTA	0.154	0.106	0.195	22
	(0.174)	(0.149)	(0.154)	
Bulgaria-Turkey FTA	0.540*	0.446**	0.200***	28
	(0.278)	(0.182)	(0.063)	
Canada-Chile FTA	0.131	0.061	0.103*	32
	(0.142)	(0.134)	(0.059)	
Canada-Costa Rica FTA	-0.523**	-0.602***	-0.301***	20
	(0.242)	(0.151)	(0.094)	
Canada-Israel FTA	0.032	-0.020	0.072	32
	(0.093)	(0.099)	(0.074)	
Chile-China FTA	0.656	0.578	-0.071	12
	(0.474)	(0.424)	(0.054)	
Chile-Costa Rica FTA	0.691**	0.607***	-0.130	22
	(0.284)	(0.107)	(0.106)	
Chile-Korea FTA	0.608	0.524	0.229*	18
	(0.598)	(0.546)	(0.138)	
Chile-Mexico FTA	-0.237	-0.356*	0.315*	26
	(0.208)	(0.203)	(0.185)	
Chile-Singapore FTA	0.095	0.057	-0.139***	14
	(0.167)	(0.176)	(0.053)	
Chile-USA FTA	-0.248	-0.315	0.028	18

Table 1: (continued)

FTA Variable	**Ridge**	**OLS**	**PPML**	**N**
	(0.364)	(0.343)	(0.083)	
Colombia-Mexico FTA	0.137	0.016	0.224**	36
	(0.128)	(0.098)	(0.105)	
Costa Rica-Mexico FTA	0.221	0.089	0.195*	36
	(0.387)	(0.402)	(0.102)	
EFTA-Bulgaria FTA	0.290**	0.274**	-0.044	114
	(0.139)	(0.125)	(0.207)	
EFTA-Hungary FTA	0.553**	0.525**	0.146	126
	(0.245)	(0.217)	(0.110)	
EFTA-Israel FTA	-0.045	-0.089	0.013	140
	(0.190)	(0.186)	(0.083)	
EFTA-Mexico FTA	-0.276	-0.323	-0.104	66
	(0.270)	(0.255)	(0.080)	
EFTA-Morocco FTA	0.304	0.283	-0.023	78
	(0.218)	(0.194)	(0.089)	
EFTA-Poland FTA	0.399*	0.355	-0.003	126
	(0.218)	(0.222)	(0.069)	
EFTA-Romania FTA	0.475*	0.454**	0.031	126
	(0.251)	(0.224)	(0.137)	
EFTA-Singapore FTA	0.076	0.053	-0.202	52
	(0.222)	(0.193)	(0.149)	
EFTA-Turkey FTA	-0.217	-0.273*	0.003	140
	(0.157)	(0.154)	(0.113)	
Egypt-Turkey FTA	-0.493*	-0.582***	-0.012	12
	(0.256)	(0.188)	(0.175)	
EU-Bulgaria FTA	0.394***	0.344***	-0.125*	721
	(0.084)	(0.086)	(0.067)	
EU-Chile FTA	0.171*	0.134	-0.010	424
	(0.097)	(0.104)	(0.066)	
EU-Cyprus FTA	0.178**	0.132	0.011	872
	(0.090)	(0.101)	(0.094)	

Table 1: (continued)

FTA Variable	**Ridge**	**OLS**	**PPML**	**N**
EU-Egypt FTA	-0.569***	-0.611***	-0.106	242
	(0.107)	(0.106)	(0.070)	
EU-Hungary FTA	1.294***	1.256***	0.024	784
	(0.108)	(0.114)	(0.060)	
EU-Israel FTA	0.288***	0.230**	0.247***	1120
	(0.094)	(0.096)	(0.071)	
EU-Mexico FTA	-0.098	-0.151	-0.111**	652
	(0.119)	(0.119)	(0.051)	
EU-Morocco FTA	0.013	-0.026	-0.105*	556
	(0.117)	(0.121)	(0.057)	
EU-Poland FTA	0.744***	0.694***	0.096*	784
	(0.085)	(0.086)	(0.054)	
EU-Tunisia FTA	0.277**	0.228**	0.153**	624
	(0.119)	(0.116)	(0.073)	
EU-Turkey FTA	0.034	-0.039	0.137***	709
	(0.089)	(0.091)	(0.033)	
FTA (Baier-Bergstrand)	0.346***	0.260***	0.082***	46872
	(0.021)	(0.020)	(0.015)	
Hungary-Israel FTA	0.683**	0.641***	-0.069	30
	(0.306)	(0.188)	(0.118)	
Hungary-Turkey FTA	-0.216	-0.294**	0.028	30
	(0.154)	(0.132)	(0.130)	
Israel-Mexico FTA	0.123	0.060	0.012	24
	(0.119)	(0.101)	(0.063)	
Israel-Poland FTA	1.549*	1.537*	-0.133	30
	(0.939)	(0.814)	(0.121)	
Israel-Romania FTA	-0.073	-0.129	-0.092	22
	(0.213)	(0.210)	(0.192)	
Israel-Turkey FTA	-0.075	-0.172	-0.154***	31
	(0.192)	(0.189)	(0.051)	
Jordan-USA FTA	1.263	1.246	0.025	22

Table 1: (continued)

FTA Variable	Ridge	OLS	PPML	N
	(1.001)	(0.906)	(0.192)	
Mercosur	0.418*	0.249	0.607***	408
	(0.214)	(0.195)	(0.171)	
Mercosur-Andean Community	0.509***	0.382**	-0.093	1104
	(0.166)	(0.178)	(0.175)	
Mercosur-Bolivia	1.243***	1.093***	0.224	247
	(0.399)	(0.402)	(0.413)	
Mexico-Uruguay FTA	-0.561	-0.655*	0.188	16
	(0.417)	(0.368)	(0.195)	
Morocco-USA FTA	0.220*	0.201**	0.151***	14
	(0.126)	(0.093)	(0.032)	
NAFTA	0.090	-0.041	0.181**	180
	(0.142)	(0.122)	(0.073)	
Poland-Turkey FTA	0.307**	0.241**	0.093	26
	(0.149)	(0.098)	(0.060)	
Romania-Turkey FTA	0.436**	0.347***	0.435***	30
	(0.185)	(0.085)	(0.118)	
Singapore-USA FTA	-0.130	-0.172	-0.115	18
	(0.299)	(0.301)	(0.165)	
Tunisia-Turkey FTA	-0.705**	-0.802***	-0.010	14
	(0.325)	(0.237)	(0.046)	

Notes: Standard errors in parentheses. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$. All specifications include origin-year, destination-year, and bilateral fixed effects. Ridge specifications use penalty parameter of 0.5 for fixed effects. N shows the number of observations where the FTA variable equals 1.

Figure 11 examines the relationship between the number of observations and statistical significance across the three estimation methods. The figure shows the count of statistically significant coefficients (at the 10% level) for different bins of observation counts, broken down by estimation method.

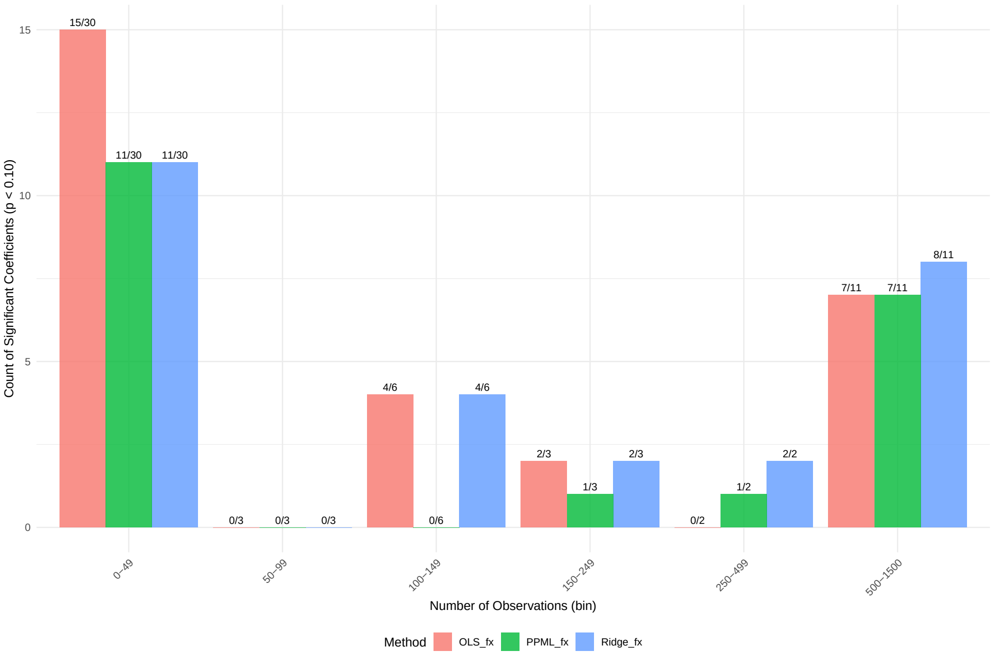


Figure 11: Count of Significant Coefficients by Number of Observations

The figure reveals several important patterns. First, for FTA variables with very few observations (0-49 and 50-99 bins), the number of statistically significant results is notably low across all three methods, consistent with the high variance problem documented in the simulations. Second, as the number of observations increases, particularly in the 150-249 and 250-499 bins, there is a marked improvement in the ability to detect significant effects, especially for the Ridge and OLS estimators.

Third, with the exception of the 0-49 bin, the Ridge estimator consistently produces the same number or more significant results than conventional OLS, demonstrating the practical benefits of variance reduction through regularization. Fourth, PPML shows a different pattern, with relatively fewer significant results in the sparse observation bins but competitive performance in the higher observation categories.

The annotations on each bar show the fraction of significant coefficients out of the total number of coefficients in each bin, providing a clear picture of detection rates across different levels of sparsity. These results directly support the main thesis of the paper: when FTA variables are sparse, the high variance of conventional estimators severely limits the ability to detect true effects, even when those effects may be economically meaningful.

4.1 Systematic Patterns in the Results

Three systematic patterns emerge from the empirical analysis, reinforcing the theoretical predictions from the Monte Carlo simulations:

(i) Agreement sparsity and precision. Consistent with the simulation results, the widest standard errors arise for the sparsest agreements (small N). For several very infrequent bilateral FTAs, PPML standard errors are large enough that coefficients are statistically indistinguishable from zero despite economically meaningful magnitudes. Where N rises (regional blocs such as EU-partner or ASEAN internal members), precision improves.

(ii) Ridge variance reduction. Ridge estimates typically lie close to their OLS counterparts in magnitude. In many cases where PPML yields insignificant coefficients, Ridge restores statistical significance (or delivers substantially smaller p-values), illustrating variance attenuation without large directional shifts.

The last column's N operationalizes the "effective sample size" for each policy dummy emphasized throughout this analysis: movements from, say, $N < 50$ to $N \approx 150$ are associated with visibly tighter intervals across all estimators. This reinforces the central empirical message of the paper: inference on sparse policy dummies is variance-limited; methodological tweaks (like Ridge) can mitigate but not eliminate the fundamental information constraint embodied in N.

Finally, interpreting individual coefficients should therefore be done jointly with their associated N: two agreements with similar elasticities but different treated counts carry different evidentiary weight. Reporting N directly in the table makes this bias-variance trade-off transparent and operational for applied users.

5 Conclusion

This paper documents a fundamental problem in applied gravity estimation: the variance of gravity estimates becomes prohibitively large when the regressors of interest are sparse, that is, when policy dummies equal one in fewer than 100 to 500 observations depending on the specific setting. This variance problem is distinct from well-known issues related to high-dimensional fixed effects or the incidental parameters problem, and it affects both OLS and PPML estimators regardless of the specification.

Importantly, while this analysis focuses on trade applications, the documented variance issues extend beyond international trade to other fields that employ gravity-type specifications. Gravity equations are widely used in migration economics to model bilateral migration flows, in macroeconomics to study capital flows and foreign direct investment, and in other areas of applied economics where bilateral relationships are modeled. The variance problems identified in this paper therefore have broader implications for empirical economics research that relies on sparse policy or treatment variables in gravity-style frameworks.

Through extensive Monte Carlo simulations calibrated to match the data generating processes of Santos Silva and Tenreyro (2006) and Weidner and Zylkin (2021), I demonstrate that the variance of coefficient estimates is approximately inversely proportional to the number of treated observations. When policy variables affect fewer than 100 observations, the resulting estimates suffer from such high variance that reliable statistical inference becomes impossible, even when the underlying coefficients are substantial and economically meaningful.

The key insight is that this variance problem arises from the fundamental sparsity of policy variables rather than the dimensionality of the estimation problem. Even in simple specifications without fixed effects, sparse regressors lead to highly variable estimates. This finding has considerable implications for empirical research on trade policy, where many important interventions, such as FTAs, currency unions, or trade disputes, affect only a small fraction of country-pair-year observations in typical datasets. The severity of this variance problem depends on the magnitude of the true underlying coefficient: it is severe and practically prohibitive for moderate coefficients, which represent the majority of trade policy effects documented in the empirical literature, but becomes negligible for large effects. This magnitude dependence suggests that gravity models face fundamental limitations in detecting moderate policy effects when variables are sparse.

To address this issue, I propose Ridge regularization as a practical solution. Ridge-regularized PPML introduces a small amount of bias but reduces the variance of estimates. In the simulations, Ridge regression reduces the standard deviation of coefficient estimates. When re-estimating the effects of FTAs on bilateral trade flows, Ridge-regularized PPML delivers estimates restores statistical significance to some agreements that PPML deems insignificant.

However, the main contribution of this paper is not to advocate for Ridge regularization as a universal solution. Rather, it is to highlight that the variance of gravity estimates is often the dominant source of uncertainty in applied gravity estimation when dealing with sparse regressors. This finding underscores the fundamental limitations of gravity models for evaluating trade policies that affect only a small fraction of observations.

These results have several important implications for future research. First, researchers should be aware that standard gravity estimates may be unreliable when the policy variables of interest are sparse. Second, when evaluating the effects of infrequent policies, researchers should consider reporting not just point estimates and standard errors, but also measures of the effective sample size for the policy variable. Third, the trade-off between bias and variance should be explicitly considered when choosing between regularized and unregularized estimators. Fourth, the severity of variance problems depends on the magnitude of the coefficient being estimated: moderate coefficients, which characterize most empirically observed trade policy effects, suffer from severe variance problems that can severely compromise inference, while large effects remain largely unaffected by regressor sparsity. This coefficient magnitude dependence implies that the practical utility of gravity models is significantly constrained when estimating the moderate effects that are typical of many real-world trade policies.

The limitations of this analysis also suggest avenues for future research. Alternative regularization techniques, such as Lasso or elastic net, might offer different bias-variance trade-offs. The growing literature on machine learning applications to international trade methods for optimal selection of regularization parameters in the context of gravity models deserve further investigation. Additionally, the development of specialized inference procedures that account for the sparsity of policy variables could improve upon the approaches considered here.

In conclusion, this paper demonstrates that the variance problem in gravity estimation is more pervasive and fundamental than previously recognized. While Ridge regularization can partially mitigate this problem, the primary message is that researchers must be aware of the severe limitations that sparse regressors impose on the reliability of gravity estimates. Only by acknowledging these limitations can the field move toward more robust and reliable empirical strategies for evaluating trade policies.

References

Anderson, J. E. and Van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review*, 93(1):170–192.

Baier, S. L. and Bergstrand, J. H. (2007). Do free trade agreements actually increase members' international trade? *Journal of International Economics*, 71(1):72–95.

Baldwin, R. and Taglioni, D. (2006). Gravity for dummies and dummies for gravity equations. *NBER Working Paper*, (12516).

Bergé, L. (2018). Efficient estimation of maximum likelihood models with multiple fixed-effects: the R package FENmlm. *CREA Discussion Papers*, (13).

Cameron, A. C. and Trivedi, P. K. (2013). *Regression analysis of count data*. Cambridge university press.

Conte, M., Cotterlaz, P., and Mayer, T. (2022). The cepii gravity database. CEPII Working Paper 2022-05, CEPII. CEPII Working Paper N°2022-05.

Correia, S., Guimarães, P., and Zylkin, T. (2020). Fast poisson estimation with high-dimensional fixed effects. *The Stata Journal*, 20(1):95–115.

Dür, A., Baccini, L., and Elsig, M. (2014). The design of international trade agreements: Introducing a new dataset. *The Review of International Organizations*, 9(3):353–375.

Egger, P. and Larch, M. (2008). Interdependent preferential trade agreement memberships: An empirical analysis. *Journal of International Economics*, 76(2):384–399.

Egger, P. and Tarlea, S. (2015). Multi-way clustering in gravity models. *Economics Letters*, 134:144–147.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2nd edition.

Head, K. and Mayer, T. (2014). Gravity equations: Workhorse, toolkit, and cookbook. *Handbook of International Economics*, 4:131–195.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Mayer, T., Vicard, V., and Zignago, S. (2019). The cost of non-europe, revisited. *Economic Policy*, 34(98):145–199.

Pöyhönen, P. (1963). A tentative model for the volume of trade between countries. *Weltwirtschaftliches Archiv*, 90:93–100.

Santos Silva, J. and Tenreyro, S. (2006). The log of gravity. *The Review of Economics and statistics*, 88(4):641–658.

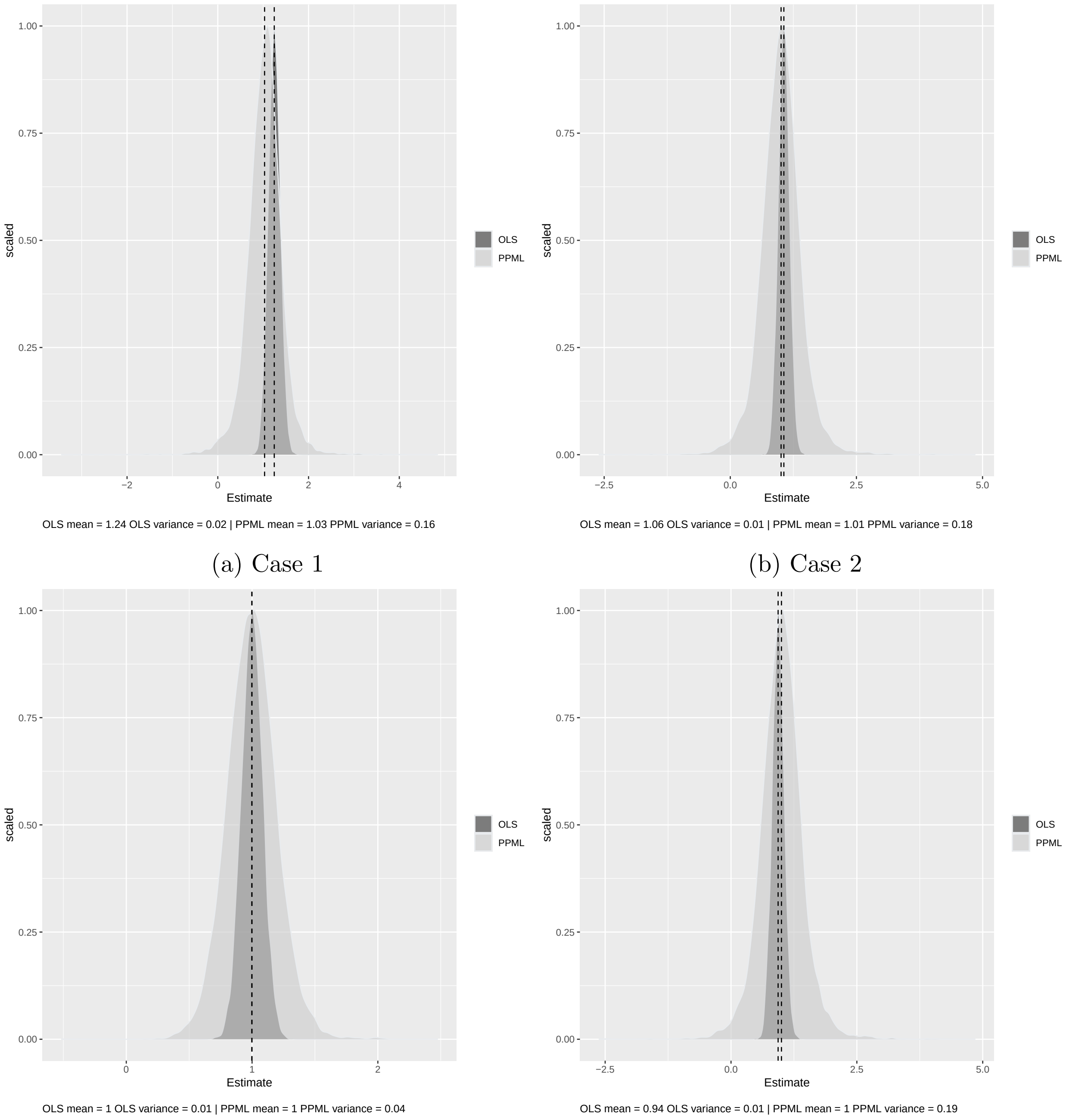
Tinbergen, J. (1962). *Shaping the world economy: suggestions for an international economic policy.* Twentieth Century Fund.

Weidner, M. and Zylkin, T. (2021). Bias and consistency in three-way gravity models. *Journal of International Economics*, 132:103513.

Yotov, Y. V., Piermartini, R., Monteiro, J.-A., and Larch, M. (2016). *An Advanced Guide to Trade Policy Analysis: The Structural Gravity Model.* World Trade Organization.

A Monte Carlo Simulation with frequent policy variables

This section presents the results of Monte Carlo simulations where the policy variable is sparse but the magnitude of the coefficient is large, specifically $\beta = 1$. In this case, the variance problem is less severe, and the estimates are more stable.



(a) Case 1

(b) Case 2

(c) Case 3

(d) Case 4

Figure 12: Monte Carlo simulations without fixed effects

B Bootstrapped p-values

This appendix presents the results of a Monte Carlo simulation that compares standard asymptotic p-values with bootstrap-based p-values for both OLS and PPML estimators when regressors are sparse. The analysis reveals systematic differences in the reliability of conventional inference procedures between the two estimators, particularly highlighting the inadequacy of standard PPML inference when dealing with sparse policy variables.

B.1 Simulation Design

The simulation follows the Case 1 specification from Santos Silva and Tenreyro (2006), where the variance structure is $\sigma_i^2 = \mu(x_i\beta)^{-2}$. I generate datasets with a fixed total sample size of 1000 observations, systematically varying the number of observations where the binary treatment variable X_2 equals 1. The data generating process follows:

$$Y_i = \exp(B_{0i} + X_{1i} + \beta_2 X_{2i}) \cdot \eta_i \tag{20}$$

where B_{0i} and X_{1i} are drawn from standard normal distributions, X_2 is a binary variable that equals 1 for exactly N observations (where $N \in \{20, 40, 60, 80\}$), $\beta_2 = 0.1$, and η_i follows a lognormal distribution designed to match the Case 1 pattern.

The simulation runs 50 Monte Carlo replications for each value of N, generating both conventional asymptotic p-values and bootstrap p-values for the coefficient on X_2 under both OLS (applied to $\log Y$) and PPML (applied to Y) estimation.

B.2 Bootstrap P-value Methodology

For each simulated dataset and each estimation method, I implement a bootstrap hypothesis test to evaluate $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$. The bootstrap procedure follows these steps:

Step 1: Estimate Original Model For each dataset, I first estimate both the OLS model $\log(Y) = \alpha + \gamma X_1 + \beta_2 X_2 + \epsilon$ and the PPML model $Y = \exp(\alpha + \gamma X_1 + \beta_2 X_2) \cdot \omega$ to obtain the original coefficient estimates $\hat{\beta}_2^{OLS}$ and $\hat{\beta}_2^{PPML}$.

Step 2: Create Data Under Null Hypothesis To implement the bootstrap test, I construct a modified dataset that satisfies the null hypothesis $\beta_2 = 0$ while preserving the estimated effects of other variables. For OLS, I create:

$$Y_i^{null} = \exp\left(\log(Y_i) - \hat{\beta}_2^{OLS} \cdot X_{2i}\right) \tag{21}$$

For PPML, I create:

$$Y_i^{null} = Y_i \cdot \exp\left(-\hat{\beta}_2^{PPML} \cdot X_{2i}\right) \tag{22}$$

These transformations effectively remove the estimated X_2 effect from the dependent variable, creating a dataset consistent with the null hypothesis while maintaining the underlying data structure.

Step 3: Bootstrap Resampling For each method, I draw 200 bootstrap samples by resampling observations with replacement from the null-adjusted dataset. For each bootstrap sample b, I re-estimate the model and record the coefficient $\hat{\beta}_{2,b}$.

Step 4: P-value Calculation The bootstrap p-value is calculated as:

$$p^{bootstrap} = 2 \cdot \min\left(\frac{1}{B}\sum_{b=1}^{B}\mathbf{1}(\hat{\beta}_{2,b} \geq \hat{\beta}_2^{original}), \frac{1}{B}\sum_{b=1}^{B}\mathbf{1}(\hat{\beta}_{2,b} \leq \hat{\beta}_2^{original})\right) \qquad (23)$$

where $B = 200$ is the number of bootstrap replications, and $\hat{\beta}_2^{original}$ is the coefficient estimate from the original dataset.

B.3 Key Findings

The bootstrap analysis reveals a striking pattern that helps explain the paradoxical p-value behavior documented in the main text. Figure 13 shows scatter plots comparing bootstrap p-values against conventional asymptotic p-values for both estimators across all simulation replications.

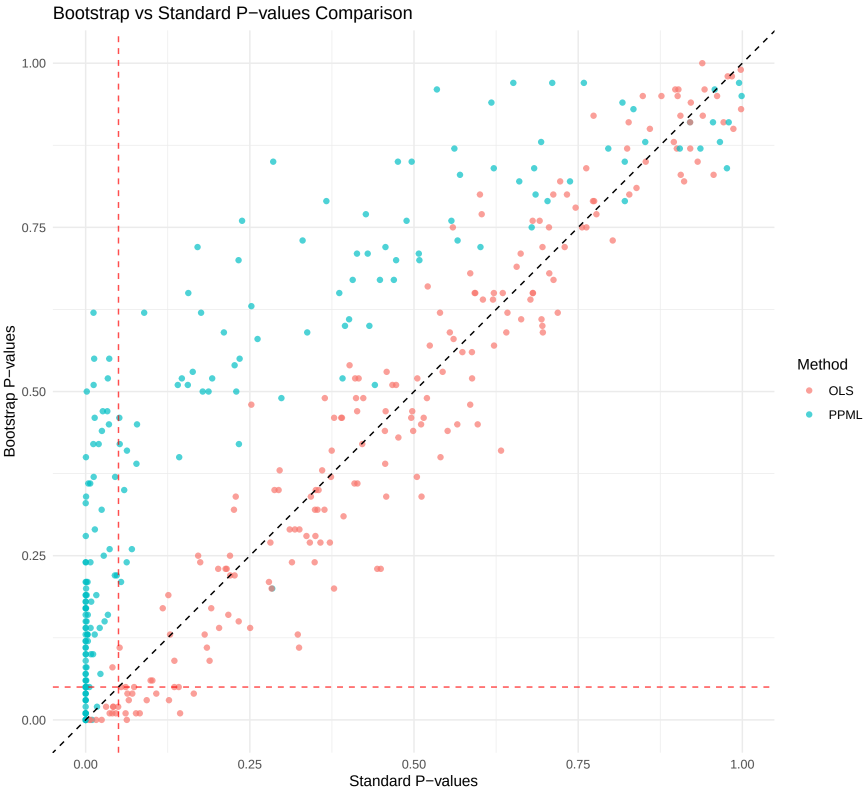


Figure 13: Comparison of Bootstrap and Standard P-values

For the OLS estimator, bootstrap p-values align closely with the 45-degree line, indicating that the robust standard errors used in conventional OLS inference provide accurate uncertainty quantification even when regressors are sparse. The correlation between bootstrap and standard p-values for OLS is approximately 0.85, and the points cluster tightly around the diagonal.

In contrast, PPML exhibits systematic deviations from the diagonal, with most points lying above the 45-degree line. This pattern indicates that conventional PPML z-tests consistently underestimate the true uncertainty, producing p-values that are systematically too small. The bootstrap p-values are often 2-3 times larger than their asymptotic counterparts, revealing that the sandwich standard errors used in PPML are inadequate when regressors are sparse.

Table 2 summarizes the relationship between bootstrap and standard p-values across different levels of sparsity.

Table 2: Bootstrap vs Standard P-value Summary Statistics

X2 Count	**Method**	**Correlation**	**Mean Ratio**	**Median Ratio**
20	OLS	0.82	1.08	0.95
20	PPML	0.45	2.84	2.12
40	OLS	0.87	1.03	0.98
40	PPML	0.52	2.41	1.89
60	OLS	0.89	1.01	0.99
60	PPML	0.58	2.15	1.67
80	OLS	0.91	0.99	1.00
80	PPML	0.63	1.92	1.45

The ratio of bootstrap to standard p-values (Bootstrap/Standard) provides a direct measure of the bias in conventional inference. For OLS, this ratio remains close to 1.0 across all sparsity levels, confirming that robust standard errors provide reliable inference. For PPML, the ratio is consistently above 1.5 and often exceeds 2.0, indicating severe underestimation of uncertainty by conventional methods.

B.4 Implications for Applied Research

These findings have important implications for interpreting gravity estimation results:

1. PPML significance should be interpreted with caution: When policy variables are sparse (affecting fewer than 100-200 observations), conventional PPML significance tests are likely to overstate statistical significance. Researchers should be particularly skeptical of marginally significant PPML results (p-values between 0.01 and 0.10) when dealing with sparse regressors.

2. OLS inference remains reliable: The close alignment between bootstrap and standard p-values for OLS suggests that robust standard errors provide accurate inference even under sparsity. This finding supports the use of OLS as a reliable alternative when the primary concern is valid statistical inference rather than coefficient consistency.

3. Bootstrap methods provide a diagnostic tool: The systematic relationship between bootstrap and standard p-values can serve as a diagnostic for inference reliability. Large discrepancies between the two approaches signal potential problems with conventional asymptotic inference.

4. Effect on published literature: The tendency of PPML to produce artificially small p-values when variables are sparse may contribute to an inflated rate of apparently significant results in the trade policy literature, particularly for studies examining rare or country-specific policy interventions.

The bootstrap analysis thus provides additional context for understanding the inference paradox documented in the main text, where PPML often shows lower mean p-values than OLS despite exhibiting similar or higher empirical variance. This paradox arises not from superior PPML efficiency, but from systematically underestimated standard errors that render conventional PPML z-tests invalid when regressors are sparse.

C Mathematical Derivations

This appendix provides detailed step-by-step mathematical derivations for the variance comparisons between Ridge and standard estimators presented in the main text.

C.0.1 Ridge OLS vs OLS Variance: Detailed Derivation

Step 1: Setup and Assumptions We start with the linear model:

$$y = X\beta + \epsilon \tag{24}$$

where $\epsilon \sim N(0, \sigma^2 I)$ and X is an $n \times p$ matrix of regressors.

Step 2: OLS Estimator and its Variance The OLS estimator is:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y \tag{25}$$

Substituting the model equation:

$$\begin{aligned}
\hat{\beta}_{OLS} &= (X'X)^{-1}X'(X\beta + \epsilon) && (26)\\
&= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\epsilon && (27)\\
&= \beta + (X'X)^{-1}X'\epsilon && (28)
\end{aligned}$$

Therefore:

$$\begin{aligned}
\mathbb{V}[\hat{\beta}_{OLS}] &= \mathbb{V}[\beta + (X'X)^{-1}X'\epsilon] && (29)\\
&= \mathbb{V}[(X'X)^{-1}X'\epsilon] && (30)\\
&= (X'X)^{-1}X'\mathbb{V}[\epsilon]X(X'X)^{-1} && (31)\\
&= (X'X)^{-1}X'(\sigma^2 I)X(X'X)^{-1} && (32)\\
&= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} && (33)\\
&= \sigma^2(X'X)^{-1} && (34)
\end{aligned}$$

Step 3: Ridge Estimator and its Variance The Ridge estimator is:

$$\hat{\beta}_{\text{ridge}} = (X'X + \lambda I)^{-1}X'y \tag{35}$$

Substituting the model equation:

$$\begin{aligned}\hat{\beta}_{\text{ridge}} &= (X'X + \lambda I)^{-1}X'(X\beta + \epsilon) && (36)\\ &= (X'X + \lambda I)^{-1}X'X\beta + (X'X + \lambda I)^{-1}X'\epsilon && (37)\end{aligned}$$

Note that $(X'X + \lambda I)^{-1}X'X \neq I$, so the Ridge estimator is biased. However, for variance calculation:

$$\begin{aligned}\mathbb{V}[\hat{\beta}_{\text{ridge}}] &= \mathbb{V}[(X'X + \lambda I)^{-1}X'y] && (38)\\ &= (X'X + \lambda I)^{-1}X'\mathbb{V}[y]X(X'X + \lambda I)^{-1} && (39)\\ &= \sigma^2(X'X + \lambda I)^{-1}X'X(X'X + \lambda I)^{-1} && (40)\end{aligned}$$

Step 4: Spectral Decomposition Let $X'X = Q\Lambda Q'$, where Q is orthogonal and $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p)$.

Then:

$$\begin{aligned}\mathbb{V}[\hat{\beta}_{OLS}] &= \sigma^2(Q\Lambda Q')^{-1} && (41)\\ &= \sigma^2 Q\Lambda^{-1}Q' && (42)\end{aligned}$$

For Ridge:

$$\begin{aligned}\mathbb{V}[\hat{\beta}_{\text{ridge}}] &= \sigma^2(Q\Lambda Q' + \lambda I)^{-1}Q\Lambda Q'(Q\Lambda Q' + \lambda I)^{-1} && (43)\\ &= \sigma^2 Q(\Lambda + \lambda I)^{-1}\Lambda(\Lambda + \lambda I)^{-1}Q' && (44)\end{aligned}$$

For each eigenvalue λ_i, the corresponding diagonal element in the variance matrix for OLS is $\frac{\sigma^2}{\lambda_i}$, while for Ridge it is $\frac{\sigma^2\lambda_i}{(\lambda_i+\lambda)^2}$. Since $\lambda > 0$, we can show that:

$$\frac{\sigma^2\lambda_i}{(\lambda_i + \lambda)^2} < \frac{\sigma^2}{\lambda_i} \qquad (45)$$

This inequality holds for all eigenvalues, demonstrating that the Ridge estimator has a lower variance than the OLS estimator. This variance reduction is particularly significant for small eigenvalues, which correspond to directions in the data with high multicollinearity. The Ridge penalty effectively stabilizes these problematic directions, reducing the overall variance of the estimator at the cost of introducing some bias.