Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Double Degree programme Econometrics & Economics

Intuitive Inference: Enhancing Post-Double Selection Inference on Treatment Effects with Intuition-Based Penalties

Sven van Holten Charria (525868)

Abstract

This thesis examines the 'Post-Double Selection' methodology introduced by Belloni et al. (2014b) for achieving valid inference on treatment effects in high-dimensional datasets. While Belloni et al. (2014b) and Wüthrich and Zhu (2023) suggest the ex-post addition of economically intuitive variables to reduce bias, they do not provide a robust methodology for this integration. This study addresses this gap by incorporating economically intuitive variables directly into the selection process, overcoming the limitations of traditional ex-post additions for confounding variable selection, particularly in high-noise environments. Monte Carlo simulations demonstrate that this integrated approach significantly enhances treatment effect estimation, showing marked improvement over the traditional Post-Double Selection method under certain conditions. An empirical re-examination of the effect of abortion on crime rates, as studied by Donohue and Levitt (2001), highlights the practical benefits of this approach, leading to a more comprehensive control selection and meaningful economic interpretations.

Supervisor:	Stan Koobs
Second assessor:	Jeffrey Durieux
Date final version:	2024-06-30

Contents

1 Introduction

Albert Einstein is widely quoted as saying, "The intuitive mind is a sacred gift, and the rational mind is a faithful servant." However, in contemporary econometric research, this perspective is undergoing a reversal. Since the introduction of the first simple econometric variable selection algorithm, the stepwise regression, the practice of variable selection has shifted from rational economic reasoning and intuition towards complex statistical models (Desboulets, 2018). Nowadays, most explanatory variable selection in simple and complex models tends to be guided more by minimizing a loss function or a metric than by economic intuition. In forecasting, the shift toward econometric rationality is driven by the pursuit of improved in-sample and out-of-sample results, wherein model selection aims solely to enhance accuracy (Berrevoets et al., 2024; Chernozhukov et al., 2018). Major economic institutes and government agencies increasingly rely on black-box mechanisms to uncover and exploit the benefits of increasingly high-dimensional data. For instance, researchers at the ECB are using random forest models to forecast inflation (Chinn et al., 2023; Lenza et al., 2023), and researchers at the IMF are using neural networks to forecast tail risk (Sakurai & Chen, 2024).

The shift to machine learning methods is emphasized by the current combinatorial explosion in empirical econometric practice, where the (baseline) control set is enriched with larger sets of interactions and other complex transformations. This is done to leverage higher computational power and model flexibility to achieve potential efficiency gains (Wüthrich & Zhu, 2023). Although modern machine learning models are designed to identify only those covariates with the most potent predictive ability, these approaches, often driven by regularization, induce model sparsity well, provided a sparse linear form can effectively capture the outcome. Empirically, this is often the case, resulting in more accurate forecasts with smaller variances (Belloni et al., 2014a). However, as regularization introduces bias, these machine learning models need to be refined to exploit the potential information in high-dimensional data and provide better-informed answers to causal questions, such as the inference of causal or treatment effects, while minimizing their side effects (Belloni & Chernozhukov, 2011). Hence, there is a continuous strong drive to further develop high-dimensional causal econometric models in the academic and practitioner fields.

A recent approach in this discussion is the 'Post-Double Selection', which employs a double selection step for asymptotically valid inference on treatment effects in a high-dimensional environment (Belloni et al., 2014a). Empirical results demonstrate significant reductions in bias through the use of a regularization model that includes a double (sparsity-inducing) selection step. The methodology relies on conducting two preliminary Lasso regressions: one of the outcome variable on the control set and another of the treatment on the control set. The main objective is to choose sufficient confounding controls (or simply controls) such that conditioning on these selected controls renders the treatment variable exogenous (Belloni et al., 2017). A final Ordinary Least Squares (OLS) regression of the outcome variable on the treatment and the selected controls from the preliminary step substantially reduces bias in estimating the treatment effect compared to a single selection step and similar variants. Following this methodology, the Post-Double Selection provides valid uniform inference across a broad range of models.

At first sight, the Post-Double Selection method appears to be yet another data-driven

mechanism where minimizing a loss function yields an 'optimal' set of controls, and the economic intuition behind control selection is discarded. However, in the literature, there is precedent for reintroducing economic intuition into the model. Researchers are even recommended to *"always augment the union of the selected controls with an 'amelioration set' of controls motivated by economic theory and prior knowledge"* (Wüthrich & Zhu, 2023). For the purpose of this thesis, following Desboulets (2018), economically intuitive variables are defined as raw variables with a clear economic interpretation and are typically expected to act as confounding variables for the relevant causal problem. The addition of economically intuitive controls ex-post for the final OLS regression is reasoned by arguing that combining a formal, rigorous approach with complementary economic intuition can significantly enhance the estimation of treatment effects, especially when the treatment is exogenous to the controls (Belloni et al., 2014a). Furthermore, even when this assumption does not hold, the amelioration set is seen as a potentially significant addition for ensuring valid inference by conditioning on the necessary controls, thereby making the treatment plausibly exogenous (Belloni et al., 2017).

In the existing literature, the inclusion of the amelioration set is typically limited to an ex-post addition to the final OLS regression (Belloni et al., 2014a; Gillen et al., 2014; Wüthrich & Zhu, 2023). However, this simplistic approach disregards proper selection criteria and lacks robustness. Under unfavourable conditions, the ex-post addition of the amelioration set could even worsen causal inference. If the controls in the amelioration set are not described by the Data Generating Process (DGP), this could lead to increased variance. Additionally, multicollinearity may arise if the controls in the amelioration set are strongly correlated with previously selected controls, thereby increasing the standard deviations of parameter estimates. Whilst the treatment effect is not directly affected, the precision of any post-hoc economic interpretation based on the coefficient estimates is affected.

Empirically, the necessity for an amelioration set is especially high in problems with noisy data, where the true effects of fundamental variables are overshadowed by noise. In such a setting, perfect recovery of the real representation of the DGP, even asymptotically, cannot longer be assumed (Wainwright, 2019). As a result, the Post-Double Selection method could yield a subset in which complex transformed variables dominate over economically intuitive variables, potentially better fitting the noisy data. If these controls are chosen through a spurious relationship, their addition could prevent the detection of important confounding variables within the Post-Double Selection and cause bias by missing underlying trends (Belloni et al., 2016). In this context, the amelioration set can mitigate the bias due to the missing confounding variables, albeit against a higher variance. This thesis, therefore, aims to provide an integrated alternative to the ex-post addition of the amelioration set by applying lower penalization to economically intuitive variables within Post-Double Selection Lasso regressions. This leads to the following research question:

> *To what extent can integrated amelioration set penalization be used in Post-Double Selection to reduce bias and improve inference in a high-dimensional dataset with many controls?*

With respect to existing literature, the main findings of this thesis are interesting for several reasons. First, different implementations of data-dependent and data-independent amelioration

set penalties are proposed, producing competitive inferential results for treatment effects in approximately sparse models. These methods are more robust than those currently described in the literature as the information on the amelioration set is integrated throughout the selection procedure instead of added ex-post, exploiting available information to fine-tune the penalization weights. Additionally, the methods accommodate both homoscedastic and heteroscedastic error distributions and are tested across various levels of residual system noise. Second, several data-dependent and data-independent amelioration set penalties are evaluated through Monte Carlo simulations, yielding a better treatment effect estimation than Post-Double Selection under certain conditions. This differential is particularly significant when the treatment variable and its confounding variables are highly correlated and the relationship is subject to low residual noise levels. This finding is applied in an empirical replication examining the effect of abortion on crime rates, which features such a data structure (Donohue & Levitt, 2001). The most basic amelioration set penalty, which removes penalization of the amelioration set in the Post-Double Selection, acts as a hybrid between regular Post-Double Selection and Post-Double Selection with the ex-post added amelioration set: an intermediate number of economically intuitive controls and an intermediate total number of controls are selected, yielding a methodology that provides a solid middle ground between economic interpretation and econometric robustness.

The thesis is structured as follows: Section 2 provides a brief overview of the current literature on Post-Double Selection. Section 3 covers the relevant methodological background necessary to understand the methods used in this research. Section 4 discusses the data used in the research. Next, Section 5 describes and explains the steps required to perform valid usage of the Post-Double Selection and the amelioration set penalties. Section 6 presents the results of the Monte Carlo simulations and offers a practitioner's recommendation for using the amelioration set penalties. Section 7 applies these findings to an empirical case. Finally, Section 8 presents the conclusions drawn, the limitations of the research, and ideas for future research.

Notation. Support(.) represents the number of non-zero elements in a vector. The average expectation operator is defined as $\mathbb{E}_n[f] \coloneqq \mathbb{E}_n\left[f(.)\right] \coloneqq \sum_{i=1}^{n} f(.)/n$. The norms used are the ℓ_2-norm, denoted by $||.||_2$, the ℓ_1-norm, denoted by $||.||_1$, and the ℓ_∞-norm, denoted by $||.||_\infty$, representing the maximal element of a vector.

2 Related Literature

On a fundamental level, this thesis builds directly on the existing literature of machine learning for causal inference, first introduced by Belloni, Chernozhukov and Hansen (2013). In this seminal paper, the Post-Double Selection method is introduced with the Lasso estimator as a means of achieving approximate sparsity. An extension paper provides a formal proof for uniformly valid inference on treatment effects after selection in a setting with high-dimensional controls (Belloni et al., 2014a). Further papers extend the double-Lasso method to various applications, such as high-dimensional panel models (Belloni et al., 2016), high-dimensional approximately sparse quantile regression models (Belloni et al., 2019), and principled variable selection (Urminsky et al., 2019). Advancements generalize the impact of regularization bias and overfitting of the parameter of interest using Neyman-orthogonal moments and efficient data-

splitting, in a method known as *debiased-ML*, of which the Post-Double Selection is a specific case (Chernozhukov et al., 2018, 2022). All of these papers mention ℓ_1-regularization, specifically the Lasso estimator, as a good means to approximate sparsity. This thesis will primarily focus on the original methodology of the Post-Double Selection and extend the literature by providing a robust improvement to the control selection.

2.1 Data-Driven Penalization

The ℓ_1-regularization parameter plays a critical role in the Lasso estimator, balancing the tradeoff between variance and bias. Too many unnecessary controls may be included if the penalty is not conservative enough, leading to higher variance. Conversely, a penalty that is too conservative can increase omitted variable bias. Bickel et al. (2009) first proposed a choice for the Lasso penalty based on Gaussian error assumptions and data dimensionality. Building on this, Belloni and Chernozhukov (2013) introduced data-independent penalty loadings for Post-Lasso under Gaussian errors, and Belloni et al. (2012) introduced data-driven penalty loadings accommodating non-Gaussian errors. Finally, Belloni, Chernozhukov and Hansen (2013) developed an algorithm to estimate the residual variance in the data, providing a framework for optimal penalty loadings under non-Gaussian, heteroscedastic errors. The thesis will extend the methodology for non-Gaussian homo- and heteroscedastic errors by accommodating beliefs based on economic intuition, adding an extra dimension to the estimation of penalty loadings.

In addition, an important feature of data-driven penalties is their co-dependence in determining optimal loadings: each loading is fit to its control based on all available information. Therefore, an adjusted loading for one control will affect the loading of another. Therefore, this thesis provides a useful insight into the behaviour of data-driven penalty loadings under beliefs, which are not necessarily data-driven but do translate to lower penalization.

2.2 Practisioner's Recommendations

Considering the existing literature on Post-Double Selection, the theoretical focus of Post-Double Selection is often overshadowed by diverse empirical applications. [1] Belloni et al. (2014a) stands out as one of the few theoretical papers that proposes practical recommendations for using causal inference estimators across data conditions. They analyze the performance of Post-Lasso and Post-Double Selection estimators against the infeasible Oracle benchmark across various combinations of first and second-stage R^2 values [2] and coefficient designs, to guide the appropriate use of Post-Double Selection in causal inference analysis. The authors find that Post-Lasso performs poorly across all combinations of R^2 and designs, managing only to control bias when the treatment is uncorrelated with the control (first-stage $R^2 = 0$). Post-Lasso also generally performs better in a scenario where the confounding variables in the first-stage differ from those in the second-stage regression, occasionally even outperforming Post-Double Selection under conditions of low first- and second-stage R^2. This situation arises when fundamental

[1] See for instance, Qiu et al. (2020) for an epidemiological application on the transmission of Covid-19, Dhar et al. (2022) for an economic application on adolescent gender attitudes, and Hangartner et al. (2021) for a psychological application on hiring discrimination.

[2] Referring to the correlation between the treatment variable and the design matrix, and between the outcome and the design matrix, respectively.

variables are significantly distorted by noise and the identities of the confounding variables are uncertain. Amelioration penalties aim to precisely address this challenge by providing supplementary information to better capture confounding variables under low first- and second-stage R^2. Otherwise, the Post-Double Selection estimator performs similarly to the Oracle across all combinations of R^2 and achieves its best performance under high first-stage R^2 when the identities of the confounding variables are well known. This thesis aims to extend recommendations for practitioners by conducting an updated analysis using amelioration penalties across different designs and error distributions. In addition, this thesis will consider varying DGPs where the confounding variables in the first stage are not the same as those in the second stage.

2.3 Bayesian Lasso

The idea of imposing a lower level of penalization for controls that are considered fundamental can be explored through the lens of the Bayesian Lasso, introduced by Park and Casella (2008). The paper suggests that the behaviour of the Lasso estimator can be interpreted through a Bayesian posterior distribution, assuming that the priors on the regression parameters follow independent LaPlace distributions. More specifically, the Lasso penalty estimates are equivalent to the mode of the posterior distributions. The shape of the LaPlace distribution, peaking sharply at the origin, produces a similar shrinkage effect as regularization with ℓ_1-penalty, and in practice, the solutions of both methods should be identical. Therefore, the results of this thesis can be extended to Bayesian statistics by translating adjusted penalties in the amelioration set into different shaped priors for their respective Bayesian counterparts. For instance, the prior for a variable that is believed to be economically intuitive, under discrete amelioration set penalties, can be modelled as a discrete posterior distribution with $\mathbb{P}(X = 0) = 1$. Although this thesis is not directly related to the Bayesian Lasso, exploring this connection could be an interesting topic for future research.

3 Theoretical Framework

This section provides an overview of the relevant theoretical background necessary to understand the choices made for the penalties in the amelioration set. First, the theory behind approximately sparse models and its contextualization to causal inference is presented. Next, the Lasso, Post-Lasso, indirect Post-Lasso, and Post Double Selection estimators are introduced alongside the relevant penalties to ensure validity. Lastly, a brief theoretical check is conducted on the regularity conditions of the Post Double Selection, to ensure inferential validity of the amelioration set penalties.

3.1 Approximately Sparse Models

In an econometric model where the number of regressors p is relatively large compared to the number of observations n, conducting a simple OLS regression is impossible. However, if a small subset of regressors $s \subset p$ can approximately capture the main features of the regression, identifying this subset and performing regression using only these regressors offers a feasible approach to OLS regression (Belloni & Chernozhukov, 2011). This intuition behind a model that can be

approximated by a small subset of its regressors describes the structure of an *'Approximately Sparse Model' (ASM)* and motivates the consideration of regularization techniques to identify s. Formally, a model is approximately sparse if $p \gg n$ and $s \ll p$.

The following model is introduced to provide intuition into the approximation of a high-dimensional model through an ASM (Belloni, Chernozhukov & Hansen, 2013):

$$y_i = f(z_i) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n, \tag{3.1}$$

where y_i is the outcome variable, z_i is a vector of primary regressors that are significant drivers of the regression function, and ε_i are i.i.d. errors. The regression function $f(z_i)$ is unknown, implying that the exact relationship between z_i and y_i is also unknown. To approximate the regression function, the design matrix $X = [x_1, ..., x_n]'$ is introduced, where each $x_i = P(z_i)$. Here, $P(z_i)$ represents an unknown mapping of z_i, possibly involving simple and complex transformations (constants, splines, interactions, exponents), or even creating an ambiguous relationship. The resulting approximation of the approximately sparse model through the design matrix is as follows:

$$y_i = x_i'\beta_0 + r_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.2}$$

where β_0 represents the coefficients and r_i the approximation errors. Assuming there exists a set of unknown regressors s such that a linear combination of these regressors produces sufficiently low approximation errors, the model is considered approximately sparse. This equivalence is characterized by a linear combination of regressors with support$(\beta_0) = s \ll p$, resulting in relatively small r_i. However, identifying the specific elements of β_0 is challenging, and researchers typically introduce a high-dimensional design matrix where $p \gg n$ in an effort to uncover these complex underlying relationships.[3] In practice, pursuing the sparse identification of a high-dimensional model is only beneficial if it is believed that y_i can be well-approximated using a much smaller subset of regressors (Belloni & Chernozhukov, 2011).

3.2 Approximately Sparsity for Inference

This approach of identifying the driving regressors in models with high-dimensional controls using approximately sparse models can be extended to causal inference (Belloni et al., 2014a). Consider the following partially linear model:

$$y_i = d_i'\alpha_0 + g(z_i) + u_i, \quad \mathbb{E}[u_i|z_i, d_i] = 0, \tag{3.3}$$

$$d_i = m(z_i) + v_i, \quad \mathbb{E}[v_i|z_i] = 0, \tag{3.4}$$

where y_i is the outcome variable, d_i is the treatment variable, α_0 is the coefficient of interest capturing the treatment effect, z_i are the controls, and ζ_i and ν_i are zero-mean errors. The functions $g(\cdot)$ and $m(\cdot)$ are unknown and describe the relationships of the controls on the outcome variable and the treatment variable, respectively. Similar to $f(z_i)$ in (3.2), the unknown nature and potentially complex structure of these functions hinder the ability to identify the relevant

[3] Consider the empirical case in Section 7, where a baseline control set of seven regressors expands to include 284 non-multicollinear controls. In the end, however, only an average of 10 controls were selected in the Post-Double Selection.

controls in the partially linear model. However, if the relevant controls in $g(z_i)$ and $m(z_i)$ are correctly identified and assuming sufficiently low approximation errors, the exogeneity of d_i could be assumed. In practice, uncovering the identities of z_i is challenging, and sparse approximations provide a method to approximate $g(z_i)$ and $m(z_i)$. Consider the following sparse approximation of the partially linear model (Belloni et al., 2014a):

$$y_i = d_i'\alpha_0 + \underbrace{x_i'\beta_{g0} + r_{gi}}_{g(z_i)} + u_i, \tag{3.5}$$

$$d_i = \underbrace{x_i'\beta_{m0} + r_{mi}}_{m(z_i)} + v_i, \tag{3.6}$$

where $x_i'\beta_{g0}$ and $x_i'\beta_{m0}$ are the sparse approximations for $g(z_i)$ and $m(z_i)$, and r_{gi} and r_{gm} are the respective approximation errors. Conditional on sufficiently small approximation errors and other regularity conditions, some of which are discussed in Section 3.6, valid causal inference can be performed on the treatment effect α_0 (Belloni et al., 2014a). Therefore, the challenge remains to find a feasible and valid method to construct such an approximately sparse model given a high-dimensional design matrix.

3.3 One-stage Lasso Estimation

The Lasso estimator provides a simple, theoretically guaranteed, and computationally feasible solution to induce sparsity through ℓ_1-regularization. The estimator shrinks some coefficients and sets others to zero without requiring prior knowledge of the structure of the DGP (Tibshirani, 1996):

$$\hat{\beta}_L \quad = \quad \min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^p} \mathbb{E}_n\left[\left(y_i - d_i\alpha - x_i'\beta\right)^2\right] + \frac{\lambda}{n}\|\beta\|_1, \tag{3.7}$$

where $\|\beta\|_1 = \sum_{j=1}^{p}|\beta_j|$, and λ is a specified penalty vector. Note that $\lambda_\alpha = 0$ in all relevant Lasso estimations to restrict the penalization and ensure the selection of the treatment variable. Although ℓ_1-regularization helps Lasso reduce overfitting, the shrinkage of coefficients to zero can introduce significant bias. The Post-Lasso estimator attempts to alleviate Lasso's shrinkage bias by performing an OLS regression on the variables selected from an initial Lasso selection. This estimator is found to perform at least as well as a single Lasso estimation but with lower bias on the treatment effect (Belloni & Chernozhukov, 2013):

$$\hat{\beta}_{PL} \quad = \quad \min_{\alpha\in\mathbb{R},\beta\in\mathbb{R}^p} \mathbb{E}_n\left[\left(y_i - d_i\alpha - \sum_{j=1}^{p} x_{i,j}\beta_j\right)^2\right] \quad : \ \beta_j = 0 \ \text{ if } \ \hat{\beta}_{L,j} = 0, \quad \forall j. \tag{3.8}$$

The Indirect Post-Lasso is an adaptation of the Post-Lasso, conditioning on the indirect effect of d_i on $x_{i,j}$ in the first-stage estimation to reduce the possibility of missing confounding variables. Next, an OLS is performed on the union of the controls selected in the first stage and the

treatment variable (Belloni & Chernozhukov, 2011):

$$\hat{\beta}_d = \min_{\beta \in \mathbb{R}^p} \mathbb{E}_n \left[\left(d_i - x_i' \beta \right)^2 \right] + \frac{\lambda}{n} \|\beta\|_1, \tag{3.9}$$

$$\hat{\beta}_{IPL} = \min_{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p} \mathbb{E}_n \left[\left(y_i - d_i \alpha - \sum_{j=1}^{p} x_{i,j} \beta_j \right)^2 \right] \quad : \ \beta_j = 0 \ \text{ if } \ \hat{\beta}_{d,j} = 0, \quad \forall j. \tag{3.10}$$

This estimator should intuitively recognize the confounding variables better, shrinking the OVB more effectively than the Post-Lasso. However, if the correlation between d_i and $x_{i,j}$ is negligible, and the correlation between y_i and $x_{i,j}$ is high, the treatment effect might become difficult to identify. Therefore, the single-stage Lasso might benefit from a second-stage estimation to address this pitfall.

3.4 Two-stage Lasso Estimation

Although the Lasso estimator is restricted to estimating a single model at a time, it can be used sequentially to accommodate partial linear models. For instance, the Post-Double Selection estimator, which sparsely approximates the partial linear model described in (3.3) and (3.4), is based on a two-stage estimation: a first-stage Lasso estimation of d_i on $x_{i,j}$ followed by a second-stage Lasso estimation of y_i on $x_{i,j}$. A final OLS regression is then performed of y_i on d_i and the union of the selected variables from the first- and second-stage Lasso estimations (Belloni et al., 2014a):

$$\hat{\beta}_x = \min_{\beta \in \mathbb{R}^p} \mathbb{E}_n \left[\left(y_i - x_i' \beta \right)^2 \right] + \frac{\lambda}{n} \|\beta\|_1. \tag{3.11}$$

Let $\hat{I}_1$ denote the controls selected by (3.9) and $\hat{I}_2$ denote the controls selected by (3.11). Let the union be $\hat{I}$. The final OLS regression is as follows:

$$\hat{\beta}_{PD} = \min_{\beta \in \mathbb{R}^p} \mathbb{E}_n \left[\left(y_i - \sum_{j=1}^{p} x_{i,j} \beta_j \right)^2 \right] \quad \text{subject to} \quad \beta_j = 0 \quad \text{if} \quad j \notin \hat{I} = \{ \hat{I}_1 \cup \hat{I}_2 \}. \tag{3.12}$$

This procedure is essentially an adaptation of Frisch-Waugh's "partialling out," in which y_i and d_i are cleaned of the effects of the confounding $x_{i,j}$'s through their inclusion in the OLS. After d_i is cleaned such that it is exogenous to the $x_{i,j}$'s, the true treatment effect of d_i on y_i can be observed (Heij et al., 2004). Note that for the estimated treatment effect $\hat{\alpha}_0$ to equal α_0 in the finite-sample, sufficient relevant confounding variables should be included in the final regression to ensure sufficiently low approximation errors r_{gi} and r_{mi}, such that exogeneity can be assumed for the treatment variable.

The main strength of the two-stage Post Double Selection estimation over a single-stage Lasso or Post-Lasso lies in the additional estimations of d_i on $x_{i,j}$ and y_i on $x_{i,j}$. Intuitively, there are two scenarios in which the two-stage estimation will outperform the one-stage estimation. First, if d_i and $x_{i,j}$ are strongly correlated but y_i and $x_{i,j}$ are not, the single-stage estimation might fail to capture this relationship and not condition on the confounding $x_{i,j}$ due to the low correlation with the outcome variable, whereas the two-stage estimation would account for this

through the first-stage estimation. Second, if both y_i and $x_{i,j}$, as well as d_i and $x_{i,j}$, are strongly correlated, the single-stage estimation might fail to distinguish the treatment effect from $x_{i,j}$, whereas the two-stage estimation would account for this through the second-stage estimation by estimating y_i on $x_{i,j}$ without conditioning on d_i.

3.5 Lasso-Penalization

So far, in the theoretical framework, the regularization behaviour of the Lasso estimator has only depended on the value of λ. However, under non-Gaussian, non-heteroscedastic error conditions, equivariance needs to be regulated through an additional parameter, Ψ. In this section, the theory behind the penalty choices is provided to ensure correct ℓ_1-regularization in the Lasso estimator, along with a brief introduction to Ω, the parameter that will regulate the amelioration set penalization. Additionally, it is important to note that for the Lasso to behave correctly, all regressors in the design matrix should be normalised.

3.5.1 λ-Penalty: Regularization

In the context of forecasting, cross-validation is a popular method for choosing λ, but its focus on minimizing prediction error is not suitable for causal inference (Belloni et al., 2014b; Wüthrich & Zhu, 2023). Instead, the main function of λ should be to dominate the 'effective noise' in the system through regularization while maintaining the regularization bias as small as possible (Belloni & Chernozhukov, 2011). Furthermore, λ directly regulates the sparsity in the Lasso function: a higher penalty generally results in more sparsity due to the added cost of an additional non-zero β-coefficient.

The following definitions will improve clarity in the derivation of the optimal λ: the criterion function of the Lasso estimator is defined as $\hat{Q}(\beta) = \mathbb{E}_n\left[\left(y_i - x_i'\beta\right)^2\right]$, and the score, $\mathbf{S}$, is defined as the gradient of the criterion function evaluated at the true parameter value β_0: $\mathbf{S} = \nabla\hat{Q}(\beta_0)$ (Newey, 1994). The derivation proceeds as follows:

$$\begin{aligned}
\mathbf{S} &= \nabla\hat{Q}(\beta_0) && (3.13)\\
&= \nabla\mathbb{E}_n\left[\left(y_i - x_i'\beta_0\right)^2\right] && (3.14)\\
&= 2\cdot\mathbb{E}_n\underbrace{\left[y_i - x_i'\beta_0\right]}_{\varepsilon_i}\cdot\frac{\partial}{\partial\beta_0}\left(\mathbb{E}_n\left[y_i - x_i'\beta_0\right]\right) \qquad (\textit{Chain rule}) && (3.15)\\
&= -2\mathbb{E}_n\left[x_i\varepsilon_i\right] \overset{\text{sym.}}{=} 2\mathbb{E}_n\left[x_i\varepsilon_i\right] \overset{\text{d.}}{=} 2\sigma\mathbb{E}_n\left[x_i g_i\right], && (3.16)
\end{aligned}$$

where g_i is an i.i.d. $\mathcal{N}(0,1)$ distributed random variable, and σ describes the standard deviation of the residuals. The final equivalence describes the convergence in the distribution of the residuals to normality, although σ remains unobserved in the problem and can only be estimated by $\hat{\sigma}$. [4] Therefore, $\mathbb{E}_n\left[x_i g_i\right]$ can be estimated through simulations using only the design matrix and random sampling from the standard normal distribution. In addition, note that λ dominating the effective noise can be conservatively simplified to λ having to dominate the most noisy

[4] An iterative procedure allows for a refined estimation of σ using Post-Lasso iterations (Belloni & Chernozhukov, 2011). This procedure will be elaborated on in Section 5.

regressors across all observations, choosing a confidence level of $(1-\alpha)$:

$$\lambda \quad > \quad c \cdot \mathbf{\Lambda} \quad \text{for} \quad \mathbf{\Lambda} \coloneqq n \left\| \mathbf{S} \right\|_{\infty} = 2n \left\| \mathbb{E}_n \left[x_i \varepsilon_i \right] \right\|_{\infty}, \tag{3.17}$$

for a constant $c > 1$, such that the optimal feasible λ is estimated as follows: [5]

$$\hat{\lambda} \quad \geq \quad 2cn \left\| \mathbb{E}_n \left[x_i \varepsilon_i \right] \right\|_{\infty} \quad \overset{\text{d.}}{=} \quad 2cn\hat{\sigma} \left\| \mathbb{E}_n \left[x_i g_i \right] \right\|_{\infty}, \tag{3.18}$$

The standard Lasso-estimation can be rewritten as:

$$\hat{\beta}_{\Lambda} \quad = \quad \min_{\beta \in \mathbb{R}^p} \hat{Q}(\beta_0) + 2c\sigma \cdot \left\| \mathbb{E}_n \left[x_i g_i \right] \right\|_{\infty} \cdot \|\beta\|_1 \tag{3.19}$$

$$= \quad \min_{\beta \in \mathbb{R}^p} \hat{Q}(\beta_0) + c\sigma \cdot \mathbf{\Lambda}(1-\alpha|X) \cdot \|\beta\|_1. \tag{3.20}$$

The latter notation is useful as the $\mathbf{\Lambda}(1-\alpha|X)$, also called the X-dependent penalty term, remains constant given a fixed regressor matrix. Using this penalty term over an alternative, data-independent Lambda, is recommended because the former adapts to the design matrix X by construction and is less conservative than the latter (Belloni & Chernozhukov, 2011). This is due to the stricter asymptotic bounds of the data-driven penalty. The data-driven penalty parameters, c and α, are asymptotically required to converge to 1 and 0, respectively, with probability 1. Non-asymptotically, Belloni and Chernozhukov (2011) found that in finite-sample experiments, $c = 1.1$ and $\alpha = 0.10$ are sufficient. Although the cross-validated λ also relies on the data, its focus on regularization for forecasting performance does not align with minimizing the bias on a treatment variable (Chernozhukov et al., 2016).

3.5.2 Ψ-Penalty: Equivariance

The previously mentioned $\hat{\beta}_{\Lambda}$ is only valid under the assumption of homoscedastic, normally distributed errors and normalised data. However, to handle data with heteroscedastic, non-Gaussian errors, a correction is introduced via Ψ: (Belloni et al., 2014a)

$$\hat{\beta}_{\Psi} \quad = \quad \min_{\beta \in \mathbb{R}^p} \hat{Q}(\beta_0) + 2c \cdot \left\| \mathbb{E}_n \left[x_i \varepsilon_i \right] \right\|_{\infty} \cdot \|\hat{\Psi}\beta\|_1, \tag{3.21}$$

where $\hat{\Psi} = \text{diag}(\hat{\psi}_1, \ldots, \hat{\psi}_p)$ is a diagonal matrix of penalty loadings designed to impose equivariance in the ℓ_1-regularization penalty term allowing for valid heteroscedastic treatment, and $\|\hat{\Psi}\beta\|_1 = \sum_{j=1}^{p} |\hat{\psi}_j \beta_j|$ (Chernozhukov et al., 2016). The following derivation shows the necessary Ψ to ensure the correct behaviour of (3.21). The derivation starts at an equivalent form of

[5] Belloni and Chernozhukov (2011) found that the combination of $\alpha = \{0.05, 0.1\}$ and $c = 1.1$ produces strong results. Belloni, Chernozhukov and Hansen (2013) and Belloni et al. (2014a), amongst other applications of the Post Double Selection estimator, continue to use those values.

(3.18) with the inclusion of Ψ (Chernozhukov et al., 2015):

$$\frac{\hat{\lambda}\Psi}{n} \geq 2c\left\|\mathbb{E}_n\left[x_i\varepsilon_i\right]\right\|_\infty = 2c\left\|\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i\right\|_\infty \tag{3.22}$$

$$\frac{\hat{\lambda}}{\sqrt{n}} \geq 2c\left\|\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i}{\Psi}\right\|_\infty \tag{3.23}$$

where $\mathbb{E}_n\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i = 0\right]$, such that $\mathbb{V}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i\right) = \mathbb{E}_n\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2\varepsilon_i^2\right]$. Therefore, by setting $\Psi = \mathbb{E}_n\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2\varepsilon_i^2\right]$, the RHS of (3.23) is standardized, such that $\lambda\Psi \overset{\text{d.}}{=} 2cn\sigma \cdot \left\|\mathbb{E}_n\left[x_i g_i\right]\right\|_\infty$. Under the correction Ψ, the Lasso estimator exhibits correct behaviour under heteroscedastic, non-Gaussian errors, similar to (3.18)

3.5.3 Ω-Penalty: Amelioration set

The amelioration set penalties are introduced through a $(p \times 1)$ vector Ω that acts as a scaling parameter of λ. Let $\hat{I}_3$ denote the indices of the amelioration set, such that:

$$\Omega = \begin{cases} \omega_i \in [0,1]\,, \; i \in \hat{I}_3 \\ 1, \text{ otherwise} \end{cases}$$

$$\hat{\beta}_\Omega = \min_{\beta\in\mathbb{R}^p} \hat{Q}(\beta_0) + \frac{\lambda\cdot\Omega}{n}\|\hat{\Psi}\beta\|_1, \tag{3.24}$$

where $\Omega = [\omega_1, \ldots, \omega_k, 1, \ldots, 1]$, and $\omega_i \in \{0, 1\}$ accommodates the scaling of 'k' a priori selected regressors $X_{amel} = [x_1, \ldots, x_k]$ for $X = [x_1, \ldots, x_k, x_{k+1}, \ldots, x_p]$, with $k \ll p$. The regressors excluded from the X_{amel} set with $\omega_i = 1$ will not receive different treatment compared to (3.21). However, note that although in the formula it appears their behaviour remains unchanged, the lower penalization of the amelioration set will still affect their bias compared to before: a lower penalization of X_{amel} in the ℓ_1-regularization will result in a lower tendency for these coefficients to be zero and will exhibit lower downside regularization bias compared to the non-amelioration set. If $\omega_i = 0$ is imposed, there is no induced regularization bias, but instead, there will be an upward bias as the downside bias weighing of the regularized regressors will push the estimate of the non-regularized regressor upwards. [6] The specific values of Ω under different 'beliefs' will be further explored in the methodology.

The amelioration set penalties can be linked to ASMs, where beliefs supplied through different values of ω_i can help uncover the 'true' structure of the partially linear models in (3.3) and (3.4). Elements within the amelioration set are considered fundamental and assumed to be included in the structural representation of the model, but they are not always selected in the sparse approximations. This can be due to several reasons, some of which are outlined in the introduction. However, the simple amalgamation of the amelioration set ex-post to guarantee their inclusion in the final OLS regression, as described in the literature, is not robust to an imperfect amelioration set that does not reflect the structural representation. By integrating

[6] I want to thank my supervisor Stan Koobs for his intuitive explanation of the '*weights*' of the regularized versus non-regularized regressors and their push-and-pull mechanisms.

the amelioration set penalties within the Lasso-estimations and approaching their selection with increased scrutiny, this extended methodology should be more robust for the valid inclusion of the amelioration set penalties than the vanilla Post-Double Selection.

3.6 Conditions

Valid inferential results for approximate sparse treatment effects in ASMs depend on the approximate sparsity and rate conditions established in Belloni et al. (2014b), among others. Under noisy conditions, consistency in variable selection is only achieved for $n > p$ in sparse high-dimensional models, among other conditions (Wainwright, 2019). In this thesis, the latter condition is violated, so perfect confounding variable selection cannot be assumed. Furthermore, the empirical literature has found the control selection behaviour of the Lasso estimator to be imperfect, causing significant omitted variable bias (Wüthrich & Zhu, 2023). This paper does, however, reach the same conclusion as (Belloni et al., 2014b), recommending the inclusion of an amelioration set motivated by economic theory to offset potential omitted variable bias in the finite sample due to imperfect Lasso behaviour.

Therefore, this thesis will not assume perfect confounding variable selection or inference in the finite sample with the Post-Double Selection or with the inclusion of amelioration set penalties. Instead, the focus will be on how well the amelioration set penalties mitigate the persisting effects of noise in the Post-Double Selection and to what extent the addition of the penalties helps in establishing valid inferential results. From an analytical viewpoint, discussing whether the amelioration set penalties, as an alternative to ex-post inclusion, will similarly adhere to the conditions is an interesting theoretical addition. The conditions will be given for the partially linear model in (3.3) and (3.4). The approximate sparsity condition (ASC) is as follows:

$$m(z_i) = x_i'\beta_{m0} + r_{mi},\ \|\beta_{m0}\|_0 \leq s,\ (\mathbb{E}_n\left[r_{mi}^2\right])^{1/2} \leq C\sqrt{s/n}, \tag{3.25}$$

$$g(z_i) = x_i'\beta_{g0} + r_{mi},\ \|\beta_{g0}\|_0 \leq s,\ (\mathbb{E}_n\left[r_{gi}^2\right])^{1/2} \leq C\sqrt{s/n}, \tag{3.26}$$

for a sparse model with support $s \geq 1$ and an absolute constant $C \in \mathbb{R}$. The rate condition (RC) regulates that $s^2\log^2 max(p,n)/n \leq \delta_n$ for $\delta_n \in \mathbb{R}^p$, and that the size of the amelioration set should follow $\hat{s}_3 \leq C(max(1, \hat{s}_1, \hat{s}_2))$, for $\hat{s}_i = \|I_i\|_0, i \in \{1,2,3\}$.

The ASC requires that the approximation errors of $m(z_i)$ and $g(z_i)$ be sufficiently small using a small subset s of the full control set p, where the growth of the approximation errors is bounded by s/n. The amelioration set penalties on k regressors are conjectured to reflect the true sparsity s, such that their inclusion can improve the estimated sparsity $\hat{s}$ to better resemble s. If incorrect, the Lasso estimator should discard the information supplied through the amelioration set by not selecting controls in k, leading to no real change. Therefore, the condition should not be violated in either case as long as the sparse approximation does not result in a significantly worse approximation error. Monte Carlo simulations illustrate this effect of the amelioration set penalties on the ASC: a higher level of precision in the amelioration set penalties leads to a better sparse approximation, where the chosen controls converge to the real structural representation.

The introduction of the amelioration set penalties should not affect the first growth condition

in the RC. The second condition is tailored for the Post Double Selection with the ex-post addition of the amelioration set, requiring that the growth rate of the cardinality of the set $\hat{s}_3$ should not exceed that of the control set selected in the Post Double Selection, $\hat{s}_1$ and $\hat{s}_2$. Due to the integrated nature of the penalties for the amelioration set, this condition can be ignored, as such an ex-post amelioration set is not applied. Regardless, the size of the amelioration set penalties will be maintained lower or equal to the expected number of confounding variables in the DGP in both the Monte Carlo simulations and the empirical case.

4 Data

Monte Carlo simulations will accommodate the exploration of the amelioration set penalties under different conditions, such as varying error distributions, different first- and second-stage R^2 values, and different DGPs. A complete description of the testing conditions is given in Section 5. In addition, the findings of the theoretical exploration are applied in a re-examination of an empirical case covered in Belloni et al. (2014a), specifically the effect of abortion on crime rates (Donohue & Levitt, 2001), to uncover the impact of the amelioration set penalties on the estimated treatment effect and any differences in variable selection and the economic interpretations of coefficient estimates.

In their research, Donohue and Levitt (2001) evaluate the causal impact of abortion rates on three different types of crime: property, violence, and murder. The original set of controls includes lagged prisoners per capita, lagged police per capita, unemployment rate, per-capita income, poverty rate, welfare generosity, concealed weapons laws, and beer consumption per capita. Belloni et al. (2014b) find that estimating the causal impact is challenging due to non-random state-level abortion rates and the potential influence of other confounding factors related to both abortion and crime rates, providing a strong motivation for the use of Post Double Selection. They expand the set by including state-specific effects, time-specific effects, a set of control variables to account for time-varying state-level factors, and a variety of complex transformations, yielding the following model:

$$y_{cit} - y_{cit-1} = \alpha_c(a_{cit} - a_{cit-1}) + z'_{cit}k_c + g_{ct} + \eta_{cit}, \tag{4.1}$$

where y_{cit} is the crime rate and a_{cit} is the abortion rate for crime type $c \in \{\text{property, violent, murder}\}$, α_c is the treatment effect of abortion on crime, g_{ct} represents time effects, η_{cit} is the error term, and z_{cit} is an enriched set of the original controls, including higher-order terms, interactions with controls, a quadratic time trend, initial level differences, and within-state differences.[7] The z_{cit} term can be considered a high-dimensional control set that can be reduced to an ASM, for which it is assumed that an approximation using only a small number of controls yields a small enough approximation error.

The resulting data includes the same state-level information as reported by Donohue and Levitt (2001), excluding Alaska, Hawaii, and Washington, D.C., resulting in a sample of 48 cross-sectional observations on yearly state-level data from 1985 to 1997 (12 time series observations), yielding a total of 576 observations. The control set comprises 284 variables for each of the

[7]The full set of variables and transformations can be found in Belloni, Chernozhukov and Hansen (2013).

three types of crimes. This implies that for determining the state-specific effect, there are $n = 12$ observations for $p = 284$ controls, making it a case where $n \ll p$, and hence a good opportunity to evaluate the performance of Ω. In this empirical context, the amelioration set has been previously defined in the literature as the original set of 7 controls (Belloni et al., 2014a). Therefore, this same set will be used in this empirical re-examination. Given the economically intuitive nature of the amelioration set, its inclusion in the Post Double Selection will yield more straightforward and meaningful economic interpretations compared to more complex terms.

5 Methodology

In this section, the setup of the Monte Carlo simulations under different data properties is described, the varying designs of Data Generation Processes are introduced, the necessary procedures for a valid Post Double Selection are outlined, different data-independent and data-dependent amelioration set penalties are introduced, and the metrics for evaluating their performance are presented.

5.1 Monte Carlo Simulations

Given the theoretical focus of this paper and the difficulty in measuring the specific effect of the amelioration set penalties on the complex interactions within the control selection, Monte Carlo simulations are used. Each estimation is run for 1,000 simulations to ensure a sufficiently large sample for valid Monte Carlo properties.

5.1.1 DGP Setup

Following Belloni, Chernozhukov and Hansen (2013), the structure of the DGP is based on the partially linear models (3.3) and (3.4):

$$y_i = d_i'\alpha_0 + x_i'\beta_0 + u_i, \quad u_i \sim \mathcal{N}(o, \sigma_u^2) \tag{5.1}$$

$$d_i = x_i'\eta_0 + v_i, \quad v_i \sim \mathcal{N}(o, \sigma_v^2) \tag{5.2}$$

for a fixed design matrix $X = [x_1, \ldots, x_p]$ containing $p = 200$ regressors across $n = 100$ observations, $x \sim \mathcal{N}(0, \Sigma)$ with $\Sigma_{kj} = (0.5)^{|j-k|}$ for $j, k \in \{1, \ldots, p\}$, and the treatment effect $\alpha_0 = 1$. The design matrix is normalised such that $\mathbb{E}_n[x_{ij}^2] = 1$ for $j \in \{1, \ldots, p\}$, ensuring the equivariance condition is met for the Lasso to produce valid control selection under homoscedastic errors. Each simulation draws new x_i's, u_i's, and v_i's from their respective distributions.
By construction, this partially linear model cannot be estimated using OLS, making it an ideal testing ground for evaluating the sparsity approximation abilities of Post-Double Selection and its extensions in finite samples. Three distinct DGP designs are considered to evaluate control

selection in finite samples. [8] Design 1 is a Linear decay:

$$\beta_0 = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 0, 0, 0, 0, 0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 0, \ldots, 0\right)', \tag{5.3}$$

$$\eta_0 = \left(1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{8}, \frac{1}{9}, \frac{1}{10}, 0, \ldots\ldots\ldots\ldots, 0\right)'. \tag{5.4}$$

Design 2 is a Quadratic decay:

$$\beta_0 = \left(\frac{1}{1}, \frac{1}{2^2}, \frac{1}{3^2}, \frac{1}{4^2}, \frac{1}{5^2}, 0, 0, 0, 0, 0, 1, \frac{1}{2^2}, \frac{1}{3^2}, \frac{1}{4^2}, \frac{1}{5^2}, 0, \ldots, 0\right)', \tag{5.5}$$

$$\eta_0 = \left(1, \frac{1}{2^2}, \frac{1}{3^2}, \frac{1}{4^2}, \frac{1}{5^2}, \frac{1}{6^2}, \frac{1}{7^2}, \frac{1}{8^2}, \frac{1}{9^2}, \frac{1}{10^2}, 0, \ldots\ldots\ldots\ldots, 0\right)'. \tag{5.6}$$

and Design 3 is an extended Quadratic decay:

$$\beta_0 = \left(\frac{1}{1}, \frac{1}{2^2}, \frac{1}{3^2}, \frac{1}{4^2}, \frac{1}{5^2}, 0, 0, 0, 0, 0, 1, \frac{1}{2^2}, \frac{1}{3^2}, \frac{1}{4^2}, \frac{1}{5^2}, , \ldots, \frac{1}{200^2}\right)', \tag{5.7}$$

$$\eta_0 = \left(1, \frac{1}{2^2}, \frac{1}{3^2}, \frac{1}{4^2}, \frac{1}{5^2}, \frac{1}{6^2}, \frac{1}{7^2}, \frac{1}{8^2}, \frac{1}{9^2}, \frac{1}{10^2}, \ldots\ldots\ldots\ldots, \frac{1}{200^2}\right)'. \tag{5.8}$$

The three designs are chosen for their different extents of approximating a sparse representation. Design 1 has the clearest representation, with a sparse design that uses sizable coefficients in s. Design 2 has a similarly sparse representation but with significantly lower coefficients in s. Design 3 deviates slightly from Design 2, including continuously decreasing, yet possibly significant, coefficients throughout the full structural representation. For Design 3, the approximate sparsity condition $s \ll p$ is violated. Due to their structures, it is expected that in Design 1, the vanilla Post-Double Selection will capture the exact number of confounding variables; in Design 2, the estimator will capture only some of the confounding variables; and in Design 3, the estimator will capture too many. Therefore, the amelioration set penalties are expected to produce the greatest gains in the latter two designs.

5.1.2 Error Distribution

The amelioration set penalties are evaluated under both homoscedastic and heteroscedastic error terms, as in practice, perfect homoscedasticity is not assumed. Under the assumption of homoscedasticity, $\sigma_u = \sigma_v = 1$ and $\sigma_{uv} = 0$. Under the assumption of heteroscedasticity, the following specification is considered (Belloni, Chernozhukov & Hansen, 2013):

$$\sigma_d = \sqrt{\frac{(1 + x_i'\beta_0)^2}{\frac{1}{n}\sum_{i=1}^n (1 + x_i'\beta_0)^2}}, \qquad \sigma_y = \sqrt{\frac{(1 + \alpha_0 d_i + x_i'\beta_0)^2}{\frac{1}{n}\sum_{i=1}^n (1 + \alpha_0 d_i + x_i'\beta_0)^2}}, \tag{5.9}$$

where averages of $\sigma_d(x_i)$ and $\sigma_y(d_i, x_i)$ both converge to one. Any other heteroscedastic design can also be employed as long as their averages converge to one, ensuring a fair comparison to the homoscedastic case.

[8] The DGP designs are inspired by a combination of those used in Belloni and Chernozhukov (2011) and Belloni et al. (2014a).

5.1.3 Partial R^2 Parameterization

Recalling from the introduction, in a system with significant noise, it seems plausible that the true effects of fundamental variables are masked, and sparsity-inducing methods tend to spuriously choose controls that are primarily correlated with the noise instead of the outcome variable. To evaluate the effectiveness of using amelioration set penalties to mitigate spurious selection, they are tested under varying levels of noise. More specifically, the DGPs of the different designs are scaled such that the ratio of the explainable model variance to the total variance (also called the R^2 of a model) can be specified through a process called *'partial R^2 parametrization'* (Cinelli & Hazlett, 2020). The parameters c_y and c_d are analytically determined for the intended R^2 in both specifications of the partially linear model, such that the following holds:

$$R_d^2 = \frac{\mathbb{V}(\tilde{x_i}' c_d \eta_0)}{\mathbb{V}(\tilde{x_i}' c_d \eta_0 + v_i)} \tag{5.10}$$

$$R_y^2 = \frac{\mathbb{V}(d_i' \alpha_0 + \tilde{x_i}' c_y \beta_0)}{\mathbb{V}(d_i' \alpha_0 + \tilde{x_i}' c_y \beta_0 + u_i)} = \frac{\mathbb{V}((\tilde{x_i}' c_d \eta_0)' \alpha_0 + \tilde{x_i}' c_y \beta_0)}{\mathbb{V}((\tilde{x_i}' c_d \eta_0)' \alpha_0 + \tilde{x_i}' c_y \beta_0 + u_i)}, \tag{5.11}$$

where resulting values of c_y and c_d are specified in Appendix A. The $R^2 \in \{0, 0.2, 0.4, 0.6, 0.8\}$ values do not necessarily have to be equal, allowing for a 3-dimensional grid to show the behaviour of the different amelioration set penalties across varying levels of noise.

5.2 Post-Double Selection

This section describes the practical methodology for practitioners to correctly initialize and make necessary adjustments to the Post-Double Selection method. [9]

5.2.1 Λ-Simulations

The data-dependent λ-penalty introduced in (3.18) relies on the value of $\Lambda(1-\alpha|X)$, where $\Lambda := 2n \left\| \mathbb{E}n[x_i g_i] \right\| \infty$ is estimated through the interactions of vectors in the design matrix x_i and $g_i \sim \mathcal{N}(0,1)$. $\Lambda(1-\alpha|X)$ is determined as the $(1-\alpha)^{\text{th}}$ percentile of the simulations. The '*hdm*' package recommends 5000 simulations, but preliminary tuning showed that 3000 iterations produce negligible differences while significantly reducing computation time.

5.2.2 σ-Estimation

The other important element in determining λ to ensure correct Lasso functioning is the estimation of $\hat{\sigma}$, such that the property $\left\| \mathbb{E}n\left[x_i \varepsilon_i\right] \right\|_\infty \overset{\text{d.}}{=} \sigma \left\| \mathbb{E}n\left[x_i g_i\right] \right\|_\infty$ can be used. The difficulty lies in that σ is unobserved; hence, an iterative procedure is needed to refine the estimations. This iterative procedure, based on the data-driven penalty, is also crucial for the validity of the asymptotic properties of the post-double-selection estimator (Belloni & Chernozhukov, 2011).

[9] Note that there is already an R package called *'hdm'*, written by one of the authors based on the original paper by Belloni et al. (2014a). However, due to its inflexibility in adjusting the penalties within the pre-made function —a requirement for this thesis' analysis— the code was replicated, with some elements copied and others adjusted (Chernozhukov et al., 2016). See Appendix F for a full description of the various functions used in the code.

The Lasso penalty is directly dependent on the value of $\hat{\sigma}$; therefore, each iteration producing a different estimate affects the λ of the next iteration. This continues until a local optimum $\hat{\sigma}$ is found, and further iterations do not affect the estimate.

The original paper initialises the iterative procedure with a conservative estimate of y_i on a constant. However, this method is adjusted to use a better-informed preliminary OLS regression to estimate $\hat{\sigma}_{I_0} = \sqrt{\mathbb{E}_n[(y_i - X'\beta)]}$. The full iterative procedure to estimate $\hat{\sigma}$ using Post-Lasso iterations is given in Appendix B.

5.3 Amelioration Set Penalties

The concept of incorporating beliefs in approximating sparse representations is complex to simulate in Monte Carlo simulations and is underrepresented in the literature. Hence, the following methodology presents an experimental approach to generating 'beliefs' on which controls should be included in the amelioration set. The primary challenge is avoiding data-snooping bias, where one assumes beliefs that should not be known ex-ante. The goal is to specify a sparse DGP with a relatively small number of known confounding variables and many irrelevant controls, assuming approximate sparsity, and to create a sparse approximation based on 'beliefs' that must be specified ex-ante.

Logically, if the parametric model is known, it is easy to provide correct information to the estimator, but this would be an invalid practice. Conversely, purposefully providing the estimator with incorrect information will deliberately make the estimator perform worse than it otherwise would. [10] Therefore, the resulting methodological setup is experimental. First, the methodology of varying levels of belief accuracy is explained, followed by the different proposed penalization methods.

5.3.1 Generating Ω-beliefs

Recall that the true number of confounding variables in an ASM is denoted by s, and the elements that are 'believed' should receive a lower penalization are denoted by k. Let $\hat{k}$ be the number of controls out of k that actually receive a lower penalization through their reduction in Ω. The following setup is proposed for $p = 200$ under three levels of precision, reflecting the a priori precision of the extent of the knowledge of k on s:

Figure 1 shows the three levels of precision: high (k_h), medium (k_m), and low (k_l), in comparison to s. These levels of precision correspond to the density of precision in the beliefs. A higher level of precision increases the probability of selecting confounding variables. In comparison, a lower level of precision maintains the same probability of selecting confounding variables but also increases the probability of selecting irrelevant controls. This approach reflects the practitioners' methodology: having focused knowledge about which factors drive the DGP versus casting a wider net when such knowledge is lacking.

[10] Consider as an intuitive example trying to determine the treatment effect of an economic policy on inflation. A researcher might have economic intuition about several important confounding controls that should be included and therefore incorporates these beliefs into the Lasso estimations. However, whether these controls are truly confounding will probably be unknown, and therefore, they will not cause data snooping. When one has parametric knowledge, the lower penalization on controls known to be confounding would be considered data-snooping. There is a very fine line to thread in this experimental setup to avoid data snooping.

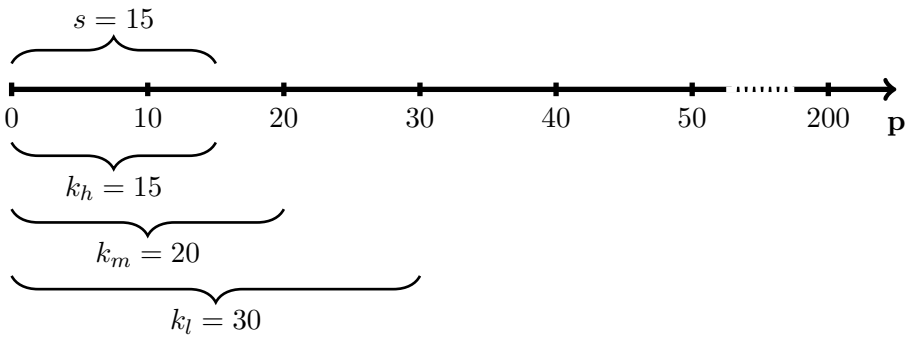


Figure 1. The domain of the amelioration set under varying levels of precision, given $p = 200$

Note that $\hat{k} < k$, as not all elements that are believed to receive a lower penalization can be chosen, allowing for more variety in sets of beliefs across the Monte Carlo simulations and hindering perfect selection of s, regardless of the precision level, therefore reducing the data-snooping. This selection mechanism is implemented as follows: each element in $\hat{k}_i$ is independently drawn from k_i for $i \in \{h, m, l\}$ using a Bernoulli distribution with a probability of 0.5. Therefore, the expected precision of choosing a correct confounding control for $\hat{k}_h$ is 1/2, for $\hat{k}_m$ is 3/8, and for $\hat{k}_l$ is 1/4.

5.3.2 Data-Independent Penalties

The first choice for Ω is the data-independent *'discrete'* penalty, which depends only on the identities of $\hat{k}$ and does not consider other available information in the data. One supporting argument for this penalty is that it counteracts Lasso's potentially spurious reliance on control selection based on correlation with noise. Instead, the penalty is based entirely on the discrete belief regarding whether a control should or should not be included in the amelioration set, guided by economic intuition.

The Discrete penalty, denoted by Ω_D, assigns a penalty of zero if the control is an element of $\hat{k}$ and a penalty of one otherwise. This penalty serves as a baseline comparison to existing literature by providing a robust alternative to the ex-post addition of the amelioration set for the final OLS regression in the Post Double Selection by integrating the amelioration set within the Lasso-estimations and forcing their selection.

A second data-independent penalty is introduced through the *'half-discrete'* penalty, denoted by Ω_{HD}. This penalty functions similarly to Ω_D but assigns a penalty of one-half if the control is an element of $\hat{k}$. It serves as a robustness check to determine whether the complete removal of penalization is necessary for the amelioration set or if a partial penalty, such as one-half, yields similar results. Exploring the effects of scaling the Discrete penalty to other values within its domain is an interesting idea for future research.

$$\Omega_D = \begin{cases} 0, \ i \in \hat{k} \\ 1, \text{ otherwise} \end{cases} \qquad \Omega_{HD} = \begin{cases} 0.5, \ i \in \hat{k} \\ 1, \text{ otherwise} \end{cases} \tag{5.12}$$

5.3.3 Data-Dependent Penalties

Data-dependent penalties are introduced to improve Ω_D by scaling the amelioration set penalties based on the data in the design matrix. The robustness of using these penalties can be argued from two perspectives: On one hand, due to their data dependence, they might succumb to the fallacy of overfitting to noise, similar to the Lasso estimator. On the other hand, selecting a small subset of potentially information-dense controls allows for the exploitation of information that irrelevant controls might have previously masked.

The adaptive penalty, denoted by Ω_A, is based on the logic of the two-step adaptive Lasso (Zou, 2006) [11]. First, an OLS regression is performed with y_i regressed on the controls in $\hat{k}$. The resulting coefficients determine the penalization: the highest coefficients receive the lowest penalization. This approach adapts the amelioration set penalties to the correlation structure of the data. The scaling is determined by taking the multiplicative inverse of the coefficients and scaling the weights between zero and one, such that the controls with the largest coefficients in the preliminary OLS are penalized the least in the Lasso (See 5.13).

The score penalty, denoted by Ω_S, uses information from the score in the problem to set the amelioration set penalties. Similar to how the λ-penalty is proportional to the maximum score: $\mathbf{S} \propto \mathbb{E}_n[x_i \varepsilon_i]$, this approach is applied to setting Ω_S based on the score of each control within $\hat{k}$. Controls with a lower score are penalized less, aligning with the logic of ℓ_1-regularization in the Lasso estimator. The same scaling method used for Ω_A is applied here. Note that the residual vector changes across iterations of estimating $\hat{\sigma}$, so Ω_S should be updated with each iteration.

$$\text{scale}(\zeta_i) := \frac{\zeta_i - \min(\zeta)}{\max(\zeta) - \min(\zeta)} \tag{5.13}$$

$$\Omega_A = \begin{cases} \omega_i \in \text{scale}(b_i^{-1}), & i \in \hat{k} \quad \text{for} \quad b_i = (X_t' X_t)' X_t' y_i, \; X_t = X' \mathbb{I}_{i \in \hat{k}} \\ 1, \text{ otherwise} \end{cases} \tag{5.14}$$

$$\Omega_S = \begin{cases} \omega_i \in \text{scale}(b_i), & i \in \hat{k} \quad \text{for} \quad b_i = \mathbb{E}_n[x_i \hat{\varepsilon}_i] \\ 1, \text{ otherwise} \end{cases} \tag{5.15}$$

where $\mathbb{I}_{i \in \hat{k}}$ is an indicator function that determines whether a regressor in the design matrix is an element of $\hat{k}$. A full sequential overview of all amelioration set penalties is presented in Appendix C.

5.4 Metrics

The metrics used to evaluate the performance of different estimators and penalties align with existing literature (Belloni & Chernozhukov, 2011; Belloni et al., 2014a). These metrics include the treatment effect's bias, the bias's standard deviation, and the rejection probabilities of the 95% confidence intervals. The confidence intervals are calculated using the jackknife standard error estimator, which is preferred over regular heteroscedastic standard error estimators due to its ability to reduce bias, especially in small samples, by systematically leaving out one observation at a time (MacKinnon & White, 1985). This method provides more robust standard

[11] I want to thank Dr. Kaspar Wüthrich for pointing out this idea in an e-mail conversation.

errors in the presence of heteroscedasticity and performs strongly under homoscedasticity, accommodating DGPs with either error distribution. In addition, the jackknife method yields more reliable inferences in finite samples, which is crucial in this research where $n = 100$ (Long & Ervin, 2000).

Therefore, in the final OLS regression of the Post-Double Selection, the residuals can be used to consistently estimate the standard error of the treatment effect. The rejection probability should ideally be 5% to indicate uniformly valid performance. Furthermore, the number of controls in $\hat{I} = \{\hat{I}_1 \cup \hat{I}_2\}$ that are an element of s, and the total number of chosen controls in $\hat{I}$, are examined to evaluate the effect of the amelioration set penalties on better-performing control selection. Ideally, the $\hat{I}$ should converge to s while the total number of selected controls remains low.

6 Simulation Results

The results of the Monte Carlo simulations are presented in three stages. First, a preliminary analysis of the different Lasso estimators is conducted to holistically evaluate their performance and provide intuition for using both the Post-Double Selection and amelioration set penalties. Second, different cases of R^2 are examined to evaluate the performance of the various amelioration set penalties and levels of precision in specific chosen cases, demonstrating that some penalties perform better under certain conditions. Third, generalized results are presented across the full grid of R^2 values to comprehensively assess the amelioration set penalties. Here, a practitioner's recommendation is provided for appropriately using the robust amelioration set penalties under varying conditions.

6.1 Lasso-Estimators for Causal Inference

Table 1 displays the results of the different Lasso-estimators performed on Design 1 with homoscedastic errors, and Figure 2 shows the distribution of the estimated treatment effect, of which the shapes align with those found in existing literature (Belloni et al., 2014a).

Table 1. Simulation results for Lasso-estimators

Estimator	Mean Bias	Std. Dev	rp(0.05)	Relevant Controls	Total Controls
Lasso	0.653	0.088	1.000	4.8	5.3
Post-Lasso	0.434	0.219	0.879	4.8	5.3
Indirect-Lasso	0.087	0.188	0.006	5.4	5.5
Post-Double Selection	-0.017	0.108	0.097	5.4	5.5
Post-Double Selection Oracle	-0.001	0.110	0.052	15.0	15.0
Oracle	-0.002	0.100	0.055	10.0	10.0

Note: The table reports performance metrics of the different Lasso-estimators, described in Section 3.3, based on 1000 simulations of Design 1 with homoscedastic errors, as detailed in Section 5.1.1, using the metrics detailed in Section 5.4.

The Lasso and Post-Lasso estimators show a significant positive bias. The Lasso bias is expected to result from a combination of ℓ_1-regularization bias and omitted variable bias (OVB). In contrast, the Post-Lasso corrects for the regularization bias and only suffers from OVB, resulting in its distinct bimodal shape. This result is clarified in Table 1, where, although

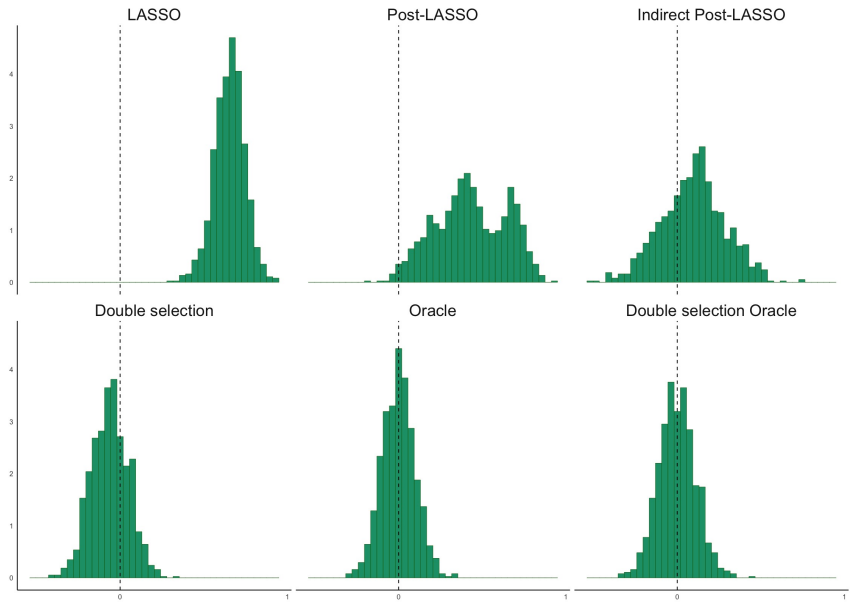


Figure 2. The finite-sample distributions of the different Lasso estimators, described in Section 3.3, are shown. These distributions are based on 1000 simulations of Design 1, as detailed in Section 5.1.1. A dotted vertical line indicates the expected value of the standard normal distribution.

Lasso and Post-Lasso show similarly low numbers of selected relevant and total controls, the regular Lasso has a higher bias than Post-Lasso. In both estimators, the bias dominates the standard error, leading to a high rejection rate and poor inferential capabilities. Note that this distribution is accentuated in this setup where $n \ll p$ under finite-sample assumptions.

The indirect Post-Lasso, an extension of the Post-Lasso with an additional Lasso selection step of d_i on x_i, shows considerably lower bias and no bimodal shape. The additional step of selecting confounding variables for d_i significantly reduces the OVB, although the standard error remains unaffected. Hence, the magnitude of the test statistic is relatively high, and the rejection probability is the lowest among all estimators and far under the 5% target. This is not a favourable result as it indicates that the standard error bounds are too conservative, leading to type II errors where incorrect inferences are unlikely to be rejected.

The Double-Selection estimator corrects this problem by reducing the variability in the estimated treatment effect through an additional Lasso selection step of y_i on x_i. By doing so, the bias is significantly reduced while also halving the standard error. An intuitive explanation for this trend across all estimators is that each additional control selection step detects more relevant confounding variables in the causality problem by conditioning on the relationships of x_i with both y_i and d_i. This relationship is shown in Table 1, where the number of selected relevant controls monotonically increases. Moreover, there are still gains to be made until the Oracle and Double Selection Oracle are reached. The table shows that a perfect selection of confounding variables leads to negligible levels of OVB and an approximately correct rejection probability of 0.05. This result strongly motivates the use of amelioration set penalties, which could assist in better selecting confounding variables to reach favourable oracle-level performance. However, note that Design 1 with homoscedastic errors is specifically designed to be an approximately sparse model with easy-to-control errors and, therefore, provides optimal conditions for Post-Double Selection.

6.2 'Precision' in the Amelioration Set

The implementation of generating 'beliefs' with varying densities of correct information has not resulted in significantly different outcomes across the three levels of precision. Appendix D shows that the differences across precision levels in all metrics are marginal and do not alter the conclusions per precision. The figures do indicate that some of the constructed hypotheses are realized. For instance, as precision increases, the number of selected total controls increases while the number of selected confounding variables converges to s. As a result, the bias decreases slightly, and the coverage is closer to the true value. These graphical results are subtle and difficult to discern. Therefore, 'medium' precision is chosen for further analyses as it represents the midpoint and the conclusions drawn are expected to apply to all precision levels.

The most obvious explanation for the methodology's failure to produce significant results is that different levels of precision only account for varying probabilities of adding more irrelevant controls, whereas the fundamental variables within the DGP are still selected with equal probabilities across all precision levels. Given the strength of Post-Double Selection in selecting correct confounding variables in ASMs, the addition of lower penalizations for irrelevant controls may not have led to significant spurious selection. One possible extension for future research is to calibrate the different precision levels based on both the number of confounding variables and irrelevant controls that are accessible for selection, with higher precision, including a higher ratio of fundamental variables to irrelevant controls.

6.3 Selected R^2 Cases

In this section, the focus shifts to the Post-Double Selection and the amelioration set penalties. Table 2 displays the results of the amelioration set penalties across different designs and error types. The best-performing penalties are identified and contextualized. The Root-Mean-Squared Error (RMSE) is used instead of mean error (bias) to allow negative and positive biases to have an additive effect rather than averaging out.

The specific cases of R^2 values are chosen concerning the existing literature (see Section 2.2), where significantly disparate behaviour is observed between the Post-Lasso and the Post-Double Selection methods based on the different levels of correlation between the confounding variables and the treatment variable, as well as the confounding variables and the outcome. This same analysis is performed. One important difference is that, in the existing literature, the confounding variables in both stages are the same, whereas this is not the case in all current designs. Therefore, these cases provide insight into how the amelioration set penalties, which are purposefully equal in both stages, affect control selection when the correlation is higher in one of the two stages.

Starting with the homoscedastic designs (A, B, and C), a striking observation is that the control estimator performs the worst across all designs and all R^2 cases. The difference is most pronounced in Cases I and III, where the second-stage $R^2 = 0$. This could be because, in the second stage, the relationship between the outcome and the control and treatment variables is difficult to establish, causing the control to struggle in identifying the confounding variables. While some correct confounding variables can be identified in the first stage, the different iden-

Table 2. Simulation Results for Selected R^2 Values

	Case I First Stage $R^2 = 0.2$ Second Stage $R^2 = 0$		*Case II* First Stage $R^2 = 0.2$ Second Stage $R^2 = 0.8$		*Case III* First Stage $R^2 = 0.8$ Second Stage $R^2 = 0$		*Case IV* First Stage $R^2 = 0.8$ Second Stage $R^2 = 0.8$	
Estimation method	RMSE	Rej. Rate	RMSE	Rej. Rate	RMSE	Rej. Rate	RMSE	Rej. Rate
A. Linear decay with homoscedastic errors								
Control	0.130	0.112	0.128	0.110	0.176	0.108	0.123	0.087
Discrete	0.109	0.064	0.120	0.066	0.133	0.073	0.105	0.046
Half-Discrete	0.118	0.086	0.122	0.086	0.149	0.072	0.115	0.066
Adaptive	0.118	0.082	0.123	0.090	0.151	0.081	0.116	0.069
Score	0.115	0.072	0.122	0.067	0.149	0.100	0.109	0.057
B. Quadratic decay with homoscedastic errors								
Control	0.118	0.090	0.116	0.073	0.168	0.101	0.112	0.071
Discrete	0.112	0.070	0.110	0.053	0.132	0.062	0.109	0.059
Half-Discrete	0.114	0.075	0.113	0.060	0.151	0.063	0.111	0.070
Adaptive	0.114	0.072	0.114	0.061	0.150	0.073	0.110	0.065
Score	0.113	0.075	0.110	0.052	0.150	0.075	0.109	0.066
C. Extended Quadratic decay with homoscedastic errors								
Control	0.116	0.087	0.121	0.091	0.160	0.081	0.112	0.067
Discrete	0.110	0.076	0.112	0.070	0.131	0.058	0.104	0.046
Half-Discrete	0.112	0.066	0.118	0.090	0.145	0.054	0.111	0.057
Adaptive	0.113	0.072	0.118	0.085	0.145	0.063	0.110	0.062
Score	0.112	0.079	0.115	0.070	0.148	0.068	0.106	0.049
D. Linear decay with heteroscedastic errors								
Control	0.186	0.078	0.186	0.077	0.201	0.081	0.176	0.084
Discrete	0.167	0.080	0.201	0.086	0.176	0.087	0.170	0.093
Half-Discrete	0.175	0.080	0.182	0.082	0.187	0.080	0.172	0.088
Adaptive	0.175	0.077	0.189	0.089	0.186	0.084	0.174	0.078
Score	0.170	0.080	0.201	0.088	0.185	0.105	0.168	0.095
E. Quadratic decay with heteroscedastic errors								
Control	0.192	0.098	0.174	0.070	0.200	0.080	0.167	0.074
Discrete	0.180	0.083	0.172	0.068	0.184	0.087	0.164	0.079
Half-Discrete	0.186	0.085	0.169	0.064	0.192	0.072	0.166	0.078
Adaptive	0.187	0.083	0.169	0.061	0.194	0.086	0.164	0.077
Score	0.183	0.083	0.173	0.061	0.186	0.077	0.165	0.075
F. Extended Quadratic decay with heteroscedastic errors								
Control	0.183	0.084	0.185	0.085	0.192	0.079	0.173	0.077
Discrete	0.170	0.079	0.184	0.076	0.178	0.076	0.168	0.081
Half-Discrete	0.175	0.070	0.181	0.078	0.187	0.067	0.172	0.087
Adaptive	0.176	0.072	0.180	0.072	0.189	0.077	0.173	0.078
Score	0.172	0.072	0.185	0.075	0.187	0.083	0.168	0.072

Note: The table reports the RMSE and the rejection probabilities of 95% confidence intervals for the Post-Double Selection (control) and the Post-Double Selection with amelioration set penalties, as described in Section 5.3, for selected sets of first-stage and second-stage R^2. A 'medium' precision level is assumed for the amelioration set penalties. The results are based on 1,000 simulations of Designs 1, 2, and 3 with homoscedastic and heteroscedastic errors, as detailed in Section 5.1.1, using the metrics outlined in Section 5.4.

tities of confounding variables in the equations of the partially linear model leave those that don't overlap unidentified in the second stage. [12] The amelioration set penalties, which assume some knowledge of the confounding variables in both stages, help mitigate the issue of missing confounding variables in the second stage.

This conclusion aligns with the observation that the Discrete penalty performs the best across all designs. Forced inclusion through zero penalization effectively helps identify additional confounding variables in both stages. The Half-Discrete penalty, which follows a similar procedure but with less intense penalization, produces slightly weaker results. However, these results also highlight a previously identified shortcoming in the methodology: all precision levels assume too perfect knowledge of the real confounding variables. For instance, if the amelioration set does not accurately reflect the DGP and includes a number of irrelevant controls with lower penalties, adding amelioration set penalties might worsen performance compared to the control.

Moreover, under homoscedastic errors, there is a trend indicating that lower absolute penalization of the amelioration set results in better outcomes. However, the question of whether the information in the data-driven penalties is valuable remains open, as scaling from 0 to 1 might not have been sufficient. Perhaps scaling the data-driven penalties from 0 to 0.5, such that the absolute penalization is closer to zero while still exploiting useful information in the data, might produce better results than the Discrete penalty. This sensitivity to scaling is left open for further research.

With heteroscedastic designs (D, E, and F), the performance of different amelioration set penalties becomes more nuanced. The RMSE strictly increases compared to the homoscedastic errors case, which is expected since the final OLS regression under heteroscedastic errors produces estimates with higher variance. While the control estimator consistently shows weak performance across all cases, the previously dominant performance of the Discrete penalty diminishes. Notably, in Case II, the Discrete penalty performs the worst among all estimators in D and nearly as poorly as the control in E and F. This is particularly interesting because, under homoscedastic errors, the Discrete penalty consistently performed the best for that case. A possible explanation is that the forced selection of the amelioration set distorted the selection of other important controls that might have been highly correlated with the amelioration set, leading to their exclusion in the Lasso-estimation. However, this argument is dubious, as it is exactly the opposite of the intended advantage of the amelioration set penalties.

An important observation is that the Discrete penalty continues to perform well under a second-stage $R^2 = 0$. This makes sense, as in situations with low correlation, the exact structure of the confounding variables is unknown, making any correct information on their identities through the amelioration set valuable. This finding is consistent with literature suggesting that the Post-Double Selection method works best when the second-stage R^2 is low and the first-stage R^2 is high (Belloni et al., 2014b). Therefore, the Discrete penalty shows promise for its integration when precise beliefs about the amelioration set exist. For this reason, the following section will specifically focus on the behaviour of the Control versus the Discrete penalty.

[12] The difference in coefficient designs between the first and second stages likely explains the difference in results of the Post-Double Selection in Belloni et al. (2014b) across different R^2 cases compared to the results in Table 2.

6.4 The Discrete penalty

The trends in design, bias and rejection probability are difficult to establish through Table 2 as both the design and the rejection probabilities do not follow a specific pattern given the selected cases. Therefore, this section introduces a three-dimensional grid in Figure 3 to offer better insight into the strength of the Discrete penalty against the Control under imperfect amelioration set selection. Appendix D displays these grids for all metrics (and all Ω's) [13] across all designs and both error specifications. This section will focus only on the heteroscedastic errors to create valid practitioner recommendations primarily aimed at realistic heteroscedastic data. Additionally, the heteroscedastic results have more visible differences in specific metrics features, allowing for a more intuitive graphical analysis. [14]

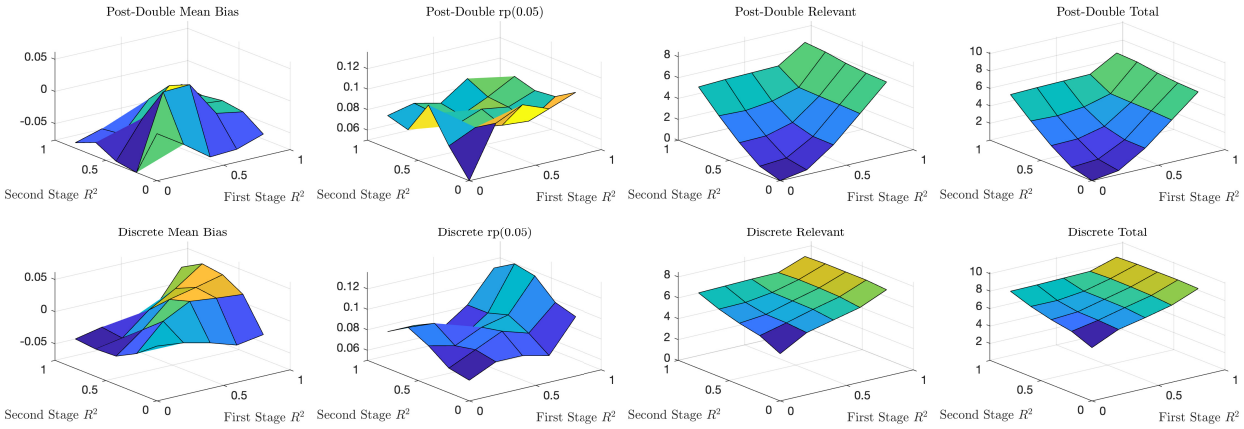
Starting with the Control, the differences between the linear, quadratic, and extended Quadratic decay designs perfectly illustrate how the Post-Double Selection method handles ASMs and sparsely approximates non-ASMs for a low enough approximation error. The difference between the Linear and Quadratic decay designs is expected, as the coefficients in the latter drop much more quickly. Therefore, some controls need not be selected for a valid approximately sparse approximation with a sufficiently low approximation error. Under the assumption that not selecting controls with coefficients below a certain threshold in the true DGP does not significantly alter the approximation error, the sparse approximation of the Linear decay will have to select more controls for a similar approximation.

This leads to several outcomes. First, under unfavourable conditions where correct variable selection is challenging (such as low R^2 in either stage), the Linear decay displays higher bias. This can be attributed to the larger coefficient sizes in the Linear decay, meaning that if a confounding variable is not selected —a likely scenario in this finite sample setup— the resulting OVB will be more significant. This is especially important given the imperfect selection of the full set of confounding variables in the amelioration set for a given level of 'precision'. Second, the larger coefficient sizes in the true DGP of the Linear decay increase the number of significantly large coefficients that need to be included in the ASM, reflected in the number of selected relevant and total controls. Furthermore, the rejection probability of the quadratic designs appears to be closer to the true value, indicating that the behaviour of the Control is more valid for quadratic designs.

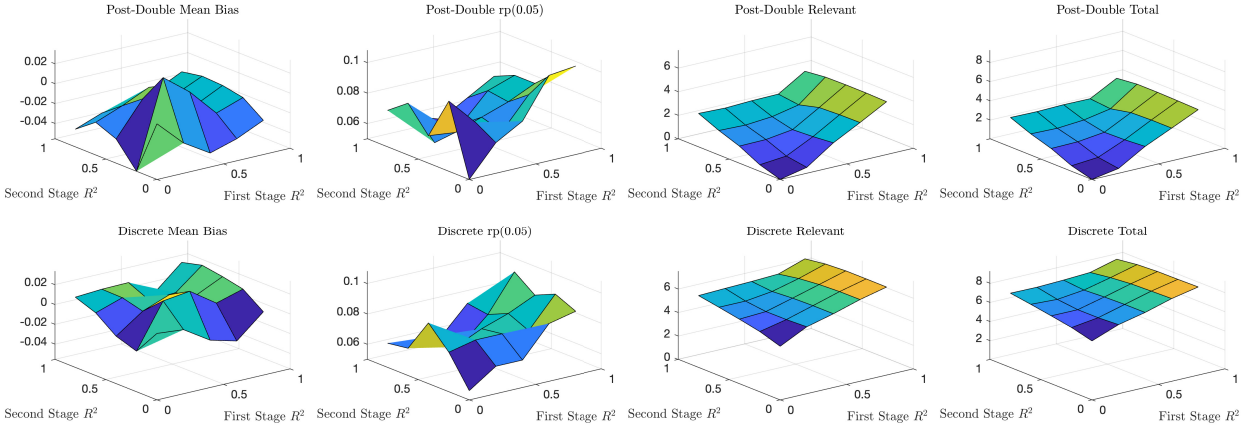
The trend is clear: Quadratic decay leads to a sparser approximation, making it less prone to missing the selection of confounding variables with significant effects in the DGP. This conclusion also explains the similar behaviour between Quadratic decay and extended Quadratic decay. It was unexpected that the extended Quadratic decay, which is not sparse by design, is approximated just as well as a regular Quadratic decay. This is likely due to the threshold idea, where a control with sufficiently small significance in the true DGP is ignored because of its minimal impact on the approximation error. Considering that some controls were already ignored in the Quadratic decay, it is not implausible that all other controls with even smaller

[13] Further analysis of the other amelioration set penalties is an interesting extension to uncover the effects of data-driven versus data-independent penalties. However, considering the scope of this thesis, only the Discrete penalty can feasibly be evaluated in detail.

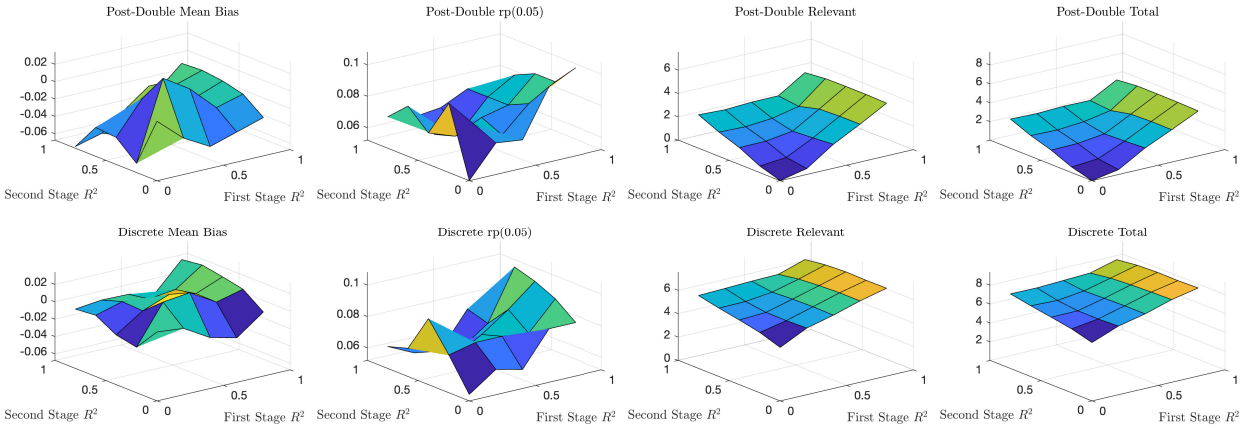
[14] The values of the bias and rejection probability in the heteroscedastic designs have significantly more mountainous, less uniform shapes. This accommodates a more graphically intuitive comparison across the designs.



(a) Design 1 (Linear decay) with Heteroscedastic Errors



(b) Design 2 (Quadratic decay) with Heteroscedastic Errors



(c) Design 3 (extended Quadratic decay) with Heteroscedastic Errors

Figure 3. The figure reports the mean bias, the rejection probabilities of 95% confidence intervals, the number of selected controls in the true DGP, and the total number of selected controls for both the Post-Double Selection (Control) and the Post-Double Selection with amelioration set penalties methods. These penalties are described in Section 5.3 and are evaluated across a grid of first-stage and second-stage R^2. A 'medium' precision level is assumed for the amelioration set penalties. The results are based on 500 simulations of Designs 1, 2, and 3, which include heteroscedastic errors as detailed in Section 5.1.1.

effects are similarly ignored.

The implementation of the Discrete penalty has asymmetrical effects on the Linear and Quadratic decay designs. The bias flattens and converges around zero in the Quadratic decay designs across all values R^2. One explanation is that for all values of R^2, the Discrete penalty forces increased relevant and control selection by removing the Lasso-penalty on select controls, which is reflected in the flat near-uniform surface at a high base-line level in Figure 3. Increased (correct) control selection could have led to a reduced level of OVB.

The most interesting finding is the bias of the Discrete penalty on the Linear decay, which seems to show little improvement over the control in terms of absolute bias. The bias stays at a near-zero level for moderate levels of first-stage and second-stage R^2. However, especially for a high value of second-stage R^2, the bias is volatile to the change in first-stage R^2. This behaviour is more moderate given a lower level of first-stage R^2. The surface strongly resembles that of the Post-Lasso in a study in the literature conducted on a DGP variant of the extended Quadratic decay by Belloni et al. (2014a), which has also been described in Section 2.2. In that case, the Post-Lasso had trouble under a high first-stage R^2 to select all relevant confounding controls in the second-stage estimation. This could be the same here, where the Discrete penalties, which are equal for the first-stage and second-stage equations, hinder the selection of the true confounding variables by wrongly forcing irrelevant controls. Given the higher coefficient values of those controls in the Linear decay and the impact of missing the selection of a relevant control on the bias, this could be the reason for the turbulent bias plot.

Therefore, based on the analysis of the Monte Carlo simulations, the following recommendation for practitioners is offered: data-independent Discrete penalties can yield significant improvements across different ASMs, with performance varying according to the first and second-stage correlations. The optimal results are achieved with a moderate to high first-stage R^2 and a low second-stage R^2. Under favorable conditions, the bias is significantly reduced across all correlations, while under unfavorable conditions, it does not worsen compared to the control.

7 Empirical Case: The Effect of Abortion Rate on Crime

Building on the practitioner's recommendation, the empirical case of Donohue and Levitt (2001) is revisited to evaluate the effect of the amelioration set penalties, particularly the Discrete penalty, on the treatment effect, control selection, and the economic interpretations of the economically intuitive controls. Table 3 presents the different estimations of the treatment effect, while Table 4 shows the coefficient estimates of the economically intuitive variables. These are defined as the set of controls used in the original research. [15] In the appendix, Table E.1 shows the cardinalities of selected relevant and total controls, and Tables E.2 to E.4 show the effect of amelioration set penalties on control selection.

To improve readability, the terms will be referred to as follows: Post-Double Selection will be called the 'Control', Post-Double Selection with the ex-post addition of the amelioration set

[15] This follows the methodology used in previous re-examinations of Donohue and Levitt (2001) using the Post-Double Selection (Belloni et al., 2014b)

will be called the 'Amel-Control', and Post-Double Selection with the Discrete penalty for the amelioration set will be called the 'Discrete'.

The R^2 values of the different first-stage and second-stage regressions across crime types—violent crime, property crime, and murder—have been established in the literature (Belloni et al., 2014b). Respectively, the first-stage R^2 values are 0.8420, 0.6116, and 0.7781, while the second-stage R^2 values are 0.0251, 0.1179, and 0.0039. Based on the practitioner's recommendation, these values correspond to regions of the R^2 grid where the Discrete can be a valuable addition to the Control, assuming some degree of precision in identifying the amelioration set. In the empirical case, it is therefore assumed that the set of economically intuitive variables includes some significant information on the true confounding variables.

Table 3. Estimated treatment effect of abortion on different crimes

	Violent Crime		Property Crime		Murder	
	Effect	Std. Error	Effect	Std. Error	Effect	SE
Donohue and Levitt (2001)	-0.129	0.024	-0.091	0.018	-0.121	0.047
Belloni et al. (2014b)	-0.104	0.107	-0.030	0.055	-0.125	0.151
Belloni et al. (2014b) + ex-post Amel.	-0.082	0.106	-0.031	0.057	-0.068	0.200
Control	-0.158	0.120	-0.024	0.043	-0.117	0.417
Control + ex-post Amel.	-0.133	0.122	-0.028	0.045	-0.103	0.434
Discrete	-0.113	0.120	-0.013	0.044	-0.063	0.421

Note: The table reports the estimated treatment effects of abortion on three types of crimes in the empirical study, including results in literature, for the Post-Double selection, and the Post-Double Selection with the ex-post addition of the amelioration (Amel.) set, under heteroscedastic errors. Standard error estimates are produced using the jackknife method.

The replication results align with those found in the literature, showing similar values for the Control and Amel-Control. As in previous studies, the estimated treatment effects from the original research are significantly different for property crime and slightly for violent crime. All standard errors for the Control and its extensions are large, with 95% confidence intervals spanning a broad positive and negative range. Therefore, drawing valid economic conclusions from these results with sufficient certainty is challenging.

Table E.1 reveals an important observation: none of the Control estimations selects economically intuitive variables. Instead, more complex interactions and time-dependent controls are chosen. For property crime, the Discrete chooses one less variable than the Control but does choose five economically intuitive variables. For murder, the Discrete selects four additional controls, all of which are economically intuitive. Due to the disparity in control selection, the Discrete shows a larger deviation from the Control, with the treatment effect being approximately halved for both property crime and murder. However, since the true effect is unknown, it is difficult to assess this result and conclude whether a higher number of economically intuitive penalties is favourable.

Another interesting observation is that the treatment variable estimates of the Amel-Control surprisingly fall between those of the Control and the Discrete. Therefore, the Discrete penalty, which should be a less extreme and more robust alternative to the Amel-Control, selects controls considerably differently. To uncover the true treatment effect, the treatment variable should be assumed to be exogenous within the large control set, which could be better accommodated

through the alternative selection method of the Discrete penalties. Thus, examining the selected variables across different methods may provide better insights into the correct confounding variable selection.

Several trends on control selection emerge across Tables E.2 to E.4. First, as a robustness check, the variable selection between Belloni et al. (2014b) and the Control is largely similar. The Control generally includes all the controls used in the literature and adds a few extra ones, mostly interactions with time, which already comprise most of the variable selection. Second, as expected, the variables selection between the Amel-Control and the Discrete is very similar. For example, only one variable was selected out of 18 for violent crimes. However, the treatment effect estimates differ significantly, ranging from -0.133 to -0.113. Similarly, for murder, the Control selects three additional controls for a total of 15, while the ex-post amelioration set halves the estimated effect from -0.103 to -0.063. From here, the question arises as to how such a minor difference in Control selection can significantly affect the estimated effect and whether the controls excluded by the integration of the Discrete undermine the exogeneity assumptions of the treatment variable. The disparity in control selection is most pronounced for property crime. The Amel-Control set selects 24 controls, while the Discrete penalty selects a subset of 16 of those controls. Of the eight missing controls, only two are economically intuitive, illustrating the Discrete's tendency to prioritize economically intuitive variables over others.

Table 4. Coefficient estimates of economically intuitive variables

	Violent Crime		Property Crime		Murder	
	Ex-Post	Discrete	Ex-Post	Discrete	Ex-Post	Discrete
D (treatment effect)	-0.133	-0.114	-0.028	-0.013	-0.103	-0.063
Dinc	16.4	23.1	-4.74	-	-230	-
Dpov	0.161	0.152	0.069	0.092	-0.112	0.064
Dafdc	0.034	0.040	-0.016	-0.020	-0.362	-0.368
Dbeer	0.086	0.164	-0.047	-0.099	0.267	-0.109
Dpolice	-0.019	-0.017	-0.026	-0.031	0.337	0.334
Dprison	-0.048	-	-0.019	-0.019	-0.020	-

Note: The table reports the coefficient estimates of the economically intuitive variables in the Post-Double Selection with the ex-post addition of the amelioration set and the Post-Double Selection with the Discrete penalty. A '-' indicates that the variable was not selected.

The coefficient estimates of the selected controls presented in Table 4 allow for a final economic interpretation. Generally, the differences in coefficient estimates are slight, with the maximum deviation being by a factor of two. The only outlier is *'Dinc'*, which represents the change in income. This variable shows a significantly higher estimate but is not selected in two of the Discrete estimations. This is remarkable behaviour since the penalty in the Lasso-estimation is set to zero, yet the Control is not selected, indicating its low confounding potential. This demonstrates the advantage of the amelioration set penalties, which discard such variables and avoid including potentially spurious variables with outlier effects. The table also displays changes in sign for variables such as *'Dbeer'* and *'Dpov'* in the murder estimation. Given that *'Dinc'* was forced into the ex-post amelioration set estimation, its non-selection in the Discrete penalty naturally creates a different interpretation of the economically intuitive variables. Whether this difference in interpretation more closely resembles reality remains unanswered.

8 Conclusion

The Post-Double Selection method effectively reduces bias in causal inference on treatment effects, particularly in settings with high-dimensional controls. Selecting spurious confounding variables can lead to significant bias in noisy conditions and finite samples. This thesis proposes a methodology to mitigate this bias by conditioning on beliefs about the true structural representation through a targeted set of regressors, known as the amelioration set. Various data-dependent and data-independent penalties are evaluated for their effectiveness in reducing bias and improving inference. The Discrete penalty, a robust adaptation of current literature that integrates information on the amelioration set into the Post-Double Selection methodology, has shown potential for enhancing inference across various designs, error distributions, and correlations between the outcome, treatment, and controls. This finding assumes sufficient knowledge of economically intuitive variables driving the DGP. When applied to empirical situations, the Discrete penalty functions as a hybrid of the Post-Double Selection and an extended version with an ex-post amelioration set addition.

Acknowledging the limitations of this research, the Monte Carlo simulation results may rely too heavily on accurate knowledge of the amelioration set, potentially overstating results through data-snooping. Therefore, analysing scenarios with incorrect information would be valuable to determine the robustness of the Post-Double Selection method when faced with a poorly chosen amelioration set. The treatment of error distributions has also been polariseds: simulations either assumed homoscedasticity with appropriate equivariance correction or heteroscedasticity with the same correction. However, the scenario in which the wrong type of correction is applied has not been considered. This is particularly relevant because, in practice, the error distribution is very sensitive and difficult to determine when using such a complex method. In addition, the high rejection probabilities across the simulations suggest that the error might not have been conservative enough. Hence, an idea for future research is to investigate the usage of a fatter-tailed distribution.

The potential of the integrated amelioration set offers an opportunity for its integration into Bayesian statistics through the Bayesian Lasso. The Discrete penalty has proven strong and is easily implemented by modifying the priors for specific controls. A further theoretical extension would be to find conditions under which the integrated amelioration set produces uniformly valid results, a manageable adaptation from the existing proof of the Post-Double Selection.

References

Belloni, A., Chen, D., Chernozhukov, V., & Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, *80*(6), 2369–2429.

Belloni, A., Chernozhukov, V., Fernández-Val, I., & Hansen, C. (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica*, *85*(1), 233–298.

Belloni, A., & Chernozhukov, V. (2011). High dimensional sparse econometric models: An introduction. In P. Alquier, E. Gautier & G. Stoltz (Eds.), *Inverse problems and high-dimensional estimation* (pp. 121–156, Vol. 203). Springer Berlin Heidelberg.

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*(2).

Belloni, A., Chernozhukov, V., & Hansen, C. (2013, June 20). Supplementary appendix for "inference on treatment effects after selection amongst high-dimensional controls". Retrieved June 27, 2024, from http://arxiv.org/abs/1305.6099

Belloni, A., Chernozhukov, V., & Hansen, C. (2014a). Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, *81*(2), 608–650.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014b). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29–50.

Belloni, A., Chernozhukov, V., Hansen, C., & Kozbur, D. (2016). Inference in high-dimensional panel models with an application to gun control. *Journal of Business & Economic Statistics*, *34*(4), 590–605.

Belloni, A., Chernozhukov, V., & Hansen, C. B. (2013). Inference for high-dimensional sparse econometric models. In *Advances in economics and econometrics: Tenth world congress: Volume 3: Econometrics* (pp. 245–295, Vol. 3). Cambridge University Press.

Belloni, A., Chernozhukov, V., & Kato, K. (2019). Valid post-selection inference in high-dimensional approximately sparse quantile regression models. *Journal of the American Statistical Association*, *114*(526), 749–758.

Berrevoets, J., Kacprzyk, K., Qian, Z., & van der Schaar, M. (2024, February 14). Causal deep learning. Retrieved June 19, 2024, from http://arxiv.org/abs/2303.02186

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, *37*(4), 1705–1732.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), C1–C68.

Chernozhukov, V., Hansen, C., & Spindler, M. (2016, August 1). Hdm: High-dimensional metrics. Retrieved May 17, 2024, from http://arxiv.org/abs/1608.00354

Chernozhukov, V., Hansen, C., & Spindler, M. (2015). Valid post-selection and post-regularization inference: An elementary, general approach. *Annual Review of Economics*, *7*(1), 649–688.

Chernozhukov, V., Newey, W. K., & Singh, R. (2022). Automatic debiased machine learning of causal and structural effects. *Econometrica*, *90*(3), 967–1027.

Chinn, M. D., Meunier, B., & Stumpner, S. (2023, June). Nowcasting world trade with machine learning: A three-step approach. Retrieved June 27, 2024, from https://www.nber.org/papers/w31419

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(1), 39–67.

Desboulets, L. D. D. (2018). A review on variable selection in regression analysis. *Econometrics*, *6*(4), 45.

Dhar, D., Jain, T., & Jayachandran, S. (2022). Reshaping adolescents' gender attitudes: Evidence from a school-based experiment in india. *American Economic Review*, *112*(3), 899–927.

Donohue, J. J., III, & Levitt, S. D. (2001). The impact of legalized abortion on crime. *The Quarterly Journal of Economics*, *116*(2), 379–420.

Gillen, B. J., Shum, M., & Moon, H. R. (2014). Demand estimation with high-dimensional product characteristics. In *Bayesian model comparison* (pp. 301–323, Vol. 34). Emerald Group Publishing Limited.

Hangartner, D., Kopp, D., & Siegenthaler, M. (2021). Monitoring hiring discrimination through online recruitment platforms. *Nature*, *589*(7843), 572–576.

Heij, C., Boer, P. D., Franses, P. H., Kloek, T., & Dijk, H. K. V. (2004). *Econometric methods with applications in business and economics*. Oxford University PressOxford.

Lenza, M., Moutachaker, I., & Paredes, J. (2023). Density forecasts of inflation: A quantile regression forest approach. *Available at SSRN 4511273*.

Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(3), 217–224.

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*(3), 305–325.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica*, *62*(6), 1349–1382.

Park, T., & Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686.

Qiu, Y., Chen, X., & Shi, W. (2020). Impacts of social and economic factors on the transmission of coronavirus disease 2019 (COVID-19) in china. *Journal of Population Economics*, *33*(4), 1127–1172.

Sakurai, Y., & Chen, Z. (2024). Forecasting tail risk via neural networks with asymptotic expansions. https://www.imf.org/-/media/Files/Publications/WP/2024/English/wpiea2024099-print-pdf.ashx

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, *58*(1), 267–288.

Urminsky, O., Hansen, C., & Chernozhukov, V. (2019, September 13). The double-lasso method for principled variable selection. Retrieved June 19, 2024, from https://osf.io/2pema

Wainwright, M. J. (2019). Sparse linear models in high dimensions. In *High-dimensional statistics: A non-asymptotic viewpoint* (pp. 194–235). Cambridge University Press.

Wüthrich, K., & Zhu, Y. (2023). Omitted variable bias of lasso-based inference methods: A finite sample analysis. *The Review of Economics and Statistics*, *105*(4), 982–997.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429.

A Derivations Partial R^2 Parameterization

The following derivation is based off of the supplementary material (Belloni, Chernozhukov & Hansen, 2013). Introduce the partially linear model:

$$y_i = d_i'\alpha_0 + x_i'(c_y\beta_0) + \sigma_y(d_i, x_i)u_i \tag{A.1}$$

$$d_i = x_i'(c_d\eta_0) + \sigma_d(x_i)v_i \tag{A.2}$$

where $X = [x_1, \ldots, x_p]$ is a $(n \times p)$ matrix with Variance-Covariance matrix Σ, $\eta_0 = (\eta_{0,1}, \ldots, \eta_{0,p})'$ and $\beta_0 = (\beta_{0,1}, \ldots, \beta_{0,p})'$ are $(p \times 1)$ vectors, $u_i \sim \mathcal{N}(0,1)$, $v_i \sim \mathcal{N}(0,1)$. σ_y and σ_d are dependent on the error specification of the Design. Constants $c_d \in \mathbb{R}$ and $c_y \in \mathbb{R}$ are introduced to scale the R^2 of Equations (A.2) and (A.1) respectively.

A.1 First-Stage Parametrization

$$R_d^2 \coloneqq \frac{\mathbb{V}(\text{Explained Variance in Equation (A.2)})}{\mathbb{V}(\text{Total Variance in Equation (A.2)})} \tag{A.3}$$

$$= \frac{\mathbb{V}(x_i'(c_d\eta_0))}{\mathbb{V}(x_i'(c_d\eta_0) + \sigma_d(x_i)v_i)} \tag{A.4}$$

$$= \frac{c_d^2\eta_0'\Sigma\eta_0}{c_d^2\eta_0'\Sigma\eta_0 + \mathbb{V}(\sigma_d(x_i)v_i)} \tag{A.5}$$

$$\tag{A.6}$$

which can be rewritten as:

$$c_d = \sqrt{\frac{\sigma_d(x_i)v_i \cdot R_d^2}{(1 - R_d^2)\eta_0'\Sigma\eta_0}} \tag{A.7}$$

A.2 Second-Stage Parametrization

$$R_y^2 \coloneqq \frac{\mathbb{V}(\text{Explained Variance in Equation (A.1)})}{\mathbb{V}(\text{Total Variance in Equation (A.1)})} \tag{A.8}$$

$$= \frac{\mathbb{V}(x_i'(c_d\eta_0 + c_y\beta_0))}{\mathbb{V}(x_i'(c_d\eta_0 + c_y\beta_0) + \sigma_y(d_i, x_i)u_i)} \tag{A.9}$$

$$= \frac{(c_d\eta_0 + c_y\beta_0)'\mathbb{V}(x_i)(c_d\eta_0 + c_y\beta_0)}{(c_d\eta_0 + c_y\beta_0)'\mathbb{V}(x_i)(c_d\eta_0 + c_y\beta_0) + \mathbb{V}(\sigma_y(d_i, x_i)u_i)} \tag{A.10}$$

where the numerator can be expanded as:

$$(c_d\eta_0 + c_y\beta_0)'\mathbb{V}(x_i)(c_d\eta_0 + c_y\beta_0) = c_d^2\eta_0'\Sigma\eta_0 + 2c_dc_y\eta_0'\Sigma\beta_0 + c_y^2\beta_0'\Sigma\beta_0 \tag{A.11}$$

such that

$$R_y^2 = \frac{c_d^2\eta_0'\Sigma\eta_0 + 2c_dc_y\eta_0'\Sigma\beta_0 + c_y^2\beta_0'\Sigma\beta_0}{c_d^2\eta_0'\Sigma\eta_0 + 2c_dc_y\eta_0'\Sigma\beta_0 + c_y^2\beta_0'\Sigma\beta_0 + \mathbb{V}(\sigma_y(d_i, x_i)u_i)} \tag{A.12}$$

$$0 = (R_y^2 - 1)c_y^2\beta_0'\Sigma\beta_0 + (R_y^2 - 1)2c_dc_y\eta_0'\Sigma\beta_0 + (R_y^2 - 1)c_d^2\eta_0'\Sigma\eta_0 + \mathbb{V}(\sigma_y(d_i, x_i)u_i) \tag{A.13}$$

which can be written as a quadratic model $ax^2 + bx + c = 0$:

$$a = (R_y^2 - 1)\beta_0' \Sigma \beta_0 \tag{A.14}$$

$$b = (R_y^2 - 1)2c_d \eta_0' \Sigma \beta_0 \tag{A.15}$$

$$c = (R_y^2 - 1)c_d^2 \eta_0' \Sigma \eta_0 + \mathbb{V}(\sigma_y(d_i, x_i)u_i) \tag{A.16}$$

which can be solved using the quadratic formula. Note that the second stage parametrization is directly dependent on the value of c_d, meaning it should be performed in the specified order.

B Post-Lasso $\hat{\sigma}$-Estimation

The iterative Post-Lasso estimation of $\hat{\sigma}$ is adjusted from the methodology of Belloni and Chernozhukov (2011), accomodating for the inclusion of Ω:

Algorithm 1 Estimation of σ using Post-Lasso iterations

Input:
Positive number ψ
Small constant $\nu > 0$ (tolerance level)
Constant $K > 1$ (upper bound on the number of iterations)
Output:
Estimate of σ
Initialization:
Set $k = 0$
Set initial $\hat{\sigma} = \psi \hat{\sigma}_{I_0}$
Iterative Procedure:
while convergence not achieved and $k < K$ **do**
Compute the Post-Lasso estimator $\hat{\beta}$ based on $\lambda = 2c\hat{\sigma}_k \Lambda(1 - \gamma | X)$, $\Psi = 1$ (homoscedastic) or $\sqrt{\mathbb{E}_n[x_i^2 \varepsilon_i^2]}$ (heteroscedastic), and (if specified) Ω.
Set $\hat{\sigma}_{k+1} = \hat{Q}(\hat{\beta})$
if $|\hat{\sigma}_{k+1} - \hat{\sigma}_k| \leq \nu$ **or** $k \geq K$ **then**
Set $\bar{\sigma} = \hat{\sigma}_{k+1}$
Break
else
Increment k by 1 ($k = k + 1$)
end if
end while
Final Output:
Report $\bar{\sigma}$ as the estimate of σ

where ψ was recommended in the *'hdm'*-package to be 0.75. Preliminary tuning has indicated that this value of ψ gives similarly strong results for homoscedastic and heteroscedastic data. The constant c is recommended to be 1.1.

C Overview of Ω-penalties

$$\Omega_C = \Big\{1,\ i \in p \tag{C.1}$$

$$\Omega_D = \begin{cases} 0,\ i \in \hat{k} \\ 1, \text{ otherwise} \end{cases} \tag{C.2}$$

$$\Omega_{HD} = \begin{cases} 0.5,\ i \in \hat{k} \\ 1, \text{ otherwise} \end{cases} \tag{C.3}$$

$$\Omega_A = \begin{cases} \omega_i \in \text{scale}(b_i^{-1}), \quad i \in \hat{k} \ \text{ for } \ b_i = (X_t' X_t)' X_t' y_i,\ X_t = X' \mathbb{I}_{i \in \hat{k}} \\ 1, \text{ otherwise} \end{cases} \tag{C.4}$$

$$\Omega_S = \begin{cases} \omega_i \in \text{scale}(b_i), \quad i \in \hat{k} \ \text{ for } \ b_i = \mathbb{E}_n[x_i \hat{\varepsilon}_i] \\ 1, \text{ otherwise} \end{cases} \tag{C.5}$$

where the subscripts are defined as follows: C = Control, D = Disfrete, HD = Half-Discrete, A = Adaptive, S = Score.

D Monte Carlo Simulation Results on Amelioration Set Penalties

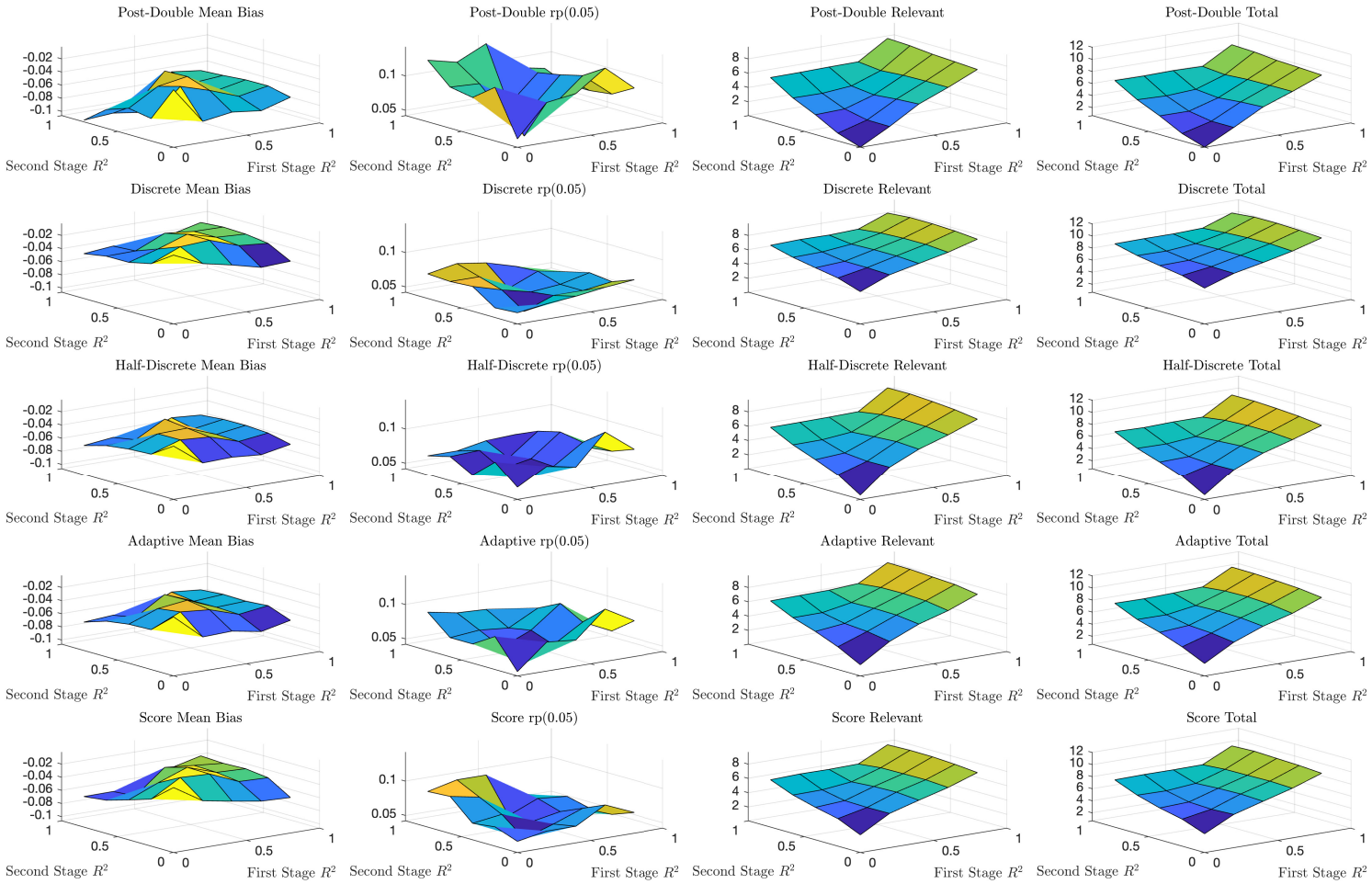


Figure D.1. The figure reports the mean bias, the rejection probabilities of 95% confidence intervals, the number of selected controls in the true DGP, and the total number of selected controls for both the Post-Double Selection (control) and the Post-Double Selection with amelioration set penalties methods. These penalties are described in Section 5.3 and are evaluated across a grid of first-stage and second-stage R^2. A 'medium' precision level is assumed for the amelioration set penalties. The results are based on 500 simulations of **Design 1**, which includes **homoscedastic errors** as detailed in Section 5.1.1.

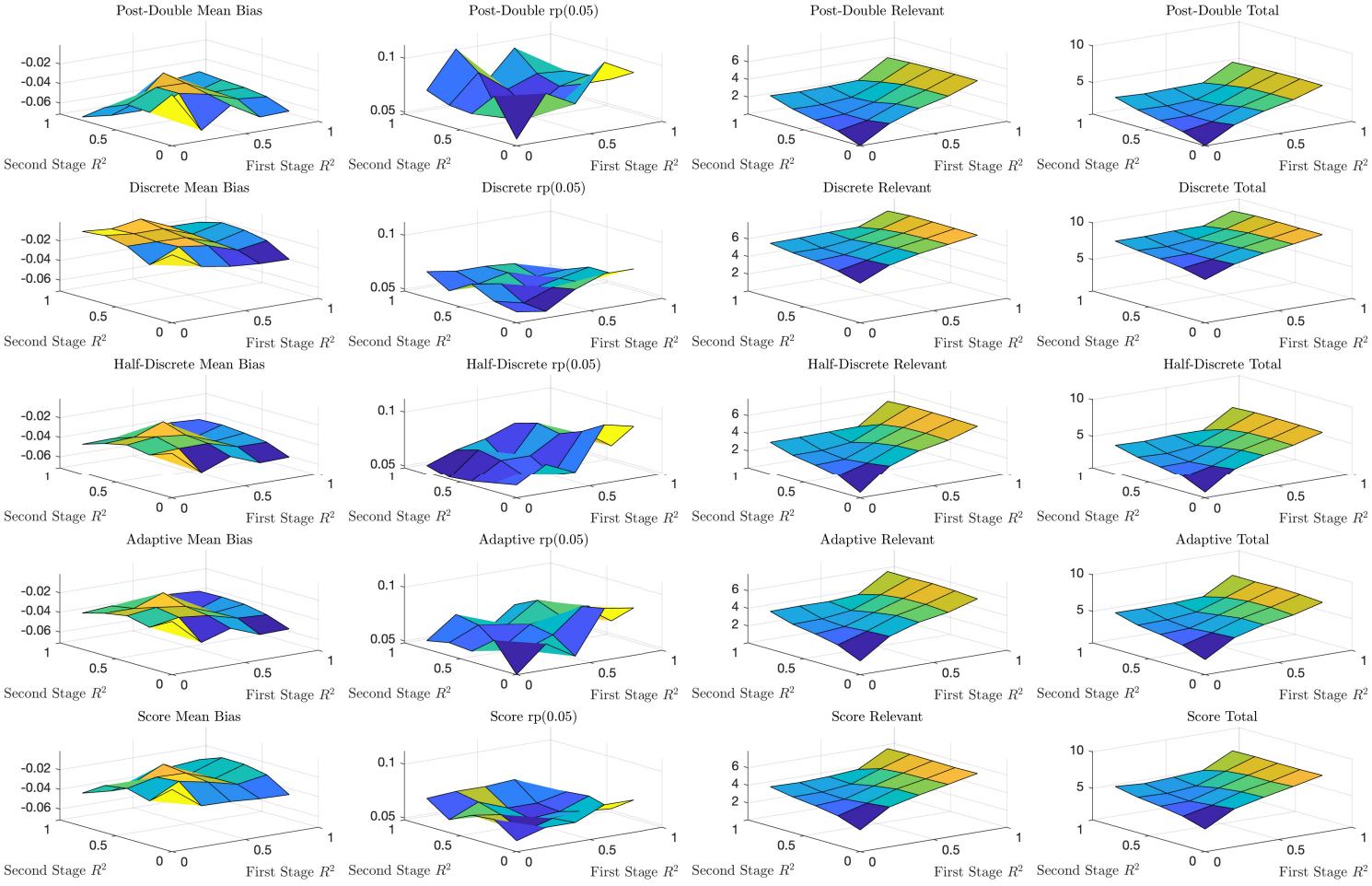


Figure D.2. The figure reports the mean bias, the rejection probabilities of 95% confidence intervals, the number of selected controls in the true DGP, and the total number of selected controls for both the Post-Double Selection (control) and the Post-Double Selection with amelioration set penalties methods. These penalties are described in Section 5.3 and are evaluated across a grid of first-stage and second-stage R^2. A 'medium' precision level is assumed for the amelioration set penalties. The results are based on 500 simulations of **Design 2**, which includes **homoscedastic errors** as detailed in Section 5.1.1.

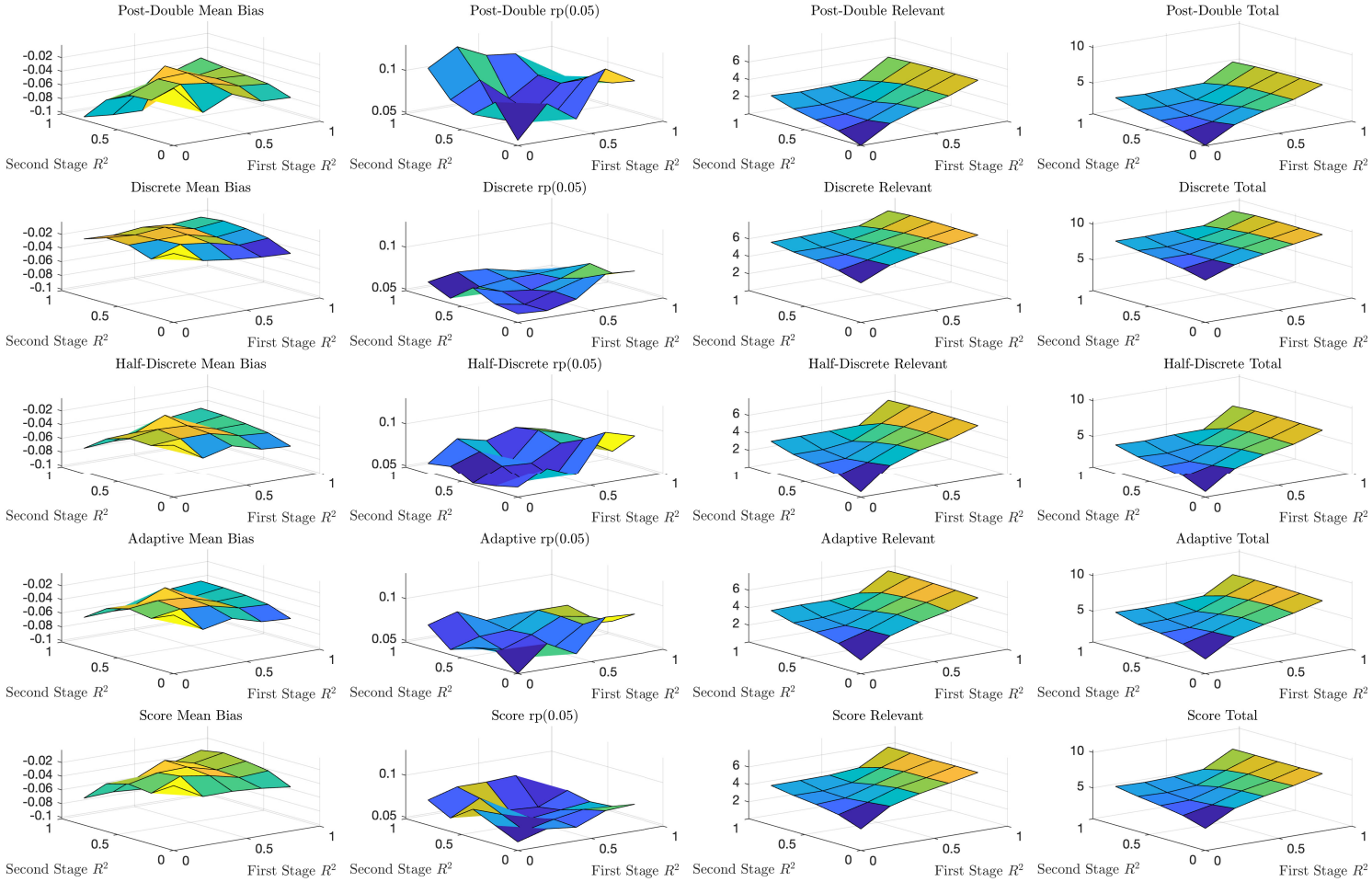


Figure D.3. The figure reports the mean bias, the rejection probabilities of 95% confidence intervals, the number of selected controls in the true DGP, and the total number of selected controls for both the Post-Double Selection (control) and the Post-Double Selection with amelioration set penalties methods. These penalties are described in Section 5.3 and are evaluated across a grid of first-stage and second-stage R^2. A 'medium' precision level is assumed for the amelioration set penalties. The results are based on 500 simulations of **Design 3**, which includes **homoscedastic errors** as detailed in Section 5.1.1.

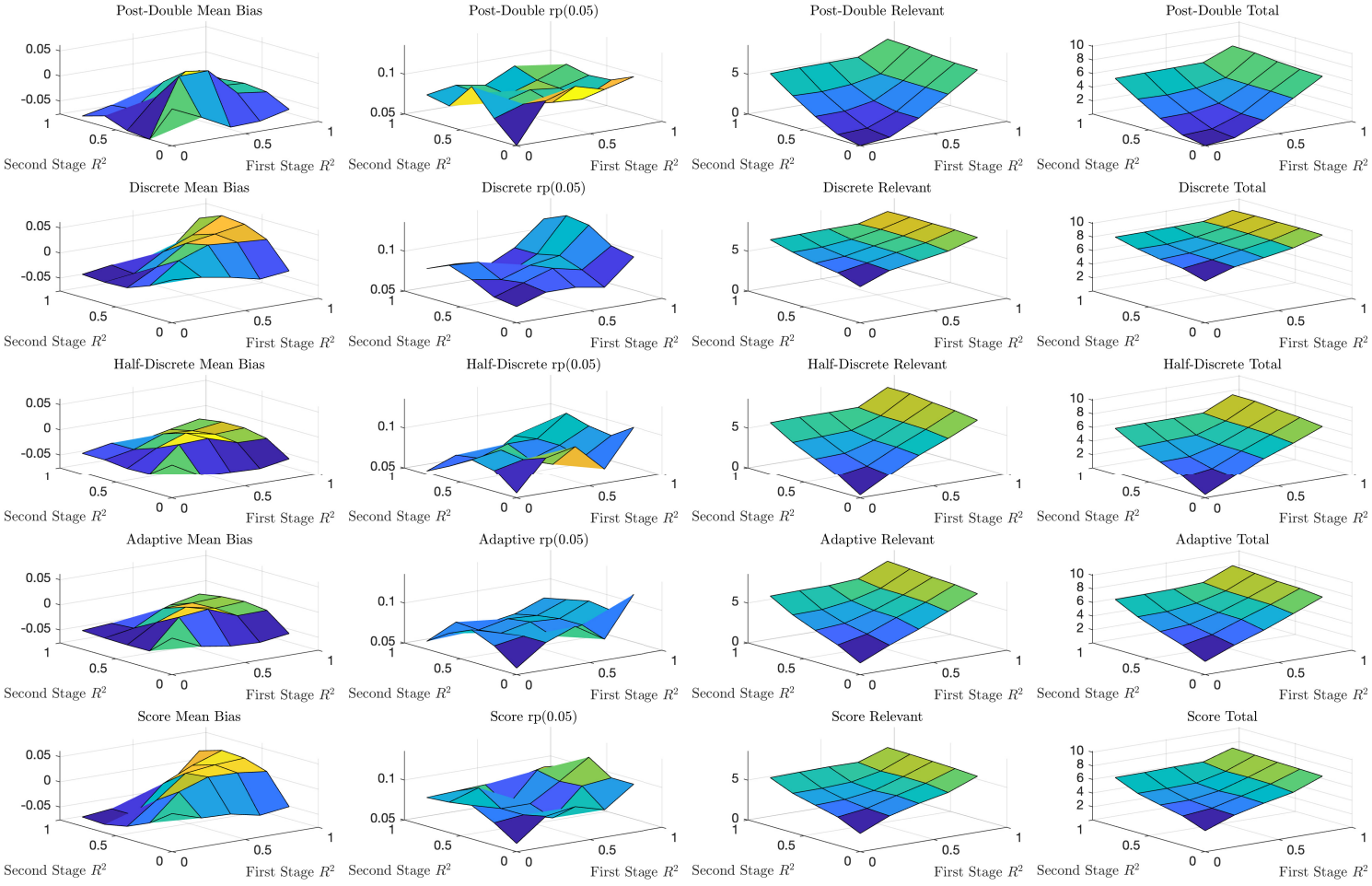


Figure D.4. The figure reports the mean bias, the rejection probabilities of 95% confidence intervals, the number of selected controls in the true DGP, and the total number of selected controls for both the Post-Double Selection (control) and the Post-Double Selection with amelioration set penalties methods. These penalties are described in Section 5.3 and are evaluated across a grid of first-stage and second-stage R^2. A 'medium' precision level is assumed for the amelioration set penalties. The results are based on 500 simulations of **Design 1**, which includes **heteroscedastic errors** as detailed in Section 5.1.1.

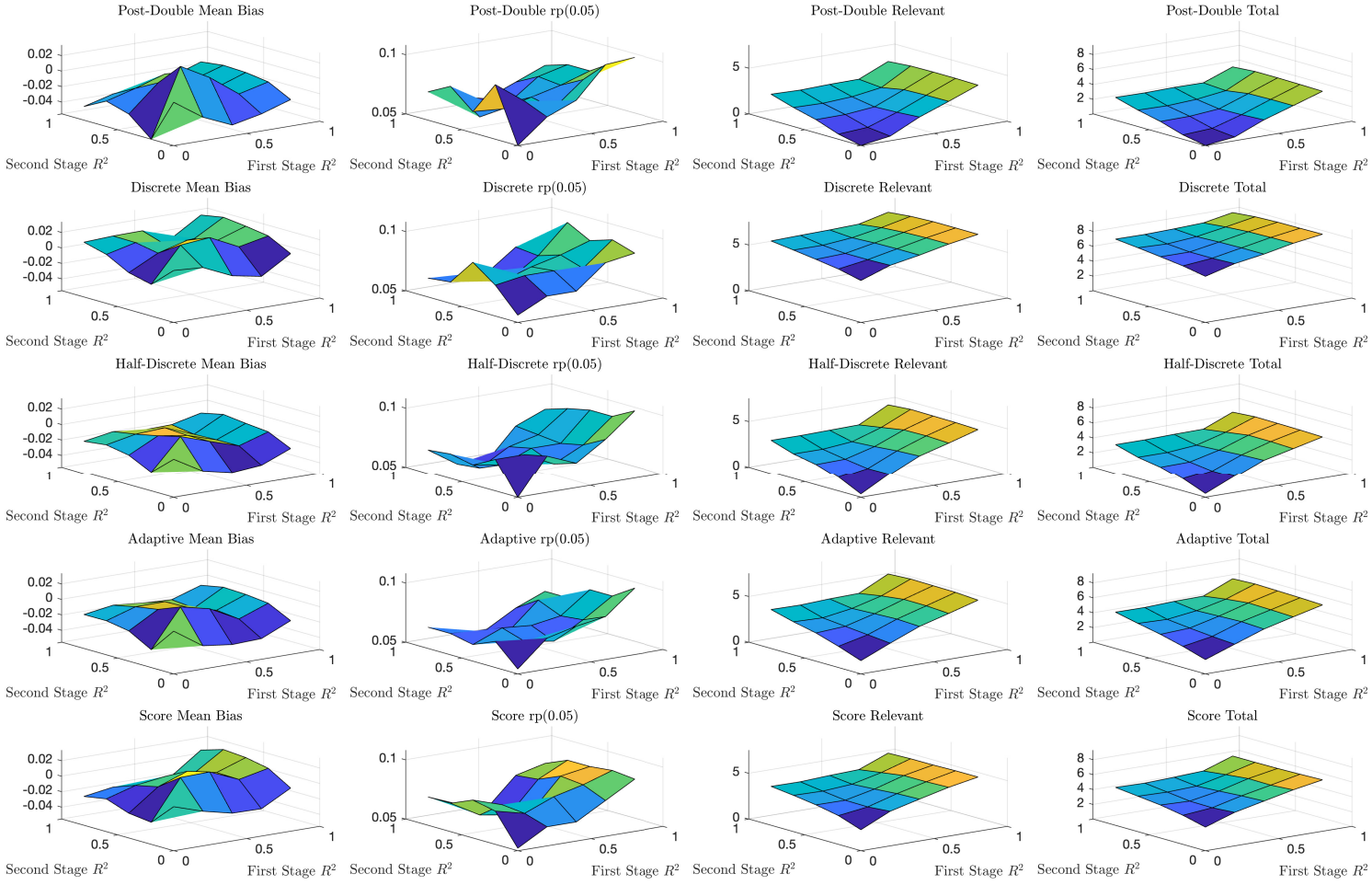


Figure D.5. The figure reports the mean bias, the rejection probabilities of 95% confidence intervals, the number of selected controls in the true DGP, and the total number of selected controls for both the Post-Double Selection (control) and the Post-Double Selection with amelioration set penalties methods. These penalties are described in Section 5.3 and are evaluated across a grid of first-stage and second-stage R^2. A 'medium' precision level is assumed for the amelioration set penalties. The results are based on 500 simulations of **Design 2**, which includes **heteroscedastic errors** as detailed in Section 5.1.1.

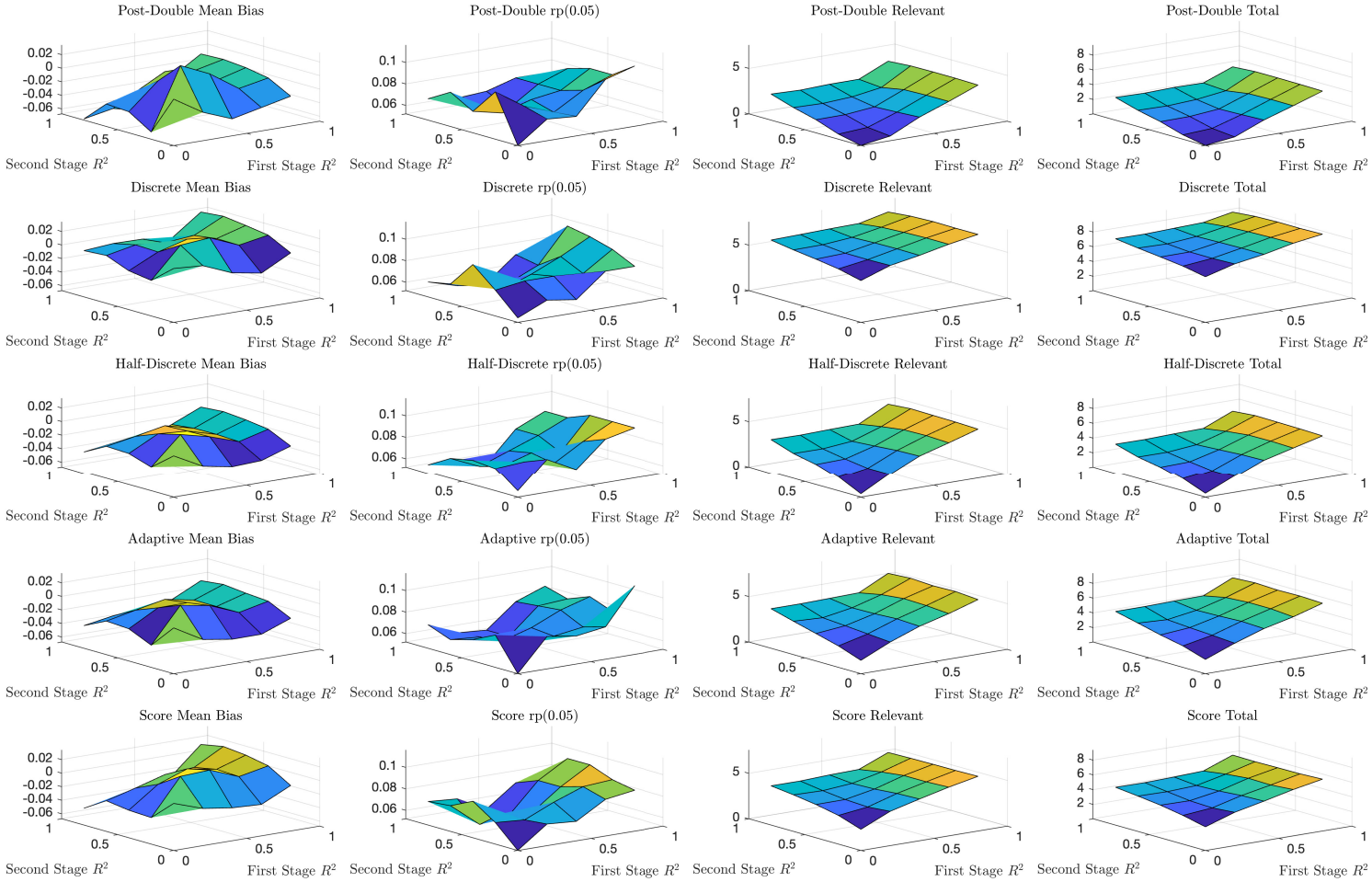


Figure D.6. The figure reports the mean bias, the rejection probabilities of 95% confidence intervals, the number of selected controls in the true DGP, and the total number of selected controls for both the Post-Double Selection (control) and the Post-Double Selection with amelioration set penalties methods. These penalties are described in Section 5.3 and are evaluated across a grid of first-stage and second-stage R^2. A 'medium' precision level is assumed for the amelioration set penalties. The results are based on 500 simulations of **Design 3**, which includes **heteroscedastic errors** as detailed in Section 5.1.1.

E Empirical Variable Selection

Table E.1. Cardinalities of selected relevant and total controls in the empirical case

	Violence		Property		Murder	
	Relevant	Total	Relevant	Total	Relevant	Total
Control	0	15	0	17	0	8
Control + ex-post Amel	7	20	7	24	7	15
Discrete	6	20	5	16	4	12
Half-Discrete	0	12	1	13	0	8
Adaptive	2	13	2	14	0	8
Score	3	15	2	13	2	10

Note: Relevant and total number selected for the control, the control with the ex-post addition of the amelioration set, and all amelioration set penalties in the empirical case under the assumption of heteroscedastic errors are described in Section 7. Relevant control selection refers to the selection of economically intuitive variables.

Table E.2. Control selection per amelioration set penalty: Violence

Control	**Belloni**	**Control**	**Control + Amel**	**Discrete**	**Half-Discrete†**	**Adaptive**	**Score**
Lprison	✓	✓	✓	✓	✓	✓	✓
Lur	✓	✓	✓	✓	✓	✓	✓
Dbeer0 × t	✓	✓	✓	✓	✓	✓	✓
incBar	✓	✓	✓	✓	✓	✓	✓
incBar × t	✓	✓	✓	✓	✓	✓	✓
xV0	✓	✓	✓	✓	✓	✓	✓
Dinc0^2 × t	✓						
LprisonBar × t	✓						
Lpolice		✓	✓	✓	✓	✓	✓
Dinc0 × t		✓	✓	✓	✓	✓	✓
$\text{Lbeer0} \times \text{t}^2$		✓	✓	✓	✓	✓	✓
$\text{Lprison0}^2 \times \text{t}^2$		✓	✓	✓	✓	✓	✓
Linc0× t		✓	✓	✓	✓		✓
Dur		✓	✓	✓			
Dpov		✓	✓	✓			✓
Dinc0		✓	✓	✓			
Linc0		✓	✓		✓	✓	
Dinc			✓	✓		✓	✓
Dbeer			✓	✓		✓	
Dpolice			✓	✓			✓
Dafdc			✓	✓			
Dprison			✓				
$\text{Dbeer}^2 \times \text{t}^2$				✓			✓

Note: Selected controls in the empirical case of the study on the effect of abortion on violent crime, including the control, the control with the ex-post addition of the amelioration set, and all amelioration set penalties under the assumption of heteroscedastic errors, are described in Section 7. The amelioration set is constructed from economically intuitive variables.

Table E.3. Control selection per amelioration set penalty: Property

Control	**Belloni**	**Control**	**Control + Amel**	**Discrete**	**Half-Discrete**	**Adaptive**	**Score**
Lprison	✓	✓	✓	✓	✓	✓	✓
Dinc0	✓	✓	✓	✓	✓	✓	✓
Linc0	✓	✓	✓	✓	✓	✓	✓
Dbeer0 × t	✓	✓	✓	✓	✓	✓	✓
incBar	✓	✓	✓	✓	✓	✓	✓
xP0	✓	✓	✓	✓	✓	✓	✓
Linc	✓	✓	✓		✓	✓	
$Dinc0^2$ × t	✓	✓	✓				
incBar × t	✓	✓	✓				
$DincBar0^2$ × t	✓	✓	✓				
afdcBar	✓	✓	✓				
$afdcBar^2$	✓	✓	✓				
Lpolice		✓	✓	✓	✓	✓	✓
Lur		✓	✓	✓	✓	✓	✓
Dinc0 × t		✓	✓	✓	✓	✓	✓
$Lprison0^2$ × ^{2}		✓	✓	✓	✓	✓	✓
$Lbeer0^2$ × t^2		✓	✓	✓	✓	✓	✓
Dur			✓	✓	✓	✓	✓
Dpolice			✓	✓		✓	
Dpov			✓	✓			✓
Dprison			✓	✓			
Dafdc			✓	✓			
Dbeer			✓				
Dinc			✓				
$Lprison0^2$ × t					✓		

Note: Selected controls in the empirical case of the study on the effect of abortion on property crime, including the control, the control with the ex-post addition of the amelioration set, and all amelioration set penalties under the assumption of heteroscedastic errors, are described in Section 7. The amelioration set is constructed from economically intuitive variables.

Table E.4. Control selection per amelioration set penalty: Murder

Control	**Belloni**	**Control**	**Control + Amel**	**Discrete**	**Half-Discrete**	**Adaptive**	**Score**
Lur	✓	✓	✓	✓	✓	✓	✓
Lprison0 × t	✓	✓	✓	✓	✓	✓	✓
Dbeer0 × t^2	✓	✓	✓	✓	✓	✓	✓
incBar× t	✓	✓	✓	✓	✓	✓	✓
xM0	✓	✓	✓	✓	✓	✓	✓
xM0 × t	✓	✓	✓	✓	✓	✓	✓
prisonBar × t	✓						
$Dur0^2$	✓						
Lprison	✓						
Linc0 × t		✓	✓	✓	✓	✓	✓
policeBar × t		✓	✓	✓	✓	✓	✓
Dpolice			✓	✓			
Dpov			✓	✓			✓
Dbeer			✓	✓			✓
Dafdc			✓	✓			
Dprison			✓				
Dur			✓				
Dinc			✓				

Note: Selected controls in the empirical case of the study on the effect of abortion on murder, including the control, the control with the ex-post addition of the amelioration set, and all amelioration set penalties under the assumption of heteroscedastic errors, are described in Section 7. The amelioration set is constructed from economically intuitive variables.

F Code Description

The description of the code is structured according to its affiliation with one of the following categories: General, Replication, Monte Carlo Simulations, Empirical Case, Results, and Computing.

1. **General Functions**
   (a) externalFunctions.R: Includes code sourced from the 'hdm' package and supplementary material (Belloni, Chernozhukov & Hansen, 2013; Chernozhukov et al., 2016). Functions include: initial values estimation for the Post-Lasso Iterations, the LassoShooting function, and the Post-Estimator function.
   (b) auxFunctions.R: Contains general functions that are always imported. Functions include: generation of matrices given parameter inputs, partial R^2 parametrization, simulation of $\Lambda(1-\alpha|X)$, jackknife standard error estimation (MacKinnon & White, 1985), and control selection metrics.
   (c) auxPenalties.R: Generates 'beliefs' on the amelioration set penalties given a precision level.
   (d) auxAlgorithms.R: Contains different Post-Lasso iteration algorithms to estimate σ, conditional on the type of amelioration set penalty and error distribution.
2. **Replication Functions**
   (a) mainReplication.R: A wrapper for all Lasso estimations to run in the replication (Lasso, Post-Lasso, Indirect Post-Lasso, Post-Double Selection, Oracle, Double-Selection Oracle). Collects all the metrics afterwards.
   (b) auxFunctionsReplication.R: A subset of functions from auxFunctions.R tailored to the replication wrapper.
3. **Simulation Functions**
   (a) mainSimulations.R: A wrapper for all amelioration set penalties simulations to run in the extension.
4. **Empirical Case Functions**
   (a) mainExtension.R: A wrapper for all amelioration set penalties applications to run on different empirical case data.
   (b) auxFunctionsExtensions.R: A subset of functions from auxFunctions.R tailored to the empirical case.
5. **Graphical Functions**
   (a) Results.R: Generates results based on simulations of the replication and the extension.
   (b) ResultsCases.R: Generates results based on different R^2 cases.
   (c) Results3D.R: Generates results based on the 3D grid, to be exported with ExportMATLAB.R.
   (d) ExportMATLAB.R: Exports the 3D grid results to a .csv file, to create graphics in MATLAB.
   (e) Various MATLAB files: Generate different figures using the MATLAB surface function.
6. **Computing Functions**
   (a) pcSims.R: Standard code snippet for distributed computing. For reference, one standard university computer can perform 1000 simulations in approximately 2 hours. Therefore, a (5x5) 3D grid for 3 precision levels, 2 error distributions, and 3 designs takes approximately 900-1000 hours. This can be best distributed among a large number of computers.