INTERNATIONAL ECONOMIC REVIEW
Vol. 0, No. 0, March 2023 DOI: 10.1111/iere.12633

GOODHART'S LAW AND MACHINE LEARNING: A STRUCTURAL PERSPECTIVE*

By Christopher A. Hennessy and Charles A. E. Goodhart

London Business School, UK; London School of Economics, UK

We develop a simple structural model to illustrate how penalized regressions generate Goodhart bias when training data are clean but covariates are manipulated at known cost by future agents. With quadratic (extremely steep) manipulation costs, bias is proportional to Ridge (Lasso) penalization. If costs depend on absolute or percentage manipulation, the following algorithm yields manipulation-proof prediction: Within training data, evaluate candidate coefficients at their respective incentive-compatible manipulation configuration. We derive analytical coefficient adjustments: slopes (intercept) shift downward if costs depend on percentage (absolute) manipulation. Statisticians ignoring manipulation costs select socially suboptimal penalization. Model averaging reduces these manipulation costs.

1. INTRODUCTION

Recent years have witnessed increased use of machine learning (ML) in allocating resources. For example, ML is used to predict corporate defaults. Scoring algorithms are used to allocate consumer credit. ML is used to predict trading patterns, risk premia, and valuations. Big data are used in forecasting insurer outlays. Browsing history is used to predict consumer behavior. Importantly, such usage of ML brings it directly into tension with *Goodhart's law* that states:

> Any observed statistical regularity will tend to collapse once pressure is placed upon it for control purposes.

In this article, we develop a structural framework for analyzing Goodhart's law in penalized regressions. How does Goodhart bias relate to underlying functional forms? Does penalization mitigate Goodhart bias? How can Goodhart bias be corrected—and how does the correction vary with functional forms? Finally, what are the welfare implications of alternative ML procedures?

We consider a statistician with clean training data. The statistician acts as a Stackelberg leader, performing penalized regression to develop a public prediction model for future agents who maximize the difference between their predicted outcome and manipulation costs. Costs can depend on absolute manipulation or relative (percentage) manipulation.

We first analyze naive prediction that fails to directly account for manipulation. With quadratic absolute manipulation costs, Goodhart bias and manipulation costs are proportional to the sum of squared regression coefficients. Bias converges to the sum of the absolute value of coefficients as cost functions become steeper. Thus, tighter Ridge/Lasso coefficient constraints represent a rough method for mitigating Goodhart bias.

*Manuscript received September 2021; revised July 2022.

We thank colleagues at London Business School and the London School of Economics for helpful suggestions. Please address correspondence to: Christopher A. Hennessy, London Business School, Regent's Park, London NW1 6DT, United Kingdom.. E-mail: *chennessy@london.edu*.

We next solve for Stackelberg equilibria in which the statistician estimates on clean data and then develops a manipulation-proof model that predicts equally well against manipulated covariates. Here, the intercept of the manipulation-proof model shifts downward to account for gaming of covariates featuring absolute manipulation costs, with slope coefficients moving downward to account for gaming of covariates featuring relative manipulation costs. As shown, these corrections emerge from a *general algorithm* for addressing Goodhart's law: (1) Fix candidate coefficients. (2) Impute to training data agents' incentive compatible manipulation. (3) Choose coefficients to minimize the objective given manipulation.

We also highlight an externality: Statisticians ignore costs incurred by agents and choose smaller penalty parameters than a planner who places weight on manipulation costs. Ultimately, the economy is trapped in an inefficient equilibrium where agents engage in wasteful data manipulation despite the manipulation being futile. However, model averaging reduces manipulation costs, with greater reductions the greater the distance between averaged models.

Frankel and Kartik (2022) consider a variant on Frankel and Kartik (2019) where agents manipulate and heterogeneity prevents elimination of bias. They show that a principal should precommit to ex post inefficiency. They consider univariate Ordinary Least Squares (OLS), whereas we speak to ML prediction with arbitrary penalization and potentially more covariates than observations. In addition, we provide analytical formulas for coefficient adjustments under a broader set of technologies. Finally, we analyze how model averaging affects welfare. Ball (2022) considers a multivariate setting in which the principal commits to underweight some covariates to deter manipulation but overweights others, so the score is correct on average.

Björkegren et al. (2020) consider absolute quadratic manipulation costs demonstrating their method with Monte Carlo simulations. We present a host of complementary analytical results: linking Ridge and Lasso penalization to Goodhart bias, analytic expressions for manipulation-proof coefficients under both absolute and relative cost functions, and cost reductions under model averaging. Conversely, they offer real-world experimental implementation.

Bruckner and Scheffner (2011) and Hardt et al. (2015) analyze binary classifier problems, whereas we analyze continuous outcomes. This gives rise to different manipulation strategies and equilibria. Our setting yields closed-form bias expressions that we map to ML penalization. Further, Bruckner and Scheffner consider only quadratic costs, whereas Hardt et al. only consider costs expressible as $\max\{0, g(x_2) - f(x_1)\}$.

Dekel et al. (2010) and Chen et al. (2018) identify mechanisms inducing truthful reporting in training data. Our article considers the opposite scenario where historical training data are clean and future data are manipulated. Eliaz and Spiegler (2019) consider a setting in which incentives would seem to be aligned because the principal's objective is to predict the agent's most preferred outcome. Nevertheless, with Lasso, the agent may have an incentive to misreport, given that his report only matters if the covariate's coefficient is not zero.

Kay and King (2020) write, "The availability of what are now called 'big data'–the very large databases permitted by the power of modern computers–increases these dangers. The existence of a historic data set does not yield a basis for calculating a future probability distribution." Our article is intended to give a simple formal articulation of such dangers expressed in the language of ML.

Fernández-Villaverde (2021) identifies three significant challenges to ML: data requirements; eliciting truthful revelation; and the critique of Lucas (1976). On the final challenge, he anticipates our approach in writing, "Thus, any variation in policy renders previous observations useless, unless we have a structural model (i.e. an economic model that is explicit about preferences, technology, and information sets) the researcher can use to recompute the optimal responses to the new policy." Essentially, our article provides an illustration of how this prescription can be made operational in ML. Here it is worth noting that, rather than taking head-on the second challenge of eliciting truthful revelation, our proposed algorithm anticipates optimal dishonest responses and undoes manipulation by adjusting intercept and/or slope coefficients depending on cost functional forms.

The article is organized as follows: Section 2 describes the setting. Section 3 (4) considers absolute (relative) manipulation costs. Section 5 considers auxiliary implications.

2. THE ECONOMIC SETTING

Agents live for one period. There is one historical cohort and one future cohort. A linear prediction model will be developed using training data from the historical cohort and applied to the future agents. Both cohorts consist of M agents with member m. The outcome for agent m is y_m. The set of covariates has dimension J, with x_{jm} denoting the *true covariate* j for agent m, whereas $\mathbf{X}$ denotes the $M \times J$ true covariate matrix. For both cohorts, the data-generating process is:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\beta + \varepsilon, \\ \varepsilon &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}). \end{aligned} \tag{1}$$

The statistician observes $\mathbf{y}$ and $\mathbf{X}$ for the historical cohort. For the future cohort, the statistician only observes the respective $\widetilde{x}_{jm}$ representing the *manipulated covariate*:

$$\widetilde{x}_{jm} \equiv x_{jm} + a_{jm}, \tag{2}$$

where a_{jm} is manipulation. There is a known convex cost C_j to manipulating each covariate j.

We consider a Stackelberg game where the statistician acts as first mover by posting a *prediction model* mapping manipulated covariates $\widetilde{\mathbf{x}}_m$ into predicted outcomes:

$$\widehat{y}_m \equiv \left[\beta^p + \sum_{j=1}^{J} \beta_j^p \widetilde{x}_{jm} \right]. \tag{3}$$

A Stackelberg equilibrium $(\mathbf{a}^*, \beta^p, \widehat{\beta})$ satisfies three criteria. First, for all possible $\mathbf{x}_j$, manipulation $\mathbf{a}^*$ is *incentive compatible*, maximizing the difference between the predicted outcome and manipulation costs:

$$\mathbf{a}^*(\beta^p) \in \quad \arg\max_{\mathbf{a}} \ \beta^p + \sum_{j=1}^{J} \beta_j^p (x_j + a_j) - \sum_{j=1}^{J} C_j(x_j, a_j). \tag{4}$$

Second, for all possible $\mathbf{x}_j$, the posted model β^p is *manipulation proof* in delivering identical predictions against manipulated covariates as that delivered by the training data model $\widehat{\beta}$ defined over true covariates:

$$\beta^p + \sum_{j=1}^{J} \beta_j^p (x_j + a_j^*) = \widehat{\beta} + \sum_{j=1}^{J} \widehat{\beta}_j x_{jm}. \tag{5}$$

Finally, $\widehat{\beta}$ is an *optimal training data model* given observed $(\mathbf{y}, \mathbf{X})$ for the historical cohort:

$$\widehat{\beta} \in \arg\min_{\mathbf{B}} \left\{ \sum_{m=1}^{M} \left(y_m - B - \sum_{j=1}^{J} B_j x_{jm} \right)^2 + P(\mathbf{B}; \lambda) \right\}. \tag{6}$$

The preceding penalized regression objective (6) can be motivated under the assumption that the statistician is a Bayesian or a frequentist. [1] For example, consider Ridge and Lasso penalization, respectively:

$$P(\mathbf{B};\lambda_R) \equiv \lambda_R \sum_{j=1}^{J} B_j^2, \tag{7}$$

$$P(\mathbf{B};\lambda_L) \equiv \lambda_L \sum_{j=1}^{J} \left|B_j\right|.$$

From a Bayesian perspective, Ridge delivers the maximum a posteriori probability given the prior

$$\beta \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma^2}{\lambda_R}\mathbf{I}\right).$$

Alternatively, Lasso delivers the posterior mode under a double exponential prior centered at zero.

Ridge and Lasso penalization can be motivated from a frequentist perspective since there exist optimal tuning parameters (λ_R, λ_L) such that Ridge and Lasso deliver lower mean-squared prediction (MSPE) error than OLS. Since OLS is BLUE, Ridge and Lasso achieve this by introducing bias to reduce estimate variance. A model is said to be overfitted if sub-optimal penalization results in low bias but high estimator variance, resulting in large MSPE.

Finally, it is worth stressing that implicit in this definition of Stackelberg equilibrium is that the penalty parameter λ in Equation (6) is not a free variable. Rather, if the statistician is a Bayesian, then λ is fixed by her prior over β. If the statistician is a frequentist, the penalty parameter λ in Equation (6) must be understood as being set at its optimal (equilibrium) value, for example, λ minimizes MSPE. [2]

3. ABSOLUTE MANIPULATION COSTS

This section considers *absolute manipulation cost functions*:

$$C_j(a) \equiv \frac{c_j}{\gamma}|a|^\gamma \text{ where } c_j > 0 \text{ and } \gamma > 1.$$

Confronted with model β^p, agent m solves program (4). From the first-order conditions, it follows:[3]

$$\beta_j^p \geq 0 \Rightarrow a_j^* = \left(\frac{\beta_j^p}{c_j}\right)^{\frac{1}{\gamma-1}}, \tag{8}$$

$$\beta_j^p < 0 \Rightarrow a_j^* = -\left(-\frac{\beta_j^p}{c_j}\right)^{\frac{1}{\gamma-1}}.$$

Note that the preceding equation implies that, ceteris paribus, agents have a stronger incentive to manipulate those covariates entering the prediction model with larger loadings.

[1] See Parviero (2017).

[2] See Parviero (2017) for a detailed discussion of empirical approaches to optimal tuning.

[3] Convex costs imply concavity of the objective function.

3.1. *Goodhart Bias.* Consider first the Goodhart bias that emerges if the statisticians were to utilize a *naive prediction model* such that:

$$\widehat{y}_m = \widehat{y}_m^{naive} \equiv \widehat{\beta} + \sum_{j=1}^{J} \widehat{\beta}_j \widetilde{x}_{jm}.$$

That is, a naive prediction model takes the coefficients obtained from the clean training data (Equation (6)) reflecting true covariates $\mathbf{x}$, and reflexively applies them to the future cohort's manipulated covariates $\widetilde{\mathbf{x}}$. From Equation (8) and the fact that with naive prediction $\beta^p = \widehat{\beta}$, we obtain the following lemma:

LEMMA 1 (ABSOLUTE COSTS). *If the future cohort faces a naive prediction model parameterized with the coefficients $\widehat{\beta}$ obtained from the historical cohort, the prediction for agent m will be*

$$\widehat{y}_m = \left[\widehat{\beta} + \sum_{j=1}^{J} \widehat{\beta}_j x_{jm}\right] + \sum_{j=1}^{J} \left(\frac{1}{c_j}\right)^{\frac{1}{\gamma-1}} \left|\widehat{\beta}_j\right|^{\left(\frac{\gamma}{\gamma-1}\right)} \tag{9}$$

and each agent incurs manipulation costs equal to

$$\frac{1}{\gamma} \sum_{j=1}^{J} \left(\frac{1}{c_j}\right)^{\frac{1}{\gamma-1}} \left|\widehat{\beta}_j\right|^{\left(\frac{\gamma}{\gamma-1}\right)}. \tag{10}$$

Notice that the square bracketed term in Equation (9) captures the prediction the naive statistician believes she is making, mapping true x_{jm} into predicted values as she did in the training data. The final term captures *Goodhart bias*.

With Lemma 1 in mind, recall that Ridge and Lasso coefficients can be expressed as:

$$\begin{aligned} \widehat{\beta}_R &\in \arg\min_{\mathbf{B}} \left\{ \sum_{m=1}^{M} \left(y_m - B - \sum_{j=1}^{J} B_j x_{jm} \right)^2 \right\} \quad s.t. \quad \sum_{j=1}^{J} B_j^2 \le t_R \\ \widehat{\beta}_L &\in \arg\min_{\mathbf{B}} \left\{ \sum_{m=1}^{M} \left(y_m - B - \sum_{j=1}^{J} B_j x_{jm} \right)^2 \right\} \quad s.t. \quad \sum_{j=1}^{J} \left|B_j\right| \le t_L. \end{aligned} \tag{11}$$

In fact, there is a one-to-one mapping between the respective penalty parameters λ in Equation (6) and the constraint tuning parameters t in Equation (11).

From Lemma 1, we have the following proposition expressing Goodhart bias in terms of Ridge and Lasso constraint tuning parameters:

PROPOSITION 1 (ABSOLUTE COSTS). *Suppose manipulation costs are quadratic and $c_j = c$ for all j. Then under naive Ridge regression:*

$$\widehat{y}_m = \widehat{\beta} + \sum_{j=1}^{J} \widehat{\beta}_j x_{jm} + \frac{t_R}{c}, \tag{12}$$

and each future cohort agent will incur manipulation costs equal to $t_R/\gamma c$. If, instead, naive Lasso regression with constraint tuning parameter t_L is employed and $c_j = 1$ for all j, then in the limit as γ approaches infinity:

$$\widehat{y}_m = \widehat{\beta} + \sum_{j=1}^{J} \widehat{\beta}_j x_{jm} + t_L. \tag{13}$$

Recall that the frequentist motivation for opting for penalized regressions is achieving lower MSPE by reducing model variance (overfitting) in exchange for bias arising from departing from OLS. However, from the preceding proposition, it follows true MSPE for the future cohort is:

$$\begin{aligned} \text{(14)} \quad MSPE &\equiv \mathbb{E}\left\{(\widehat{y} - \mathbb{E}[y])^2\right\} \\ &= \mathbb{E}\left[\left(\widehat{\beta} + \sum_{j=1}^{J} \widehat{\beta}_j x_j - \mathbb{E}[y]\right)^2\right] + 2\left(\frac{t}{c}\right)\mathbb{E}\left[\widehat{\beta} + \sum_{j=1}^{J} \widehat{\beta}_j x_j - \mathbb{E}[y]\right] + \left(\frac{t}{c}\right)^2. \end{aligned}$$

The first term immediately above captures MSPE absent data manipulation. Standard formulations of optimal tuning select t to minimize this first term. The second and third terms capture the effect of Goodhart bias (t/c) on MSPE. Evidently, with data manipulation, there is an added benefit to choosing lower coefficient constraints t, aside from reduction of variance: reduction in the Goodhart bias component of MSPE. This argument notwithstanding the next subsection shows how Goodhart bias can be directly eliminated, with no need for adjustment of tuning parameters.

3.2. *Equilibrium.* Under absolute costs, manipulation is independent of true covariates. With this in mind, we conjecture and verify a Stackelberg equilibrium in which the statistician corrects Goodhart bias by making an adjustment to the prediction model intercept while utilizing the slopes from the original penalized regression program (6). Under this conjecture, the prediction for future cohort agents will be:

$$\begin{aligned} \widehat{y}_m &= \beta^p + \sum_{j=1}^{J} \widehat{\beta}_j [x_{jm} + a^*_{jm}] \\ &= \left[\beta^p + \sum_{j=1}^{J} \left(\frac{1}{c_j}\right)^{\frac{1}{\gamma-1}} |\widehat{\beta}_j|^{\left(\frac{\gamma}{\gamma-1}\right)}\right] + \sum_{j=1}^{J} \widehat{\beta}_j x_{jm}. \end{aligned} \tag{15}$$

Notice, the preceding equation implies that manipulation proofness (5) will be achieved if the term in squared brackets is equal to the training data intercept $\widehat{\beta}$. We thus have the following proposition:

Proposition 2 (Absolute Costs). *Let $\widehat{\beta}$ be the coefficient vector obtained from the training data according to Equation (6). There is a Stackelberg equilibrium in which statisticians form predictions for the future cohort according to*

$$\widehat{y}_m = \beta^p + \sum_{j=1}^{J} \beta^p_j \widetilde{x}_{jm}, \tag{16}$$

with slopes $\beta_j^p = \widehat{\beta}_j \; \forall \; j = 1, \ldots, J$ *and intercept:*

$$\beta^p = \widehat{\beta} - \sum_{j=1}^{J} \left(\frac{1}{c_j}\right)^{\frac{1}{\gamma-1}} |\widehat{\beta}_j|^{\left(\frac{\gamma}{\gamma-1}\right)}. \tag{17}$$

Notice that the preceding proposition implies that as one moves from clean training data to future data, data manipulation will cause regression intercepts to shift downward. Further, the downward shift is increasing in the absolute value of historical slope coefficients. Intuitively, historical slopes will carry over to the future, and the incentive for future manipulation is increasing in the absolute value of the slopes. Thus, the intercept must shift down by a greater amount to counteract more severe data manipulation.

3.3. *Social Welfare and Model Tuning.* Consider a statistician who uses Ridge regression with the objective of minimizing the MSPE. As shown in Proposition 2, even with manipulation, in equilibrium, it is *as-if* the statistician observes the true covariate matrix **X**. With this in mind, MSPE is:

$$\begin{aligned} MSPE(\lambda, \beta) &= \sigma^2 tr[\mathbf{X}\mathbf{\Omega}_\lambda(\mathbf{X}'\mathbf{X})^{-1}\mathbf{\Omega}_\lambda'\mathbf{X}'] + \beta'(\mathbf{\Omega}_\lambda - \mathbf{I}_J)'\mathbf{X}'\mathbf{X}(\mathbf{\Omega}_\lambda - \mathbf{I}_J)\beta \\ \mathbf{\Omega}_\lambda &\equiv [\mathbf{I}_J + \lambda(\mathbf{X}'\mathbf{X})^{-1}]^{-1}. \end{aligned} \tag{18}$$

The privately optimal penalty parameter λ_R^* minimizes MSPE.[4] Assuming the objective in the preceding program is globally convex, the first-order condition pins down λ_R^*:

$$\frac{\partial MSPE(\lambda_R^*, \beta)}{\partial \lambda} = 0.$$

Consider instead a planner who computes social losses as a weighted sum of the MSPE plus manipulation costs. Proposition 1 allows us to write the social planner's program as

$$\begin{aligned} \lambda_R^{**} &\in \arg\min_{\lambda} \quad \omega_E MSPE(\lambda, \beta) + \omega_C \frac{t_R(\lambda)}{\gamma c} \\ &\Rightarrow \frac{\partial MSPE(\lambda_R^{**}, \beta)}{\partial \lambda} = -\frac{\omega_C}{\omega_E}\frac{t_R'(\lambda_R^{**})}{\gamma c} > 0 \\ &\Rightarrow \lambda_R^{**} > \lambda_R^*. \end{aligned} \tag{19}$$

The first > sign above follows from $t_R' < 0$. The final > sign above follows from convexity of the MSPE function. We thus have the following proposition:

PROPOSITION 3. *Suppose agents face homogeneous absolute manipulation cost functions, with the statistician minimizing MSPE under Ridge regression, whereas the social planner minimizes a weighted sum of the MSPE and manipulation costs. Then the privately optimal penalty parameter is less than the socially optimal penalty parameter.*

4. RELATIVE MANIPULATION COSTS

In many contexts, for example, financial statements, it might be more reasonable to assume that costs depend on the magnitude of manipulation relative to true covariates. With this in

[4] In practice, plug-ins (e.g., OLS) are used to estimate β. For brevity, we consider theoretical optima.

mind, this section assumes that true covariate (x_{jm}) values are positive and considers that agents face *relative manipulation cost functions*:

$$C_j(a) \equiv \frac{c_j}{\gamma}\left|\frac{a}{x}\right|^{\gamma} x. \tag{20}$$

$$c_j > 0 \text{ and } \gamma > 1.$$

4.1. *Naive Prediction.* With relative manipulation costs, the first-order conditions for each agent's program (4) imply that:[5]

$$\begin{aligned} \beta_j^p \geq 0 &\Rightarrow a_j^* = \left(\frac{\beta_j^p}{c_j}\right)^{\frac{1}{\gamma-1}} x_j \\ \beta_j^p < 0 &\Rightarrow a_j^* = -\left(-\frac{\beta_j^p}{c_j}\right)^{\frac{1}{\gamma-1}} x_j. \end{aligned} \tag{21}$$

Notice that manipulation here is similar to that under absolute manipulation costs (8), but now manipulation is scaled by the true covariate.

Once again, let us begin by considering naive prediction in which the original training data regression coefficients $\widehat{\beta}$ are carried over directly in making predictions for the future cohort. Under relative manipulation costs, we have:

$$\begin{aligned} \widehat{y}_m^{naive} &\equiv \widehat{\beta} + \sum_{j=1}^{J} \widehat{\beta}_j (x_{jm} + a_{jm}^*) \\ &= \widehat{\beta} + \sum_{j=1}^{J} \widehat{\beta}_j x_{jm} + \sum_{j=1}^{J} \left(\frac{1}{c_j}\right)^{\frac{1}{\gamma-1}} |\widehat{\beta}_j|^{\frac{\gamma}{\gamma-1}} x_{jm}. \end{aligned} \tag{22}$$

The first two terms in the preceding equation capture the prediction a naive statistician believes she is making, mapping true covariates into predicted values just as she did in the training data. The final term captures Goodhart bias. Notice that here Goodhart bias is less closely related to standard regression penalization functions, in contrast to Lemma 1.

4.2. *Equilibrium Redux.* To begin, suppose we once again *conjecture* an equilibrium in which each $\beta_j^p = \widehat{\beta}_j$. Under this conjecture, Equation (21) implies:

$$x_{jm} = \left[1 + \left(\frac{\widehat{\beta}_j^+}{c_j}\right)^{\frac{1}{\gamma-1}} - \left(-\frac{\widehat{\beta}_j^-}{c_j}\right)^{\frac{1}{\gamma-1}}\right]^{-1} \widetilde{x}_{jm}. \tag{23}$$

Substituting the preceding equation into the manipulation-proofness condition (5) yields:

$$\widehat{\beta} + \sum_{j=1}^{J} \widehat{\beta}_j x_{jm} = \widehat{\beta} + \sum_{j=1}^{J} \left[\widehat{\beta}_j + \left(\frac{1}{c_j}\right)^{\frac{1}{\gamma-1}} |\widehat{\beta}_j|^{\frac{\gamma}{\gamma-1}}\right]^{-1} \widetilde{x}_{jm}. \tag{24}$$

But notice that the preceding condition is inconsistent with our initial conjecture that the coefficients on the manipulated covariates will satisfy $\beta_j^p = \widehat{\beta}_j$. Apparently, under relative ma-

[5] Convex costs imply concavity of the objective function.

nipulation costs, we must adjust slopes to eliminate Goodhart bias—after all, here manipulation is linear in true covariates. Nevertheless, the fact that manipulation is proportional implies that a relatively simple adjustment of slope coefficients suffices to correct the bias, as shown by the following proposition:

Proposition 4 (Relative Costs). *Let $\widehat{\beta}$ be the coefficient vector obtained from the historical data according to Equation (6), and let each $\widetilde{\beta}_j$ be a real solution to*

$$\widehat{\beta}_j \equiv \widetilde{\beta}_j + \left(\frac{1}{c_j}\right)^{\frac{1}{\gamma-1}} \left|\widetilde{\beta}_j\right|^{\frac{\gamma}{\gamma-1}}. \tag{25}$$

Then there is a Stackelberg equilibrium in which statisticians form predictions for the future cohort according to

$$\widehat{y}_m = \widetilde{\beta}^p + \sum_{j=1}^{J} \widetilde{\beta}_j^p \widetilde{x}_{jm},$$

where $\widetilde{\beta}^p = \widehat{\beta}$ and $\widetilde{\beta}_j^p = \widetilde{\beta}_j \; \forall \; j = 1, \ldots, J$.

To prove the proposition, it suffices to verify that the manipulation proofness (5) condition is satisfied. Using the incentive condition (21), we have

$$\begin{aligned} \widehat{y}_m &= \widetilde{\beta}^p + \sum_{j=1}^{J} \widetilde{\beta}_j^p \widetilde{x}_{jm} \\ &= \widehat{\beta} + \sum_{j=1}^{J} \left[\widetilde{\beta}_j + \left(\frac{1}{c_j}\right)^{\frac{1}{\gamma-1}} \left|\widetilde{\beta}_j\right|^{\frac{\gamma}{\gamma-1}} \right] x_{jm} \\ &= \widehat{\beta} + \sum_{j=1}^{J} \widehat{\beta}_j x_{jm}. \end{aligned} \tag{26}$$

Intuitively, each $\widetilde{\beta}_j$ solves:

$$\widehat{\beta}_j x_{jm} = \widetilde{\beta}_j \widetilde{x}_{jm}(\widetilde{\beta}_j). \tag{27}$$

Therefore, the contribution of each manipulated covariate $\widetilde{x}_{jm}$ to the prediction model $\widehat{y}_m$ is just equal to the true covariate x_{jm} contribution to the training model's predicted $\widehat{y}_m$.

It is readily verified that Equation (25) might not have a real-valued solution, or might have multiple solutions. For example, with quadratic manipulation costs ($\gamma = 2$), Equation (25) gives rise to a quadratic equation that may have no real solution, a unique real solution, or two distinct real solutions.

Proposition 4 can also be viewed from a positive perspective, delivering a prediction regarding how regression coefficients will evolve as one moves from historical training data to future data. In particular, the proposition predicts that under relative manipulation costs, slope coefficients will shift downward, with:

$$\widetilde{\beta}_j^p = \widehat{\beta}_j - \left(\frac{1}{c_j}\right)^{\frac{1}{\gamma-1}} \left|\widetilde{\beta}_j^p\right|^{\frac{\gamma}{\gamma-1}} \leq \widehat{\beta}_j. \tag{28}$$

5. EXTENSIONS

This section considers some extensions and implications of the preceding analysis.

5.1. *Algorithmic Perspective.* Propositions 2 and 4 used a conjecture and verify approach to evaluating Goodhart bias corrections. This section considers whether there might be a more robust approach for addressing Goodhart bias, one with the potential to address the problem of strategic manipulation in other settings, for example, alternative manipulation cost functions or heterogeneity in agent-level parameters.[6]

We propose the following *general algorithm*: Make the historical training data mimic the future manipulation-riddled data by imputing to the training data the manipulated covariates $\widetilde{\mathbf{x}}$ one would observe if the training data agents were to face the candidate prediction model. As an informal test of this algorithm, we next evaluate how it would fare against both absolute and relative cost functions.

To begin, suppose agents face absolute manipulation cost functions. Now let $\mathbf{B}$ denote a candidate coefficient vector to be evaluated within the training data. From Equation (8), it follows that for each candidate coefficient vector, each training data covariate x_{jm} can be converted to an imputed manipulated covariate as follows:

$$(29) \qquad \widetilde{x}_{jm} = x_{jm} + \left(\frac{B_j^+}{c_j}\right)^{\frac{1}{\gamma-1}} - \left(-\frac{B_j^-}{c_j}\right)^{\frac{1}{\gamma-1}}.$$

Next, within the training data, coefficients mapping the imputed manipulated covariates into predicted y values can be computed by finding:

$$(30) \quad \widetilde{\beta} \equiv \min_{\mathbf{B}} \left\{ \sum_{m=1}^{M} \left(y_m - B - \sum_{j=1}^{J} B_j \widetilde{x}_{jm} \right)^2 + P(\mathbf{B}; \lambda) \right\}$$
$$= \min_{\mathbf{B}} \left\{ \sum_{m=1}^{M} \left[y_m - \left(B + \sum_{j=1}^{J} \left(\frac{1}{c_j}\right)^{\frac{1}{\gamma-1}} |B_j|^{\left(\frac{\gamma}{\gamma-1}\right)} \right) - \sum_{j=1}^{J} B_j x_{jm} \right]^2 + P(\mathbf{B}; \lambda) \right\}.$$

But notice that the solution to the preceding program is the same as that offered in Proposition 2. In particular, slopes are unaffected by manipulation, whereas the intercept is shifted downward just as Proposition 2 prescribes.

Consider next relative manipulation costs. Fixing a candidate $\mathbf{B}$ and using condition (21), imputed covariates are:

$$(31) \qquad \beta^p = \mathbf{B} \Rightarrow \widetilde{x}_{jm} = \left[1 + \left(\frac{B_j^+}{c_j}\right)^{\frac{1}{\gamma-1}} - \left(-\frac{B_j^-}{c_j}\right)^{\frac{1}{\gamma-1}} \right] x_{jm}.$$

It follows that within training data, we can generate coefficients mapping imputed manipulated covariates to predicted outcomes as follows:

$$(32) \qquad \widetilde{\beta} \equiv \min_{\mathbf{B}} \left\{ \sum_{m=1}^{M} \left(y_m - B - \sum_{j=1}^{J} B_j \widetilde{x}_{jm} \right)^2 + P(\mathbf{B}; \lambda) \right\}$$

[6] See the Online Appendix for illustrations.

$$= \min_{\mathbf{B}} \left\{ \sum_{m=1}^{M} \left(y_m - B - \sum_{j=1}^{J} \left[B_j + \left(\frac{1}{c_j} \right)^{\frac{1}{\gamma-1}} \left| B_j \right|^{\frac{\gamma}{\gamma-1}} \right] x_{jm} \right)^2 + P(\mathbf{B}; \lambda) \right\}.$$

The second line in the preceding equation leads us back to Proposition 4. In particular, one can solve the canonical minimization program (Equation (6)) on the original training data to obtain the vector $\widehat{\beta}$ of coefficients on true $\mathbf{x}$. Next, the second line in Equation (32) implies the slope adjustment (25) converting $\widehat{\beta}$ into isomorphic coefficients on manipulated covariates.

5.2. *Model Averaging and Welfare.* Bagging is a common procedure intended to reduce model variance, and with it, MSPE. Bagging splits test data into $\mathcal{B}$ bootstrap samples with replacement, performing estimation, and then taking averages over the $\mathcal{B}$ predictions $\widehat{\mathbf{y}}$. Another method of model averaging is ensemble estimation in which different procedures are applied, say Lasso in one estimation and Ridge in another. With linear models, these procedures are equivalent to averaging the constituent model regression coefficients.

It is readily verified that model averaging can reduce data manipulation costs. To illustrate, suppose agents have quadratic absolute adjustment cost functions. Now suppose the statistician estimates Equation (6) using two bootstrap samples or two alternative estimators, resulting in coefficient vectors $\widehat{\beta}$ and $\widetilde{\beta}$. Suppose also that the two estimations generate equal sum of squared regression coefficients, implying equal manipulation costs (Lemma 1). It follows that the difference between manipulation costs without model averaging (MC) and with model averaging ($\overline{MC}$) is given by

$$(33) \qquad MC - \overline{MC} = \frac{1}{2}\frac{1}{\gamma c}\left[\sum_{j=1}^{J}(\widehat{\beta}_j^2 + \widetilde{\beta}_j^2)\right] - \frac{1}{\gamma c}\sum_{j=1}^{J}\left[\frac{1}{2}(\widehat{\beta}_j + \widetilde{\beta}_j)\right]^2$$

$$= \frac{1}{4}\left(\frac{1}{\gamma c}\right)\sum_{j=1}^{J}(\widehat{\beta}_j - \widetilde{\beta}_j)^2.$$

Notice that the preceding difference is positive, provided that the prediction models are not identical. In fact, the reduction in manipulation costs is larger, the larger the distance between the slope coefficients of the two models. The maximal feasible reduction in manipulation costs would occur if the two prediction models had different active sets. With different active sets, the preceding equation implies $\overline{MC} = MC/2$. Although nonintersecting active sets is a zero probability event under Ridge, it occurs frequently under Lasso in settings with sparse data.

6. CONCLUSION

This article provides a simple analytical framework illustrating how Goodhart's law relates to basic tools of ML prediction. It was shown that tighter coefficient constraints serve not only to mitigate overfitting, as recognized in the ML literature, but also to mitigate Goodhart bias. Similarly, it was shown that, in addition to improving fit, model averaging can reduce welfare costs of data manipulation. Finally, in addition to proposing specific Goodhart bias corrections against a broad class of manipulation cost functions, it was shown that imputing incentive problems to training data represents a potentially fruitful means of developing more robust prediction models. Having only explored canonical ML prediction techniques, much remains to be done.

REFERENCES

BALL, I., "Scoring Strategic Agents," arXiv Working Paper, 2022.

BJÖRKEGREN, D., J. E. BLUMENSTOCK, and S. KNIGHT, "Manipulation-Proof Machine Learning," arXiv Working Paper, 2020.

BRUCKNER, M., and T. SCHEFFER, "Stackelberg Games for Adversarial Prediction Problems," *Journal of Machine Learning Research* 13 (2011), 2617–54.

CHEN, Y., C. PODIMATA, A. D. PROCACCIA, and N. SHAH, "Strategyproof Linear Regression in High Dimensions," Working Paper, Harvard University, 2018.

DEKEL, O., F. FISCHER, and ———, "Incentive Compatible Regression Learning," *Journal of Computing System Science* 76 (2010), 759–77.

ELIAZ, K., and R. SPIEGLER, "The Model Selection Curse," *American Economic Review: Insights*, 1 (2019), 127–40.

FERNÁNDEZ-VILLAVERDE, J., "Has Machine Learning Rendered Simple Rules Obsolete?" *European Journal of Law and Economics* 52 (2021), 251–65.

FRANKEL, A., and N. KARTIK, "Improving Information from Manipulable Data," *Journal of the European Economic Association* 20 (2022), 79–115.

———, and Kartik, N.,"Muddled Information," *Journal of Political Economy* 129 (2019), 1739–76.

GOODHART, C. A., "Problems of Monetary Management: The U.K. Experience," Papers in Monetary Economics I, Reserve Bank of Australia, 1975.

HARDT, M., N. MEGIDDO, C. PAPADIMITRIOU, and M. WOOTERS, "Strategic Classification," *Proceedings of the 7th Innovations in Theoretical Computer Science Conference* (2016), 111–22.

KAY, J., and M. KING, *Radical Uncertainty: Decision Making Beyond the Numbers* (London, United Kingom: Little, Brown Book Group, 2020).

LUCAS, R., "Econometric Policy Evaluation: A Critique," *Carnegie-Rochester Conference Series on Public Policy* 1 (1976), 19–46.

PARVIERO, R., "Improving Ridge Regression via Model Selection and Focussed Fine-Tuning," Thesis, Scuola di Economia e Statistica, Milano-Bicocca, 2017.