On Incentive-Compatible Estimators*

Kfir Eliaz† and Ran Spiegler‡

January 10, 2022

Abstract

An estimator is incentive-compatible (for a given prior belief regarding the model's true parameters) if it does not give an agent an incentive to misreport the value of his covariates. Eliaz and Spiegler (2019) studied incentive-compatibility of estimators in a setting with a single binary explanatory variable. We extend this analysis to penalized-regression estimation in a simple multi-variable setting. Our results highlight the incentive problems that are created by the element of variable selection/shrinkage in the estimation procedure.

*Eliaz gratefully acknowledges financial support from ISF grant 470/19. Spiegler acknowledges support from the Sapir Center. We also thank an associate editor and two referees for their helpful comments.

†School of Economics, Tel-Aviv University and David Eccles School of Business, University of Utah. E-mail: kfire@tauex.tau.ac.il.

‡School of Economics, Tel Aviv University; Department of Economics, University College London; and CfM. E-mail: rani@tauex.tau.ac.il.

1 Introduction

The rise of data science and the growing use of big data has led to the adoption of machine-learning techniques for the purpose of prediction and automated decision-making. For instance, online platforms rely on such methods to display content that is predicted to be attractive for the user. These methods are also starting to be used in the medical arena for automating check-up and test appointments for patients based on their medical history. The common practice is to feed the automatic prediction system data from *past* users, in order to estimate the parameters of a model that relates a user's characteristics to his best outcome. The user is then assigned the predicted outcome (say, a song or a doctor's appointment) according to his *own* personal characteristics, which he himself provides to the system - either actively by means of an explicit report or passively through his observed behavior (e.g. his internet navigation history). A natural question that arises is whether it is in the user's best interest to *truthfully* report his characteristics to the system that tries to predict his best outcome.

Obviously, if the user's data is also utilized for other purposes that may adversely affect him (e.g., being sold to third parties for purposes of marketing or price discrimination), then the user may prefer not to disclose his private characteristics. However, even if the user's private information is employed *exclusively* for the purpose of predicting his best outcome, do the special features of common machine-learning methods incentivize the user to distort his true characteristics?

While a wide variety of estimation techniques have been developed for big data, the vast majority of them involve an element of *variable selection* - i.e., they try to identify the important variables for prediction and exclude others. Variable selection (or "regularization") is typically carried out by augmenting the loss function of the estimation procedure (which is related to the distance between the estimated variable and its true underlying value) to include penalties for model complexity. A prevalent procedure knows as LASSO

(Least Absolute Shrinkage and Selection Operator (Tibshirani (1996))) is a variant on standard linear regression analysis, which adds a cost function that penalizes non-zero coefficients. A penalized-regression procedure like LASSO is considered useful in situations where users have a great number of potentially relevant characteristics, but only few of them are actually relevant for predicting the agent's best outcome (i.e., the true data-generating process is *sparse*). This feature of penalizing the inclusion of explanatory variables creates an incentive for users *not* to disclose their true characteristics to the system that employs this prediction method.

Specifically, we present a simple model of an interaction between an "agent" and a "statistician", where the latter represents an automated algorithm that gathers data about the agent and outputs an action on his behalf. The agent's ideal action is a linear function of binary personal characteristics. The parameters of this function are unknown. The statistician learns about them by means of a sample that consists of noisy observations of the ideal actions of other agents with heterogeneous characteristics. This sample is the statistician's private information - i.e., the agent is not exposed to it. However, the sample design (i.e., the number of observations for each vector of personal characteristics) is common knowledge. The statistician employs a penalized linear regression to predict the agent's ideal action as a function of his characteristics. The penalty taxes non-zero estimated coefficients. We assume it is a linear combination of the three most basic forms: L_0, L_1 (LASSO) and L_2 (Ridge).[1] The agent's characteristics are his private information, and he reports them to the statistician. The action that the statistician takes is the penalized regression's predicted output, given the reported values of the agent's personal characteristics. We take the statistician's procedure as given, without trying to "rationalize" it (see a discussion in Section 2). The agent's payoff is a standard quadratic loss function - thus coinciding with the

[1] An L_0 penalty is a fixed cost for the mere inclusion of a non-zero coefficient. An L_1 penalty is a cost for the magnitude of the coefficient in absolute value. An L_2 penalty is a cost for the squared value of the coefficient.

most basic criterion for evaluating estimators' predictive success.

We pose the following question: Fixing the statistician's procedure and the agent's prior belief over the true model's parameters, *would the agent always want to truthfully report his personal characteristics to the statistician?* When this is the case for all possible priors in some class, we say that the statistician's procedure (or "estimator") is *incentive-compatible* for this class of priors. In Eliaz and Spiegler (2019) we analyzed a simple example in which the agent has a *single* binary characteristic that he needs to report to the statistician. In this case, we showed that an incentive problem arises *only* in the presence of *asymmetrically* distributed sampling error.

This paper extends the example to the case of *multiple* binary variables. The underlying source of the incentive problem in this setting is *fundamentally different* than in the single-variable setting. In particular, the element of variable selection in the statistician's procedure can generate an incentive problem even when the statistician faces *no* sampling error, and also when the error distribution is symmetric. The reason is that the cumulative bias due to the exclusion of some variables can be so large that the agent would like to introduce a counter-bias by misreporting the value of a variable he *does* expect to be included.

We proceed to investigate whether the estimator is incentive compatible for some natural classes of the agent's prior belief over the model's true coefficients. To be able to do this analytically and tractably, we focus on normally distributed sample noise and assume that there is an equal number of sample points for each value of the agent's covariates. These assumptions also ensure that OLS *is* incentive-compatible. That is, the only source of incentive problems in our example is the penalization of model complexity. The stark OLS benchmark thus enables us to focus on the role of variable selection and shrinkage in generating incentive problems.

Our first main result is that the estimator is *not* incentive-compatible for an unrestricted class of prior beliefs. The reason is that there exist prior

beliefs that exhibit an asymmetry between variables, such that the agent would like to misreport at least one characteristic. We then show that when the agent's prior over each coefficient is *independent* and *symmetric around zero* (reflecting agnosticism regarding the effect of each variable), he has *no* incentive to misreport. Finally, and perhaps most interestingly, when the agent's prior over each coefficient is $i.i.d$ (but with a non-zero mean), the agent has no incentive to misreport only if his characteristics vector is *sufficiently balanced* - i.e., its numbers of 0's and 1's are not too different. This result has the following implication with regards to an agent's incentive to hide his navigation history - say, by "deleting cookies" from his computer - when facing an online platform that employs a penalized-regression prediction method: The agent has an incentive to delete cookies only if there are relatively *few* stored in his browsing history.

2 Related literature

Our paper joins a recent literature in computer science that studies the problem of a planner who wishes to compute and implement some function of inputs provided by agents, where these agents can manipulate the inputs to their advantage. Since this literature is growing rapidly, we mention here only a sample of notable works, organizing them according to the kind of statistical procedure that the statistician performs.

One strand of literature focuses *binary classification* procedures. Meir et al. (2012) consider an environment with a set of input vectors, where each vector is the private information of some agent. Agents report binary labels for each vector. A planner wants to design an aggregate classifier of the users' private information that minimizes the average mistake (misclassifying an input) *across users*. The problem is that each agent can misreport his information in an effort to minimize the average mistake only on *his own* inputs. Under some restrictions on the domain, the authors characterize the

optimal classifier among those that induce truth-telling.

Hardt et al. (2016) consider a population of users who independently and privately draw an input from a common distribution. There is a true underlying classifier that assigns a binary label, low or high, to each input. A planner, who wishes to assign each agent to his true label, commits to a classifier and asks each agent to report his input. Each agent can incur a cost to manipulate his report. The agent trades off the probability of being assigned the high label against the manipulation cost. The authors show that for some family of cost functions, there exists a classifier for the planner that attains a classification error that is arbitrarily close to the theoretical minimum.

In Haghtalab et al. (2020), inputs are interpreted as features and the associated label is interpreted as the agent's quality. A planner observes only a subset of the features and can assign (or classify) each vector of observable features to a probability of accepting the agent. The planner's objective is to design a classifier that maximizes the expected quality of the agents, taking into account that each agent can incur a cost to change his features so as to maximize his probability of acceptance.

Kleinberg and Raghavan (2020) study a model in which an agent has a budget of effort that he can privately allocate across actions. Each profile of efforts induces some vector of observable outputs. A planner wishes to implement a particular allocation by mapping each vector of outputs to a real valued score, taking into account that the agents choose their allocation in order to maximize their score. The authors show that under certain assumptions, a linear function of the outputs implements any desired allocation.

In Krishnaswamy et al. (2021), the key strategic decision for agents is whether to *withhold* information. In their model, there is a known distribution of agents who are characterized by a vector of binary attributes and an associated binary label. A planner, who aims to predict the label for each agent, chooses a classifier that will be applied to reported attributes by

agents. Agents can strategically omit attributes in order to maximize the chance of being assigned the high label. The authors characterize classifiers that maximizes the chances of correctly predicting the label of each agent, subject to the constraint that truthful reporting is a dominant action.

A second related strand of literature focuses on statistical procedures that, like the present paper, are based on *linear regressions*. Perote and Perote-Pena (2004), Dekel et al. (2010) and Chen et al. (2018) consider the problem of a statistician who aims to minimize a loss function over the *union* of all the samples, while each agent would like the estimation to minimize the loss function only over *their own* sample. The first paper characterizes loss functions that induce truthful reporting in a dominant strategy equilibrium when the function is linear and each agent's sample is a single observation. The second paper allows the statistician to pay the agents and also extends the analysis to a larger class of functions and to samples that can be any arbitrary distribution. The third paper extends the analysis to multi-dimensional inputs. The source of the conflict of interest between the planner and the agents in these papers is essentially the same as in Meir et al. (2012).

Other papers in this second strand examine alternative motives for the agents. Cummings et al. (2015) assume that agents may manipulate their private information in order to avoid a loss of privacy (modeled via differential privacy). The authors characterize a payment scheme that under certain assumptions, achieves the following objectives as the number of agents increases: (i) the mean squared error of the planner's estimator goes to zero, (ii) agents have no incentive to misreport and attain a positive payoff, and (iii) the total payments to agents go to zero. Caragiannis et al. (2016) study the problem of estimating the population mean of an unknown unidimensional distribution from samples that are provided by strategic agents, who wish to move the estimate as close as possible to their own value. The authors characterize the worst-case optimal truthful estimator, and show that it achieves a lower mean square error than the sample median (which is known

to be strategy-proof). Cai et al. (2018) consider a statistician who wants to estimate a function that maps each vector of inputs into a real-valued output. The data for the estimation is privately held by agents who can provide only a noisy observation of the output associated with each input, and must incur a cost to reduce this noise. The authors analyze the problem of designing a payment scheme that minimizes the sum of the mean squared error of the estimator and the total payments to the agents.

Throughout the literature summarized in this section, a running theme is the existence of an *explicit conflict of interest* between the statistician/planner and the agents who provide the data - because the latter might be concerned about their privacy, or they might have to incur a cost in order to provide a precise report, or they might have a different objective than the statistician. In contrast, our question is whether variable selection and shrinkage - which is a characteristic of prevalent machine-learning procedures - gives an incentive to misreport, *even in the absence of an explicitly modeled conflict of interest* between the two parties.

3 A Model

Let $x_1, ..., x_K$ be a collection of binary explanatory variables; $x_k \in \{0, 1\}$ for every $k = 1, ..., K$. Each variable represents a personal characteristic of an *agent*. In the context of medical decision making, a variable can represent a risk factor (obesity, smoking, etc.). Under the online-content-provision interpretation, a variable can represent whether the agent visited a particular website. Denote $X = \{0, 1\}^K$ and $x = (x_1, ..., x_K)$. In what follows, it will be convenient (as well as conventional) to add a fictitious variable x_0, which is deterministically set at $x_0 = 1$.

A *statistician* must take an action $a \in \mathbb{R}$ on behalf of the agent (e.g., dosage of some drug, or a proportion of rock versus hip-hop music in a playlist). The agent's payoff from action a is $-(a - f(x))^2$, where $f(x)$ is the

agent's ideal action as a function of x, given by

$$f(x) = \sum_{k=0}^{K} \beta_k x_k$$

The coefficients $\beta_0, ..., \beta_K$ are fixed but unknown. The value of x is the agent's private information. Before taking an action, the statistician privately gets access to a sample that consists of N observations per value of x. For every $x \in X$, the N observations are $(y_x^n)_{n=1,...,N}$, where $y_x^n = f(x) + \varepsilon_x^n$, and ε_x^n is random noise that is drawn $i.i.d$ from a *normal* distribution with mean zero and variance σ^2. Denote $\varepsilon = (\varepsilon_x^n)_{x,n}$. The observations do not involve the agent himself. We have thus described an environment with two-sided private information: the agent privately knows x, whereas the statistician privately learns the sample.

We will discuss the importance of the uniform-sample assumption in Section 3. The broader assumption that the statistician has observations for *every* value of x means that the total number of observations is large relative to the number of potentially relevant variables. It also rules out the possibility that some of the variables represent interactions among other variables. This is a limitation of our model: In practice, one motivation for estimation procedures that involve variable selection is the "big data" predicament of having more explanatory variables than observations. However, another key motivation for such procedures - namely, *an underlying belief that the true model is sparse* (i.e. $\beta_k = 0$ for most values of k) - *is* consistent with our specification.

The statistician wishes to estimate the function f - equivalently, the coefficients $\beta_0, ..., \beta_K$. He follows a penalized regression procedure that assigns costs to including explanatory variables in the regression. We assume a generalized penalty function that is additively separable in the three most common forms of penalties: a fixed cost for the mere inclusion of a non-zero coefficient (L_0 penalty), a cost for the magnitude of the coefficient in absolute value

(the LASSO or L_1 penalty) and cost for the squared value of the coefficient (the "Ridge" or L_2 penalty).[2]

Formally, given the sample $(y_x^n)_{x\in X}^{n=1,...,N}$, the statistician solves the following minimization problem,

$$\min_{b_0,...,b_K} \sum_{x\in X}\sum_{n=1}^{N}(y_x^n - \sum_{k=0}^{K} b_k x_k^n)^2 + 2^K N \sum_{k=1}^{K}\left(c_0 \mathbf{1}_{b_k\neq 0} + c_1|b_k| + c_2 b_k^2\right) \qquad (1)$$

We denote the solution to this problem by $b(\varepsilon,\beta) = (b_0(\varepsilon,\beta),...,b_K(\varepsilon,\beta))$, and refer to it as the *estimator*. The dependence on (ε,β) follows from the fact that the estimator depends on the sampled observations $(y_x^n)_{x\in X}^{n=1,...,N}$, and these observations are determined by (ε,β). Note that there are no costs associated with the intercept b_0. Note also that the penalty costs are multiplied by the number of observations, such that the cost per observation remains constant. When $c_0 = c_1 = c_2 = 0$, we are back with the OLS estimator. We sometimes refer to c_0, c_1, c_2 as *complexity costs*. We treat them as constant per observation for notational convenience, as N is taken to be fixed for almost throughout the paper.

Having estimated f, the statistician receives a report $r \in X$ from the agent. Denote $r_0 = 1$ for convenience. The statistician then takes the action $a = \sum_{k=0}^{K} b_k(\varepsilon,\beta) r_k$. The agent's expected payoff for given $\beta_0,...,\beta_K$ is therefore

$$-\mathbb{E}_\varepsilon\left[\sum_{k=0}^{K}\left(b_k(\varepsilon,\beta)r_k - \beta_k x_k\right)\right]^2 \qquad (2)$$

The quadratic loss function is a standard criterion for evaluating estimators' predictive success. Suppose $r = x$ - i.e., the agent reports truthfully. Then, $\hat{f}(x) = \sum_{k=0}^{K} b_k(\varepsilon,\beta)x_k$ is the predicted ideal action. Expression (2) can thus be written as $-\mathbb{E}_\varepsilon[\hat{f}(x) - f(x)]^2$ - i.e., the agent's expected payoff is defined by the estimator's mean squared error.

[2] A combination of LASSO and Ridge penalties is known as an "elastic net" regression.

Discussion: Why does the statistician use penalized regression?
Real-life use of penalized regression methods such as (1) is motivated by an attempt to perform well according to criteria like mean squared error. Consider the following quote from Hastie et al. (2015, p. 7):

> "There are two reasons why we might consider an alternative to the least-squares estimate. The first reason is prediction accuracy: the least-squares estimate often has low bias but large variance, and prediction accuracy can sometimes be improved by shrinking the values of the regression coefficients, or setting some coefficients to zero. By doing so, we introduce some bias but reduce the variance of the predicted values, and hence may improve the overall prediction accuracy (as measured in terms of the mean-squared error). The second reason is for the purposes of interpretation. With a large number of predictors, we often would like to identify a smaller subset of these predictors that exhibit the strongest effects."

The first reason says that in the absence of a clear prior idea of the true data-generating process, a penalized regression is a plausible method for making automatic predictions on the basis of statistical data. In this informal sense, there is no conflict of interests between the two parties in our model: the statistician follows a procedure that is considered useful for predictive success, where the criterion for predictive success coincides with the agent's expected utility given the true model.

More formal justifications for penalized-regression methods (see Ch. 11 in Hastie et al. (2015)) often show that their predictive success (measured by the mean squared error criterion) is good under some restrictions on the domain of the true parameters $\beta_0, ..., \beta_K$ - e.g., when K is large and yet the statistician is convinced that $\beta_k = 0$ for most values of k (or that $\beta_k = 0$ with high independent probability for each k). This rationale for penalized

regression is relevant for our example - see our analysis of restricted classes of prior beliefs in Section 4.4. In contrast, another common justification for penalized regression - namely, situations in which the number of covariates is large relative to the number of observations - cannot be invoked in our example, since it assumes that sample size increases exponentially with K. This is a limitation of our example.

Note that this type of justification does not amount to a complete *Bayesian* rationalization of penalized regression. Although one can justify LASSO estimates as properties of a Bayesian posterior derived from some prior (Tibshirani (1996), Park and Casella (2008), Gao et al. (2015)), these properties are not necessarily relevant for maximizing the agent's welfare. Furthermore, the priors that rationalize LASSO in this manner are rarely used in economic applications (the priors in the above-cited papers involve Laplacian distributions over parameters).

The second justification for penalized regression that the quote from Hastie et al. (2015) invokes is essentially a *bounded rationality* rationale. Dealing with large models is difficult. Both practitioners of statistical analysis and their audience benefit from a model that simplifies things by omitting most variables, hopefully leaving only a few relevant ones. The penalty function is a way of capturing this implicit cognitive constraint. Penalized regression is an instrument for mitigating false discovery when K is large. The constant c_0 can be interpreted as a "statistical significance threshold" that excludes variables whose OLS-estimated coefficient is small. In this sense, the statistician in our model (or his implicit audience) can be viewed as a boundedly rational decision maker - somewhat as in Gabaix (2014), who offers a more elaborate sparsity-based model to describe decision makers with limited ability to pay attention to multiple variables.

Finally, the complexity cost c_0 can be motivated by physical costs of obtaining personal information from the agent. Even if many personal characteristics are relevant for predicting the agent's ideal action, it is costly

to collect them from the agent (e.g. because this requires long forms), and therefore it makes sense to truncate the list of variables in order to save these implementation costs. However, while bounded rationality or physical data collection are plausible informal justifications for the relevance of complexity costs, they do not amount to strict *rationalizations* of the statistician's procedure, in the absence of an explicit model for how cognitive or physical costs are traded off against some clear ex-ante objective.

3.1 Solving for the Estimator

We begin this sub-section with some notation that will serve us for the rest of the paper. Let $\bar{y}$ and $\bar{\varepsilon}$ denote the sample averages of the dependent variable and the noise:

$$\bar{y} = \frac{1}{2^K N} \sum_{x \in X} \sum_{n=1}^{N} y_x^n \qquad \bar{\varepsilon} = \frac{1}{2^K N} \sum_{x \in X} \sum_{n=1}^{N} \varepsilon_x^n$$

In addition, $\bar{\varepsilon}_{x_k=1}$ and $\bar{\varepsilon}_{x_k=1}$ denote the average noise realization in the subsamples for which $x_k = 1$ and $x_k = 0$, respectively:

$$\bar{\varepsilon}_{x_k=1} = \frac{1}{2^{K-1} N} \sum_{x|x_k=1} \sum_{n=1}^{N} \varepsilon_x^n \qquad \bar{\varepsilon}_{x_k=0} = \frac{1}{2^{K-1} N} \sum_{x|x_k=0} \sum_{n=1}^{N} \varepsilon_x^n$$

Finally, define $\Delta_k \equiv \bar{\varepsilon}_{x_k=1} - \bar{\varepsilon}_{x_k=0}$.

We are now able to give a complete characterization of the solution to the statistician's penalized regression problem. Our convention will be that when the statistician is indifferent between including and excluding a variable, he includes it. This characterization makes use of an auxiliary estimator $\tilde{b}_k$ of

β_k defined as follows:

$$\tilde{b}_k(\varepsilon,\beta)=\begin{cases}(\beta_k+\Delta_k-c_1)/(1+2c_2) & if \quad \beta_k+\Delta_k\geq c_1\\ (\beta_k+\Delta_k+c_1)/(1+2c_2) & if \quad \beta_k+\Delta_k\leq -c_1\\ 0 & if \quad -c_1<\beta_k+\Delta_k<c_1\end{cases} \tag{3}$$

Lemma 1 *The solution to the statistician's minimization problem (1) is as follows:*

$$b_k(\varepsilon,\beta)=\begin{cases}\tilde{b}_k(\varepsilon,\beta) & if \quad (\tilde{b}_k(\varepsilon,\beta))^2\geq 2c_0\\ 0 & if \quad (\tilde{b}_k(\varepsilon,\beta))^2<2c_0\end{cases} \tag{4}$$

for every $k=1,...,K$, *and*

$$b_0(\varepsilon)=\bar{y}-\frac{1}{2}\sum_{k=1}^{K}b_k(\varepsilon,\beta)$$

Note that (4) means that when the statistician (who does *not* know β_k and does *not* observe the realized noise in each data point) performs the penalized regression and ends up including the k-th variable, the numerical estimate of β_k is determined by the true value of β_k and the realized sample noise according to the function described in (3).

The L_2 penalty factor shrinks the coefficient b_k but it does not lead to variable selection - i.e., it does not affect the statistician's decision whether to set $b_k\neq 0$. In contrast, the L_0 penalty term only leads to variable selection but it does not affect the value of b_k conditional on being non-zero. Finally, the L_1 penalty term leads to both shrinkage and variable selection. When $c_1=c_2=0$, the characterization of b_k is very simple: $b_k=\beta_k+\Delta_k$ when $(\beta_k+\Delta_k)^2\geq 2c_0$, and $b_k=0$ when $(\beta_k+\Delta_k)^2<2c_0$. When $c_0=0$, $b_k=\tilde{b}_k$.

Note that $b_k(\varepsilon,\beta)$ is only a function of $\beta_k+\Delta_k$ - i.e., it is *functionally* independent of β_j and Δ_j for all $j\neq k$ (this simplicity is due to our assumption of a uniform sample). Of course, this by itself does not imply that it

is *statistically* independent of Δ_j, $j \neq k$. However, the assumption that the sample noise is normally distributed implies that for every $k = 1, ..., K$:

$$\Delta_k \sim N(0, \frac{\sigma^2}{2^{K-2}N})$$

This gives rise to the following useful observation.

Lemma 2 *For every distinct $k, j \in \{1, ..., K\}$, Δ_k and Δ_j are statistically independent.*

This lemma implies that the estimators for different coefficients $k, j > 0$ are both functionally and statistically independent.

3.2 Incentive Compatibility

The following are the key definitions of this paper.

Definition 1 *The estimator is **incentive compatible at a given prior belief** over the true model's parameters $\beta = (\beta_0, \beta_1, ..., \beta_K)$ if the agent is weakly better off with truthful reporting of his personal characteristic, given his prior. That is,*

$$\mathbb{E}_\beta \mathbb{E}_\varepsilon \left[\sum_{k=0}^{K} (b_k(\varepsilon, \beta) - \beta_k) x_k \right]^2 \leq \mathbb{E}_\beta \mathbb{E}_\varepsilon \left[\sum_{k=0}^{K} (b_k(\varepsilon, \beta) r_k - \beta_k x_k) \right]^2$$

for every $x = (x_1, ..., x_K)$, $r = (r_1, ..., r_K)$.[3]

In this definition, the expectation operator $\mathbb{E}_\varepsilon$ is taken with respect to the given exogenous distribution over the noise realization profile (since the agent does not observe the statistician's sample). The expectation operator

[3]Recall that $r_0 = x_0 = 1$ by definition.

$\mathbb{E}_\beta$ is taken with respect to the agent's prior belief over β. Note that this definition does not rely on the explicit solution we provide for the estimator, and would therefore be well-defined in extensions of the model for which a simple closed-form solution for the estimator is unavailable.

Definition 2 *The estimator is* ***incentive compatible*** *if it is incentive compatible at every prior belief. Equivalently,*

$$\mathbb{E}_\varepsilon \left[\sum_{k=0}^{K} (b_k(\varepsilon, \beta) - \beta_k) x_k \right]^2 \leq \mathbb{E}_\varepsilon \left[\sum_{k=0}^{K} (b_k(\varepsilon, \beta) r_k - \beta_k x_k) \right]^2 \tag{5}$$

for every $\beta = (\beta_0, ..., \beta_K)$, $x = (x_1, ..., x_K)$ *and* $r = (r_1, ..., r_K)$.

Incentive compatibility means that the agent is unable to perform better by misreporting his personal characteristic, *regardless* of his beliefs over the true model's parameters. When incentive compatibility fails, there are opportunities for new firms to enter and offer the agent paid advice for how to manipulate the procedure - in analogy to the industry of "search engine optimization". Incentive compatibility eliminates the need for such an industry. In some contexts (especially online content provision), certain misreporting strategies take the form of erasing part of the agent's internet navigation history ("deleting cookies"). Such deviations are straightforward to implement, and the agent can check if it makes him better off in the long run. Definitions 1 and 2 may be interpreted as *Bayesian* and *ex-post* incentive-compatibility, where the relevant state space consists of the possible realizations of β.

The incentive compatibility requirement can be described as a collection of bias-variance trade-offs between our estimator and alternative ones. Because of the form of the agent's payoff function, his expected utility takes the form of mean square deviation of the estimator from the true model. This loss function is known to be decomposable into two terms, one capturing the bias of estimator and another its variance. Comparing the predictive

success of different estimators thus boils down to trading off the estimators' bias and variance. The incentive compatibility condition can be viewed as a bias-variance comparison between two estimators: one is the statistician's estimator, and another is an estimator that applies the statistician's procedure to r rather than x. The latter is not an estimation method that a statistician is likely to propose, but it arises naturally in our setting.

4 Analysis

Eliaz and Spiegler (2019) analyzed the case of $K = 1$ and showed that a symmetric noise distribution ensures incentive-compatibility of the estimator. Since we assume here that the noise is normally distributed, this settles the case of $K = 1$. Let us turn to analyzing the estimator's incentive compatibility when $K > 1$.

4.1 Preliminary Observations

We begin with some convenient notation. First, represent a deviation from truth-telling by the subset $M = \{k = 1, ..., K \mid r_k \neq x_k\}$. That is, M is the set of variables that the agent's reporting strategy misrepresents. Second, denote

$$w_k = 1 - 2x_k$$

This is merely a rescaling of x_k such that it gets the values -1 and 1.

The normality assumption - specifically, the property that the noise density is a well-defined, decreasing function of the distance from zero - enables a useful characterization of the ex-ante expectation of estimated coefficients. Recall that the formula for $b_k(\varepsilon, \beta)$ is purely a function of $\beta_k + \Delta_k$, and that the distribution of Δ_k is the same for all k. Therefore, we can write the

ex-ante expectation of $b_k(\varepsilon, \beta)$ as a deterministic function of β_k:

$$e(\beta_k) = \mathbb{E}_\varepsilon(b_k(\varepsilon, \beta))$$

Lemma 3 *If for every x and n, ε_x^n is i.i.d according to a normal distribution, then the function e is: (i) anti-symmetric; (ii) strictly increasing, and (iii) shrinking β_k toward zero - i.e., $e(\beta_k)\beta_k > 0$ and $0 < |e(\beta_k)| < |\beta_k|$ whenever $\beta_k \neq 0$.*

Parts (i) and (ii) only rely on symmetry of the sample noise distribution, without requiring normality. When $c_0 = 0$, part (iii) (which means that the estimator shrinks the true coefficient on average) holds whenever the sample noise has symmetric density. However, when $c_0 > 0$, it also requires the property that the density of sample noise is decreasing in the distance from zero.

The following is an alternative formulation of the inequality that underlies the definition of incentive compatibility.

Lemma 4 *A deviation M is unprofitable for given β, x if and only if*

$$\left(\sum_{k \in M} e(\beta_k) w_k\right) \left(\sum_{k=1}^{K} \beta_k w_k - \sum_{j \notin M} e(\beta_j) w_j\right) \geq 0 \tag{6}$$

This condition is convenient because it is stated entirely in terms of the expected coefficients of individual variables according to the agent's prior.

4.2 Two Benchmarks

Before we embark on our analysis, two benchmark cases will be useful. These cases do not require the assumption that the sample noise is normally distributed.

Benchmark I: Precise Measurement

Suppose that $\varepsilon_x^n = 0$ with probability one for every n, x. Consider the L_0 estimator - i.e., $c_0 > 0 = c_1 = c_2$. Then, for every k, $b_k = \beta_k$ if $(\beta_k)^2 \geq 2c_0$, and $b_k = 0$ otherwise. The subset of selected variables is given by $V = \{k = 1, ..., K \mid (\beta_k)^2 \geq 2c_0\}$. The inequality (6) can be written as

$$\left(\sum_{k \in V \cap M} \beta_k w_k\right)\left(\sum_{k \notin V - M} \beta_k w_k\right) \geq 0 \tag{7}$$

When $K = 1$, this is reduced to $0 \geq 0$ or $\beta_1^2 \geq 0$, which obviously holds.

The condition is also satisfied when $K = 2$, for the following reason. Without loss of generality, let $x = (0, 0)$ and consider the possible configurations of V and M. First, suppose that $V = M = \{1, 2\}$. Then, the inequality becomes $(\beta_1 + \beta_2)^2 \geq 0$. Second, suppose that $V = \{1, 2\}$ and $M = \{1\}$. Then, the inequality becomes $(\beta_1)^2 \geq 0$. Third, suppose that $V = M = \{1\}$. Then, the condition becomes $\beta_1(\beta_1 + \beta_2) \geq 0$. This inequality must hold because by the definition of V, $|\beta_1| \geq \sqrt{2c_0} \geq |\beta_2|$, such that $sign(\beta_1 + \beta_2) = sign(\beta_1)$. The cases of $V = \{1, 2\}, M = \{2\}$ and $V = M = \{2\}$ are essentially the same. Finally, if $V \cap M$ is empty, the condition becomes $0 \geq 0$.

However, incentive compatibility can fail when $K > 2$. To see why, suppose that $K = 3$, and let $\beta_1 = \sqrt{2c_0} + \delta$, $\beta_2 = \beta_3 = -\sqrt{2c_0} + \delta$, where $\delta > 0$ is arbitrarily small. Then, $V = \{1\}$. Suppose that the agent's characteristics are $x = (0, 0, 0)$, and that he deviates to the report $r = (1, 0, 0)$ - i.e., $M = \{1\}$. Then, $V \cap M = \{1\}$ and $V - M = \varnothing$. The condition becomes

$$\beta_1 \cdot (\beta_1 + \beta_2 + \beta_3) \geq 0$$

This inequality fails because $\beta_1 + \beta_2 + \beta_3 = -\sqrt{2c_0} + 3\delta < 0$, whereas $\beta_1 > 0$.

Thus, unlike the $K = 1$ case, precise measurement of coefficients does *not* eliminate the incentive problem due to variable selection. The reason is as

follows. When there are multiple variables, omitting some of them because their coefficients are too close to zero leads to a biased action. The bias from the omission of any single variable is small (because by definition, their true coefficients are small to begin with). However, omitting several variables can generate a large cumulative bias, such that the agent may find it profitable to counter this bias by misreporting the value of one of the variables that *are* selected.

Benchmark II: OLS
Now consider the model with non-degenerate noise, but without variable selection - i.e., $c_0 = c_1 = c_2 = 0$. This produces the OLS estimator $b_k = \beta_k + \Delta_k$ for every $k = 1, ..., K$.

Proposition 1 *The OLS estimator is incentive-compatible.*

Thus, OLS estimation does not generate an incentive problem. Note that the result does not rely on any property of the sample noise distribution beyond the assumption of zero mean. Instead, it relies on the property that $\bar{\varepsilon}_{x_k=1}$ and $\bar{\varepsilon}_{x_k=0}$ are $i.i.d$, which in turn relies on the *uniform-sample* assumption. It should be emphasized that the OLS estimator does *not* induce the Bayesian-optimal action given the agent's prior. Nevertheless, this de-facto conflict of interests does not give the agent an incentive to misreport his personal characteristics. It is easy to verify that this conclusion extends to the case of Ridge regression - i.e., $c_2 > 0 = c_0 = c_1$. Thus, variable selection is crucial for the incentive to misreport.

4.3 Failure of Incentive Compatibility

Let us now turn to the case of noisy measurement where either $c_0 > 0$ or $c_1 > 0$ or both, such that the statistician's procedure involves variable selection. The following is our first main result, which is a simple consequence

of Lemma 4 and our observation that $|\beta_k| > |e(\beta_k)|$ (see part (iii) of Lemma 3).

Proposition 2 *The estimator is not incentive-compatible for any $K > 1$.*

Proof. Suppose that the agent's prior is degenerate, with $\beta_k = 0$ for all $k > 2$. Then, $e(\beta_k) = 0$ for all $k > 2$. Consider a deviation $M = \{1\}$. The condition for its unprofitability is

$$(e(\beta_1)w_1)\,(\beta_1 w_1 + \beta_2 w_2 - e(\beta_2)w_2) \geq 0$$

Select β_1 and β_2 such that $sign(\beta_1 w_1) = -sign(\beta_2 w_2)$. Since $sign(e(\beta_1)) = sign(\beta_1)$ and $sign(\beta_2 - e(\beta_2)) = sign(\beta_2)$, we obtain that if $sign(w_2) = -sign(\beta_2)$ and $|\beta_1|$ is sufficiently small relative to $|\beta_2|$, the inequality will be violated. ■

Thus, unlike the precise-measurement benchmark case, noisy measurement means that the estimator fails incentive compatibility even when $K = 2$.

4.4 Incentive Compatibility for Restricted Classes of Priors

In the remainder of this section, we characterize incentive compatibility for three specific families of priors.

An ultra-sparse prior

Suppose that the agent believes that only one variable is relevant, say $\beta_1 > 0$, whereas $\beta_k = 0$ for all $k > 1$. Then, $e(\beta_k) = 0$ for all $k > 1$. If $1 \notin M$, the condition for the unprofitability of the deviation M trivially becomes $0 \geq 0$. If $1 \in M$, the condition is reduced to $e(\beta_1)\beta_1 \geq 0$, which holds by part (iii) in Lemma 3. This observation implies the following corollary.

Corollary 1 *The estimator is incentive-compatible at any prior over* $(\beta_1, ..., \beta_K)$ *that only assigns positive probability to profiles in which at most one coefficient is non-zero.*

Independent, symmetric priors

Suppose that the agent's prior over $(\beta_1, ..., \beta_K)$ is independent across components, such that for each $k = 1, ..., K$, the prior over β_k is symmetric around zero. This reflects the agent's agnosticism regarding the sign of the effect of each variable. We do not require the priors to be identical. Also, the agent's belief over β_0 is irrelevant. Given such a prior, the agent will report truthfully if the L.H.S of (6) is non-negative in expectation (with respect to the agent's prior) for every deviation M.

Proposition 3 *Suppose that the agent's prior over* β_k *for each* k *is independent and symmetric around zero. Then, the estimator is incentive-compatible at this prior.*

i.i.d priors

Now suppose that the agent's prior over β_k is $i.i.d$ for each k. Let β^* denote the expectation of β_k. Accordingly, e^* is the expected estimated coefficient of each variable.

In this special case incentive compatibility has a very simple structure because the most profitable deviation can be pinned down. The following notation is useful for our next result. For any $x \in X$, define $m(x)$ as the number of components $k = 1, ..., K$ for which $x_k = 1$. Define the subset $M^* \subseteq \{1, ..., K\}$ as follows:

$$M^* = \begin{cases} \{k \mid x_k = 1\} & if \quad m(x) \leq \frac{K}{2} \\ \{k \mid x_k = 0\} & if \quad m(x) > \frac{K}{2} \end{cases}$$

That is, M^* is the smaller between the set of characteristics that get the value 1 and the set of characteristics that get the value 0. Denote $m^* = |M^*|$.

Proposition 4 *Suppose that the agent's prior over β_k for each k is i.i.d. Then, the following three statements are equivalent:*
(i) The estimator is incentive-compatible at the agent's prior.
(ii) M^* is not a profitable deviation.
(iii) The following inequality holds:

$$\mathbb{E}(e(\beta)\beta) + (e^*)^2(K - m^*) + e^*\beta^*[(m^* - 1) - (K - m^*)] \geq 0$$

Suppose that there is an equal number of 1's and 0's in x - i.e., $m^* = \frac{K}{2}$. Plugging this value into the condition for incentive compatibility, we obtain the following corollary.

Corollary 2 *Suppose that the agent's prior over β_k for each k is i.i.d. When $m(x) = \frac{K}{2}$, truth-telling is optimal.*

Thus, the characteristics vectors that are most conducive to deviation from truth-telling are those that are very skewed - i.e., the number of 1's is either very small or very large. When the vector is perfectly balanced (with the same number of 0's and 1's), truth-telling is optimal. The result also implies that the x that is most conducive to violation of incentive compatibility has $m = 1$, such that the condition for profitable deviation becomes

$$\mathbb{E}(e(\beta)\beta) - e^*(\beta^* - e^*)(K - 1) < 0$$

It follows that if K is small enough, the estimator is incentive-compatible, but when K is large enough, there will be values of x for which the agent will deviate from truth-telling.

Comment: "Deleting cookies"
Suppose that the set of feasible deviations is restricted, such that the agent can only deviate downward - i.e. if $r_k \neq x_k$ then $x_k = 1$ and $r_k = 0$. One

interpretation is that every variable indicates whether a particular "cookie" is installed on the agent's computer (where K is the number of cookies on which there is data); the agent can delete cookies but he cannot manufacture a "fake cookie".[4] Suppose that the agent's prior over β_k is $i.i.d$ across k. Our previous characterization is the same, except that M^* is now forced to be $\{k \mid x_k = 1\}$, such that truthful reporting is profitable if only if

$$\mathbb{E}(e(\beta)\beta) + e^*\beta^*(m(x) - 1) - e^*(\beta^* - e^*)(K - m(x)) < 0$$

Thus, the values of x that are conducive to misreporting by deleting cookies are those in which $m(x)$ is small - i.e., when the number of cookies is small (and in particular, strictly lower than $\frac{K}{2}$). Note that in this special case, checking whether truthful reporting is optimal for the agent is simple - it suffices to compare it with the deviation of deleting all the cookies.

5 Conclusion

This paper examined the incentive compatibility of penalized regression with multiple variables. We constructed a simple example that involves binary covariates, uniform sample design and normally distributed sample noise. These simplifying assumptions played two roles. First, they ensured that the unpenalized-regression (OLS) benchmark is incentive-compatible, thus enabling us to focus on the incentive effects of variable selection and shrinkage. Second, they gave rise to simple closed-form solutions for the penalized-regression estimators, which enabled a tractable characterization of incentive compatibility for various classes of prior beliefs.

These characterizations conveyed two major insights. First, when the

[4] An incentive to delete cookies in order to manipulate an estimator bears resemblance to strategically witholding information in order to manipulate a classifier. This latter incentive was studies by Krishnaswamy et al. (2020), who propose methods for training classifiers that will make them robust to this type of manipulation.

agent's prior belief over variable coefficients displays an asymmetry across variables, there is an incentive problem because the agent may want to misreport the variable his prior belief deems relatively unimportant, in order to mitigate the bias due to the possible omission of another variable he deems more important. Second, when the agent's prior is symmetric across variables but his covariates are unbalanced (in the sense that the numbers of 1's and 0's are very different), there is an incentive problem because the agent may want to misrepresent the "minority variables" in order to counter the estimation bias of the "majority variables". We believe that these forces have broader relevance beyond our simple example, even in settings that do not allow tractable analytical characterizations, and including machine-learning methods outside the domain of penalized linear regression.

By its nature, our example leaves a number of open technical problems. First, there remains the challenge of extending our results to environments with continuously distributed covariates and general samples. Second, an important case our analysis left out is where the number of covariates exceeds the sample size, which is a typical justification for applying machine learning techniques. As mentioned above, the technical difficulty here is that there is no closed-form characterization of the penalized regression estimators. A step in this direction is carried out in Caner and Eliaz (2021). That paper, which follows up the current one, provides sufficient conditions for *asymptotic* incentive compatibility of the Lasso and general weighted Lasso estimators. Caner and Eliaz borrow tools from high-dimensional econometrics to characterize the rate at which the penalty parameter needs to change as a function of the sample size in order to preserve incentive compatibility in large samples. Surprisingly, incentive compatibility demands that the penalty parameter is not too *low*. This means that the considerations that are relevant for asymptotic incentive-compatibility of a estimator are distinct from those that pertain to its asymptotic consistency.

References

[1] Cai, Y., C. Daskalakis, and C. H. Papadimitriou (2015). Optimum statistical estimation with strategic data sources. In Proceedings of the 28th Conference on Learning Theory (COLT'15). 40.1–40.40.

[2] Caner, M. and K. Eliaz (2021). Should humans lie to machines? The incentive compatibility of lasso and general weighted lasso. Working paper.

[3] Caragiannis, I., A. D. Procaccia, and N. Shah (2016). Truthful univariate estimators. In Proceedings of the 33rd International Conference on Machine Learning (ICML'16).

[4] Chen, Y., C. Podimata, A. D. Procaccia, and N. Shah (2018). Strategyproof linear regression in high dimensions. In Proceedings of the ACM Conference on Economics and Computation (EC'18), 9–26.

[5] Cummings, R., S. Ioannidis and K. Ligett (2015). Truthful linear regression, Conference on Learning Theory, 448-483.

[6] Dekel, O., F. Fischer, and A. D. Procaccia (2010). Incentive compatible regression learning. J. Comput. Syst. Sci. 76, 8, 759–777.

[7] Eliaz, K. and R. Spiegler (2019). The model selection curse. American Economic Review: Insights 1, 127-140.

[8] Gabaix, X. (2014). A sparsity-based model of bounded rationality, Quarterly Journal of Economics 129, 1661-1710.

[9] Gao, C., van der Vaart, A. and H. Zhou (2015). A general framework for Bayes structured linear models. arXiv preprint arXiv:1506.02174.

[10] Haghtalab, N., N. Immorlica, B. Lucier and J. Wang (2020). Maximizing welfare with incentive-aware evaluation mechanisms. In 29th International Joint Conference on Artificial Intelligence.

[11] Hardt, M., N. Megiddo, C. Papadimitriou, and M. Wootters (2016). Strategic classification, in Proceedings of the ACM Conference on Innovations in Theoretical Computer Science (ITCS'16), 111–122.

[12] Hastie, T., R. Tibshirani and M. Wainwright (2015). Statistical learning with sparsity: the LASSO and Generalizations, CRC press.

[13] Kleinberg, J. and M. Raghavan (2019). How do classifiers induce agents to invest effort strategically? In Proceedings of the 2019 ACM Conference on Economics and Computation, 825–844.

[14] Krishnaswamy, A.K., H. Li, D. Rein, H. Zhang, and V. Conitzer (2020). Classification with strategically withheld data arXiv:2012.10203v1.

[15] Meir, R., A. D. Procaccia, and J. S. Rosenschein (2012). Algorithms for strategyproof classification. Artif. Intell. 186, 123–156.

[16] Park, T. and G. Casella (2008). The Bayesian lasso. Journal of the American Statistical Association 103, 681-686.

[17] Perote, J. and J. Perote-Pena (2004). Strategy-proof estimators for simple regression. Math. Soc. Sci. 47, 2, 153–176

[18] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso, Journal of the Royal Statistical Society, Series B (Methodological), 267-288.

Appendix: Proofs

Proof of Lemma 1

Fix the realization of sample noise ε and denote the set of non-zero coefficients (the set of included variables) by $V(\varepsilon) = \{k \in K \mid b_k(\varepsilon) \neq 0\}$. These coefficients are given by the solution to the first-order conditions of

$$\min_{b_0,...,b_K} \sum_{x\in X} \sum_{n=1}^{N} (y_x^n - b_0 - \sum_{k=1}^{K} b_k x_k^n)^2 + 2^K N \sum_{k=1}^{K} \left(c_0 \mathbf{1}_{b_k \neq 0} + c_1 |b_k| + c_2 b_k^2 \right)$$

where the dependence of the coefficients $b_0, ..., b_K$ on the noise realization ε is suppressed for notational ease. The first-order condition with respect to b_0 is

$$\sum_{x\in X} \sum_{n=1}^{N} (y_x^n - b_0 - \sum_{k\in V(\varepsilon)} b_k x_k^n) = 0 \tag{8}$$

while the first-order condition with respect to each b_j, $j \in V(\varepsilon)$, is

$$2 \sum_{x\in X} \sum_{n=1}^{N} x_j^n (y_x^n - b_0 - \sum_{k\in V(\varepsilon)} b_k x_k^n) = 2^K N((sign(b_j)c_1 + 2c_2 b_j) \tag{9}$$

From (8) we obtain

$$b_0 = \bar{y} - \frac{1}{2} \sum_{k\in V(\varepsilon)} b_k$$

Substituting (8) into (9) yields $\tilde{b}_j$ whenever $\beta_j + \Delta \notin (-c_1, c_1)$. When $\beta_j + \Delta \in (-c_1, c_1)$, the first-order condition is self-contradictory, and therefore we must have $\tilde{b}_j = 0$.

The remaining task is to derive $V(\varepsilon)$. Let $P = 2^K N$ denote the total number of observations. In this proof, use x_k^p and y^p to denote the values of x_k and y in observation $p \in \{1, ..., P\}$. Without loss of generality, let us compare the residual sum of squares (RSS) when the admitted coefficients

are $b_0, b_1, ..., b_m$ and when b_m is omitted. The RSS in the former case is

$$\begin{aligned} RSS(b_0, ...b_{m-1}, b_m) &= \sum_{p=1}^{P} \left(b_0 + \sum_{k=1}^{m-1} b_k x_k^p + b_m x_m^p - y^p \right)^2 \\ &= \sum_{p=1}^{P} \left(b_m x_m^p + \left(b_0 + \sum_{k=1}^{m-1} b_k x_k^p - y^p \right) \right)^2 \end{aligned}$$

while in the latter case it is

$$RSS(b_0, ...b_{m-1}) = \sum_{p=1}^{P} \left(\frac{1}{2} b_m + \left(b_0 + \sum_{k=1}^{m-1} b_k x_k^p - y^p \right) \right)^2$$

As we have already shown, the values of the coefficients $b_1, ..., b_m$ are independent of whether b_m is included. We use b_0 to denote the intercept in the regression *with* b_m.

The difference between $RSS(b_0, ...b_{m-1}, b_m)$ and $RSS(b_0, ...b_{m-1})$ is equal to

$$\sum_{p=1}^{P} \left[\left(\frac{1}{2} b_m + \left(b_0 + \sum_{k=1}^{m-1} b_k x_k^p - y^p \right) \right)^2 - \left(b_m x_m^p + \left(b_0 + \sum_{k=1}^{m-1} b_k x_k^p - y^p \right) \right)^2 \right]$$

which can be rewritten as a sum of three terms:

$$\begin{aligned} &\sum_{p=1}^{P} \left[\frac{1}{4} (b_m)^2 - (b_m x_m^p)^2 \right] + b_m \sum_{p=1}^{P} \left(b_0 + \sum_{k=1}^{m-1} b_k x_k^p - y^p \right) \\ &-2 b_m \sum_{p=1}^{P} x_m^p \left(b_0 + \sum_{k=1}^{m-1} b_k x_k^p - y^p \right) \end{aligned}$$

Each of the three terms in this sum can be further simplified as follows. First,

$$\begin{aligned}
& \sum_{p=1}^{P} \left[\frac{1}{4}(b_m)^2 - (b_m x_m^p)^2 \right] \\
= \; & (b_m)^2 \sum_{p=1}^{P} \left[\frac{1}{4} - (x_m^p)^2 \right] \\
= \; & (b_m)^2 \cdot \left[\frac{K \cdot 2^n}{4} - K \cdot 2^{n-1} \right] \\
= \; & -(b_m)^2 \cdot K \cdot 2^{n-2}
\end{aligned}$$

Second,

$$\begin{aligned}
& b_m \sum_{p=1}^{P} \left(b_0 + \sum_{k=1}^{m-1} b_k x_k^p - y^p \right) \\
= \; & b_m \sum_{p=1}^{P} \left(b_0 + \frac{1}{2} b_m + \sum_{k=1}^{m-1} b_k x_k^p - y^p - \frac{1}{2} b_m \right) \\
= \; & b_m \sum_{p=1}^{P} \left(b_0 + \frac{1}{2} b_m + \sum_{k=1}^{m-1} b_k x_k^p - y^p \right) - \frac{1}{2} b_m \sum_{p=1}^{P} b_m \\
= \; & -\frac{1}{2}(b_m)^2 \cdot N \cdot 2^K
\end{aligned}$$

where the last equality follows from observing that in the regression *without* b_m, the first-order condition with respect to b_0 implies that

$$b_0 + \frac{1}{2} b_m + \sum_{k=1}^{m-1} b_k x_k^p - y^p = 0$$

Finally,

$$\begin{aligned}
& -2b_m \sum_{p=1}^{P} x_m^p \left(b_0 + \sum_{k=1}^{m-1} b_k x_k^p - y^p \right) \\
= \; & -2b_m \sum_{p=1}^{P} x_m^p \left(b_0 + \sum_{k=1}^{m} b_k x_k^p - y^p - b_m x_m^p \right) \\
= \; & -2b_m \sum_{p=1}^{P} x_m^p \left(b_0 + \sum_{k=1}^{m} b_k x_k^p - y^p \right) + 2(b_m)^2 \sum_{p=1}^{P} (x_m^p)^2 \\
= \; & 2(b_m)^2 \cdot N \cdot 2^{K-1}
\end{aligned}$$

where the last equality follows from observing that in the regression *with* b_m, the first-order condition with respect to b_m implies that

$$\sum_{p=1}^{P} x_m^p \left(b_0 + \sum_{k=1}^{m} b_k x_k^p - y^p \right) = 0$$

Adding all three terms yields

$$(b_m)^2 \cdot N \cdot \left[-2^{K-2} - 2^{K-1} + 2^K \right] = (b_m)^2 \cdot N \cdot 2^{K-2}$$

We include b_m in $V(\varepsilon)$ if and only if this term is weakly greater than Nc_0. ■

Proof of Lemma 2

By definition,

$$\begin{aligned}
\Delta_k &= \frac{1}{2} \left[\bar{\varepsilon}_{x|x_{k=1},x_{j=0}} + \bar{\varepsilon}_{x|x_{k=1},x_{j=0}} - \bar{\varepsilon}_{x|x_{k=0},x_{j=1}} - \bar{\varepsilon}_{x|x_{k=0},x_{j=0}} \right] \\
\Delta_j &= \frac{1}{2} \left[\bar{\varepsilon}_{x|x_{k=1},x_{j=1}} + \bar{\varepsilon}_{x|x_{k=0},x_{j=1}} - \bar{\varepsilon}_{x|x_{k=1},x_{j=0}} - \bar{\varepsilon}_{x|x_{k=0},x_{j=0}} \right]
\end{aligned}$$

Thus, $\Delta_k = A + B$ and $\Delta_j = A - B$, where

$$\begin{aligned} A &= \bar{\varepsilon}_{x|x_{k=1},x_{j=1}} - \bar{\varepsilon}_{x|x_{k=0},x_{j=0}} \\ B &= \bar{\varepsilon}_{x|x_{k=1},x_{j=0}} - \bar{\varepsilon}_{x|x_{k=0},x_{j=1}} \end{aligned}$$

By definition, A and B are $i.i.d$, and therefore $\mathbb{E}(A+B)(A-B) = \mathbb{E}(A^2) - \mathbb{E}(B^2) = 0$. Therefore, $cov(\Delta_k, \Delta_j) = 0$. Since Δ_k and Δ_j are normally distributed, this also implies their statistical independence. ■

Proof of Lemma 3

Denote $c^* = (1+2c_2)\sqrt{2c_0} + c_1$. Use g to denote the (normal) density of Δ_k, and G to denote its induced cdf. By symmetry of g, $G(\Delta) + G(-\Delta) = 1$ for every Δ. For notational ease, remove the subscript from β_k. Then,

$$e(\beta) = \frac{1}{1+2c_2}\left[\int_{-\infty}^{-c^*-\beta}(\beta+\Delta+c_1)g(\Delta) + \int_{c^*-\beta}^{\infty}(\beta+\Delta-c_1)g(\Delta)\right]$$

We can rewrite $e(\beta)$ as follows:

$$e(\beta) = \frac{\beta[1 - G(c^*-\beta) + G(-c^*-\beta)] + c_1[G(-c^*-\beta) + G(c^*-\beta) - 1] - \int_{-c^*-\beta}^{c^*-\beta}\Delta g(\Delta)}{1+2c_2}$$

(i) Anti-symmetry of e (i.e., $e(-\beta) = -e(\beta)$) follows mechanically from the formula for e and the symmetry of g. □

(ii) For the purpose of this claim, we can ignore the term $1/(1+2c_2)$, and rewrite the formula for e as follows:

$$\begin{aligned} e(\beta) &= \beta + (c^*-\beta)G(c^*-\beta) - (-c^*-\beta)G(-c^*-\beta) - \int_{-\infty}^{c^*-\beta}\Delta g(\Delta) \\ &\quad + \int_{-\infty}^{-c^*-\beta}\Delta g(\Delta) - (c^*-c_1)[G(c^*-\beta) + G(-c^*-\beta)] - c_1 \end{aligned}$$

Using integration by parts, this is equal to

$$\beta + \int_{-\infty}^{c^*-\beta} G(\Delta) - \int_{-\infty}^{-c^*-\beta} G(\Delta) - (c^* - c_1)[G(c^* - \beta) + G(-c^* - \beta)] - c_1$$

hence

$$e(\beta) = \beta + \int_{-c^*-\beta}^{c^*-\beta} G(\Delta) - (c^* - c_1)[G(c^* - \beta) + G(-c^* - \beta)] - c_1 \qquad (10)$$

Now differentiate this expression with respect to β:

$$\begin{aligned} & 1 - G(c^* - \beta) + G(-c^* - \beta) + (c^* - c_1)[g(c^* - \beta) + g(-c^* - \beta)] \\ = \; & G(\beta - c^*) + G(-c^* - \beta) + (c^* - c_1)[g(c^* - \beta) + g(-c^* - \beta)] \end{aligned}$$

Each of the terms in this expression are strictly positive, hence the derivative is strictly positive. □

(*iii*) For a demonstration that $e(\beta)\beta \geq 0$, see the proof of Proposition 3 in Eliaz and Spiegler (2019). Our remaining task is thus to show that $|e(\beta)| \leq |\beta|$. The proof will rely on two properties of G: (1) $G(\Delta) + G(-\Delta) = 1$ for every Δ (due to symmetry of g); (2) G is strictly convex over $\Delta < 0$ and strictly concave over $\Delta > 0$ (due to normality of g). Suppose $\beta > 0$, without loss of generality. Denote $d(\beta) = (1+2c_2)e(\beta) - \beta$. Note that $d(\beta) \leq e(\beta) - \beta$. Substituting (10) for $(1 + 2c_2)e(\beta)$, we obtain

$$d(\beta) = \int_{-c^*-\beta}^{c^*-\beta} G(\Delta) - (c^* - c_1)[G(-c^* - \beta) + G(c^* - \beta)] - c_1$$

Define $d^0(\beta)$ as the value of $d(\beta)$ when $c_1 = 0$. That is,

$$d^0(\beta) = \int_{-c^*-\beta}^{c^*-\beta} G(\Delta) - c^*[G(-c^* - \beta) + G(c^* - \beta)]$$

Let us first prove the claim for d^0. By property (1) above, $d^0(0) = 0$.

Assume $\beta > 0$ (this is without loss of generality). The above expression for $d^0(\beta)$ can be viewed as the difference between two terms. The first term, $\int_{-c^*-\beta}^{c^*-\beta} G(\Delta)$, represents the area under G over the range $[-c^* - \beta, c^* - \beta]$. The second term, $c^*[G(c^* - \beta) + G(-c^* - \beta)]$, is the area of the trapezoid whose nodes are the points $(c^* - \beta, 0), (c^* - \beta, G(c^* - \beta)), (-c^* - \beta, 0), (-c^* - \beta, G(-c^* - \beta))$. Our task is to show that the area represented by the first term is strictly smaller than the area represented by the second term. Suppose that $\beta \geq c^*$. Then, because G is strictly convex over $\Delta < 0$, the trapezoid strictly contains the area under G in the range $[-c^* - \beta, c^* - \beta]$, which immediately implies the result for this range of values of β. Next, suppose that $\beta \in (0, c^*)$. Consider the line that connects the points $(c^* - \beta, G(c^* - \beta))$ and $(-c^* + \beta, G(-c^* + \beta))$. Thanks to property (2) above, this line lies below G when $\Delta \in [0, c^* - \beta]$ and above G when $\Delta \in [-c^* + \beta, 0]$. By property (1) above, the areas between this line and G over the two intervals $[0, c^* - \beta]$ and $[-c^* + \beta, 0]$ are equal. Now, because G is strictly convex over negative values of Δ, the line lies strictly below the side of the trapezoid that connects the nodes $(c^* - \beta, G(c^* - \beta))$ and $(-c^* - \beta, G(-c^* - \beta))$. This in turn implies that the area between this trapezoid side and G to the left of their intersection point is strictly larger than the area between the trapezoid side and G to the right of their intersection point, which proves the result for this range of values of β.

Now, observe that

$$\begin{aligned} d(\beta) &= d^0(\beta) + c_1[G(-c^* - \beta) + G(c^* - \beta) - 1] \\ &\leq d^0(\beta) + c_1[G(-c^*) + G(c) - 1] \\ &= d^0(\beta) \end{aligned}$$

where the first inequality follows from examining the case of $\beta > 0$, and the second equality follows from the symmetry of g around zero. Then, we have established that $d(\beta) \leq d^0(\beta) < 0$. Thus, $e(\beta) < \beta$. ■

Proof of Lemma 4

Throughout the proof, we use V to denote the set of selected variables given some ε - i.e.,

$$V = \{k = 1, ..., K \mid b_k(\varepsilon) \neq 0\}$$

We begin with the following lemma.

Claim 1 *The deviation M is unprofitable for given β, x if and only if*

$$\mathbb{E}_\varepsilon \left[\left(\sum_{k \in M} b_k(\varepsilon, \beta) w_k \right) \left(2\bar{\varepsilon} + \sum_{k=1}^{K} \beta_k w_k - \sum_{k \notin M} b_k(\varepsilon, \beta) w_k \right) \right] \geq 0 \qquad (11)$$

Proof. Denote $z_k = r_k - x_k$. Inequality (5) can be rewritten as:

$$\begin{aligned} & \mathbb{E}_\varepsilon \left[b_0(\varepsilon, \beta) + \sum_{k=1}^{K} b_k(\varepsilon, \beta) x_k - \beta_0 - \sum_{k=1}^{K} \beta_k x_k \right]^2 \\ \leq \; & \mathbb{E}_\varepsilon \left[b_0(\varepsilon, \beta) + \sum_{k=1}^{K} b_k(\varepsilon, \beta) x_k + \sum_{k=1}^{K} b_k(\varepsilon, \beta) z_k - \beta_0 - \sum_{k=1}^{K} \beta_k x_k \right]^2 \end{aligned}$$

This inequality can be simplified into

$$\mathbb{E}_\varepsilon \left(\sum_{k=1}^{K} b_k(\varepsilon, \beta) z_k \right) \left(\sum_{k=1}^{K} b_k(\varepsilon, \beta) z_k + 2b_0(\varepsilon, \beta) + 2 \sum_{k=1}^{K} b_k(\varepsilon, \beta) x_k - 2\beta_0 - 2 \sum_{k=1}^{K} \beta_k x_k \right) \geq 0$$

Then, (5) can be rewritten as

$$\mathbb{E}_\varepsilon \left[\left(\sum_{k \in V} b_k(\varepsilon, \beta) z_k \right) \left(\sum_{k \in V} b_k(\varepsilon, \beta) z_k + 2b_0(\varepsilon, \beta) + 2 \sum_{k \in V} b_k(\varepsilon, \beta) x_k - 2\beta_0 - 2 \sum_{k=1}^{K} \beta_k x_k \right) \right] \geq 0$$

Note that for each $k \in M \cap V$, $z_k = 1 - 2x_k$, while for each $k \in V - M$,

$z_k = 0$. Note also that

$$b_0(\varepsilon, \beta) = \beta_0 + \frac{1}{2}\sum_{k=1}^{K} \beta_k + \bar{\varepsilon} - \frac{1}{2}\sum_{k \in V} b_k(\varepsilon, \beta)$$

Hence, we can rewrite the above inequality as follows:

$$\mathbb{E}_\varepsilon \left\{ \left[\sum_{k \in M \cap V} b_k(\varepsilon, \beta)(1 - 2x_k) \right] \left[2\bar{\varepsilon} + \sum_{k=1}^{K} \beta_k (1 - 2x_k) - \sum_{k \in V - M} b_k(\varepsilon, \beta)(1 - 2x_k) \right] \right\} \geq 0$$

Since $w_k = 1 - 2x_k$ and $b_k(\varepsilon, \beta) = 0$ for each $k \notin V$, the above inequality is equivalent to (11). ■

Fix a profile of realized coefficients $b = (b_1, ..., b_K)$. Our first step is to show that $\mathbb{E}(\bar{\varepsilon} \mid b) = 0$. We already observed that $E(\Delta_k \bar{\varepsilon}) = 0$ for any $k = 1, ..., K$. Because both Δ_k and $\bar{\varepsilon}$ are normally distributed with mean zero, this means that $\bar{\varepsilon}$ and Δ_k are statistically independent for all $k = 1, ..., K$. Since b is purely a function of $\Delta_1, ..., \Delta_K$, it follows that $\bar{\varepsilon}$ is independent of b. Since $\mathbb{E}(\bar{\varepsilon}) = 0$, we conclude that $\mathbb{E}(\bar{\varepsilon} \mid b) = 0$ for any b, hence $\mathbb{E}(\bar{\varepsilon} \mid V) = 0$ for any V. This means that inequality (11) can be simplified into

$$\sum_V \Pr(V) \mathbb{E}_\varepsilon \left[\left(\sum_{k \in V \cap M} b_k(\varepsilon, \beta) w_k \right) \left(\sum_{k=1}^{K} \beta_k w_k - \sum_{k \in V - M} b_k(\varepsilon, \beta) w_k \right) \mid V \right] \geq 0$$

Our next step is to characterize $\Pr(V)$, namely the probability that the set of variables V is selected. Recall that whether or not $b_k(\varepsilon, \beta) \neq 0$, and the distribution of $b_k(\varepsilon, \beta)$, conditional on it being non-zero, depend only on Δ_k and the parameters of the model (the true coefficients and the costs). Because all Δ_k are mutually independent, the probability that $k \in V$ is independent, and denoted $\lambda_k = \Pr(\beta_k + \Delta_k)^2 > c^*$ (where c^* is defined as in the previous

proof). Therefore,

$$\Pr(V) = \prod_{k \in V} \lambda_k \prod_{j \notin V} (1 - \lambda_j) \tag{12}$$

This enables us to further simplify the condition for the unprofitability of the deviation:

$$\begin{aligned} &\sum_{k=1}^{K} \beta_k w_k \sum_{k \in M} \lambda_k w_k \mathbb{E}_\varepsilon(b_k(\varepsilon, \beta) \quad | \quad k \in V) \\ -\sum_{k \in M} \sum_{j \notin M} &\lambda_k \lambda_j w_k w_j \mathbb{E}_\varepsilon(b_k(\varepsilon, \beta) b_j(\varepsilon, \beta) \quad | \quad \{k, j\} \subseteq V) \geq 0 \end{aligned}$$

Because we have established that b_k and b_j are statistically independent whenever $k \neq j$,

$$\mathbb{E}_\varepsilon(b_k(\varepsilon, \beta) b_j(\varepsilon, \beta) \mid \{k, j\} \subseteq V) = \mathbb{E}_\varepsilon(b_k(\varepsilon, \beta) \mid k \in V) \mathbb{E}_\varepsilon(b_j(\varepsilon, \beta) \mid j \in V)$$

Furthermore, observe that $\lambda_k \mathbb{E}_\varepsilon(b_k(\varepsilon, \beta) \mid k \in V)$ is equal to $\mathbb{E}_\varepsilon(b_k(\varepsilon, \beta))$, namely the *ex-ante* expectation of b_k - which we have denoted by $e(\beta_k)$. Therefore, we can further simplify the inequality into

$$\left(\sum_{k \in M} e(\beta_k) w_k \right) \left(\sum_{k=1}^{K} \beta_k w_k - \sum_{j \notin M} e(\beta_j) w_j \right) \geq 0$$

This completes the proof. ■

Proof of Proposition 1

Plug $b_k(\varepsilon, \beta) = \beta_k + \Delta_k$ into Condition (11):

$$\mathbb{E}_\varepsilon \left(\sum_{k \in M} (\beta_k + \Delta_k) w_k \right) \left(2\bar{\varepsilon} + \sum_{k=1}^{K} \beta_k w_k - \sum_{k \notin M} (\beta_k + \Delta_k) w_k \right) \geq 0$$

The L.H.S can be elaborated as follows:

$$2\sum_{k\in M}\beta_k w_k\mathbb{E}(\bar{\varepsilon})+\sum_{k\in M}2w_k\mathbb{E}(\Delta_k\bar{\varepsilon})+\left(\sum_{k\in M}\beta_k w_k\right)^2+\sum_{k\in M}(w_k)^2\beta_k\mathbb{E}(\Delta_k)$$
$$-\left(\sum_{k\in M}\beta_k w_k\right)\left(\sum_{j\notin M}w_j\mathbb{E}(\Delta_j)\right)-\mathbb{E}\left(\sum_{k\in M}\Delta_k w_k\right)\left(\sum_{j\notin M}\Delta_j w_j\right)$$

The first term is equal to zero because $\mathbb{E}(\bar{\varepsilon})=0$. Likewise, the fourth and fifth terms are equal to zero because $\mathbb{E}(\Delta_k)=0$ for every k. The last term is equal to zero because $\mathbb{E}(\Delta_k\Delta_j)=0$ whenever $k\neq j$. As to the second term, Finally, recall that for every k, we can write

$$\begin{aligned}\Delta_k &= \bar{\varepsilon}_{x_k=1}-\bar{\varepsilon}_{x_k=0}\\ 2\bar{\varepsilon} &= \bar{\varepsilon}_{x_k=1}+\bar{\varepsilon}_{x_k=0}\end{aligned}$$

such that

$$\mathbb{E}(\Delta_k\bar{\varepsilon})=\mathbb{E}(\bar{\varepsilon}_{x_k=1}+\bar{\varepsilon}_{x_k=0})(\bar{\varepsilon}_{x_k=1}-\bar{\varepsilon}_{x_k=0})=\mathbb{E}\left[(\bar{\varepsilon}_{x_k=1})^2-(\bar{\varepsilon}_{x_k=0})^2\right]$$

which is equal to zero because $\bar{\varepsilon}_{x_k=1}$ and $\bar{\varepsilon}_{x_k=0}$ are *i.i.d.* It follows that the only non-zero term on the L.H.S of the condition is

$$\left(\sum_{k\in V_1}\beta_k w_k\right)^2$$

which is obviously non-negative. ■

Proof of Proposition 3

Denote $\beta_M=(\beta_k)_{k\in M}$, $\beta_{-M}=(\beta_k)_{k\notin M}$. Because of the independence across

components, the L.H.S of (6) can be written as

$$\mathbb{E}_{\beta_M}\left[\left(\sum_{k\in M} e(\beta_k)w_k\right)\left(\sum_{k\in M}\beta_k w_k\right)\right]$$
$$-\mathbb{E}_{\beta_M}\left(\sum_{k\in M} e(\beta_k)w_k\right)\mathbb{E}_{\beta_{-M}}\left(\sum_{j\notin M}(e(\beta_j)-\beta_j)w_j\right)$$

Recall that e is an anti-symmetric function. Therefore, $e(\beta)-\beta$ is also anti-symmetric. Combined with the symmetry around zero of the prior over each β_j, $\mathbb{E}_{\beta_j}(e(\beta_j)-\beta_j)w_j = 0$ for every j. Recall that $w_k \in \{-1,1\}$, such that $(w_k)^2 = 1$. The inequality thus becomes

$$\begin{aligned}
&\mathbb{E}_{\beta_M}\left[\left(\sum_{k\in M} e(\beta_k)w_k\right)\left(\sum_{k\in M}\beta_k w_k\right)\right] \\
= \;& \mathbb{E}_{\beta_M}\left[\sum_{k\in M} e(\beta_k)\beta_k + \sum_{k,j\in M,k\neq j} e(\beta_k)\beta_j w_k w_j\right] \\
= \;& \sum_{k\in M}\mathbb{E}(e(\beta_k)\beta_k) + \sum_{k,j\in M,k\neq j} w_k w_j \mathbb{E}(e(\beta_k))\mathbb{E}(\beta_j) \geq 0
\end{aligned}$$

Because $\mathbb{E}(\beta_j) = 0$ for every j, this inequality is reduced to

$$\sum_{k\in M}\mathbb{E}(e(\beta_k)\beta_k) \geq 0$$

Recall that $sign[e(\beta)] = sign(\beta)$ for every β, hence this inequality holds. ■

Proof of Proposition 4

Given the independence assumption, a deviation M is profitable if

$$\mathbb{E}_{\beta_M}\left[\left(\sum_{k\in M} e(\beta_k)w_k\right)\left(\sum_{k\in M}\beta_k w_k\right)\right] - \mathbb{E}_{\beta_M}\left(\sum_{k\in M} e(\beta_k)w_k\right)\mathbb{E}_{\beta_{-M}}\left(\sum_{j\notin M}(e(\beta_j)-\beta_j)w_j\right)$$

is strictly negative, as in the previous example. Denote $m = |M|$. Using the $i.i.d$ assumption, we can simplify the terms. The first term is

$$\begin{aligned} & \mathbb{E}_{\beta_M}\left[\left(\sum_{k\in M} e(\beta_k)w_k\right)\left(\sum_{k\in M}\beta_k w_k\right)\right] \\ = & \sum_{k\in M}\mathbb{E}(e(\beta_k)\beta_k) + \sum_{k,j\in M,k\neq j} w_k w_j \mathbb{E}(e(\beta_k))\mathbb{E}(\beta_j) \\ = & m\mathbb{E}(e(\beta)\beta) + e^*\beta^* \sum_{k,j\in M,k\neq j} w_k w_j \end{aligned}$$

The second term is

$$\begin{aligned} & \mathbb{E}_{\beta_M}\left(\sum_{k\in M} e(\beta_k)w_k\right)\mathbb{E}_{\beta_{-M}}\left(\sum_{j\notin M}(e(\beta_j)-\beta_j)w_j\right) \\ = & ((e^*)^2 - e^*\beta^*)\sum_{k\in M} w_k \sum_{j\notin M} w_j \end{aligned}$$

The condition then becomes

$$m\mathbb{E}(e(\beta)\beta) + e^*\left[\beta^* \sum_{k,j\in M,k\neq j} w_k w_j + (\beta^* - e^*)\sum_{k\in M} w_k \sum_{j\notin M} w_j\right] < 0 \quad (13)$$

Define M to be *homogenous* if $w_k = w_j$ for every $k, j \in M$. Suppose that M is not homogenous - i.e., there exist $k, j \in M$ such that $w_k = 1$ and $w_j = -1$. Let us consider two cases. First, suppose $m = 2$. Then, $\sum_{k\in M} w_k = 0$ and $\sum_{k,j\in M,k\neq j} w_k w_j = -1$, such that (13) is reduced to

$$\mathbb{E}(e(\beta)\beta) - e^*\beta^* < 0$$

Because e is strictly increasing in β, this contradicts Chebyshev's algebraic inequality. Therefore, M is unprofitable, a contradiction. Second, suppose

that $m > 2$. Consider the deviation $M' = M - \{k, j\}$. Then:

$$\begin{aligned} |M'| &= m - 2 \\ \sum_{i \in M'} w_i &= \sum_{i \in M} w_i \\ \sum_{i,h \in M', i \neq h} w_i w_h &= \sum_{i,h \in M, i \neq h} w_i w_h + 1 \end{aligned}$$

such that as a result of the deviation, the L.H.S of (13) decreases by $2\mathbb{E}(e(\beta)\beta) - 2e^*\beta^*$, which we have established to be weakly positive. We can repeat this argument until we obtain a homogenous deviation M'' that is at least as profitable as M.

It follows that if there is a profitable deviation M, we can set it to be homogenous without loss of generality. Inequality (13) becomes

$$m\mathbb{E}(e(\beta)\beta) + e^* \left[\beta^* m(m-1) - (\beta^* - e^*)m(K-m)\right] < 0$$

We have already established that $e(\beta)\beta \geq 0$ and $0 < |e^*| < |\beta^*|$. Therefore, $e^*\beta^* > 0$ and $e^*(\beta^* - e^*) > 0$. The L.H.S of the inequality thus unambiguously increases with m. There are two candidates for a homogenous deviation: $\{k \mid w_k = 1\}$ or $\{k \mid w_k = -1\}$. Therefore, the more profitable of them is the smaller one, namely M^*. ■