Incentive-Compatible Estimators*

Kfir Eliaz† and Ran Spiegler‡

January 14, 2018

Abstract

We study a model in which a "statistician" takes an action on behalf of an agent, based on a random sample involving other people. The statistician follows a penalized regression procedure: the action that he takes is the dependent variable's estimated value given the agent's disclosed personal characteristics. We ask the following question: Is truth-telling an optimal disclosure strategy for the agent, given the statistician's procedure? We discuss possible implications of our exercise for the growing reliance on "machine learning" methods that involve explicit variable selection.

*We thank Yoav Binyamini, Assaf Cohen, Rami Atar, Lorens Imhof, Benny Moldovanu, Ron Peretz and especially Martin Cripps for helpful conversations. We are also grateful to seminar audiences at BRIQ, DICE and the Warwick Economic Theory conference, for their useful comments.

†School of Economics, Tel-Aviv University and Economics Dept., Aarhus University. E-mail: kfire@post.tau.ac.il.

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

1 Introduction

In recent years, actions in ever-expanding domains are taken on our behalf by automatic systems that rely on machine-learning tools. Consider the case of online content provision. A website obtains information about a user's personal characteristics. Some of these characteristics are actively provided by the user himself; others are obtained by monitoring his navigation history. The website then feeds these characteristics into a predictive statistical model, which is estimated on a sample consisting of observations of other users. The estimated model then outputs a prediction of the user's ideal content. In domains like autonomous driving or medical decision making, AI systems are mostly confined to issuing recommendations for a human decision maker. In the future, however, it is possible that decisions in such domains will be entirely based on machine learning.

How should users interact with such a procedure? In particular, should they truthfully share personal characteristics with the automatic system? Of course, in the presence of a conflict of interests between the two parties - e.g., when an online content provider operating the automatic system has a distinct political or commercial agenda - the user might be better off if he misreports his characteristics, deletes "cookies" from his computer or adopts incognito browsing. This is a familiar situation of communication under misaligned preferences, which seems amenable to economists' standard model of strategic information transmission as a game of incomplete information (with a common prior).

However, suppose that there is no conflict of interests between the two parties - i.e., the objective behind the machine-learning algorithm is to make the best prediction of the user's ideal action. But how do such systems perform this prediction task in reality? Consider a basic tool like LASSO (Tibshirani (1996)).[1] This is a variant on standard linear regression analy-

[1] Least Absolute Shrinkage and Selection Operator.

sis, which adds a cost function that penalizes non-zero coefficients. It is considered useful in situations where users have a great number of potentially relevant characteristics that could influence their ideal action. The procedure involves both variable selection (i.e. choosing which of the many variables will enter the regression) and estimation of the selected variables' coefficients. The predicted action for an agent with a particular vector of personal characteristics x is the dependent variable's estimated value at x.

A penalized-regression procedure like LASSO is not fundamentally Bayesian. Indeed, it is an extension of a familiar classical-statistics procedure. Although it is possible to justify LASSO estimates as properties of a Bayesian posterior derived from some prior (Tibshirani (1996), Park and Casella (2008)), these properties are not necessarily relevant for maximizing the user's welfare. Furthermore, there is no reason to assume that the prior that rationalizes LASSO coincides with the user's actual prior beliefs. Thus, neither the preferences nor the prior beliefs they involve are necessarily the ones an economic modeler would like to attribute to the user in a plausible model of the interaction. This observation could be extended to many machine-learning predictive methods. If we want to model human interaction with such algorithms, some departure from the standard Bayesian framework with common priors seems to be required. Put differently, if one were to analyze a model with common priors, where a benevolent Bayesian decision maker tries to take the optimal action for an agent with unknown characteristics, then for almost all prior beliefs, the decision maker's behavior will not be mimicked by a familiar machine-learning procedure.

Motivated by this observation, we present a model of an interaction between an "agent" and a "statistician" - the latter is a stand-in for 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. Specifically, he obtains N samples points for each configuration of agent characteristics. This sample is the statistician's private information - i.e., the agent is not exposed to it. 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). 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. The agent's payoff is a standard quadratic loss function - thus coinciding with the most basic criterion for evaluating estimators' predictive success.

We ask 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, we say that the statistician's procedure (or "estimator") is *incentive-compatible*. Thus, in line with the methodological observation above, we do not think of the statistician as a Bayesian decision maker who shares the agent's prior, observes a signal (i.e., the sample) and takes an action that maximizes the agent's expected payoff according to the Bayesian posterior belief. Instead, we take the penalized regression method *as given* and ask whether it creates an incentive for the agent to misreport his personal characteristics.

As mentioned above, variable selection is a key feature of penalized-regression methods. It also turns out to be crucial for our main question. When the statistician's procedure involves no variable selection (i.e., it is OLS), it is incentive-compatible. This result relies on the assumption that the statistician obtains the same number of observations for each characteristics vector. Introducing variable selection can create an incentive problem.

(Thus, our uniform-sample assumption focuses serves to focus our attention on the effects of variable selection.)

We begin our analysis of this problem with the case of a single explanatory variable - i.e., the agent's reporting decision involves ticking only one yes/no box. We show that the statistician's procedure gives rise to a "*variable selection curse*". Because the agent's report only matters when the variable is selected to be relevant, he should only care about the distribution of the variable's estimated coefficient conditional on the "pivotal event" in which the variable is selected. As the terminology suggests, the logic is reminiscent of pivotal-thinking phenomena like the winner's curse in auction theory (Milgrom and Weber (1982)) or the swing voter's curse in the theory of strategic voting (Pesendorfer and Feddersen (1996)). One can construct distributions of the sample noise for which the estimated coefficient conditional on the pivotal event is so biased that the agent is better off introducing a counter-bias by misreporting his personal characteristic. Furthermore, the variable selection curse does not disappear with large samples: If the noise distribution is asymmetric, the statistician's procedure can fail incentive compatibility even asymptotically. In contrast, we show that when the sample noise is symmetrically distributed, the estimator is incentive-compatible in the single-variable case.

Next, we consider multiple explanatory variables. In this case, variable selection can generate an incentive problem even if the statistician faces no sampling error. The reason is that the cumulative bias due to the exclusion of multiple variables can be so large that the agent would like to introduce a counter-bias by misreporting the value of an included variable. We then introduce normally distributed sample noise. This makes the problem tractable and we are able to obtain simple conditions for the procedure's robustness to misreporting for various classes of the agent's priors regarding the model's true coefficients. First, the procedure is not incentive-compatible because there exist prior beliefs for which the agent would like to misreport at least one characteristic. Second, we 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, when the agent's prior over each coefficient is $i.i.d$ (but with non-zero mean), the agent has no incentive to misreport only if the profile of his personal characteristics is sufficiently balanced - i.e., its number of 0's and 1's is not too different. This result has an implication for the question of whether the agent has an incentive to "delete cookies" from his computer when facing a penalized-regression system: the agent has a disincentive to delete cookies only if has a sufficient number of them.

The lesson from our analysis is that the variable selection aspect of penalized-regression procedures creates an incentive problem. This has potentially broader implications for the evaluation of machine-learning algorithms. Even when they are good at predicting an agent's ideal action on average, his cooperation with the algorithm depends on other statistical properties - e.g., the bias of estimated coefficients *conditional* on being non-zero. Integrating incentive compatibility into the evaluation of estimation and prediction methods is an interesting project for future research.

2 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. 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 some distribution with zero mean. 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 assumption of a uniform sample (N observations for each value of x) in Section 3.1. 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.

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 $(y_x^n)_{x=0,1}^{n=1,...,N}$, the statistician solves the following minimization problem,

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

We denote the solution to this problem by $b(\varepsilon) = (b_0(\varepsilon), ..., b_K(\varepsilon))$, and refer to $(b(\varepsilon))_\varepsilon$ as the *estimator*. 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*.

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 $a = \sum_{k=0}^{K} b_k(\varepsilon) r_k$. The agent's expected payoff for given $\beta_0, ..., \beta_K$ is therefore

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

Discussion

The agent's preferences are given by a quadratic loss function. This is also a standard criterion for evaluating the predictive success of estimators. Suppose that $r = x$ - i.e., the agent submits a truthful report of his personal characteristic. Then, $\hat{f}(x) = \sum_{k=0}^{K} b_k(\varepsilon)x_k$ is the predicted ideal action for the agent. 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.

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):

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

> "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 to be useful for predictive success, where the criterion for predictive success coincides with the agent's expected utility given the true model. The standard formalization of this description assumes the statistician has well-defined preferences that coincide with the agent's and rationalize his procedure. In the Introduction, we explained the difficulty to rationalize the statistician's procedure in these terms. Formal justifications for penalized-regression methods in the literature (e.g. 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 $\beta_0, ..., \beta_K$, without going all the way to a complete Bayesian rationalization.

The second justification for penalized regression that the quote invokes is essentially a *bounded rationality* rationale. Dealing with large models is difficult, and users of statistical analysis 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. In this sense, our model falls into the bounded rationality literature - it describes interaction between a Bayesian-rational agent and a boundedly rational decision maker.

2.1 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, $\bar{\varepsilon}_k^1$ and $\bar{\varepsilon}_k^0$ denote the average noise realization in the subsamples for which $x_k = 1$ and $x_k = 0$, respectively:

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

Finally, define $\Delta_k = \bar{\varepsilon}_k^1 - \bar{\varepsilon}_k^0$.

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:

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

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

$$b_k(\varepsilon) = \begin{cases} \tilde{b}_k(\varepsilon) & if \quad (\tilde{b}_k(\varepsilon))^2 \geq 2c_0 \\ 0 & if \quad (\tilde{b}_k(\varepsilon))^2 < 2c_0 \end{cases} \tag{3}$$

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

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

Thus, $b_k(\varepsilon)$ 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 achieved thanks to the assumption of a uniform sample.) Of course, this does not imply that it is *statistically* independent of Δ_j, $j \neq k$. 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 $c_0 = 0$, $b_k = \tilde{b}_k$.

2.2 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,*

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

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. The expectation operator $\mathbb{E}_\beta$ 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,*

$$\mathbb{E}_\varepsilon \left[\sum_{k=0}^{K} (b_k(\varepsilon) - \beta_k) x_k \right]^2 \leq \mathbb{E}_\varepsilon \left[\sum_{k=0}^{K} (b_k(\varepsilon) r_k - \beta_k x_k) \right]^2 \quad (4)$$

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

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. How should we interpret this requirement, given that we do not necessarily want to think of the agent as being sophisticated enough to think in these terms? One interpretation is that lack of incentive compatibility is merely a *normative* statement about the agent's welfare - namely, given our model of how the statistician takes actions on the agent's behalf, it would be advisable for him to misrepresent his personal characteristics. Furthermore, 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 theoretically eliminates the need for such an industry. In the context of the online content provision story, some misreporting strategies take the form of

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

"deleting cookies". This deviation is straightforward to implement, and the agent can check if it makes him better off in the long run.

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.

3 Analysis: The Single Variable Case

We begin our analysis in the case of a single explanatory variable - i.e. $K = 1$. Although there is something ironic about single-variable analysis of machine learning methods, we follow here the tradition of microeconomic theory and start with the simplest version of our model. Indeed, key aspects of the incentive-compatibility problem will be manifest even in this simple case. Furthermore, a few results in this section will also be relevant in the multi-variable case. Note that in the single-variable case, the linear form of f is without loss of generality because x_1 is a binary variable. Throughout this section, we abuse notation and remove the subscripts from x_1 and Δ_1.

3.1 Two Benchmarks

There are two factors that *jointly* give rise to an incentive compatibility problem: sample noise and variable selection. In this sub-section we establish that neither factor generates an incentive problem on its own in the single-variable case.

First, suppose that the statistician makes perfectly precise measurements - that is, $\varepsilon_x^n = 0$ by definition for every x, n. In this case, it is easy to see that if $c_0 = c_1 = c_2 = 0$, the statistician's objective function coincides with the agent's payoff for any given β. However, the introduction of complexity cost creates a de-facto conflict of interests between the two parties, because the statistician ends up choosing an action that maximizes a different deterministic payoff function than the agent's. Nevertheless, the following simple result establishes that this by itself does not give the agent a reason to misreport his personal characteristic.

Claim 1 *Suppose that $\varepsilon_x^n = 0$ with probability one for every x, n. Then, the estimator is incentive compatible.*

Proof. The agent can perfectly predict b_0, b_1 as a function of β_0, β_1. Suppose that β_1 is such that $b_1 = 0$. Then, the agent's report has no effect on the statistician's action, and the incentive-compatibility condition holds trivially. Now suppose that β_1 is such that $b_1 > 0$. Given the characterization of b_1, it must be the case that $\beta_1 - c_1 \geq 0$. The statistician's action as a function of the agent's report is b_0 if $r = 0$, and $b_0 + b_1$ if $r = 1$, where

$$\begin{aligned} b_0 &= \beta_0 + \frac{1}{2}\beta_1 - \frac{1}{2}b_1 = \beta_0 + \frac{1}{2}\beta_1 - \frac{1}{2}(\beta_1 - c_1)/(1 + 2c_2) \\ b_0 + b_1 &= \beta_0 + \frac{1}{2}\beta_1 - \frac{1}{2}b_1 + b_1 = \beta_0 + \frac{1}{2}\beta_1 + \frac{1}{2}(\beta_1 - c_1)/(1 + 2c_2) \end{aligned}$$

When $x = 0$, the agent's ideal action is β_0. Because $\beta_1 - c_1 \geq 0$, the action b_0 is closer to the ideal point than the action $b_0 + b_1$. Therefore, truthful reporting is optimal for the agent. Likewise, when $x = 1$, the agent's ideal action is $\beta_0 + \beta_1$. Because $\beta_1 - c_1 \geq 0$, the action $b_0 + b_1$ is closer to the ideal point than the action b_0. Therefore, truthful reporting is optimal for the agent.

A similar calculation establishes incentive compatibility when $b_1 < 0$. ■

Suppose next that the statistician faces sample noise and employs standard OLS. The next result shows that incentive compatibility holds in this case. Although it is a special case of a result we will prove in Section 4.2, we present the proof because it sheds light on the incentive-compatibility problem in the single-variable case.

Claim 2 *If $c_0 = c_1 = c_2 = 0$, then the estimator is incentive-compatible.*

Proof. The coefficient b_1 is included in the regression for all realizations of ε_0 and ε_1. Suppose $x = 1$ and the agent contemplates whether to report $r(x) = 0$. In this case inequality (4) can be simplified into

$$\mathbb{E}_{\varepsilon_0,\varepsilon_1}[(b_1(\varepsilon))^2 + 2b_1(\varepsilon) \cdot (b_0(\varepsilon) - \beta_0 - \beta_1)] \leq 0$$

Plugging in the expressions for $b_0(\varepsilon)$ and $b_1(\varepsilon)$ given by (3), this inequality reduces to

$$\mathbb{E}_{\varepsilon_0,\varepsilon_1}[-(\beta_1)^2 + 2\beta_1\bar{\varepsilon}_0 + (\bar{\varepsilon}_1)^2 - (\bar{\varepsilon}_0)^2] \leq 0 \qquad (5)$$

Since ε_0 and ε_1 are $i.i.d$ with mean zero, this inequality immediately holds for all β_1. An analogous argument shows that an agent with $x = 0$ will not benefit from reporting $r(x) = 1$. Therefore, the OLS estimator is incentive-compatible. ■

Intuitively, when the statistician uses OLS, his estimates are unbiased. Therefore, although his action deviates from the Bayesian-optimal response to his sample, the deviation is not systematic and therefore the agent would not want to create a bias by misreporting. However, this intuition is misleading because it crucially relies on the *uniform sample* - i.e., the assumption that the statistician draws the same number of observations from $x = 0$ and from $x = 1$ (even if their proportions in the population are uneven).

To see this, suppose there are N_0 observations with $x = 0$ and $N_1 \neq N_0$ observations with $x = 1$. Assume first that $N_0 > N_1$. Then, $\mathbb{E}_{\varepsilon_0,\varepsilon_1}(\bar{\varepsilon}_1)^2 >$

$\mathbb{E}_{\varepsilon_0,\varepsilon_1}(\bar{\varepsilon}_0)^2$. When β_1 is small, inequality (5) will fail - i.e., an agent with $x = 1$ will prefer to report $r = 0$. Likewise, when $N_0 < N_1$, an agent with $x = 0$ will prefer to report $r = 1$ when β_1 is small. Thus, *heteroskedasticity* (i.e., differences between observations with $x = 0$ and observations with $x = 1$) creates an incentive problem, because of the bias-variance trade-off that characterizes the agent's reporting decision. If β_1 is small, the bias due to misreporting is relatively small, and may be overweighed by the reduced variance due to the larger sample taken for the value of x that the agent pretends to be. Thus, uniform samples are *necessary* for incentive compatibility, because they imply homoskedasticity. Partly for this reason, we insist on uniform samples throughout the paper (the other reason is tractability).

3.2 The Variable Selection Curse

We now turn to the case of noisy measurement and non-zero complexity costs. The following examples illustrate that incentive compatibility can fail in this case. For expositional simplicity, we consider only the L_0 penalty (i.e., $c_0 > 0 = c_1 = c_2$) and let $N = 1$ (hence, we suppress the observation superscripts of x, y and ε).

Example 1: *Bernoulli noise*
Suppose the noise follows a Bernoulli probability distribution that assigns probability $p > 0.5$ to -1 and probability $1-p$ to $d = p/(1-p) > 1$. Consider an agent with $x = 1$. If this agent reports $r = 0$, this misrepresentation violates incentive compatibility if there is some β_1 for which

$$\mathbb{E}_{\varepsilon}\left[b_0(\varepsilon) + b_1(\varepsilon) - \beta_0 - \beta_1\right]^2 > \mathbb{E}_{\varepsilon}\left[b_0(\varepsilon) - \beta_0 - \beta_1\right]^2$$

Because the agent's misrepresentation matters only in the "pivotal event" in which $b_1(\varepsilon) \neq 0$, this inequality can be rewritten as

$$\mathbb{E}_{\varepsilon_0,\varepsilon_1}[-(\beta_1)^2 + 2\beta_1\varepsilon_0 + (\varepsilon_1)^2 - (\varepsilon_0)^2 \mid (\beta_1 + \varepsilon_1 - \varepsilon_0)^2 \geq 2c_0] > 0 \qquad (6)$$

For every $\beta_1 > 0$ we can find a range of values for c_0 such that $(\beta_1+\varepsilon_1-\varepsilon_0)^2 \geq 2c_0$ only when $\varepsilon_1 = d$ and $\varepsilon_0 = -1$. In this case (6) is reduced to $\beta_1 < d-1$. Therefore, every pair of positive numbers (β_1, c_0) that satisfies the inequalities

$$\begin{aligned} -(d+1) &< \sqrt{2c_0} - \beta_1 < d+1 \\ \beta_1 &< d-1 \end{aligned}$$

will violate incentive compatibility.

The intuition for this violation of incentive compatibility is as follows. An agent with $x = 1$ focuses only on the pivotal event in which his report matters - i.e. $\{\varepsilon \mid b_1(\varepsilon) \neq 0\}$. This event is largely determined by the difference in noise realizations, $\varepsilon_1 - \varepsilon_0$. For a range of values of β_1 and c_0, $\varepsilon_1 - \varepsilon_0 = d+1$ with probability one conditional on the pivotal event. This produces such a biased estimate of b_1 that the agent prefers to shut down the pivotal event, by pretending to be $x = 0$. $\square$

Example 1 illustrates a feature we refer to as the "*variable selection curse*", in the spirit of the "winner's curse" and "swing voter's curse". Like these very familiar phenomena, the variable selection curse involves statistical inferences from a "pivotal event". Here, the pivotal event is the inclusion of a variable in the regression. The agent's decision whether to misreport his personal characteristic is relevant only if the statistician chooses to include the variable in his regression. Misreporting will change the statistician's action by $b_1(\varepsilon)(r-x)$. Therefore, the agent only cares about the distribution of $b_1(\varepsilon)$ conditional on the event $\{\varepsilon \mid b_1(\varepsilon) \neq 0\}$. This distribution can be so skewed that the agent will prefer to introduce a bias in the opposite direction by misreporting.

The following example shows that the variable selection curse can occur for more realistic noise realizations.

Example 2: *Exponential noise*
Suppose the observations on $x \in \{0,1\}$ take the form $y_x = \beta_0 + \beta_1 x_1 + \eta_x$,

where η_0 and η_1 are drawn *i.i.d* from the *exponential distribution* with decay parameter 1. One story behind this specification is that $f(x) = \beta_0 + \beta_1 x$ is the ideal dosage of some medication when the agent is treated *immediately* after a medical incident (e.g., stroke). The personal characteristic x is a medical indicator that may be relevant for the ideal dosage. However, the statistician's sample consists of observations in which medical treatment was *delayed*. Delay dampens the effect of a given dose, and therefore leads to an exaggerated measurement of the required dosage. The amount of delay in any given observation is unknown, but it is known to be exponentially distributed (e.g., because it represents the arrival time of emergency care).

Note that the expectation of η is 1. Define $\varepsilon = \eta - 1$ and $\beta_0' = \beta_0 + 1$, such that the above specification can be rewritten as $y_x = \beta_0' + \beta_1 x_1 + \varepsilon_x$, in order to be consistent with our model. The incentive-compatibility inequality for an agent with $x = 1$ reduces to

$$\int_{\varepsilon_0} \int_{\varepsilon_1 | (\beta_1 + \varepsilon_1 - \varepsilon_0)^2 \geq 2c_0} e^{-(\varepsilon_0 + 1)} e^{-(\varepsilon_1 + 1)} [-(\beta_1)^2 + 2\beta_1 \varepsilon_0 + (\varepsilon_1)^2 - (\varepsilon_0)^2] d\varepsilon_0 d\varepsilon_1 \leq 0$$

This double integral can be computed analytically, but the solution does not seem to be elegant. It can be evaluated numerically for various values of β_1, c_0 and shown that the inequality can be violated - for instance, when $c_0 = 2$ and $\beta_1 = 0.25, 0.5, 0.75, 1$.

The intuition is similar to that of Example 1. When the noise distribution has a long tail on one side and a short tail on the other, a high complexity cost c_0 implies that the pivotal event in which the explanatory variable is included in the regression consists of far-out tail realizations of ε_1. As a result, the estimate of β_1 is heavily biased, such that if the true value of β_1 is not too big, the agent is better off misreporting. $\square$

3.2.1 Does the Curse Vanish as $N \to \infty$?

So far, our analysis was conducted for a given sample size N. A natural question is whether the incentive-compatibility problem we identified disappears as N grows large. To explore this question, return to Example 1, where we saw that when $N = 1$, there exists a set of parameters (β_1, c_0) for which incentive compatibility fails. We now ask whether this set vanishes as $N \to \infty$. We continue to assume $c_1 = c_2 = 0$ and restrict attention to the case of $\beta_1 > 0$ - both entail no loss of generality.

Recall that for every $x = 0, 1$ and every observation $n = 1, ..., N$, ε_x^n is drawn from the Bernoulli distribution that assigns probability p to -1 and probability $1-p$ to $d = p/(1-p)$. Let $\bar{\varepsilon}_x^N$ denote the average noise realization over all the N observations for x.

Recall that the pivotal event $\{\varepsilon \mid b_1(\varepsilon) \neq 0\}$ can be rewritten as

$$\left\{\varepsilon \mid \bar{\varepsilon}_1^N - \bar{\varepsilon}_0^N \notin \left(\sqrt{2c_0} - \beta_1, \sqrt{2c_0} + \beta_1\right)\right\} \tag{7}$$

Our goal is find the set of parameters (β_1, c_0) for which incentive compatibility is violated in the $N \to \infty$ limit.

We begin by finding the limit distribution over $(\bar{\varepsilon}_0^N, \bar{\varepsilon}_1^N)$, conditional on the event (7). Since $\lim_{n\to\infty} \bar{\varepsilon}_1^N = \lim_{n\to\infty} \bar{\varepsilon}_0^N = 0$, the pivotal event occurs with zero probability in the $N \to \infty$ limit. Therefore, we need tools from Large Deviation Theory (Ch. 11 in Cover and Thomas (2006)) in order to characterize the conditional limit distribution. To make use of these tools, some preliminary notation is in order. First, combine the two samples $(\varepsilon_0^1, ..., \varepsilon_0^N)$ and $(\varepsilon_1^1, ..., \varepsilon_1^N)$ into one composite sample $(\eta^1, ..., \eta^N)$, such that for every n, $\eta^n = (\varepsilon_1^n, \varepsilon_0^n)$. Thus, η^n is drawn $i.i.d$ according to the following distribution π:

$$\begin{aligned}
\pi_{-1,-1} &= \Pr(-1,-1) = p^2 \\
\pi_{-1,d} &= \Pr(-1,d) = p(1-p) = \Pr(d,-1) = \pi_{d,-1} \\
\pi_{d,d} &= \Pr(d,d) = (1-p)^2
\end{aligned}$$

That is, the two components of the composite sample are statistically independent. Second, denote by $s_{i,j}$ the empirical frequency of the realization (i,j) in this composite sample. For instance, $s_{-1,d} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{1}(\eta^n = (-1,d))$. Then,

$$\begin{aligned}
\bar{\varepsilon}_1^N &= (s_{d,-1} + s_{d,d}) \cdot d + (s_{-1,d} + s_{-1,-1}) \cdot (-1) \\
\bar{\varepsilon}_0^N &= (s_{-1,d} + s_{d,d}) \cdot d + (s_{d,-1} + s_{-1,-1}) \cdot (-1)
\end{aligned}$$

The pivotal event can thus be redefined in terms of a subset of empirical frequencies $s = (s_{-1,-1}, s_{-1,d}, s_{d,-1}, s_{d,d})$:

$$R^N = \left\{ s^N \mid (s_{d,-1} - s_{-1,d}) \notin \left(\frac{\sqrt{2c_0} - \beta_1}{d+1}, \frac{\sqrt{2c_0} + \beta_1}{d+1} \right) \right\}$$

For any empirical distribution s, let $D(s||\pi)$ the relative entropy of s with respect to π:

$$D(s||\pi) = \sum_{i,j \in \{-1,d\}} s_{i,j} \ln \left(\frac{s_{i,j}}{\pi_{i,j}} \right) \tag{8}$$

Lemma 2 *In the $N \to \infty$ limit, the distribution over s^N conditional on $s^N \in R^N$ assigns probability one to the unique s that minimizes $D(s||\pi)$ subject to the constraint*

$$s_{d,-1} - s_{-1,d} = \frac{\sqrt{2c_0} - \beta_1}{d+1}$$

The proof relies on basic tools from Large Deviation Theory. By plugging the values of $\bar{\varepsilon}_1$ and $\bar{\varepsilon}_0$ that solve the constrained minimization problem given by Lemma 2 into the inequality that represents a violation of incentive compatibility (inequality (6)), we obtain the following characterization.

Proposition 1 *The set of parameters $\beta_1 > 0$ and c_0, d for which incentive compatibility is violated in the $N \to \infty$ limit is given by*

$$\beta_1 < \frac{c_0}{\sqrt{2c_0} + \frac{2d}{d-1}} \tag{9}$$

Thus, the incentive compatibility problem of Example 1 does not vanish when the sample is large. (On the other hand, a large sample does not make the problem worse: It can also be shown that if incentive compatibility holds for $N = 1$, it must also hold in the $N \to \infty$ limit.) Moreover, because $d > 1$, the R.H.S of (9) increases with d and c_0. That is, the more skewed the underlying noise distribution and the larger the complexity cost, the larger the set of prior beliefs for which incentive compatibility is violated in the $N \to \infty$ limit. When $d \to 1$ - i.e., when the noise distribution approaches symmetry - the R.H.S of (9) converges to zero, such that incentive compatibility is violated in a large sample only for arbitrarily small β_1. That is, the incentive compatibility problem disappears when the noise becomes symmetric. The next sub-section explores this theme.

The reason that large samples do not fix the incentive compatibility problem is that the agent's reasoning hinges on the pivotal event in which the variable is included. Therefore, even if the estimator is asymptotically well-behaved in the traditional statistician's sense, the relevant question for incentive compatibility is whether it is well-behaved conditional on the pivotal event. This event becomes very unlikely in a large sample for a large range of values of β_1 and c_0. Therefore, the relevant toolkit is Large Deviation

Theory rather than standard asymptotic analysis. And as it turns out, when the noise distribution is skewed, the average sample noises $\bar{\varepsilon}_0$ and $\bar{\varepsilon}_1$ do not vanish conditional on the pivotal event.

3.3 Symmetric Noise

A common feature of Examples 1 and 2 was the asymmetry of the noise distribution. The following result shows that this is not an accident: symmetric noise ensures incentive compatibility of the statistician's procedure. For convenience, we consider the case in which the distribution of ε_x^n is described by a well-defined density function.

Proposition 2 *If ε_x^n is symmetrically distributed around zero, then the estimator is incentive-compatible.*

Proof. Consider the deviation from $x = 1$ to $r = 0$. This deviation matters only if $b_1(\varepsilon) \neq 0$. Conditional on this event, incentive compatibility requires the following inequality to hold for all β_0, β_1:

$$\mathbb{E}_{\varepsilon_0,\varepsilon_1}[(b_1(\varepsilon))^2 + 2b_1(\varepsilon)(b_0(\varepsilon) - \beta_0 - \beta_1) \mid b_1(\varepsilon) \neq 0] \leq 0$$

By plugging in the expression for $b_0(\varepsilon)$ given by (3), this inequality reduces to

$$\mathbb{E}_{\varepsilon_0,\varepsilon_1}[b_1(\varepsilon)(-\beta_1 + \bar{\varepsilon}_0 + \bar{\varepsilon}_1) \mid b_1(\varepsilon) \neq 0] \leq 0$$

for all β_1.

Fix $b_1(\varepsilon)$ at some value $b_1^* \neq 0$. Define $\mathcal{E}(b^*) = \{(\bar{\varepsilon}_0, \bar{\varepsilon}_1) : b_1(\varepsilon) = b^*\}$. Suppose $\mathcal{E}(b^*)$ is non-empty. Then, $(u, v) \in \mathcal{E}(b^*)$ implies that $(-v, -u) \in \mathcal{E}(b^*)$. This follows immediately from the fact that $b_1(\varepsilon)$ is linear in $\bar{\varepsilon}_1 - \bar{\varepsilon}_0$. Because ε_0^n and ε_1^n are *i.i.d* and symmetrically distributed around zero, the sample averages (u, v) and $(-v, -u)$ have the same probability. This implies that for any given $b_1^* \neq 0$,

$$\mathbb{E}_{\varepsilon_0,\varepsilon_1}[b_1(\varepsilon)(\bar{\varepsilon}_0 + \bar{\varepsilon}_1)|b_1(\varepsilon) = b_1^*] = 0$$

Therefore, showing that the deviation from $x = 1$ to $r = 0$ is unprofitable reduces to showing that

$$\beta_1 \mathbb{E}_{\varepsilon_0,\varepsilon_1}[b_1(\varepsilon) \mid b_1(\varepsilon) \neq 0] \geq 0$$

which simplifies further to

$$\beta_1 \mathbb{E}_{\varepsilon_0,\varepsilon_1}(b_1(\varepsilon)) \geq 0$$

Suppose without loss of generality that $\beta_1 > 0$. We will show that $\mathbb{E}_{\varepsilon_0,\varepsilon_1}(b_1(\varepsilon)) \geq 0$. Let G and g denote the *cdf* and density of Δ that are induced by the distribution of ε_x^n. Since ε_x^n is symmetrically distributed around zero, so is Δ. This is easily seen by noticing that by the symmetry of ε_x^n, $\Pr[(\varepsilon_0^n, \varepsilon_1^n) = (u, v)] = \Pr[(\varepsilon_0^n, \varepsilon_1^n) = (-u, -v)]$, which implies that $\Pr(\Delta = u - v) = \Pr(\Delta = v - u)$. We need to show that

$$\int_{-\infty}^{-c_1-\beta_1} (\beta_1 + \Delta + c_1)g(\Delta) + \int_{c_1-\beta_1}^{\infty} (\beta_1 + \Delta - c_1)g(\Delta) \geq 0$$

Denote $t = \beta_1 + c_1$, $s = \beta_1 - c_1$, and observe that $t + s > 0$ and $t - s > 0$. By the symmetry of g, the inequality we need to show becomes

$$= \int_{-\infty}^{-t} (t + \Delta)g(\Delta) + \int_{-s}^{\infty} (s + \Delta)g(\Delta) = tG(-t) + sG(s) + \int_{s}^{t} \Delta g(\Delta) \geq 0$$

Applying integration by parts and using the symmetry of g yields

$$\begin{aligned}
tG(-t) &= -\int_{-\infty}^{-t} \Delta g(\Delta) - \int_{-\infty}^{-t} G(\Delta) = \int_{t}^{\infty} \Delta g(\Delta) - \int_{-\infty}^{-t} G(\Delta) \\
sG(s) &= \int_{-\infty}^{s} \Delta g(\Delta) + \int_{-\infty}^{s} G(\Delta)
\end{aligned}$$

It follows that

$$tG(-t) + sG(s) + \int_{s}^{t} \Delta g(\Delta) = \int_{-\infty}^{\infty} \Delta g(\Delta) + \int_{-\infty}^{s} G(\Delta) - \int_{-\infty}^{-t} G(\Delta)$$

Note that $\int_{-\infty}^{\infty} \Delta g(\Delta) = \mathbb{E}_{\varepsilon_0,\varepsilon_1}(\bar{\varepsilon}_1 - \bar{\varepsilon}_0) = 0$. Hence, the inequality we need to prove reduces to

$$\int_{-\infty}^{s} G(\Delta) - \int_{-\infty}^{-t} G(\Delta) \geq 0$$

which holds because $s > -t$.

An analogous argument shows that deviation from $x = 0$ to $r = 1$ is unprofitable. ■

Thus, under symmetric noise, the statistician's procedure does not generate an incentive compatibility problem. The reason is that symmetric noise imposes a limit on the extent of the variable selection curse.

4 The Multi-Variable Case

In this section we turn to analyzing the estimator's incentive compatibility when $K > 1$. 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 following is an alternative formulation of the inequality that underlies the definition of incentive compatibility. Although it lacks a transparent interpretation, it will be useful in the sequel.

Lemma 3 *The deviation M is unprofitable for given β, x if and only if*

$$\mathbb{E}_{\varepsilon}\left[\left(\sum_{k\in M} b_k(\varepsilon)w_k\right)\left(2\bar{\varepsilon}+\sum_{k=1}^{K}\beta_k w_k-\sum_{k\notin M} b_k(\varepsilon)w_k\right)\right]\geq 0 \tag{10}$$

The next lemma will be important for the analysis in this section.

Lemma 4 *For every distinct $k, j \in \{1, ..., K\}$, $\mathbb{E}(\Delta_k \Delta_j) = 0$.*

Thus, the random variables Δ_k and Δ_j are uncorrelated, for any distinct k, j.

4.1 Benchmark I: Precise Measurement

As in the single-variable model, one basic benchmark is when the true coefficients are measured with full precision. Thus, 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 (10) can be written as

$$\left(\sum_{k\in V\cap M}\beta_k w_k\right)\left(\sum_{k\notin V-M}\beta_k w_k\right)\geq 0 \tag{11}$$

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 single-variable 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.

This example demonstrates that variable selection generates a new incentive problem in the multi-variable case. It is different from the variable selection curse identified in Section 3, because it can exist even in the absence of sampling error. In particular, it does not arise from pivotal thinking. The reason the agent may want to misreport x_1 in the example is that $b_2 = b_3 = 0$ - i.e., precisely the event that is irrelevant for the variable selection curse. Instead, the motive behind the deviation is an *externality* between variables: the bias due to misreporting one component counters the cumulative bias due to omitting the other variables.

4.2 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 3 *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. However, as mentioned in Section 3.1, it does depend on the property that $\bar{\varepsilon}_k^1$ and $\bar{\varepsilon}_k^0$ 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 Incentive Compatibility under Normal Noise

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. We already saw in Section 3 that there is an important distinction between symmetric and asymmetric noise. In this sub-section, we strengthen the specification of the noise distribution and assume that it is *normal* with mean zero and variance σ^2. Therefore,

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

The known property that Δ_k and Δ_j are uncorrelated now implies the following important lemma.

Lemma 5 *For any $k \neq j$, Δ_k and Δ_j are statistically independent.*

The normality assumption - specifically, the property that the noise density is a well-defined, decreasing function of the distance from zero - also enables a useful characterization of the ex-ante expectation of estimated coefficients. Recall that the formula for $b_k(\varepsilon)$ 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 $b_k(\varepsilon)$ as a deterministic function of β_k:

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

Lemma 6 *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., $0 < |e(\beta_k)| < |\beta_k|$ whenever $\beta_k \neq 0$.*

We are now able to refine condition (10) for the unprofitability of a given deviation.

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

$$\left(\sum_{k\in M} e(\beta_k)w_k\right)\left(\sum_{k=1}^{K} \beta_k w_k - \sum_{j\notin M} e(\beta_j)w_j\right) \geq 0 \qquad (12)$$

This condition is a considerable simplification of (10), because it is stated entirely in terms of the expected coefficients of individual variables according to the agent's prior. This simplification is attained thanks to the assumption of normally distributed noise, which makes the distribution over estimated coefficients of individual variables not only functionally but also *statistically* independent.

The following result is a simple consequence of Proposition 4.

Proposition 5 *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 $sign(e(\beta_2) - \beta_2) = -sign(\beta_2)$, we obtain that if and $|\beta_1|$ is sufficiently small relative to $|\beta_2|$, the inequality will be violated. ∎

Unlike the precise-measurement case, noisy measurement means that the estimator fails incentive compatibility even when $K = 2$. This failure occurs despite our restriction to a normal (and therefore symmetric) noise distribution. This restriction ensured incentive compatibility in the $K = 1$ case. However, in the $K = 1$ case, the only possible motive to misreport was the variable selection curse, the extent of which was limited by symmetric noise. In contrast, the $K > 1$ case introduces the externality across variables, which does not rely on pivotal-event arguments and therefore survives the restriction to normal noise distributions.

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

A sparse prior

To see the relation between Proposition 4 and the condition for incentive-compatibility in the single-variable cases, 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 $e(\beta_1)\beta_1 \geq 0$ - as in the single-variable case analyzed in Section 3. And since the normal noise distribution is symmetric, we know from Section 3.3 that this inequality holds. 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 (12) is non-negative in expectation (with respect to the agent's prior) for every deviation M.

Proposition 6 *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 7 *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. Specifically, 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; the agent can delete cookies but he cannot manufacture a "fake cookie". 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 $\frac{K}{2}$).

5 Conclusion

Interactions between humans and machines that follow statistical procedures are becoming ubiquitous, giving rise to interesting questions for economists. The question we tackled in this paper was whether the human decision maker should act cooperatively toward the machine, when the machine employs a non-Bayesian statistical procedure that is considered good at predicting the agent's ideal action. We demonstrated that the variable-selection element of this procedure creates non-trivial incentive issues.

Our exercise exposed a methodological challenge. The standard economic model of interactive decision making is based on the Bayesian, common-prior paradigm. However, the actual behavior of machine decision makers is often hard to reconcile with this paradigm. Therefore, modeling strategic interactions that involve machines requires us to depart from the conventional modeling framework, toward an approach that admits decision makers who act as non-Bayesian statisticians. Such approaches are familiar to us from the bounded rationality literature (e.g., Osborne and Rubinstein (1998), Spiegler (2006), Cherry and Salant (2016)). Further study of human-machine interactions is thus likely to generate new ideas for modeling interactions that involve boundedly rational, human decision makers.

References

[1] Cherry, J. and Y. Salant (2006), Statistical Inference in Games, mimeo.

[2] Cover, T. and J. Thomas (2006), Elements of Information Theory, second edition, Wiley.

[3] Feddersen, T. and W. Pesendorfer (1996), The Swing Voter's Curse, American Economic Review 86, 408-424.

[4] Hastie, T., R. Tibshirani and M. Wainwright (2015), Statistical Learning with Sparsity: the LASSO and Generalizations, CRC press.

[5] Milgrom, P. and R. Weber (1982), A Theory of Auctions and Competitive Bidding, Econometrica, 1089-1122.

[6] Osborne, M. and A. Rubinstein (1998), Games with Procedurally Rational Players, American Economic Review 88, 834-847.

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

[8] Spiegler, R. (2006), The Market for Quacks, Review of Economic Studies 73, 1113-1131.

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

Appendix: Omitted 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

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

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

$$\sum_{x\in X} \sum_{n=1}^{N} (y_x^n - b_0 - \sum_{k\in V(\varepsilon)} b_k x_k^n) = 0 \tag{13}$$

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

$$2 \sum_{x\in X} \sum_{n=1}^{N} x_j^n (y_x^n - b_0 - \sum_{k\in V(\varepsilon)} b_k x_k^n) = [sign(b_j)] N c_1 + 2 N c_2 b_j \tag{14}$$

From (13) we obtain

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

Substituting (13) into (14) 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 $P = N \cdot 2^K$ 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:

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

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

Denote

$$b_l = \frac{\sqrt{2c_0} - \beta_1}{d+1} \qquad b_h = \frac{\sqrt{2c_0} + \beta_1}{d+1}$$

Recall that we are restricting attention to a range of parameters such that

$-1 < b_l < b_h < 1$. We can partition the pivotal event R^N into two closed intervals: $[-1, b_l]$ and $[b_h, 1]$. Because $\beta_1 > 0$, $|b_l| < |b_h|$.

The relative entropy function $D(s||\pi)$ is strictly convex in s and attains a unique unconstrained minimum of zero at $s = \pi$. Furthermore, because $\pi_{-1,d} = \pi_{d,-1}$, $D(s||\pi)$ treats $s_{-1,d}$ and $s_{-d,1}$ symmetrically. Therefore, for any $b \in [-1, 1]$, the minimum of $D(s||\pi)$ subject to $s_{-1,d} - s_{-d,1} = b$ is equal to the minimum of $D(s||\pi)$ subject to $s_{d,-1} - s_{-1,d} = b$, such that the minimum of $D(s||\pi)$ subject to $s_{d,-1} - s_{-1,d} = b$ is strictly increasing with $|b|$. Therefore, the minimum of $D(s||\pi)$ subject to $s_{d,-1} - s_{-1,d} \in [-1, b_l]$ is strictly below the minimum of $D(s||\pi)$ subject to $s_{d,-1} - s_{-1,d} \in [b_h, 1]$. By Sanov's Theorem (see Theorem 11.4.1 in Cover and Thomas (2006, p. 362)), the probability of the event $[-1, b_l]$ is arbitrarily higher than the probability of the event $[b_h, 1]$ as $N \to \infty$. Therefore, we can take the pivotal event to be $[-1, b_l]$. Furthermore, by the conditional limit theorem (Theorem 11.6.2 in Cover and Thomas (2006, p. 371)), in the $N \to \infty$ limit, the probability that $s_{d,-1} - s_{-1,d} = b_l$ conditional on the event $s_{d,-1} - s_{-1,d} \in [-1, b_l]$ is one.

It follows that the objective function is $D(s||\pi)$ and the constraints are

$$
\begin{aligned}
s_{d,-1} - s_{-1,d} &= \frac{\sqrt{2c_0} - \beta_1}{d+1} \\
s_{-1,-1} + s_{-1,d} + s_{d,-1} + s_{d,d} &= 1
\end{aligned}
$$

Writing down the Lagrangian, the first-order conditions with respect to $(s_{i,j})$ are (λ_1 and λ_2 are the multipliers of the first and second constraints):

$$
\begin{aligned}
1 + \ln s_{-1,-1} - \ln p^2 - \lambda_2 &= 0 \\
1 + \ln s_{d,d} - \ln(1-p)^2 - \lambda_2 &= 0 \\
1 + \ln s_{d,-1} - \ln p(1-p) - \lambda_1 - \lambda_2 &= 0 \\
1 + \ln s_{-1,d} - \ln p(1-p) + \lambda_1 - \lambda_2 &= 0
\end{aligned}
$$

These equations imply

$$\begin{aligned}
s_{d,-1}s_{-1,d} &= s_{d,d}s_{-1,-1} \\
\frac{s_{-1,-1}}{s_{d,d}} &= d^2
\end{aligned}$$

Recall that

$$\begin{aligned}
d &= \frac{p}{1-p} \\
\bar{\varepsilon}_1 &= (s_{d,-1}+s_{d,d})(d+1)-1 \\
\bar{\varepsilon}_0 &= (s_{-1,d}+s_{d,d})(d+1)-1
\end{aligned}$$

This implies that in the $N \to \infty$ limit, the distribution over ε conditional on the pivotal event assigns probability one to

$$\begin{aligned}
\bar{\varepsilon}_0 &= -\frac{1}{2}(\sqrt{2c_0}-\beta_1)-\frac{d}{d-1}+\frac{1}{2}\sqrt{(\sqrt{2c_0}-\beta_1)^2+\frac{4d^2}{(d-1)^2}} \\
\bar{\varepsilon}_1 &= \frac{1}{2}(\sqrt{2c_0}-\beta_1)-\frac{d}{d-1}+\frac{1}{2}\sqrt{(\sqrt{2c_0}-\beta_1)^2+\frac{4d^2}{(d-1)^2}}
\end{aligned}$$

which immediately gives the result for $s_{d,-1}-s_{-1,d}$. ■

Proof of Lemma 3

Denote $z_k = r_k - x_k$. Inequality (4) can be rewritten as:

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

This inequality can be simplified into

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

Then, (4) can be rewritten as

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

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

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

Hence, we can rewrite the above inequality as follows:

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

Since $w_k = 1 - 2x_k$ and $b_k(\varepsilon) = 0$ for each $k \notin V$, the above inequality is equivalent to (10). ■

Proof of Lemma 4

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 $\mathbb{E}(A+B)(A-B) = \mathbb{E}(A^2) - \mathbb{E}(B^2) = 0$. ■

Proof of Proposition 3

Plug $b_k(\varepsilon) = \beta_k + \Delta_k$ into Condition (10):

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

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

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

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

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

such that

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

which is equal to zero because $\bar{\varepsilon}_k^1$ and $\bar{\varepsilon}_k^0$ 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 Lemma 6

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. For notational ease, remove the subscript from β_k. Then,

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

It is immediately evident that the value of c_2 is irrelevant for this result. Therefore, set $c_2 = 0$ for notational simplicity. We can rewrite $e(\beta)$ as follows:

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

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

(ii) 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

$$e(\beta) = \beta + \int_{-c^*-\beta}^{c^*-\beta} G(\Delta) - (c^* - c_1)[G(c^* - \beta) + G(-c^* - \beta)] - c_1 \qquad (15)$$

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*) The proof relies on two properties of G: (1) $G(\Delta) + G(-\Delta) = 1$ for every Δ; (2) G is strictly convex over $\Delta < 0$ and strictly concave over $\Delta > 0$. Denote $d(\beta) = e(\beta) - \beta$. Substituting (15) for $e(\beta)$ yields

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

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

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

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, $e(\beta) < \beta$. Anti-symmetry of e then ensures that $e(\beta) - \beta > -\beta$. ■

Proof of Proposition 4

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

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

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 (10) can be simplified into

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

Our next step is to characterize $\Pr(V)$, namely the probability that the set of variables V is selected. Recall that whether or not $b_k(\varepsilon) \neq 0$, and the distribution of $b_k(\varepsilon)$, 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,

$$\Pr(V) = \prod\nolimits_{k \in V} \lambda_k \prod\nolimits_{j \notin V} (1 - \lambda_j) \tag{16}$$

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

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

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

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

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

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

■

Proof of Proposition 6

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 (12) 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 7

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

$$
m\mathbb{E}(e(\beta)\beta) + e^*\left[\beta^* \sum_{k,j\in M, k\neq j} w_k w_j + (\beta^* - e^*)\sum_{k\in M} w_k \sum_{j\notin M} w_j\right] < 0 \tag{17}
$$

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 (17) 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 (17) 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 (17) 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^*. ■