The Model Selection Curse*

Kfir Eliaz† and Ran Spiegler‡

November 17, 2018

Abstract

A "statistician" takes an action on behalf of an agent, based on the agent's self-reported personal data and a sample involving other people. The action that he takes is an estimated function of the agent's report. The estimation procedure involves model selection. We ask the following question: Is truth-telling optimal for the agent given the statistician's procedure? We analyze this question in the context of a simple example that highlights the role of model selection. We suggest that our simple exercise may have implications for the broader issue of human interaction with "machine learning" algorithms.

*This paper is extracted from a significantly longer working paper titled "Incentive-Compatible Estimators" (Eliaz and Spiegler (2018)). We thank Susan Athey, Yoav Binyamini, Assaf Cohen, Rami Atar, Lorens Imhof, Annie Liang, Benny Moldovanu, Ron Peretz and especially Martin Cripps for helpful conversations. We are also grateful to seminar and conference audiences at Aarhus, Bocconi, DICE, UCL, Brown, Yale, BRIQ, Warwick and ESSET, the editor and the referees of this journal for their useful comments.

†School of Economics, Tel-Aviv University and Economics Dept., Columbia 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 automated 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 online 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 the online content provider has a distinct political or commercial agenda - the user might be better off if he misreports his characteristics or deletes "cookies" from his computer. 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 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 actual systems perform this prediction task? Consider a very basic textbook tool like LASSO[1] (Tibshirani (1996)). This is a variant on standard linear regression analysis, which adds a cost function that penalizes non-zero coefficients. The proce-

[1] Least Absolute Shrinkage and Selection Operator

dure involves both model selection (i.e. choosing which of many available 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. Such a procedure is considered useful when users have many potentially relevant characteristics relative to the sample size, and especially when we can expect few of them to be relevant for predicting the agent's ideal action (i.e., the true data-generating process is *sparse*).

However, LASSO is not a fundamentally Bayesian procedure. Although one can justify its 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 user's welfare. Furthermore, there is no reason to assume that the prior that rationalizes LASSO in this manner coincides with the user's actual prior beliefs (the priors in the above-cited papers involve Laplacian distributions over parameters). Thus, neither the preferences nor the priors that take part in the Bayesian foundation for LASSO 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 that are far more elaborate than the simple textbook example of penalized regression. If we want to model human interaction with such algorithms, some departure from the standard Bayesian framework with common priors seems warranted. 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 perfectly mimicked by a familiar machine-learning procedure. Our approach in this paper is to take the statistician's procedure as *given* (rather than trying to provide a formal rationalization for it) and examine the user's strategic response to it.

Machine-learning algorithms can be extremely complicated. Neverthe-

less, in this paper we follow the tradition of using simple "toy" models to get insight into complex phenomena. Economists have developed models in this tradition to study the behavior of large organizations or the macroeconomy; surely these are more complex than the most intricate machine-learning algorithm. Accordingly, our model is perhaps the simplest that can capture the key element we wish to address - namely, how the element of model selection in machine-learning algorithms affect users' self-reporting decision.

Specifically, we present a model of an interaction between an "agent" and a "statistician" - the latter is a stand-in for an automated system that obtains personal data from the agent and outputs an action on his behalf. The agent has a single binary personal characteristic x, which is his private information. The agent has an ideal action, which is a function of x. This function is unknown. The statistician learns about it by obtaining noisy observations of *other* agents' ideal actions. This sample constitutes the statistician's private information. It is *small*, consisting of *one* observation for each value of x. The statistician follows a "penalized regression" procedure: the estimated coefficients of his model minimize a combination of the Residual Sum of Squares and a cost function that combines two common forms of penalties: A fixed penalty for the mere inclusion of the explanatory variable x (L_0 penalty) and a penalty for the absolute value of the variable's coefficient (L_1 penalty or LASSO). The procedure's element of model selection in this simple example consists of the decision whether to admit x as a predictor of the agent's ideal action.

With one binary characteristic and two sample points, this environment is as far from "big data" as one could imagine. Nevertheless, it shares a crucial feature with a typical big-data predicament that motivates machine-learning methods: the sample size is roughly the same as the number of potential explanatory variables, such that an estimation procedure that does not involve selection or shrinkage risks over-fitting (e.g., see Hastie et al. (2015)). Indeed, an unpenalized regression would *perfectly* fit the data. As a result,

the estimator would have high variance and its predictive performance could be poor, relative to an estimator that excludes x or shrinks its coefficient. Thus, the merit of our simple example is that it manages to capture in a tractable manner the over-fitting problem.

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, we say that the statistician's procedure (or "estimator") is *incentive-compatible*. Our analysis identifies an aspect of the problem that creates a misreporting incentive. Because the agent's report of x only matters when this variable is selected by the statistician's procedure, he should only care about the distribution of the variable's estimated coefficient conditional on the "*pivotal event*" in which the variable's coefficient is not zero. 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.

We refer to this effect as the "*model selection curse*". As the term suggests, the logic is reminiscent of pivotal-reasoning 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 (Feddersen and Pesendorfer (1996)). The model selection curse does not disappear with large samples: When 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.

Related literature

Our paper joins a small literature that has begun exploring incentive issues that emerge in the context of classical-statistics procedures. Cummings et al. (2015) study agents with privacy concerns who strategically report their

personal data to an analyst who performs a linear regression. Caragiannis et al. (2016) consider the problem of estimating a sample mean when the agents who provide the sample observations want to bias the mean close to their value. Hardt et al. (2016) consider the problem of designing the most accurate classifier when the input to the classifier is provided by a strategic agent who faces a cost of lying. Chassang et al. (2012) argue for a modification of randomized controlled trials when experimental subjects take unobserved actions that can affect treatment outcomes. Banerjee et al. (2017) rationalize norms regarding experimental protocols (especially randomization) by modeling experimenters as ambiguity-averse decision makers. Spiess (2018) studies the design of estimation procedures that involve model selection when the statistician and the social planner have conflicting interests (e.g., when the statistician has a preference for reporting large effects).

2 A Model

An *agent* has a privately known, binary personal characteristic $x \in \{0, 1\}$. In the context of medical decision making, x can represent a risk factor (e.g. smoking). In the context of online content provision, it can indicate whether the agent visited a particular website. A *statistician* must take an action $a \in \mathbb{R}$ on the agent's behalf. The agent's payoff from action a is $-(a - f(x))^2$, where $f(x) \in \mathbb{R}$ is the agent's ideal action as a function of x. It will be convenient to write $f(0) = \beta_0$ and $f(1) = \beta_0 + \beta_1$, such that β_1 captures the effect of x on the agent's ideal action. The parameter profile $\beta = (\beta_0, \beta_1)$ is unknown.

Before taking an action, the statistician privately observes a noisy signal about f. Specifically, for each $x = 0, 1$, he obtains a *single* observation $y_x = f(x) + \varepsilon_x$, where ε_0 and ε_1 are drawn $i.i.d$ from some distribution with zero mean. Denote $\varepsilon = (\varepsilon_0, \varepsilon_1)$. 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 has private access to the sample (y_0, y_1).

Equipped with the sample (y_0, y_1), the statistician follows a "penalized regression" procedure for estimating β. That is, he solves the following minimization problem,

$$\min_{b_0,b_1} \qquad \sum_{x=0,1} (y_x - b_0 - b_1 x)^2 + C(b_1) \tag{1}$$

The first term is the standard Residual Sum of Squares, whereas the second term is a cost associated with b_1; the intercept b_0 entails no cost. (Of course, given our simple set-up, referring to the procedure as a "penalized regression" is a bit of an exaggeration.) The solution to (1) is denoted $b(\varepsilon, \beta) = (b_0(\varepsilon, \beta), b_1(\varepsilon, \beta))$. We refer to $(b(\varepsilon, \beta))_\varepsilon$ as the *estimator*. The dependence on (ε, β) follows from the fact that the estimator is a function of (y_0, y_1), which in turn is determined by (ε, β).

We assume the penalty function

$$C(b_1) = c_0 \mathbf{1}_{b_1 \neq 0} + c_1 |b_1|$$

where $c_0, c_1 \geq 0$. This is a linear combination of the two common penalties mentioned in the introduction, L_0 and L_1.[2] Assume that when the statistician is indifferent between including and excluding x, he includes it.

In the absence of the penalty C, the solution to (1) is $b_0 = y_0$, $b_1 = y_1 - y_0$, such that the Residual Sum of Squares is zero. In other words, the estimator *perfectly* fits the data. As a result, the estimator's predictive performance will tend to be poor - relative to an estimator that sets $b_0 = \frac{1}{2}(y_0 + y_1)$, $b_1 = 0$ - when the true value of β_1 is relatively small.

Having estimated f, the statistician receives a report $r \in X$ from the agent. The statistician then takes the action $a = b_0 + b_1 r$. The agent's

[2] Adding an L_2 (Ridge) term $c_2(b_1)^2$ would not change any of the results in the paper.

expected payoff for a given β is therefore

$$-\mathbb{E}_{\varepsilon}\left[(b_0(\varepsilon,\beta)+b_1(\varepsilon,\beta)r-\beta_0-\beta_1 x)\right]^2 \tag{2}$$

This expression can also be written as

$$-\mathbb{E}_{\varepsilon}[\hat{f}(r)-f(x)]^2$$

where $\hat{f}(r)=b_0(\varepsilon,\beta)+b_1(\varepsilon,\beta)r$ is the estimated model's value at the agent's self-report r.

Note that the agent's preferences are given by a quadratic loss function. This is also a standard criterion for evaluating estimators' predictive success. Suppose that $r=x$ - i.e., the agent submits a truthful report of his personal characteristic. Then, the agent's expected payoff coincides with the estimator's mean squared error.

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 f (i.e. the parameters β) if the agent is weakly better off truthfully reporting his personal characteristic, given his prior. That is,*

$$\mathbb{E}_{\beta}\mathbb{E}_{\varepsilon}\left[\hat{f}(x)-f(x)\right]^2 \leq \mathbb{E}_{\beta}\mathbb{E}_{\varepsilon}\left[\hat{f}(r)-f(x)\right]^2$$

for every $x, r \in \{0,1\}$.

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.

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

$$\mathbb{E}_{\varepsilon}\left[\hat{f}(x) - f(x)\right]^2 \leq \mathbb{E}_{\varepsilon}\left[\hat{f}(r) - f(x)\right]^2 \tag{3}$$

for every true model f and every $x, r \in \{0, 1\}$.

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 a purely *normative* statement about the agent's welfare - namely, given how the statistician takes actions on the agent's behalf, it would be advisable for the agent to misreport. 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 online content provision, deviating from $x = 1$ to $r = 0$ can be interpreted as "deleting a cookie". This deviation is straightforward to implement, and the agent can check if it leads to better content match in the long run.

The agent's expected payoff 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. Incentive-compatibility can thus be viewed as a collection of bias-variance comparisons 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 real-life statistician is likely to propose, but it arises naturally in our setting.

3 Analysis

We first derive a complete characterization of the estimator.

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

$$b_1(\varepsilon,\beta) = \begin{cases} \beta_1+\varepsilon_1-\varepsilon_0-c_1 & if \quad \beta_1+\varepsilon_1-\varepsilon_0-\sqrt{(c_1)^2+2c_0} \geq 0 \\ \beta_1+\varepsilon_1-\varepsilon_0+c_1 & if \quad \beta_1+\varepsilon_1-\varepsilon_0+\sqrt{(c_1)^2+2c_0} \leq 0 \\ 0 & otherwise \end{cases} \tag{4}$$

and

$$b_0(\varepsilon,\beta) = \frac{1}{2}\left[y_0+y_1-b_1(\varepsilon,\beta)\right]$$

The proof is mechanical and relegated to the supplementary appendix. Note that L_0 penalty leads to model selection without affecting the value of b_1 conditional on being non-zero. The L_1 penalty term leads to both shrinkage and selection.

Let us now turn to incentive-compatibility. Two factors create a problem in this regard: sample noise and model selection. Neither factor is problematic on its own, as the following pair of observations establishes.

Claim 1 *Suppose that $\varepsilon = (0,0)$ with probability one. Then, the estimator is incentive compatible.*

Proof. Suppose that β_1 is such that $b_1 = 0$. Then, the agent's report has no effect on the statistician's action, and incentive-compatibility holds trivially. Now suppose β_1 is such that $b_1 > 0$. Given the characterization of b_1, we must have $\beta_1 - c_1 \geq 0$. 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

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

When $x = 0$ ($x = 1$), the agent's ideal action is β_0 ($\beta_0 + \beta_1$), and since $\beta_1 - c_1 \geq 0$, the action b_0 ($b_0 + b_1$) is closer to the ideal point than $b_0 + b_1$ (b_0). Thus, honesty is optimal for the agent. A similar calculation establishes incentive-compatibility when $b_1 < 0$. ■

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

Proof. When $c_0 = c_1 = 0$, we have $b_1 = (\beta_1 + \varepsilon_1 - \varepsilon_0)$. Suppose $x = 1$ and the agent contemplates whether to report $r = 0$. In this case inequality (3) can be simplified into

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

Plugging in the expressions for $b_0(\varepsilon, \beta)$ and $b_1(\varepsilon, \beta)$ given by (4), this inequality reduces to

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

This inequality holds for all β_1 because ε_0 and ε_1 are $i.i.d$ with mean zero. An analogous argument shows that an agent with $x = 0$ will not benefit from reporting $r = 1$. ■

Thus, sampling noise and model selection are both necessary to produce violations of incentive-compatibility in our simple set-up. This finding should not be taken for granted. First, even in the absence of sampling noise, the penalty C creates a wedge between the statistician's objective function and the agent's utility. Therefore, it is not obvious a priori that this de-facto conflict of interest does not give the agent an incentive to misreport. Second, as long as the agent's prior over β_1 is not diffuse, the zero-penalty estimator does not produce actions that maximize his subjective expected utility. This, too, creates a de-facto conflict of interests between the two parties, which nevertheless does not give the agent a sufficient incentive to misreport. One might think that the *unbiasedness* of the zero-penalty estimator

explains Claim 2. However, this intuition is misleading because the agent's utility function involves a *bias-variance* trade-off. As a result, Claim 2 breaks down when the statistician draws different numbers of observations for $x = 0$ and $x = 1$: the agent may be willing to experience a biased action due to misreporting because it will reduce its variance.

Our next result establishes that incentive compatibility is an issue in the presence of noisy measurement and non-zero penalty. For expositional convenience, we restrict attention to the case of $c_1 = 0$. However, the result can easily be extended to arbitrary $(c_0, c_1) > (0, 0)$.

Proposition 2 *Let $c_0 > c_1 = 0$. Then, there exists a distribution over sample noise for which the estimator is not incentive-compatible.*

Proof. Construct the following sample noise distribution. For each x, let

$$\varepsilon_x = \begin{cases} -1 & \text{with probability } p \\ \\ d = p/(1-p) & \text{with probability } 1-p \end{cases}$$

where $p > \frac{1}{2}$, such that $d > 1$. Consider an agent with $x = 1$ who reports $r = 0$. This misreporting violates incentive-compatibility if there is some β_1 for which

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

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

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

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

A key feature of the above example is the asymmetry in the noise distribution. Our next result shows that this is a crucial feature: Symmetric noise ensures incentive-compatibility of the statistician's procedure. For convenience, we consider the case in which the distribution of ε_x is described by a well-defined density function. The result is stated for arbitrary $c_0, c_1 \geq 0$.

Proposition 3 *If ε_x 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 $b_1(\varepsilon, \beta) \neq 0$. Incentive-compatibility thus requires the following inequality to hold for all β_0, β_1:

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

Plugging the expression for $b_0(\varepsilon)$ given by (4), this inequality reduces to

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

Fix $b_1(\varepsilon,\beta)$ at some value $b_1^* \neq 0$. Define $\mathcal{E}(b^*) = \{(\varepsilon_0,\varepsilon_1) : b_1(\varepsilon,\beta) = b_1^*\}$. Suppose $\mathcal{E}(b^*)$ is non-empty. Then, $(u,v) \in \mathcal{E}(b_1^*)$ implies that $(-v,-u) \in \mathcal{E}(b^*)$. This follows immediately from the fact that $b_1(\varepsilon,\beta)$ is defined by the difference $\varepsilon_1 - \varepsilon_0$. Because ε_0 and ε_1 are $i.i.d$ and symmetrically distributed around zero, the realizations (u,v) and $(-v,-u)$ have the same probability. This implies that for any given $b_1^* \neq 0$,

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

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

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

which simplifies further to

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

Suppose without loss of generality that $\beta_1 > 0$. We will show that $\mathbb{E}_{\varepsilon_0,\varepsilon_1}(b_1(\varepsilon,\beta)) \geq 0$. Denote $\Delta = \varepsilon_1 - \varepsilon_0$. Let G and g denote the *cdf* and density of Δ. Since ε_0 and ε_1 are symmetrically distributed around zero, g is symmetric. Denote

$$c^* = \sqrt{(c_1)^2 + 2c_0}$$

We need to show that

$$\int_{-\infty}^{-c^*-\beta_1} (\beta_1+\Delta+c_1)g(\Delta) + \int_{c^*-\beta_1}^{\infty} (\beta_1+\Delta-c_1)g(\Delta) \geq 0 \qquad (7)$$

Denote $t = \beta_1 + c^*$, $s = \beta_1 - c_1$, and observe that because $c^* \geq c_1 \geq 0$, $t + s > 0$ and $t - s > 0$. By the symmetry of g, (7) becomes

$$= \int_{-\infty}^{-t} (t+\Delta)g(\Delta) + \int_{-s}^{\infty} (s+\Delta)g(\Delta) = tG(-t) + sG(s) + \int_{s}^{t} \Delta g(\Delta) \geq 0 \quad (8)$$

Applying integration by parts and the symmetry of g, (8) becomes

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

Since $\int_{-\infty}^{\infty} \Delta g(\Delta) = \mathbb{E}_{\varepsilon_0,\varepsilon_1}(\varepsilon_1 - \varepsilon_0) = 0$, 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. ■

The intuition behind this result is that symmetric noise curbs the model selection curse: although model selection implies that b_1 is a biased estimate of β_1, the bias is too small to give the agent the incentive to introduce the counter-bias that results from misreporting.

4 Does the Curse Vanish with Large Samples?

So far, we focused on a sample with two observations, hence, one may think that the model selection curse is a small-sample phenomenon. In this section we show that this need not be the case. Extend our model by assuming that for each $x = 0, 1$, the statistician obtains N observations of the form $y_x^n = f(x) + \varepsilon_x^n$, $n = 1, ..., N$, where ε_x^n is $i.i.d$ with mean zero across all x, n.

The statistician's problem is essentially the same:

$$\min_{b_0,b_1} \sum_{x=0,1} \sum_{n=1}^{N} (y_x^n - b_0 - b_1 x_k^n)^2 + N\left(c_0 \mathbf{1}_{b_1 \neq 0} + c_1 |b_1|\right)$$

The entire model and its analysis are unchanged, except that now $\varepsilon = (\varepsilon_0^n, \varepsilon_1^n)_{n=1,...,N}$; and in the solution for the estimator (4), ε_x is replaced with the average sample noise $\bar{\varepsilon}_x = \frac{1}{N}\sum_{i=1}^{n} \varepsilon_x^n$. Denote $\varepsilon = (\varepsilon_x^n)_{x=0,1;n=1,...,N}$.

Returning to the *Bernoulli-noise example* from the previous section, we investigate whether the set of parameters that violate incentive compatibility vanishes as $N \to \infty$. We continue to assume $c_1 = 0$ and restrict attention to the case of $\beta_1 > 0$ - both are without loss of generality. Note that c_0 is constant per observation, we address this issue at the end of this section.

Suppose that for every $x = 0, 1$ and every observation $n = 1, ..., N$, ε_x^n is independently drawn from the Bernoulli distribution that assigns probability $p > \frac{1}{2}$ 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 $x \in \{0, 1\}$. The pivotal event $\{\varepsilon \mid b_1(\varepsilon, \beta) \neq 0\}$ can be written as

$$\left\{\varepsilon \mid \bar{\varepsilon}_1(N) - \bar{\varepsilon}_0(N) \notin \left(-\sqrt{2c_0} - \beta_1, \sqrt{2c_0} - \beta_1\right)\right\} \tag{9}$$

Our goal is find the set of parameters for which incentive-compatibility is violated in the $N \to \infty$ limit.

Proposition 4 *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*

$$\beta_1 < \frac{c_0}{\sqrt{2c_0} + \frac{2d}{d-1}} \tag{10}$$

Proof. We first find the limit distribution over $(\bar{\varepsilon}_0(N), \bar{\varepsilon}_1(N))$, conditional on the event (9). To do this, it helps to 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:

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

Denoting by $s_{i,j}$ the empirical frequency of the realization (i,j) in this composite sample allows us to redefine the pivotal event 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:

$$D(s||\pi) = \sum_{i,j \in \{-1,d\}} s_{i,j} \ln \left(\frac{s_{i,j}}{\pi_{i,j}} \right) \tag{11}$$

Denote

$$\theta_l = \frac{-\sqrt{2c_0} - \beta_1}{d+1} \qquad \theta_h = \frac{\sqrt{2c_0} - \beta_1}{d+1}$$

We will now show that 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 $s_{d,-1} - s_{-1,d} = \theta_h$. Recall that we are restricting attention to a range of parameters such that $-1 < \theta_l < \theta_h < 1$. We can partition the pivotal event R^N into two closed intervals: $[-1, \theta_l]$ and $[\theta_h, 1]$. Because $\beta_1 > 0$, $|\theta_l| < |\theta_h|$.

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 $\theta \in [-1,1]$, the minimum of $D(s||\pi)$ subject to $s_{-1,d} - s_{-d,1} = \theta$ is equal to the minimum of $D(s||\pi)$ subject to $s_{d,-1} - s_{-1,d} = \theta$, such that the

minimum of $D(s||\pi)$ subject to $s_{d,-1} - s_{-1,d} = \theta$ is strictly increasing with $|\theta|$. Therefore, the minimum of $D(s||\pi)$ subject to $s_{d,-1} - s_{-1,d} \in [\theta_h, 1]$ is strictly below the minimum of $D(s||\pi)$ subject to $s_{d,-1} - s_{-1,d} \in [-1, \theta_l]$. By Sanov's Theorem (see Theorem 11.4.1 in Cover and Thomas (2006, p. 362)), the probability of the event $[\theta_h, 1]$ is arbitrarily higher than the probability of the event $[-1, \theta_l]$ as $N \to \infty$. Therefore, we can take the pivotal event to be $[\theta_h, 1]$. 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 $s_{d,-1} - s_{-1,d} = \theta_h$ conditional on the event $s_{d,-1} - s_{-1,d} \in [\theta_h, 1]$ 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

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

Now, since

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

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

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

Plugging these values into (6)) produces the result. ■

Thus, the incentive-compatibility problem in the Bernoulli-noise example does not vanish when the sample is large. Moreover, 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. 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's unconditional distribution is asymptotically well-behaved, the relevant question for incentive-compatibility is whether it is well-behaved *conditional on the pivotal event*.

Recall that our original assumption of only two observations captured (in a highly stylized fashion) the idea that model selection can avert over-fitting. When we continue to assume a single explanatory variable and raise N, the over-fitting problem is attenuated and the role of model selection diminishes. Indeed, practitioners of penalized regression adjust penalty parameters to

sample size, such that $c_0, c_1 \to 0$ as $N \to \infty$. The key question is therefore whether the *rate* by which c_0 or c_1 decrease with N is *fast enough* to outweigh the model selection curse. To answer this question, one needs to characterize the condition for incentive-compatibility for arbitrary values of N, c_0, c_1. This is an open question that we leave for future work.

Since the probability of the pivotal event decreases with n, the payoff consequence of misreporting vanishes in the $N \to \infty$ limit, such that the agent becomes almost indifferent between reporting and misreporting (as is indeed the case in models of strategic voting in large elections). If we were to extend our analysis to account for the strategic reasoning of *all* the individuals - including those in the statistician's sample - the equilibrium outcome could stray far from the sincere-reporting benchmark. Exploring this problem, too, is left for future research.

5 Conclusion

Interactions between humans and machines that follow statistical procedures are becoming ubiquitous, giving rise to interesting questions for economists. Our question is whether human decision makers should act cooperatively toward a machine that employs a non-Bayesian statistical procedure that aims at good predictions. We demonstrated, via a toy example, that the element of model selection in the procedure creates non-trivial incentive issues.

Our little exercise exposed a major 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. We addressed this challenge by examining the agent's response to a $fixed$ statistical procedure with a given specification of its parameters. One would like to *endogenize* these choices. However, given that the procedure is fundamentally non-Bayesian, capturing this endogenization with a well-defined ex-ante optimization prob-

lem is not obvious. Incorporating incentive-compatibility as a criterion for selecting prediction methods is therefore conceptually challenging.

In general, modeling strategic interactions that involve machine learning requires us to depart from the conventional Bayesian 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) and Liang (2018)). Further study of human-machine interactions is likely to generate new ideas for modeling interactions that involve boundedly rational *human* decision makers.

References

[1] Banerjee, A., S. Chassang, S. Montero and E. Snowberg (2017). "A Theory of Experimenters," NBER Working Paper No. 23867.

[2] Caragiannis, I, Ariel D. Procaccia and N. Shah (2016): "Truthful Univariate Estimators," *Proceedings of the 33rd International Conference on Machine Learning* **48**.

[3] Chassang, S., P. Miquel and E. Snowberg (2012). "Selective trials: A Principal-Agent Approach to Randomized Controlled Experiments," *American Economic Review* **102**, 1279-1309.

[4] Cherry, J. and Y. Salant (2006). "Statistical Inference in Games," Northwestern University Working Paper.

[5] Cover, T. and J. Thomas (2006). *Elements of Information Theory*, Second Edition, Wiley.

[6] Cummings, R., S. Ioannidis and K. Ligett (2015). "Truthful Linear Regression," *Conference on Learning Theory*, 448-483.

[7] Eliaz, K. and R. Spiegler (2018). "Incentive-Compatible Estimators," Tel-Aviv University Working Paper.

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

[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] Hardt, M., N. Megiddo, C. Papadimitriou and J. Wootters (2016): "Strategic Classification," *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, 111-122

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

[12] Liang, A. (2018): "Games of Incomplete Information Played by Statisticians," University of Pennsylvania Working Paper.

[13] Milgrom, P. and R. Weber (1982). "A Theory of Auctions and Competitive Bidding," *Econometrica* **50**, 1089-1122.

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

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

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

[17] Spiess, J. (2018). "Optimal Estimation when Researcher and Social Preferences are Misaligned," Harvard University Working Paper.

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

Appendix: Proof of Proposition 1

Fix the realization of sample noise ε. The coefficients b_0 and b_1 are given by the solution to the first-order conditions of

$$\min_{b_0,b_1} \quad \sum_{x=0,1} (y_x - b_0 - b_1 x)^2 + c_0 \mathbf{1}_{b_1 \neq 0} + c_1 |b_1|$$

where the dependence of the coefficients b_0 and b_1 on the noise realization ε is suppressed for notational ease.

The first-order condition with respect to b_0 is

$$(y_0 - b_0) + (y_1 - b_0 - b_1) = 0 \tag{12}$$

while the first-order condition with respect to b_1 when $b_1 \neq 0$ gives

$$2(y_1 - b_0 - b_1) = sign(b_1) c_1 \tag{13}$$

In particular, $2(y_1 - b_0 - b_1) = c_1$ when $b_1 > 0$, and $2(y_1 - b_0 - b_1) = -c_1$ when $b_1 < 0$.

From (12) we obtain

$$b_0 = \frac{1}{2}(y_0 + y_1 - b_1)$$

Plugging this into (13), we obtain the following characterization of b_1 conditional on it being non-zero:

$$b_1 = \begin{cases} \beta_1 + \varepsilon_1 - \varepsilon_0 - c_1 & if \quad b_1 > 0 \\ \beta_1 + \varepsilon_1 - \varepsilon_0 + c_1 & if \quad b_1 < 0 \end{cases}$$

This means in particular that when $\beta_1 + \varepsilon_1 - \varepsilon_0 \in (-c_1, c_1)$, $b_1 = 0$.

To complete the characterization of when $b_1 \neq 0$, we compute the difference between the Residual Sum of Squares (RSS) when the coefficient b_1 is

admitted and when it is omitted. First,

$$RSS(b_1 \neq 0) = (b_0 - y_0)^2 + (b_0 + b_1 - y_1)^2$$

where b_0 and b_1 are given by (12) and (13). In contrast, when b_1 is omitted, $b_0 = \frac{1}{2}(y_0 + y_1)$, such that

$$RSS(b_1 = 0) = \left(\frac{1}{2}y_0 + \frac{1}{2}y_1 - y_0\right)^2 + \left(\frac{1}{2}y_0 + \frac{1}{2}y_1 - y_1\right)^2 = \frac{1}{2}(y_1 - y_0)^2$$

It follows that

$$\begin{aligned} RSS(b_1 &= 0) - RSS(b_1 \neq 0) = \frac{1}{2}(y_1 - y_0)^2 - (b_0 - y_0)^2 - (b_0 + b_1 - y_1)^2 \\ &= b_1[y_1 - y_0 - \frac{1}{2}b_1] \\ &= [y_1 - y_0 - sign(b_1)c_1][y_1 - y_0 - \frac{1}{2}(y_1 - y_0 - sign(b_1)c_1)] \\ &= \frac{1}{2}(y_1 - y_0)^2 - \frac{1}{2}(c_1)^2 \\ &= \frac{1}{2}(\beta_1 + \varepsilon_1 - \varepsilon_0)^2 - \frac{1}{2}(c_1)^2 \end{aligned}$$

The condition for $b_1 \neq 0$ is

$$RSS(b_1 = 0) - RSS(b_1 \neq 0) \geq c_0$$

i.e.

$$(\beta_1 + \varepsilon_1 - \varepsilon_0)^2 \geq (c_1)^2 + 2c_0$$

This concludes the proof.