Manipulable Data, Goodhart's Law, and Credit Risk Prediction

Andrea Gamba
WBS-University of Warwick

Christopher A. Hennesy
LBS, CEPR, ECGI

Abstract

We analyse default risk parameter estimation when agents can manipulate, at cost, a covariate used in credit scoring. A qualified version of Goodhart's law obtains: When a scoring model uses coefficients from clean historical data, intercept and/or slope coefficients shift down subsequently, if the coefficient on the clean covariate is not zero. Conversely, if the clean covariate coefficient is zero, parameter estimates remain stable over time regardless of the models posted. We then characterize Nash models. If the clean covariate coefficient is not zero, so Goodhart's critique applies, intercept and/or slope coefficients of any Nash model must undershoot clean data counterparts, and the Nash slope coefficient cannot be zero. Finally,we consider Stackelberg leader models, showing that under some technologies it can be optimal to encourage manipulation by posting a model with an upward biased slope.

1 Introduction

Since Altman (1968), historical relationships between accounting ratios and bankruptcy events have been used to estimate corporate default probabilities. Since then, econometric techniques have become increasingly sophisticated, with recent research using machine learning to predict household and corporate defaults, e.g. Tian, Yu and Guo (2015), Tian and Yu (2017), Barboza, Kimura and Altman (2017), Chen and Xiang (2017), Chan-Lau (2017), and Zhou (2021).

Such use of econometric models to estimate default probabilities brings them into direct conflict with *Goodhart's law*, which states that:[1]

[1] See Goodhart (1975) and the related formal work of Lucas (1976).

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

In particular, Fair Isaac Corporation (2018) notes that, "In markets where credit risk scoring models are regulated and scrutinized, there is a strong requirement for the models, and the credit decisions derived from them, to be explainable. The impact each variable has on the credit score must be traceable (transparent), clearly explained and palatable (understandable and acceptable) to lenders, regulators and consumers." Faced with such models, borrowers have an obvious incentive to game data. Indeed, echoing Goodhart's law, Mark Zandi, chief economist at Moody's Analytics, expresses concern that, "The scoring models may not be telling us the same thing that they have historically, because people are so focused on their scores and working hard to get them up, mucking with their relationship to the underlying credit risk." Consistent with this argument, Liu, et al. (2010) and Caton, et al. (2011) document earnings management prior to bond flotations.

Although the problem of data manipulation in credit risk assessment is oft-noted, it is not clear how to model it formally, to say nothing of the more important challenge of how to address the problem econometrically. With this in mind, this paper develops an analytical framework for assessing and addressing Goodhart's law in the context of econometric models of credit risk. How do we incorporate data manipulation into a formal statement of the econometrician's program? How does Goodhart's law relate to the underlying technologies available to borrowers as they contemplate costly data manipulation? What tools are available for addressing Goodhart's law in the context of credit risk assessment, and how well do these tools perform?

The setting considered is as follows. In the spirit of Goodhart (1975), an econometrician has access to clean historical data drawn from a prior cohort of borrowers who did not face a default prediction model. As shown, this clean data allows the econometrician to correctly estimate the intercept and slope parameters (a, b) determining each agent's latent repayment probability, $F(a + bx)$, with $x \geq 0$, $b \geq 0$, and F taking standard functional forms, specifically, linear, logistic (logit) or normal (probit).

Taking into account the historical parameters (a, b), the econometrician posts her preferred default prediction model with intercept and slope parameters $(\widetilde{\alpha}, \widetilde{\beta})$. Each agent in the future cohort submits a manipulated covariate $\widetilde{x}$, with repayment probability computed according to $F(\widetilde{\alpha} + \widetilde{\beta}\widetilde{x})$. In our binary setup, each agent makes a yes/no decision regarding whether to manipulate their covariate upward by $\delta > 0$, with the manipulation cost being an i.i.d. shock. In the alternative setup, borrowers face a quadratic manipulation cost function.

In order to illustrate Goodhart's law, we first consider a naive econometrician who announces to the future strategic borrowers that their respective repayment probabilities will be assessed according to $F(a + b\widetilde{x})$. Ex post, the econometrician tests for model instability by coming up with new parameter estimates $(\widehat{\alpha}, \widehat{\beta})$, applying OLS (linear F) or MLE (logit/probit) to the manipulated data generated by this cohort of strategic borrowers. Consistent with Goodhart's law, model parameters are indeed unstable, with $(\widehat{\alpha}, \widehat{\beta}) \neq (a, b)$ unless $b = 0$.

After illustrating Goodhart's law, we examine alternative econometric responses, confining attention to logit and probit models, with the econometrician choosing the intercept and slope parameters of her posted model to maximize the expected log-likelihood ratio. We consider first Nash equilibria: Nash model parameters (α^*, β^*) are optimal given the borrowers' covariates, and vice-versa. Phrased differently, Nash econometric models are immune to Goodhart's law, with (α^*, β^*) representing stable ML estimates for cohorts of borrowers responding to (α^*, β^*). Historical clean data coefficients (a, b) are Nash if and only if $b = 0$. If $b > 0$, a Nash econometric model must feature a lower intercept and/or slope than (a, b), and the Nash slope $\beta^* \neq 0$. If borrowers have the binary manipulation technology and b is sufficiently large, there can be a Nash equilibrium in which $\beta^* = b$, all borrowers manipulate, and the intercept $\alpha^* = a - b\delta$. In this all-manipulate equilibrium, the expected likelihood ratio reaches that attained if manipulation were impossible. Conveniently, this fixed point would be found after performing a single round of ex post estimation on borrowers who face the historical coefficients (a, b).

Next, we analyze Stackelberg default prediction models. In contrast to Nash models, where the econometrician treats the distribution of covariates as given, a Stackelberg model parameterization $(\alpha^{**}, \beta^{**})$ accounts for the effect of the posted model on borrower incentives. We find Nash and Stackelberg models are (numerically) similar if borrowers face quadratic manipulation costs. However, if borrowers have the binary manipulation technology, the Stackelberg model differs markedly from the Nash model. In particular, if b is sufficiently low, a Stackelberg leader econometrician nudges borrowers toward homogeneity and low probabilities of manipulation by posting a slope $\beta^{**} < b$. By way of contrast, if b is sufficiently high, a Stackelberg leader econometrician finds it optimal to nudge all borrowers toward manipulating their covariate upward by posting slope $\beta^{**} > b$. If all borrowers can be induced to manipulate in this way, a simple downward adjustment of the intercept suffices to restore predictive power to its level in economies with zero manipulation.

Our paper draws much inspiration from recent work of Frankel and Kartik (2023), and to a lesser extent Hennessy and Goodhart (2023).[2] Frankel and Kartik consider a more abstract setting with linear benefits, quadratic manipulation costs, and a parameterized covariance structure. Here they show a Stackelberg principal finds it optimal to make allocations less sensitive to the covariate. In contrast, our Stackelberg principal sometimes finds it optimal to increase this sensitivity under particular manipulation technologies. Another point of contrast is that much of our analysis entails comparative static analysis on the clean data slope b, with their model treating the special case of $b = 1$. However, the primary difference is that we consider a specific setting, credit risk prediction via MLE, in a setting with an endogenous manipulation gain.

Another closely related paper is that of Rajan, Seru and Vig (2010). They demonstrate a complementary variation on Goodhart's law in credit markets: An increase in securitization rates over time will weaken lender incentives to collect soft information, implying that historical estimates of default probabilities will undershoot prospective default probabilities. Of course one point of contrast is that we focus on borrower moral hazard, not lender moral hazard. However, the more important point of contrast methodologically is that we cast all of our analysis in an explicit econometric framework, inspired by Frankel and Kartik (2023) and Hennessy and Goodhart (2023).

[2] Hennessy and Goodhart (2023) consider machine learning in linear settings with manipulation.

Our all-manipulate econometric equilibrium under binary manipulation is in the spirit of the signal jamming equilibrium of Stein (1989). In his model, manager's can engage in clandestine borrowing at a cost to inflate earnings. In equilibrium, the market infers the extent of incentive-compatible earnings inflation, so that stocks are priced correctly.

Eliaz and Spiegler (2019) consider a setting in which the agent is part of the training data. By construction, they consider a setting in which incentives would seem to be aligned in that the objective of the principal is to predict the agent's most preferred outcome. Nevertheless, they identify the following problem that is likely to be acute under Lasso estimation with sparsity: The agent may have an incentive to misreport given that his report only matters in the event that the covariate's coefficient is not zero.

Björkegren, Blumenstock and Knight (2020) consider the special case of absolute quadratic manipulation costs demonstrating their method with Monte Carlo simulations. In addition, they offer a real-world implementation in a field experiment in Kenya. Brückner and Scheffner (2011) and Hardt et al. (2015) also analyze agents who can manipulate covariates. Brückner and Scheffner consider only quadratic manipulation costs while Hardt et al. only consider costs expressible as $\max\{0, g(x_2) - f(x_1)\}$ for some (f, g) which includes linear manipulation costs but excludes quadratic costs and other standard distance measures.

There is another line of research in computer science focusing on strategic manipulation of training data, e.g. Dekel et al. (2010) and Chen et al. (2018). This literature contemplates statistical inference combined with mechanism design, with the core idea being to identify mechanisms that will induce truthful reporting in the training data.

The remainder of the paper is as follows. Section 2 analyzes incentives for data manipulation. Section 3 examines Goodhart's law in the context of linear probability models. Section 4 examines Goodhart's law in the context of logit and probit models. Section 5 examines potential econometric remedies to the problem of manipulable data in credit risk prediction. Section 6 considers a multivariate setup. We conclude with suggestions for future lines of research.

2 Data Manipulation Incentives

We begin this section be describing the institutional setting in which the econometrician and borrowers interact. We then show that data manipulation incentives have a number of intuitive properties.

2.1 Institutional Setting

We consider a single lender relying on model-based loan pricing, as described below. Loan amounts are normalized at 1. The outcome y is a binary random variable, with $y = 1$ denoting debt repayment and $y = 0$ denoting default. In the event of default, the lender recovers l, where $0 \leq l < 1$. The risk-free rate is $r \geq 0$.

The true covariate $x \geq 0$ is a random variable with atomless cumulative distribution function H and probability density function h. The conditional expectation of y given x, or equivalently the probability of debt repayment conditional upon x, is of known functional form:

$$\mathbb{E}[y|x] = \Pr[y = 1|x] = F(a + bx), \tag{1}$$

where $b \geq 0$, and F is a known cumulative distribution function. That is, the conditional expectation function falls into a known parametric family. The argument, call it z, to which the c.d.f. F is applied is denoted the *credit score,* e.g. a borrower's true credit score is $z = a + bx$.

To illustrate, suppose x denotes the log of the ratio of firm asset value (V) to debt face value maturing at time T. In the log-normal economy considered by Merton (1974), the debt repayment probability takes the form:

$$\Pr[y = 1|x] = \mathcal{N}\left[\underbrace{\frac{(\mu_V - \frac{1}{2}\sigma_V^2)T}{\sigma_V\sqrt{T}}}_{=a} + \underbrace{\left(\frac{1}{\sigma_V\sqrt{T}}\right)}_{=b}x\right]. \tag{2}$$

In the spirit of Goodhart (1975), we assume the econometrician has access to clean training data consisting of (y, x) pairs collected from some historical cohort–a cohort that had no incentive to manipulate since their data was not being used in setting interest rates.[3] The question pondered

[3] This is to fix ideas and setting. Alternatively, one can simply assume the econometrician knows (a, b).

by Goodhart is the extent to which statistical regularities gleaned from historical data will tend to break down if those regularities are used in allocating resources, with our specific interest being the pricing of credit.

The mandate of the lender is to make loans to future borrowers subject to an institutional constraint that the model-implied expected return is equal to the risk-free rate. More specifically, letting hats denoted predicted values, the interest rate ι on any future loan must satisfy:

$$\begin{aligned} 1+r &= \widehat{\Pr}[y=1](1+\iota)+\left[1-\widehat{\Pr}[y=1]\right] l \\ &\Rightarrow \quad \iota = l+\frac{1+r-l}{\widehat{\Pr}[y=1]}-1. \end{aligned} \tag{3}$$

The underlying friction in our parable economy is that the lender cannot observe the true covariate values for future borrowers. Rather, the lender observes a reported covariate $\widetilde{x}$ that is potentially manipulated upward from x. For ease of exposition, $\widetilde{x}$ is labeled the *manipulated covariate*, although $\widetilde{x}=x$ is a potential outcome in some cases, as described below.

Data manipulation is costly. We consider that manipulation costs take one of two forms. In the *binary manipulation setup*, borrowers choose whether to manipulate or not, with manipulation being of fixed size $\delta>0$. Here manipulation comes at an idiosyncratic cost c, with c being a privately observed random variable having an atomless differentiable cumulative distribution G with $g \equiv G'$ being strictly positive on domain $[0, c_{\max}]$. In the *quadratic manipulation cost setup*, the borrower's cost of manipulation is $cm^2/2$ where $m \equiv \widetilde{x}-x$, with $c>0$ being a privately observed i.i.d. random variable with cumulative distribution Γ and density function γ. Noteworthy is the fact that noise in c, in addition to latent x, will give rise to two dimensions of private information, complicating inference and default prediction.

The econometrician's *default prediction model* (DPM) states how $\widetilde{x}$ is mapped to an assessed repayment probability:

$$\text{DPM: } \widehat{\Pr}[y=1|\widetilde{x}] \equiv F(\widetilde{\alpha}+\widetilde{\beta}\widetilde{x}). \tag{4}$$

Within this setting, we consider two classes of strategic agents: borrowers and an econometrician. The task of the econometrician is to specify the coefficients $(\widetilde{\alpha}, \widetilde{\beta})$ of the posted econometric model, perhaps relying on parameter estimates derived from the clean historical data (y, x). Borrowers respond to the posted DPM by reporting a covariate $\widetilde{x}$.[4] Since $b \geq 0$, we confine attention to posted econometric models featuring $\widetilde{\beta} \geq 0$ which implies optimal manipulation is non-negative ($m \geq 0$).

2.2 Optimal Data Manipulation

The objective of borrowers is to minimize the sum of debt service plus data manipulation costs. With this in mind, consider first determination of debt service. Substituting equation (4) into equation (3), the model-implied interest rate schedule is:

$$\iota(m, x, \widetilde{\alpha}, \widetilde{\beta}) = l + \frac{1 + r - l}{\underbrace{F[\widetilde{\alpha} + \widetilde{\beta}(x + m)]}_{\equiv \widehat{F}}} - 1. \tag{5}$$

For tractability, let

$$\Omega(z) \equiv [F(z)]^{-1} \Rightarrow \iota(m, x, \widetilde{\alpha}, \widetilde{\beta}) = l + (1 + r - l)\Omega[\widetilde{\alpha} + \widetilde{\beta}(x + m)] - 1. \tag{6}$$

The following lemma, relegated to the appendix, establishes some useful properties of Ω for three standard classes of default prediction models: linear probability, logit, and probit.

Lemma 1. *Let* $\Omega(z) \equiv [F(z)]^{-1}$ *where* $F(z) \equiv e^z(1 + e^z)^{-1}$ *or* $F(z) \equiv \mathcal{N}(z)$*. Then* Ω *is strictly decreasing and strictly convex on* $\Re$*. If* $F(z) \equiv \min\{1, \max\{0, z\}\}$*, then* Ω *is strictly decreasing and strictly convex on* $(0, 1)$.

Consider first the binary manipulation setup and let Ψ denote the reduction in the interest rate that results from data manipulation of size δ, when the posted DPM features parameters $(\widetilde{\alpha}, \widetilde{\beta})$. We have:

$$\begin{aligned} \Psi(\delta, x, \widetilde{\alpha}, \widetilde{\beta}) &\equiv \iota(0, x, \widetilde{\alpha}, \widetilde{\beta}) - \iota(\delta, x, \widetilde{\alpha}, \widetilde{\beta}) \qquad (7) \\ &= (1 + r - l)\left(\Omega(\widetilde{\alpha} + \widetilde{\beta}x) - \Omega(\widetilde{\alpha} + \widetilde{\beta}x + \widetilde{\beta}\delta)\right). \end{aligned}$$

[4] Of course, in Nash equilibrium, the econometrician and borrowers move simultaneously.

A borrower will manipulate if and only if $c \leq \Psi$. Thus, the implied probability of data manipulation is given by:

$$\Pi \equiv G\left[\Psi(\delta, x, \widetilde{\alpha}, \widetilde{\beta})\right] = G\left[(1+r-l)\left(\Omega(\widetilde{\alpha}+\widetilde{\beta}x) - \Omega(\widetilde{\alpha}+\widetilde{\beta}x+\widetilde{\beta}\delta)\right)\right]. \tag{8}$$

We have the following comparative statics:

$$\begin{aligned}
\frac{\partial \Pi}{\partial x} &= \widetilde{\beta}(1+r-l)\left[\Omega'(\widetilde{\alpha}+\widetilde{\beta}x) - \Omega'(\widetilde{\alpha}+\widetilde{\beta}x+\widetilde{\beta}\delta)\right] g\left[\Psi(\delta, x, \widetilde{\alpha}, \widetilde{\beta})\right] \qquad (9)\\
\frac{\partial \Pi}{\partial \widetilde{\alpha}} &= (1+r-l)\left[\Omega'(\widetilde{\alpha}+\widetilde{\beta}x) - \Omega'(\widetilde{\alpha}+\widetilde{\beta}x+\widetilde{\beta}\delta)\right] g\left[\Psi(\delta, x, \widetilde{\alpha}, \widetilde{\beta})\right]\\
\frac{\partial \Pi}{\partial \widetilde{\beta}} &= (1+r-l)\left[\begin{array}{c} x\left(\Omega'(\widetilde{\alpha}+\widetilde{\beta}x) - \Omega'(\widetilde{\alpha}+\widetilde{\beta}x+\widetilde{\beta}\delta)\right) \\ -\delta\Omega'(\widetilde{\alpha}+\widetilde{\beta}x+\widetilde{\beta}\delta) \end{array}\right] g\left[\Psi(\delta, x, \widetilde{\alpha}, \widetilde{\beta})\right].
\end{aligned}$$

When Ω is decreasing and convex, as is the case for standard functions F (Lemma 1), the first two comparative statics immediately above are negative: Manipulation probability is decreasing in both $\widetilde{\alpha}$ and x. Intuitively, the incentive to manipulate decreases with a borrower's *baseline credit score* $z = \widetilde{\alpha} + \widetilde{\beta}x$. More formally, this baseline effect can be understood as arising from the fact that the interest rate ι (equation (5)) is a decreasing *convex* function of the imputed repayment probability $\widehat{F}$. Starting at a higher initial $\widehat{F}$, incremental increases in $\widehat{F}$ have a smaller effect on the interest rate.

Returning to equation (9), the effect of an increase in $\widetilde{\beta}$ on manipulation probability is ambiguous. This is because two competing effects are operative. Consider, if $x = 0$, the only effect present is that with higher $\widetilde{\beta}$, each unit of manipulation has a larger effect on the credit score since $\partial z/\partial m = \widetilde{\beta}$. This effect encourages manipulation. However, for borrowers with $x > 0$, an increase in $\widetilde{\beta}$ raises the baseline credit score, which discourages manipulation.

Consider next borrower incentives for data manipulation when they instead face quadratic costs. Optimal manipulation satisfies:

$$m^*(x, \widetilde{\alpha}, \widetilde{\beta}, c) \in \arg\min_m \frac{1}{2}cm^2 + \iota(m, x, \widetilde{\alpha}, \widetilde{\beta}).$$

The first-order condition here is that the borrower equates the marginal reduction in her interest rate with marginal manipulation costs, or:

$$cm^* + \iota_m(m^*, x, \widetilde{\alpha}, \widetilde{\beta}) = 0 \Rightarrow cm^* + (1 + r - l)\widetilde{\beta}\Omega'[\widetilde{\alpha} + \widetilde{\beta}(x + m^*)] = 0. \tag{10}$$

If Ω is indeed convex (Lemma 1), the second-order condition for a local minimum will be satisfied with

$$c + \iota_{mm}(m^*, x, \widetilde{\alpha}, \widetilde{\beta}) = c + (1 + r - l)\widetilde{\beta}^2\Omega''[\widetilde{\alpha} + \widetilde{\beta}(x + m)] > 0.$$

Applying the implicit function theorem to the first-order condition, we obtain the following comparative statics:

$$\begin{aligned}
\frac{\partial m^*}{\partial x} &= -\frac{(1 + r - l)\widetilde{\beta}^2\Omega''[\widetilde{\alpha} + \widetilde{\beta}(x + m^*)]}{c + \iota_{mm}(m^*, x, \widetilde{\alpha}, \widetilde{\beta})} \\
\frac{\partial m^*}{\partial \widetilde{\alpha}} &= -\frac{(1 + r - l)\widetilde{\beta}\Omega''[\widetilde{\alpha} + \widetilde{\beta}(x + m^*)]}{c + \iota_{mm}(m^*, x, \widetilde{\alpha}, \widetilde{\beta})} \\
\frac{\partial m^*}{\partial \widetilde{\beta}} &= -\frac{(1 + r - l)\left[\widetilde{\beta}(x + m^*)\Omega''(\widetilde{\alpha} + \widetilde{\beta}(x + m^*)) + \Omega'(\widetilde{\alpha} + \widetilde{\beta}(x + m^*))\right]}{c + \iota_{mm}(m^*, x, \widetilde{\alpha}, \widetilde{\beta})} \\
\frac{\partial m^*}{\partial c} &= -\frac{m^*}{c + \iota_{mm}(m^*, x, \widetilde{\alpha}, \widetilde{\beta})}.
\end{aligned}$$

These comparative statics results are analogous to those obtained above in the binary setup. In particular, let us consider settings with Ω convex. Then the first two comparative statics are negative: Manipulation is decreasing in the true covariate x and the posted model intercept $\widetilde{\alpha}$.

The effect of an increase in $\widetilde{\beta}$ on manipulation is ambiguous. On one hand, with higher $\widetilde{\beta}$, a given manipulation increment has a larger effect on the credit score, since $\partial z/\partial m = \widetilde{\beta}$. However, starting at a given $\widetilde{x} > 0$, an increase in $\widetilde{\beta}$ raises the imputed repayment probability $\widehat{F}$. This makes the interest rate (equation (5)) less sensitive to further increases in $\widehat{F}$, discouraging manipulation at the margin.

The following proposition summarizes results from this section.

Proposition 1. *Suppose the posted model features slope coefficient* $\widetilde{\beta} > 0$. *If* $F(z) \equiv e^z(1 + e^z)^{-1}$ *or* $F(z) \equiv \mathcal{N}(z)$, *(the probability of) manipulation is strictly decreasing in the true covariate* x, *as well as the posted intercept* $\widetilde{\alpha}$. *If* $F(z) \equiv \min\{1, \max\{0, z\}\}$, *(the probability of) manipulation is weakly decreasing in* x and $\widetilde{\alpha}$. *The effect of an increase in* $\widetilde{\beta}$ *is ambiguous.*

3 Goodhart's Law in Linear Probability Models

This section analyzes whether and how Goodhart's Law would manifest itself if the true data generating process was the linear probability model (LPM), with

$$\mathbb{E}[y|x] = \Pr[y = 1|x] = a + bx. \tag{11}$$

Anticipating, the LPM is an attractive starting point since many arguments can be phrased in terms of intuitive objects, such as covariances. Moreover, as we show, many results for the LPM carry over to settings in which F is non-linear, although arguments and proofs differ.

To begin, consider the mean squared prediction error (MSPE) generated when a univariate function $\upsilon(\cdot)$ is applied to the measured covariate $\widetilde{x}$. We have:[5]

$$\begin{aligned}
\mathbb{E}\left[(y-\upsilon(\widetilde{x}))^2\right] &= \mathbb{E}\left\{[(y-\mathbb{E}(y|\widetilde{x}))+(\mathbb{E}(y|\widetilde{x})-\upsilon(\widetilde{x}))]^2\right\} \qquad (12)\\
&= \mathbb{E}\{(y-\mathbb{E}(y|\widetilde{x}))^2+(\mathbb{E}(y|\widetilde{x})-\upsilon(\widetilde{x}))^2+2(y-\mathbb{E}(y|\widetilde{x}))(\mathbb{E}(y|\widetilde{x})-\upsilon(\widetilde{x}))]\}\\
&= \mathbb{E}\{(y-\mathbb{E}(y|\widetilde{x}))^2\}+\mathbb{E}\left[(\mathbb{E}(y|\widetilde{x})-\upsilon(\widetilde{x}))^2\right]+2\mathbb{E}\left[(y-\mathbb{E}(y|\widetilde{x}))\mathbb{E}(y|\widetilde{x})\right]-2\mathbb{E}\left[(y-\mathbb{E}(y|\widetilde{x}))\upsilon(\widetilde{x})\right]\\
&= \mathbb{E}\left[[\mathbb{E}(y|\widetilde{x})-y]^2\right]+\mathbb{E}\left[[\upsilon(\widetilde{x})-\mathbb{E}(y|\widetilde{x})]^2\right].
\end{aligned}$$

The final line above results from the final terms in the penultimate line being equal to zero due to orthogonality of the prediction error $y - \mathbb{E}(y|\widetilde{x})$ to any univariate function of $\widetilde{x}$.

Equation (12) shows the MSPE obtained by applying a default prediction model (function) υ to a measured covariate $\widetilde{x}$ can be viewed as consisting of two components. The first component is the natural loss arising from using the specific covariate $\widetilde{x}$ as a basis for prediction. The second component results from any gap between the model υ and the conditional expectation function. It follows that, if an econometrician were to treat the distribution of $\widetilde{x}$ as fixed, as they would in Nash equilibrium, the optimal function υ would be the conditional expectation function. However, a Stackelberg leader econometrician would want to account for the effect of υ on the distribution of observed covariates.

[5] Frankel and Kartik (2023) present analogous expressions for their setup.

Suppose now that borrowers respond optimally to a posted prediction model υ, generating a covariate report $\widetilde{x}_\upsilon$. From equation (12) it follows:

$$\mathbb{E}\left[(y-\upsilon(\widetilde{x}_\upsilon))^2\right] = \mathbb{E}\left[[\mathbb{E}(y|\widetilde{x}_\upsilon)-y]^2\right] + \mathbb{E}\left[[\upsilon(\widetilde{x}_\upsilon)-\mathbb{E}(y|\widetilde{x}_\upsilon)]^2\right]. \tag{13}$$

The preceding decomposition stresses an important point. Changes in the default prediction model have a direct effect on the loss metric, as captured by the second term, the distance between the prediction model υ and the conditional expectation function. However, there is a second indirect effect attributable to the fact that changes in the posted model υ will change the observed data and their natural predictive power, as captured by the first term.

As an extreme case, suppose data manipulation is not possible, with υ being affine, $\upsilon(x)=\alpha+\beta x$. We then have:

$$\mathbb{E}\left[(y-\alpha-\beta x)^2\right] = \mathbb{E}\left[[\mathbb{E}(y|x)-y]^2\right] + \mathbb{E}\left[[(\alpha+\beta x)-(a+bx)]^2\right]. \tag{14}$$

Notice, if data manipulation is impossible, there is no incentive-based tradeoff in selecting the parameters (α,β), since the first term is a fixed quantity representing natural loss coming from predicting y based upon x. Consequently, here MSPE is minimized by setting $(\alpha,\beta)=(a,b)$. We have the following remark, which is also shown (below) to apply to ML estimators in our parable economy.

Remark 1. *In an economy without data manipulation, the econometric procedure (OLS/MLE) recovers the structural parameters* (a,b) *and each loan is correctly priced if interest rates are set according to equation* (5).

Consider now the consequences of data manipulation. Recall $\widetilde{x}$ denotes the covariate that would be reported by a borrower who faced a model with posted coefficients $(\widetilde{\alpha},\widetilde{\beta})$, hence the common tilde superscript. Suppose that, ex post, an econometrician were to collect data on the realized $(y,\widetilde{x})$ pairs, and then estimate a linear prediction model with coefficients (α,β). Noting that $y=y^2$, with y and $\widetilde{x}$ being independent conditional upon x, the resulting MSPE can be expressed as:

$$\mathbb{E}\left[(y-\alpha-\beta\widetilde{x})^2\right] = \alpha^2+(1-2\alpha)\mathbb{E}[y]+2\alpha\beta\mathbb{E}\left[\widetilde{x}\right]+\beta^2\mathbb{E}[\widetilde{x}^2]-2\beta\mathbb{E}\{\mathbb{E}\left[y|x\right]\mathbb{E}\left[\widetilde{x}|x\right]\}. \tag{15}$$

Applying the law of iterated expectations to the final term in the preceding equation, we find:

$$\begin{aligned}\mathbb{E}\{\mathbb{E}\left[y|x\right]\mathbb{E}\left[\widetilde{x}|x\right]\} &= \mathbb{E}\left\{(a+bx)\mathbb{E}\left[\widetilde{x}|x\right]\right\} \\ &= a\mathbb{E}\left[\widetilde{x}\right]+b\mathbb{E}\left[x\widetilde{x}\right].\end{aligned} \tag{16}$$

Substituting the preceding expression into equation (15), we find that the MSPE can be expressed parametrically as follows:

$$MSPE(\underbrace{\alpha,\beta}_{\text{Candidate}};\underbrace{\widetilde{\alpha},\widetilde{\beta}}_{\text{Posted}},\underbrace{a,b}_{\text{DGP}})=\alpha^2+(1-2\alpha)[a+b\mathbb{E}(x)]+2(\alpha-a)\beta\mathbb{E}\left[\widetilde{x}\right]+\beta^2\mathbb{E}[\widetilde{x}^2]-2\beta b\mathbb{E}\left[x\widetilde{x}\right]. \tag{17}$$

Care must be taken in interpreting the preceding equation since the probability distribution of the $\widetilde{x}$ varies with the parameters of the posted model $(\widetilde{\alpha},\widetilde{\beta})$.

An econometrician given ex post access to realized $(y,\widetilde{x})$ pairs would obtain OLS coefficients $(\widehat{\alpha}_{ols},\widehat{\beta}_{ols})$ that satisfy:

$$(\widehat{\alpha}_{ols},\widehat{\beta}_{ols})\in\arg\min_{\alpha,\beta}\ MSPE(\alpha,\beta;\widetilde{\alpha},\widetilde{\beta},a,b). \tag{18}$$

The preceding objective function is strictly concave, and the first-order conditions (FOCs below) are:

$$\begin{aligned}0 &= 2\widehat{\alpha}-2[a+b\mathbb{E}(x)]+2\widehat{\beta}\mathbb{E}[\widetilde{x}] \\ 0 &= 2(\widehat{\alpha}-a)\mathbb{E}[\widetilde{x}]+2\widehat{\beta}\mathbb{E}\left[\widetilde{x}^2\right]-2b\mathbb{E}[x\widetilde{x}].\end{aligned} \tag{19}$$

Throughout the analysis, let $\widehat{\beta}^{sw}_{ols}$ denote the OLS coefficient arising from a regression of s on w, for arbitrary (s,w). From the preceding FOCs we have:

$$\begin{aligned}\widehat{\alpha}_{ols} &= a+b\mathbb{E}[x]-\widehat{\beta}\mathbb{E}[\widetilde{x}]. \\ \widehat{\beta}_{ols} &= b\times\underbrace{\frac{\mathbb{E}[x\widetilde{x}]-\mathbb{E}[x]\mathbb{E}[\widetilde{x}]}{\mathbb{E}\left[\widetilde{x}^2\right]-(\mathbb{E}[\widetilde{x}])^2}}_{\equiv\widehat{\beta}^{x\widetilde{x}}_{ols}}.\end{aligned} \tag{20}$$

Inspection of equation (20) leads to the following proposition.

Proposition 2. *Consider a linear probability model for default and suppose the true covariate x has no explanatory power in predicting default in clean historical data $(b=0)$. Then regardless of* *the parameters $(\widetilde{\alpha}, \widetilde{\beta})$ of the model posted, the OLS/MSPE coefficient estimates derived from the resulting manipulated data will be $(\widehat{\alpha}_{ols}, \widehat{\beta}_{ols}) = (a, b) = (a, 0)$.*

Proposition 2 informs us that $b = 0 \Rightarrow \widehat{\beta} = 0$, regardless of the parameters of the posted model. Phrased colloquially, the econometrician cannot get something from nothing. Intuitively, if $b = 0$, the random variables (x, c) privately observed by borrowers, as well as their incentive compatible manipulation m, are uninformative about default risk. The proposition also informs us that some statistical regularities $(b = 0)$ observed in clean historical data will remain robust over time, even if borrowers are induced to manipulate by a posted model featuring $\widetilde{\beta} > 0$.

Let us now consider Goodhart's law when the econometrician utilizes a linear probability model. In particular, suppose the econometrician recovers the parameters (a, b) using clean training data, and then naively informs future strategic borrowers that repayment probabilities will be computed as $F(a+b\widetilde{x})$. Finally, let us suppose the econometrician re-estimates the intercept and slope parameters ex post using OLS (or MLE in the context of logit and probit models) using the resulting manipulated data drawn from this strategic cohort. Let us dub the resulting estimates, *Goodhart estimates*, since this is the thought experiment Goodhart contemplated (but certainly not an econometric practice he advocated). Formally, in the present context we have

$$\text{Goodhart Estimates: } (\widehat{\alpha}_{ols}, \widehat{\beta}_{ols}) \in \arg\min_{\alpha,\beta} MSPE(\underbrace{\alpha, \beta}_{\text{Candidate}}\ ; \underbrace{a, b}_{\text{Posted}}, \underbrace{a, b}_{\text{Historical}}). \tag{21}$$

From Proposition 2 it follows that if $b = 0$, the Goodhart estimates will be equal $(a, 0)$. That is, if $b = 0$, coefficient estimates will remain stable over time. In fact, this claim holds *a fortiori* since posting a model with slope $b = 0$ induces zero manipulation, so the Goodhart estimate must be the equal to the (true) coefficients that obtain in clean historical data. Conversely, we know that if the econometrician posts a model with a positive slope, there will be a positive measure of manipulation. That is:

$$\widetilde{\beta} = b > 0 \Rightarrow \mathbb{E}[\widetilde{x}] > \mathbb{E}[x]. \tag{22}$$

Combining the preceding equation with equation (20), we have the following proposition.

Proposition 3. *Consider a linear probability model for default, with the posted model featuring intercept and slope parameters set at their value* (a, b) *under clean historical data. The OLS/MSPE estimates derived from the resulting data will remain equal to* (a, b) *if and only if the unmanipulated covariate has no explanatory power* $(b = 0)$*. If the unmanipulated covariate has explanatory power* $(b > 0)$*, then* $\widehat{\alpha}_{ols} < a$ *and/or* $\widehat{\beta}_{ols} < b$*.*

Before proceeding to numerical analysis, it would be useful to obtain further analytical insight about the behavior of the Goodhart estimates that arise when $b > 0$. To this end, consider the extreme case of a setting where $\widetilde{x}$ and x are perfectly correlated. We know:[6]

$$\rho_{\widetilde{x}x} = 1 \Rightarrow \widehat{\beta}_{ols} = \frac{b}{\widehat{\beta}^{\widetilde{x}x}_{ols}}. \tag{23}$$

Next recall that $\widetilde{x} \equiv x + m$, with the stochastic image of m decreasing in x (Proposition 1). This implies

$$\widehat{\beta}^{\widetilde{x}x}_{ols} \leq 1. \tag{24}$$

Taken together, the two preceding equations suggest that the Goodhart slope will tend to overshoot b in settings with high correlation between $\widetilde{x}$ and x. When the Goodhart slope does indeed overshoot b, Proposition 3 informs us that the estimated intercept must adjust downwards (sharply) to counter the universal ratings inflation that would otherwise result, since $\widehat{\beta}\widetilde{x} > bx$ for all borrowers.

Further insight into coefficient behavior can be gained through numerical analysis. In terms of calibration our objective is to make results comparable across the different pair-wise combinations of model for default prediction (linear versus logit) and manipulation technology (binary versus quadratic cost). To achieve this goal, the calibration must be done for each combination, thus giving four distinct sets of parameter choices. For convenience, we set the risk-free rate to zero, $r = 0$ and the return in the low state at $l = 0.5$. The remaining parameter values depend upon the manipulation technology.

For binary manipulation decisions, we set $\delta = 0.2$, and a probability distribution of manipulation cost, G, uniform in $[0, c_{\max}]$, where $c_{\max} = 0.03$ is set to ensure that at $x = x_{\max}$ manipulation

[6] Using $\rho_{\widetilde{x}x} = 1 \Rightarrow \widehat{\beta}^{\widetilde{x}x}_{ols} = 1/\widehat{\beta}^{x\widetilde{x}}_{ols}$.

probability is lower than 1. For manipulation based on a quadratic cost function, we assume a deterministic cost parameter $c = 1$.

For the linear prediction model, the probability distribution of x is uniform on $[0, 1.8]$, which ensures a non-negative probability for all x. For the DGP, we set $a = 0.5$, and allow b to go up to a maximum value $\overline{b} = 0.25$. This ensures that the survival probabilities are similar, for the same x, for the linear and the logit model. The choice of maximum x is the same under the quadratic cost technology, because at $\overline{b} = 0.25$ the average $m^* \approx \delta$. In this sense, the models feature similar manipulation scales. For the logit prediction model of credit risk, x has uniform distribution on $[0, 1]$, with $a = 0$.

Figures 1-4 plot outputs associated with Goodhart estimates (equation (21)) that obtain for alternative values of the clean historical data slope coefficient b.

Figure 1 considers the binary manipulation setup. A few points are worth noting. First, consistent with Proposition 2, the estimated intercept always undershoots the true intercept. In fact, the estimated intercept declines monotonically with the posted model's slope $\widetilde{\beta} = b$, countering the increase in manipulation. By way of contrast, the ratio of the fitted OLS slope coefficient to b is non-monotonic, sometimes undershooting b and sometimes overshooting b.

Finally, and perhaps most interesting, the bottom panel of Figure 1 shows the gap between the respective MSPE under manipulated versus the clean data MSPE benchmark is non-monotonic in b. To understand this, note that data manipulation itself does not necessarily increase the MSPE. Rather, it is heterogeneity in manipulation across borrowers that increases the MSPE here. As the posted b increases sufficiently, the behavior of borrowers becomes increasingly similar, with most borrowers manipulating by δ with probability 1, as illustrated in Figure 2. This causes the MSPE to fall with b in some regions.

Figures 3 and 4 present analogs for the quadratic manipulation cost setup. Here the estimated OLS slope coefficient overshoots the historical slope b for all $b > 0$. To understand this, we note that, at least for the parameter values considered, the correlation between x and $\widetilde{x}$ is close to 1. Therefore, equations (24) and (23) suggest overshooting will be present. To counter the effect of ever-increasing manipulation, the estimated intercept decreases in the posted model slope b. In contrast with the binary manipulation setup, we see in Figure 4 the MSPE increases monotonically in b, due to the increase in incentive compatible manipulation, which here only serves to confound inference.

4 Manipulable Data in Logit and Probit Models

The remainder of the paper assumes the true data generating process is given by equation (1), with F being the logistic or normal distribution, as in logit and probit models.

4.1 Maximum Likelihood Estimation

Consider first an empirical likelihood function L and corresponding log likelihood $\mathcal{L}$ in an economy without data manipulation:

$$\begin{aligned} L &\equiv \prod_{i=1}^{I}(F(\alpha+\beta x_i))^{y_i}(1-F(\alpha+\beta x_i))^{1-y_i} && (25)\\ \mathcal{L} &\equiv \sum_{i=1}^{I} y_i \ln(F(\alpha+\beta x_i)) + (1-y_i)\ln(1-F(\alpha+\beta x_i)). && (26)\end{aligned}$$

Using the law of iterated expectations, an expected log likelihood function for an arbitrarily large i.i.d. sample can be computed as:

$$\begin{aligned} \frac{1}{I}\mathbb{E}\left[\mathbb{E}(\mathcal{L}|\mathbf{x})\right] &= \frac{1}{I}\int\limits_X \left[\mathbb{E}\left(\sum_{i=1}^{I} y_i \ln F(\alpha+\beta x_i) + (1-y_i)\ln(1-F(\alpha+\beta x_i))|\mathbf{x}\right)\right] h(x)dx \quad (27)\\ &= \frac{1}{I}\int\limits_X \left[\sum_{i=1}^{I}\mathbb{E}\left\{y_i \ln F(\alpha+\beta x_i) + (1-y_i)\ln(1-F(\alpha+\beta x_i))|\mathbf{x}\right\}\right] h(x)dx\\ &= \int\limits_X \left\{\ln F(\alpha+\beta x)\mathbb{E}(y|x) + \ln(1-F(\alpha+\beta x))[1-\mathbb{E}(y|x)]\right\} h(x)dx. \end{aligned}$$

Substituting the conditional expectation function into the preceding equation, we obtain the following expected log likelihood function for economies without data manipulation:

$$\mathcal{L} = \int\limits_X \left\{F(a+bx)\ln F(\alpha+\beta x) + [1-F(a+bx)]\ln[1-F(\alpha+\beta x)]\right\} h(x)dx. \quad (28)$$

Consider next the maximizer of the preceding expected log likelihood function. The MLE intercept and slope $(\widehat{\alpha}, \widehat{\beta})$ satisfy the following FOCs:

$$\begin{aligned} \int\limits_X \left[\frac{F(a+bx)}{F(\widehat{\alpha}+\widehat{\beta}x)} - \frac{1-F(a+bx)}{1-F(\widehat{\alpha}+\widehat{\beta}x)}\right] f(\widehat{\alpha}+\widehat{\beta}x)h(x)dx &= 0 \qquad (29) \\ \int\limits_X \left[\frac{F(a+bx)}{F(\widehat{\alpha}+\widehat{\beta}x)} - \frac{1-F(a+bx)}{1-F(\widehat{\alpha}+\widehat{\beta}x)}\right] xf(\widehat{\alpha}+\widehat{\beta}x)h(x)dx &= 0. \end{aligned}$$

Note, the preceding FOCs are satisfied at $(\widehat{\alpha}, \widehat{\beta}) = (a, b)$. That is, in the absence of data manipulation, the expected log likelihood is maximized by utilizing an econometric model parameterized at the true coefficients.[7] Phrased differently, given an infinite sample of clean training data, the econometrician can use MLE to recover the deep structural parameters determining credit risk and each individual loan can be priced correctly (Remark 1).

Consider instead the expected log likelihood function given manipulated covariates $\widetilde{x}$ that emerge in response to posted model parameters $(\widetilde{\alpha}, \widetilde{\beta})$. Applying the law of iterated expectations we have:[8]

$$\begin{aligned} \frac{1}{I}\mathbb{E}\left[\mathbb{E}(\mathcal{L}|x)\right] &= \frac{1}{I}\int\limits_X \left[\sum_{i=1}^{I} \mathbb{E}\left[y_i \ln(F(\alpha+\beta\widetilde{x}_i)) + (1-y_i)\ln(1-F(\alpha+\beta\widetilde{x}_i))|x\right]\right] h(x)dx \qquad (30) \\ &= \int\limits_X \{\mathbb{E}\left[y\ln(F(\alpha+\beta\widetilde{x})) + (1-y)\ln(1-F(\alpha+\beta\widetilde{x}))|x\right]\}h(x)dx \\ &= \int\limits_X \{\mathbb{E}(y|x)\mathbb{E}[\ln(F(\alpha+\beta\widetilde{x}))|x] + (1-\mathbb{E}(y|x))\mathbb{E}[\ln(1-F(\alpha+\beta\widetilde{x}))|x]\}h(x)dx \\ &= \int\limits_X \{F(a+bx)\mathbb{E}[\ln(F(\alpha+\beta\widetilde{x}))|x] + (1-F(a+bx))\mathbb{E}[\ln(1-F(\alpha+\beta\widetilde{x}))|x]\}h(x)dx. \end{aligned}$$

Recall, in the binary setup, $\widetilde{x} \in \{x, x+\delta\}$. Therefore, equation (30) implies that here the expected

[7] See Conniffe (1987) for a related discussion of the merits of MLE.

[8] Note that conditional upon x, m and $\widetilde{x}$ are uninformative about y.

log likelihood function is:

$$\mathcal{L}(\underbrace{\alpha,\beta}_{\text{Candidate}};\underbrace{\widetilde{\alpha},\widetilde{\beta}}_{\text{Posted}},\underbrace{a,b}_{\text{DGP}}) = \int\limits_X \left[\begin{array}{c} F(a+bx)[1-G(\Psi(\delta,x,\widetilde{\alpha},\widetilde{\beta}))]\ln F(\alpha+\beta x) \\ +F(a+bx)G(\Psi(\delta,x,\widetilde{\alpha},\widetilde{\beta}))\ln F(\alpha+\beta x+\beta\delta) \\ +[1-F(a+bx)][1-G(\Psi(\delta,x,\widetilde{\alpha},\widetilde{\beta}))]\ln[1-F(\alpha+\beta x)] \\ +[1-F(a+bx)]G(\Psi(\delta,x,\widetilde{\alpha},\widetilde{\beta}))\ln[1-F(\alpha+\beta x+\beta\delta)] \end{array}\right] h(x)dx. \tag{31}$$

For borrowers facing quadratic manipulation costs, optimal manipulation takes the form $m(x,\widetilde{\alpha},\widetilde{\beta},c)$. Substituting this into equation (30), we obtain the following expected log likelihood function:

$$\mathcal{L}(\alpha,\beta;\widetilde{\alpha},\widetilde{\beta},a,b) = \int\limits_C\int\limits_X \left[\begin{array}{c} F(a+bx)\ln[F(\alpha+\beta x+\beta m(x,\widetilde{\alpha},\widetilde{\beta},c))] \\ +[1-F(a+bx)]\ln[1-F(\alpha+\beta x+\beta m(x,\widetilde{\alpha},\widetilde{\beta},c))] \end{array}\right] h(x)\gamma(c)dxdc. \tag{32}$$

With the preceding two objective functions in mind, let us define the MLE estimator $(\widehat{\alpha},\widehat{\beta})$ for an economy with true structural parameters (a,b), with data being generated by borrowers who face the posted model $(\widetilde{\alpha},\widetilde{\beta})$. We have:

$$MLE:(\widehat{\alpha},\widehat{\beta})\in\arg\max_{\alpha,\beta}\quad\mathcal{L}(\underbrace{\alpha,\beta}_{\text{Candidate}};\underbrace{\widetilde{\alpha},\widetilde{\beta}}_{\text{Posted}},\underbrace{a,b}_{\text{DGP}}). \tag{33}$$

Differentiating equation (31), we have the following FOCs for intercept and slope in the binary manipulation setup:

$$\begin{array}{rcl} \mathcal{L}_1(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) & = & \int\limits_X \left[\begin{array}{c} \left[\frac{F(a+bx)}{F(\widehat{\alpha}+\widehat{\beta}x)}-\frac{1-F(a+bx)}{1-F(\widehat{\alpha}+\widehat{\beta}x)}\right][1-G(\Psi(\delta,x,\widetilde{\alpha},\widetilde{\beta}))]f(\widehat{\alpha}+\widehat{\beta}x) \\ +\left[\frac{F(a+bx)}{F(\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}\delta)}-\frac{1-F(a+bx)}{1-F(\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}\delta)}\right]G(\Psi(\delta,x,\widetilde{\alpha},\widetilde{\beta}))f(\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}\delta) \end{array}\right] h(x)dx = 0 \qquad (34) \\ \mathcal{L}_2(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) & = & \int\limits_X \left[\begin{array}{c} \left[\frac{F(a+bx)}{F(\widehat{\alpha}+\widehat{\beta}x)}-\frac{1-F(a+bx)}{1-F(\widehat{\alpha}+\widehat{\beta}x)}\right][1-G(\Psi(\delta,x,\widetilde{\alpha},\widetilde{\beta}))]f(\widehat{\alpha}+\widehat{\beta}x)x \\ +\left[\frac{F(a+bx)}{F(\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}\delta)}-\frac{1-F(a+bx)}{1-F(\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}\delta)}\right]G(\Psi(\delta,x,\widetilde{\alpha},\widetilde{\beta}))f(\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}\delta)(x+\delta) \end{array}\right] h(x)dx = 0. \end{array}$$

Notice the FOCs immediately above are similar to those arising when manipulation is impossible (equation (29)), but now at each point x the econometrician weights the manipulated covariate report $x+\delta$ by the probability of manipulation at that point.

Differentiating equation (32) we have the following FOCs for the intercept and slope in the quadratic manipulation cost setup:

$$\begin{aligned}
\mathcal{L}_1(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) &= \int_C\int_X \begin{bmatrix} \frac{F(a+bx)}{F[\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}m(x,\widetilde{\alpha},\widetilde{\beta},c)]} \\ -\frac{1-F(a+bx)}{1-F[\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}m(x,\widetilde{\alpha},\widetilde{\beta},c)]} \end{bmatrix} f[\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}m(x,\widetilde{\alpha},\widetilde{\beta},c)]h(x)\gamma(c)dxdc = 0 \qquad (35)\\
\mathcal{L}_2(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) &= \int_C\int_X \begin{bmatrix} \frac{F(a+bx)}{F[\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}m(x,\widetilde{\alpha},\widetilde{\beta},c)]} \\ -\frac{1-F(a+bx)}{1-F[\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}m(x,\widetilde{\alpha},\widetilde{\beta},c)]} \end{bmatrix} (x+m)f[\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}m(x,\widetilde{\alpha},\widetilde{\beta},c)]h(x)\gamma(c)dxdc = 0.
\end{aligned}$$

Notice the FOCs immediately above are identical in form to those arising when manipulation is impossible (equation (29)), but with the manipulated covariate $\widetilde{x}$ taking the place of x.

In light of the preceding FOCs, it is useful to consider the special case in which $b = 0$. Note, regardless of the parameters of the posted model, $(\widetilde{\alpha},\widetilde{\beta})$, MLE performed on the manipulated data will return the true coefficients $(a, 0)$, with

$$\begin{aligned}
\mathcal{L}_1(a,0;\widetilde{\alpha},\widetilde{\beta},a,0) &= 0 \qquad (36)\\
\mathcal{L}_2(a,0;\widetilde{\alpha},\widetilde{\beta},a,0) &= 0.
\end{aligned}$$

Thus, we have the analog of Proposition 2, but now in the context of MLE estimation.

Proposition 4. *Suppose the true covariate x has no explanatory power in predicting default in clean historical data ($b = 0$). Then regardless of the parameters $(\widetilde{\alpha},\widetilde{\beta})$ of the model posted, the MLE coefficient estimates derived from the resulting manipulated data will be $(\widehat{\alpha},\widehat{\beta}) = (a, b) = (a, 0)$.*

4.2 Goodhart's Law in Logit and Probit Models

To illustrate how Goodhart's law would manifest itself if logit or probit is employed, suppose as above that the econometrician recovers the deep parameters (a, b) using clean historical training data. Suppose also that, in light of Remark 1, the lender decides that, for the next cohort of borrowers, it will set interest rates according to (5), with the posted default prediction model coefficients set at $(\widetilde{\alpha},\widetilde{\beta}) = (a, b)$. Estimating on the resulting manipulated data, the econometrician would obtain:

$$\text{Goodhart Estimates} = (\widehat{\alpha}, \widehat{\beta}) \in \arg\max_{\alpha,\beta} \; \mathcal{L}(\alpha, \beta; \underbrace{a, b}_{\text{Posted}}, \underbrace{a, b}_{\text{Historical}}). \tag{37}$$

Parameter instability of the sort suggested by Goodhart (1975) is easily shown if the clean covariate has predictive power ($b > 0$). In particular, suppose $b > 0$ and consider any candidate coefficients (α, β) such that $\alpha \geq a$ and $\beta \geq b > 0$. For the respective cases of binary manipulation and quadratic manipulation costs, the FOCs cannot be satisfied, with

$$\begin{aligned}
\mathcal{L}_1(\alpha, \beta; a, b, a, b) &= \int\limits_X \left[\begin{array}{c} \left[\frac{F(a+bx)}{F(\alpha+\beta x)} - \frac{1-F(a+bx)}{1-F(\alpha+\beta x)}\right] [1 - G(\Psi(\delta, x, a, b))] f(\alpha + \beta x) \\ + \left[\frac{F(a+bx)}{F(\alpha+\beta x+\beta\delta)} - \frac{1-F(a+bx)}{1-F(\alpha+\beta x+\beta\delta)}\right] G(\Psi(\delta, x, a, b)) f(\alpha + \beta x + \beta\delta) \end{array} \right] h(x)dx < 0 \\
\mathcal{L}_1(\alpha, \beta; a, b, a, b) &= \int\limits_C \int\limits_X \left[\begin{array}{c} \frac{F(a+bx)}{F[\alpha+\beta x+\beta m(x,a,b,c)]} \\ -\frac{1-F(a+bx)}{1-F[\alpha+\beta x+\beta m(x,a,b,c)]} \end{array} \right] f[\alpha + \beta x + \widehat{\beta} m(x, a, b, c)] h(x)\gamma(c) dx dc < 0.
\end{aligned}$$

The preceding two inequalities, along with Proposition 4, establishes the following proposition, the MLE analog of Proposition 3.

Proposition 5. *Consider a logit/probit model for default, with the posted model featuring intercept and slope parameters set at their value (a, b) under clean historical data. The MLE estimates $(\widehat{\alpha}, \widehat{\beta})$ derived from the resulting data will remain equal to (a, b) if and only if the unmanipulated covariate has no explanatory power ($b = 0$). If the unmanipulated covariate has explanatory power ($b > 0$), then $\widehat{\alpha} < a$ and/or $\widehat{\beta} < b$.*

The next section, detailing potential econometric responses to Goodhart's law, will provide additional formal results. However, before proceeding to this analysis, it will be useful to get a sense of the consequences of Goodhart's law by way of numerical analysis.

Figures 5 and 6 consider Goodhart estimates (37) in the context logit estimation when borrowers have the binary manipulation technology.[9] See Figures 1 and 2 for analogs in the context of linear probability models. Consistent with Proposition 5, for all $b > 0$ there is always some amount of undershooting, primarily in the estimated intercept, in order to counter the effect of manipulation.

[9] Analogous figures for probit are virtually identical.

The estimated slope coefficient undershoots and then overshoots the historical slope b, as was the case for OLS. Finally, the bottom panel of Figure 6 shows that once again, the performance target is non-monotone in b. Apparently, in the binary manipulation setup, increases in b can bring about increases in the likelihood function arising when ex post estimation is performed on manipulated data pairs $(y, \widetilde{x})$ that arise in response to posted models parameterized according to historical clean data estimates (a, b).

Figures 7 and 8 consider Goodhart estimates in the context of logit estimation when borrowers face quadratic manipulation costs. These figures mirror the numerical results for the OLS setup contained in Figures 3 and 4. In particular, we again see overshooting of the historical slope coefficient, with the intercept term decreasing in b in order to counter the effect of higher manipulation. The net effect is a steady decline in the likelihood ratio when ex post estimation is performed on manipulated data pairs $(y, \widetilde{x})$.

5 Econometric Responses

Having illustrated how Goodhart's law will manifest itself in logit and probit estimation, this section examines potential econometric remedies. The first subsection considers Nash econometric responses and the second subsection considers Stackelberg econometric responses.

5.1 Nash Default Prediction

Proposition 5 shows that if the true covariate has explanatory power $(b > 0)$ there will be inconsistency between a posted econometric model obtained from clean historical covariates and the MLE estimates that would be obtained if ex post estimation was then performed on the manipulated covariates generated by borrowers facing the (a, b) model. This is the sort of parameter instability contemplated by Goodhart.

By way of contrast, a *fixed point model* (α^*, β^*) is internally consistent in that:

$$(\alpha^*, \beta^*) \in \arg\max_{\alpha,\beta} \;\; \mathcal{L}(\underbrace{\alpha, \beta}_{\text{Candidate}} ; \underbrace{\alpha^*, \beta^*}_{\text{Posted}}, \underbrace{a, b}_{\text{DGP}}). \tag{38}$$

That is, a fixed point is an econometric model that maximizes the expected log likelihood given the distribution of manipulated covariates that would be generated by borrowers responding to that model. Phrased differently, a fixed point constitutes the *Nash model* of a game in which borrowers submit covariates, with the econometrician making a simultaneous choice of intercept and slope coefficients.

Differentiating equation (31), we have the following FOCs for the Nash model intercept and slope in the binary manipulation setting:

$$\begin{aligned}
0 &= \mathcal{L}_1(\alpha^*, \beta^*; \alpha^*, \beta^*, a, b) \qquad\qquad (39)\\
&= \int\limits_X \left[\begin{array}{c} \left[\frac{F(a+bx)}{F(\alpha^*+\beta^*x)} - \frac{1-F(a+bx)}{1-F(\alpha^*+\beta^*x)}\right] [1 - G(\Psi(\delta, x, \alpha^*, \beta^*))] f(\alpha^* + \beta^* x) \\ + \left[\frac{F(a+bx)}{F(\alpha^*+\beta^*x+\beta^*\delta)} - \frac{1-F(a+bx)}{1-F(\alpha^*+\beta^*x+\beta^*\delta)}\right] G(\Psi(\delta, x, \alpha^*, \beta^*)) f(\alpha^* + \beta^* x + \beta^* \delta) \end{array} \right] h(x)dx \\
0 &= \mathcal{L}_2(\alpha^*, \beta^*; \alpha^*, \beta^*, a, b) \\
&= \int\limits_X \left[\begin{array}{c} \left[\frac{F(a+bx)}{F(\alpha^*+\beta^*x)} - \frac{1-F(a+bx)}{1-F(\alpha^*+\beta^*x)}\right] [1 - G(\Psi(\delta, x, \alpha^*, \beta^*))] f(\alpha^* + \beta^* x)x \\ + \left[\frac{F(a+bx)}{F(\alpha^*+\beta^*x+\beta^*\delta)} - \frac{1-F(a+bx)}{1-F(\alpha^*+\beta^*x+\beta^*\delta)}\right] G(\Psi(\delta, x, \alpha^*, \beta^*)) f(\alpha^* + \beta^* x + \beta^* \delta)(x + \delta) \end{array} \right] h(x)dx.
\end{aligned}$$

Differentiating equation (32) we have the following FOCs for the intercept and slope of a Nash model when borrowers face quadratic manipulation costs:

$$\begin{aligned}
0 &= \mathcal{L}_1(\alpha^*, \beta^*; \alpha^*, \beta^*, a, b) \qquad\qquad (40)\\
&= \int\limits_C \int\limits_X \left[\begin{array}{c} \frac{F(a+bx)}{F[\alpha^*+\beta^*x+\beta^*m(x,\alpha^*,\beta^*,c)]} \\ - \frac{1-F(a+bx)}{1-F[\alpha^*+\beta^*x+\beta^*m(x,\alpha^*,\beta^*,c)]} \end{array} \right] f[\alpha^* + \beta^* x + \beta^* m(x, \alpha^*, \beta^*, c)] h(x) \gamma(c) dx dc \\
0 &= \mathcal{L}_2(\alpha^*, \beta^*; \alpha^*, \beta^*, a, b) \\
&= \int\limits_C \int\limits_X \left[\begin{array}{c} \frac{F(a+bx)}{F[\alpha^*+\beta^*x+\beta^*m(x,\alpha^*,\beta^*,c)]} \\ - \frac{1-F(a+bx)}{1-F[\alpha^*+\beta^*x+\beta^*m(x,\alpha^*,\beta^*,c)]} \end{array} \right] [x + m(x, \alpha^*, \beta^*, c)] f[\alpha^* + \beta^* x + \beta^* m(x, \alpha^*, \beta^*, c)] h(x) \gamma(c) dx dc.
\end{aligned}$$

From these FOCs, we obtain the following propositions.

Proposition 6. *A posted econometric model with intercept and slope parameters (a, b) (derived from clean historical data) represents an MLE fixed point if and only if the unmanipulated covariate* *has no explanatory power* $(b = 0)$*. If the unmanipulated covariate has explanatory power* $(b > 0)$*, any fixed point model* (α^*, β^*) *features* $\alpha^* < a$ *and/or* $\beta^* < b$*.*

Consider first the sufficiency component of the first stated claim. If $b = 0$, then posting $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$ results in zero manipulation. And we know that here

$$\widetilde{x} = x \Rightarrow (\widehat{\alpha}, \widehat{\beta}) = (a, b) = (a, 0) \equiv (\widetilde{\alpha}, \widetilde{\beta}). \tag{41}$$

Since the necessity component of the first stated claim follows from the second stated claim, we need only establish that claim. To this end, consider $b > 0$, and note that the FOCs immediately above cannot be satisfied if the bracketed term is negative for all x. That is, from the FOCs it follows that:

$$\begin{aligned} \beta^* &\geq b > 0 \Rightarrow \alpha^* < a \qquad (42) \\ \alpha^* &\geq a \Rightarrow \beta^* < b. \end{aligned}$$

Intuitively, since borrowers manipulate their covariates upwards, an MLE estimator must respond by shifting down the intercept and/or slope.

The following proposition is also easily verified.

Proposition 7. *If the unmanipulated covariate has explanatory power, any MLE fixed point will assign explanatory power to the manipulated covariate.*

To prove the preceding proposition, consider $b > 0$ but suppose to the contrary that there exists a fixed point model featuring $\beta^* = 0$. Posting $\beta^* = 0$ results in zero manipulation, but with $\widetilde{x} = x$ the MLE estimate is $\widehat{\beta} = b > 0$, contradicting the fixed point claim.

5.2 Fixed Point Convergence

Proposition 6 is informative about the number of iterations necessary for fixed-point convergence when the econometrician has adaptive expectations. In particular, consider an econometrician who behaves adaptively in the sense of posting to the current cohort of borrowers the coefficients that emerged from estimation within the prior cohort data, with the first round of estimation being on clean historical data. Proposition 6 implies that if $b = 0$, convergence is immediate: All estimation rounds result in the same estimates $(a, 0)$ as those obtained from the clean historical data cohort.

Suppose next $b > 0$. Proposition 6 informs us that here the Goodhart estimates $(\widehat{\alpha}, \widehat{\beta})$ that result from estimating within the first strategic borrower cohort, who faced posted parameters (a, b), will necessarily differ from (a, b). Continuing with the adaptive iteration, if the econometrician were to post before the *second* strategic cohort the Goodhart estimates $(\widehat{\alpha}, \widehat{\beta})$ obtained from the first, one would expect different coefficients to emerge. After all, the first strategic cohort faced parameters (a, b) and the second will face parameters $(\widehat{\alpha}, \widehat{\beta}) \neq (a, b)$, with the inequality following from Proposition 5. Phrased differently one does not generally expect the Goodhart estimates to be a fixed point since this would imply that two distinct pairs of posted coefficients give rise to identical ex post estimates:

$$\begin{aligned}(a, b) &\Rightarrow (\widehat{\alpha}, \widehat{\beta}) \neq (a, b) \\ (\widehat{\alpha}, \widehat{\beta}) &\Rightarrow (\widehat{\alpha}, \widehat{\beta}).\end{aligned} \tag{43}$$

But note, in order to obtain a contradiction here, one would need to establish one-to-oneness of the preceding correspondence mapping posted coefficients to ex post estimates, a property that does not necessarily hold. For example, in the binary manipulation setup, if $c_{\max}$ is sufficiently small, a continuum of distinct sets of posted coefficients can induce all borrowers to manipulate. But if all borrowers manipulate by δ, the manipulation is perfectly corrected by lowering the intercept by $b\delta$, with:

$$F[a - b\delta + b(x + \delta)] = F(a + bx). \tag{44}$$

Indeed, this line or argument allows one to establish potential equality of Goodhart estimates and fixed points. We have the following proposition.

Proposition 8. *Suppose borrowers are able to manipulate covariates upward by δ in exchange for an idiosyncratic stochastic cost, with the covariate bounded above at $x_{\max}$ and the maximum possible* *manipulation cost*

$$c_{\max} \leq \Psi(\delta, x_{\max}, a - b\delta, b).$$

Then $(a - b\delta, b)$ is a fixed point. At this fixed point the zero manipulation likelihood ratio is achieved. Moreover, if $c_{\max}$ satisfies the stronger condition

$$c_{\max} \leq \Psi(\delta, x_{\max}, a, b),$$

posting the clean cohort coefficients (a, b) to the first strategic cohort will result in their generating Goodhart estimates equal to the fixed point $(a - b\delta, b)$.

We conclude this subsection with some numerical comparative statics on how Nash models change with the clean data slope parameter b. More specifically, Figures 9 and 10 consider the binary manipulation and quadratic manipulation cost setups, respectively, contrasting Nash econometric models (fixed points) with the corresponding Goodhart estimates (equation (37)). Consistent with Proposition 8, in the binary manipulation setup (Figure 9), Nash and Goodhart coefficients are identical if b is sufficiently large. For lower values of b, the Nash and Goodhart coefficients are very similar, converging to each other at the fixed point $(a, 0)$ if $b = 0$, consistent with Proposition 6. Apparently, if one were to first post a model with parameters set at their historical value (a, b), and then followed up with a single round of ex post estimation on the manipulated data, one would arrive at parameter estimates that are close to being internally consistent fixed points–at least in environments where borrowers have binary manipulation technologies.

By way of contrast, Figure 10 shows that Goodhart estimates differ from the Nash parameter estimates if borrowers instead face the quadratic manipulation cost technology. More specifically, for higher values of b, we see that Goodhart estimates differ from (a, b) by a wide margin. That is, borrowers responding to (a, b) generate Goodhart coefficients $(\widehat{\alpha}, \widehat{\beta})$ that are far from (a, b). Notice, in order for Goodhart estimates to be a fixed point here, it must be that condition (43) is satisfied, despite a large distance between (a, b) and $(\widehat{\alpha}, \widehat{\beta})$. Apparently, in the quadratic manipulation cost setup, a large difference in posted models manifests itself in large differences in ex post estimates. Thus, Nash estimates differ from Goodhart estimates by a wide margin for high values of b.

5.3 Stackelberg Default Prediction

As stressed by Frankel and Kartik (2023), if the econometrician has commitment power, a Nash econometric model can generally be improved upon. After all, in solving for her Nash strategy, the econometrician takes the distribution of manipulated covariates as given. By way of contrast, with commitment power, the econometrician can act as a Stackelberg leader and post optimal coefficients taking into account the effect of posted coefficients on borrower covariate reports. For example, in the setting considered by Frankel and Kartik, the Stackelberg econometrician always finds it optimal to post a model with slope $\beta < b$, in order to discourage manipulation.

A *Stackelberg model* $(\alpha^{**}, \beta^{**})$ satisfies

$$(\alpha^{**}, \beta^{**}) \in \arg\max_{\alpha,\beta} \mathcal{L}(\underbrace{\alpha, \beta}_{\text{Estimated}}; \underbrace{\alpha, \beta}_{\text{Posted}}, \underbrace{a, b}_{\text{DGP}}). \tag{45}$$

At this point it is instructive to contrast the Nash econometric program in equation (38) with the preceding Stackelberg program. Since, in the Nash program, the econometrician takes the covariates as given, it is as if she takes the posted model as given. In stark contrast, in the Stackelberg program, the econometrician properly views herself as determining the posted model, and with it, the distribution of manipulated covariates.

Recall, the FOCs for any MLE estimator, including Nash econometric models, are $\mathcal{L}_1 = \mathcal{L}_2 = 0$. These FOCs ensure optimal prediction "given the data." By way of contrast, the FOCs for Stackelberg econometric models are:

$$\begin{aligned}
\mathcal{L}_1(\alpha^{**}, \beta^{**}; \alpha^{**}, \beta^{**}, a, b) &= -\mathcal{L}_3(\alpha^{**}, \beta^{**}; \alpha^{**}, \beta^{**}, a, b) \\
\mathcal{L}_2(\alpha^{**}, \beta^{**}; \alpha^{**}, \beta^{**}, a, b) &= -\mathcal{L}_4(\alpha^{**}, \beta^{**}; \alpha^{**}, \beta^{**}, a, b).
\end{aligned}$$

That is, when the econometrician has commitment power, she is willing (*and able*) to sacrifice a bit on ex post prediction power in order to increase prediction power ex ante. She does so by taking into account the effects of the posted econometric model on borrower behavior, and these effects are captured by the partial derivatives $\mathcal{L}_3$ and $\mathcal{L}_4$.

In the context of the binary manipulation setup, we have the following FOC for the Stackelberg intercept:

$$\int\limits_X \left[\begin{array}{c} \left[\frac{F(a+bx)}{F(\alpha^{**}+\beta^{**}x)} - \frac{1-F(a+bx)}{1-F(\alpha^{**}+\beta^{**}x)}\right][1-G(\Psi(\delta,x,\alpha^{**},\beta^{**}))]f(\alpha^{**}+\beta^{**}x) \\ +\left[\frac{F(a+bx)}{F(\alpha^{**}+\beta^{**}x+\beta^{**}\delta)} - \frac{1-F(a+bx)}{1-F(\alpha^{**}+\beta^{**}x+\beta^{**}\delta)}\right] G(\Psi(\delta,x,\alpha^{**},\beta^{**}))f(\alpha^{**}+\beta^{**}x+\beta^{**}\delta) \end{array} \right] h(x)$$

$$= \quad -\int\limits_X \left[\begin{array}{c} F(a+bx)\ln\left[\frac{F(\alpha^{**}+\beta^{**}x+\beta^{**}\delta)}{F(\alpha^{**}+\beta^{**}x)}\right] \\ -[1-F(a+bx)]\ln\left[\frac{1-F(\alpha^{**}+\beta^{**}x+\beta^{**}\delta)}{1-F(\alpha^{**}+\beta^{**}x)}\right] \end{array} \right] g[\Psi(\delta,x,\alpha^{**},\beta^{**})]\Psi_3(\delta,x,\alpha^{**},\beta^{**})h(x)dx.$$

And the FOC for the Stackelberg slope is:

$$\int\limits_X \left[\begin{array}{c} \left[\frac{F(a+bx)}{F(\alpha^{**}+\beta^{**}x)} - \frac{1-F(a+bx)}{1-F(\alpha^{**}+\beta^{**}x)}\right][1-G(\Psi(\delta,x,\alpha^{**},\beta^{**}))]f(\alpha^{**}+\beta^{**}x)x \\ +\left[\frac{F(a+bx)}{F(\alpha^{**}+\beta^{**}x+\beta^{**}\delta)} - \frac{1-F(a+bx)}{1-F(\alpha^{**}+\beta^{**}x+\beta^{**}\delta)}\right] G(\Psi(\delta,x,\alpha^{**},\beta^{**}))f(\alpha^{**}+\beta^{**}x+\beta^{**}\delta)(x+\delta) \end{array} \right] h(x)dx$$

$$= \quad -\int\limits_X \left[\begin{array}{c} F(a+bx)\ln\left[\frac{F(\alpha^{**}+\beta^{**}x+\beta^{**}\delta)}{F(\alpha^{**}+\beta^{**}x)}\right] \\ +[1-F(a+bx)]\ln\left[\frac{1-F(\alpha^{**}+\beta^{**}x+\beta^{**}\delta)}{1-F(\alpha^{**}+\beta^{**}x)}\right] \end{array} \right] g[\Psi(\delta,x,\alpha^{**},\beta^{**})]\Psi_4(\delta,x,\alpha^{**},\beta^{**})h(x)dx.$$

In the context of quadratic manipulation costs, the FOC for the Stackelberg intercept is:

$$\int\limits_C\int\limits_X \left[\begin{array}{c} \frac{F(a+bx)}{F[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]} \\ -\frac{1-F(a+bx)}{1-F[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]} \end{array} \right] f[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]h(x)\gamma(c)dxdc$$

$$= \quad -\int\limits_C\int\limits_X \left[\begin{array}{c} \frac{F(a+bx)}{F[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]} \\ -\frac{1-F(a+bx)}{1-F[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]} \end{array} \right] f[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]\beta^{**}m_2(x,\alpha^{**},\beta^{**},c)h(x)\gamma(c)$$

Or, more compactly:

$$0 = \int\limits_C\int\limits_X \left[\begin{array}{c} \left(\frac{F(a+bx)}{F[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]} - \frac{1-F(a+bx)}{1-F[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]}\right) \\ f[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)][1+\beta^{**}m_2(x,\alpha^{**},\beta^{**},c)]h(x)\gamma(c) \end{array} \right] dxdc. \qquad (47)$$

The FOC for the Stackelberg slope is:

$$\int\limits_C\int\limits_X \left[\begin{array}{c} \frac{F(a+bx)}{F[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]} - \\ \frac{1-F(a+bx)}{1-F[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]} \end{array} \right] f[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)][x+m(x,\alpha^{**},\beta^{**},c)]h(x)\gamma(c)dx$$

$$= \quad -\int\limits_C\int\limits_X \left[\begin{array}{c} \frac{F(a+bx)}{F[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]} - \\ \frac{1-F(a+bx)}{1-F[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]} \end{array} \right] f[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]\beta^{**}m_3(x,\alpha^{**},\beta^{**},c)h(x)\gamma(c)dx$$

Or more compactly:

$$0 = \int\limits_{C}\int\limits_{X} \left[\begin{array}{c} \left(\frac{F(a+bx)}{F[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]} - \frac{1-F(a+bx)}{1-F[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)]} \right) \\ f[\alpha^{**}+\beta^{**}x+\beta^{**}m(x,\alpha^{**},\beta^{**},c)][x+m(x,\alpha^{**},\beta^{**},c)+\beta^{**}m_3(x,\alpha^{**},\beta^{**},c)]h(x)\gamma(c)dxdc \end{array} \right]. \tag{48}$$

The following two propositions follow directly upon inspecting the preceding four FOCs.

Proposition 9. *A posted econometric model with intercept and slope parameters (a, b) (derived from clean historical data) represents a Stackelberg model if the unmanipulated covariate has no explanatory power $(b = 0)$.*

Proposition 10. *If borrowers are able to manipulate covariates upward by δ in exchange for an idiosyncratic stochastic cost, a posted model with parameters set at $(a - b\delta, b)$ will be a Stackelberg model if the covariate is bounded above at some finite $x_{\max}$ and $c_{\max} \leq \Psi(\delta, x_{\max}, a-b\delta, b)$. Moreover, at this point, the zero-manipulation likelihood ratio is reached.*

Figures 11 and 12 contrast Stackelberg and Nash model coefficients, in the binary and quadratic cost setups, respectively. It is startling to note that in the quadratic cost setup, Stackelberg and Nash coefficients are virtually identical. By way of contrast, there is a large difference between Stackelberg and Nash coefficients in the binary manipulation setup.

The intuition behind Figure 11, depicting the binary manipulation technology, is as follows. If b is low, the candidate Stackelberg model is also likely to have a low slope coefficient, being in the neighborhood of b. In this case, the probability of manipulation is relatively low for most borrowers. Here the Stackelberg econometrician tries to nudge all borrowers toward homogeneity in the direction of not manipulating. This is accomplished by posting a model with a relatively low value of β. By way of contrast, for high values of b, the candidate Stackelberg model is likely to have a high slope coefficient in the neighborhood of b. In this case, the probability of manipulation is relatively high for most borrowers. Here the Stackelberg econometrician tries to nudge all borrowers toward homogeneity in the direction manipulating with probability 1. This is accomplished by posting a model with a relatively high value of β.

6 Multivariate Model

In the interest of analytical tractability, attention has been confined to estimating coefficients of a univariate econometric model. Indeed, it is well-known that little can be said about effects arising from measurement error in more than one regressor (see Greene (1997)), let alone endogenous manipulation that depends upon the coefficients of the regression model per Goodhart's law. Nevertheless, results analogous to those presented above can be obtained in a multivariate setting provided that manipulation is confined to a single regressor. To take the simplest case, consider OLS/MSPE estimation of the following linear probability model:

$$\Pr[y = 1|x, w] = \mathbb{E}[y|x, w] = a + bx + kw. \tag{49}$$

Suppose the covariate $w \geq 0$ is non-manipulable, and consider any technology, such as those considered above, such that a positive measure of borrowers will report $\widetilde{x} > x$ in response to any posted coefficient $\widetilde{\beta} > 0$.

The objective is to find coefficients:

$$(\widehat{\alpha}, \widehat{\beta}, \widehat{\kappa}) \in \arg \min_{(\alpha,\beta,\kappa)} \mathbb{E}\left[(y - \alpha - \beta\widetilde{x} - \kappa w)^2\right]. \tag{50}$$

Noting that $y = y^2$ here, the preceding expression for the MSPE can be written as:

$$\begin{aligned} MSPE = \alpha^2 + (1-\alpha)\mathbb{E}[y] + 2\alpha\beta\mathbb{E}[\widetilde{x}] + 2\alpha\kappa\mathbb{E}[w] + 2\kappa\beta\mathbb{E}[w\widetilde{x}] \\ +\beta^2\mathbb{E}[\widetilde{x}^2] + \kappa^2\mathbb{E}[w^2] - 2\kappa\mathbb{E}[wy] - 2\beta\mathbb{E}[\widetilde{x}y]. \end{aligned} \tag{51}$$

Focusing on the final term in the preceding equation, we note that conditional independence of $\widetilde{x}$ and y implies:

$$\begin{aligned} \mathbb{E}\left\{\mathbb{E}[\widetilde{x}y|x, w]\right\} &= \mathbb{E}\left\{\mathbb{E}[\widetilde{x}|x, w]\mathbb{E}[y|x, w]\right\} \\ &= \mathbb{E}\left\{\mathbb{E}[\widetilde{x}|x, w](a + bx + kw)\right\} \\ &= a\mathbb{E}[\widetilde{x}] + b\mathbb{E}[\widetilde{x}x] + k\mathbb{E}[\widetilde{x}w]. \end{aligned} \tag{52}$$

And similarly, conditional independence of w and y implies:

$$\begin{aligned}
\mathbb{E}\left\{\mathbb{E}[wy|x,w]\right\} &= \mathbb{E}\left\{\mathbb{E}[w|x,w]\mathbb{E}[y|x,w]\right\} \\
&= \mathbb{E}\left\{w(a+bx+kw)\right\} \\
&= a\mathbb{E}[w]+b\mathbb{E}[xw]+k\mathbb{E}[w^2].
\end{aligned}$$

Substituting the two preceding equalities into equation (51) allows us to rewrite the MSPE as follows:

$$\begin{gathered}
MSPE = \alpha^2 + (1-2\alpha)\{a+b\mathbb{E}[x]+k\mathbb{E}[w]\} + 2\alpha\beta\mathbb{E}[\widetilde{x}] + 2\alpha\kappa\mathbb{E}[w] + 2\kappa\beta\mathbb{E}[w\widetilde{x}] + \beta^2\mathbb{E}[\widetilde{x}^2] + \kappa^2\mathbb{E}[w^2] \\
-2\beta\{aE[\widetilde{x}]+bE[\widetilde{x}x]+kE[\widetilde{x}w]\} - 2\kappa\{aE[w]+bE[xw]+kE[w^2]\}.
\end{gathered}$$

From the FOC for the intercept, we find:

$$\widehat{\alpha} = a + b\mathbb{E}[x] + k\mathbb{E}[w] - \widehat{\beta}\mathbb{E}[\widetilde{x}] - \widehat{\kappa}\mathbb{E}[w]. \tag{53}$$

The FOC for $\widehat{\beta}$ is:

$$0 = \widehat{\alpha}\mathbb{E}[\widetilde{x}] + \widehat{\kappa}\mathbb{E}[w\widetilde{x}] + \widehat{\beta}\mathbb{E}[\widetilde{x}^2] - \{a\mathbb{E}[\widetilde{x}] + b\mathbb{E}[\widetilde{x}x] + k\mathbb{E}[\widetilde{x}w]\}. \tag{54}$$

Substituting the expression for the intercept (53) into the FOC for $\widehat{\beta}$ we obtain:

$$\begin{aligned}
0 \;=\; & a\mathbb{E}[\widetilde{x}] + b\mathbb{E}[x]\mathbb{E}[\widetilde{x}] + k\mathbb{E}[w]\mathbb{E}[\widetilde{x}] - \widehat{\beta}\left(\mathbb{E}[\widetilde{x}]\right)^2 - \widehat{\kappa}\mathbb{E}[w]\mathbb{E}[\widetilde{x}] \qquad (55) \\
& +\widehat{\kappa}\mathbb{E}[w\widetilde{x}] + \widehat{\beta}\mathbb{E}[\widetilde{x}^2] - \{a\mathbb{E}[\widetilde{x}] + b\mathbb{E}[\widetilde{x}x] + k\mathbb{E}[\widetilde{x}w]\}.
\end{aligned}$$

Rearranging terms we obtain:

$$\widehat{\beta} = b \times \beta_{ols}^{x\widetilde{x}} + (k-\widehat{\kappa})\beta_{ols}^{w\widetilde{x}}. \tag{56}$$

Finally, the FOC for $\widehat{\kappa}$ is:

$$0 = \widehat{\alpha}\mathbb{E}[w] + \widehat{\beta}\mathbb{E}[w\widetilde{x}] + \widehat{\kappa}\mathbb{E}[w^2] - \{aE[w] + bE[xw] + kE[w^2]\}. \tag{57}$$

Substituting the expression for the intercept (53) into the FOC for $\widehat{\kappa}$ we obtain:

$$\widehat{\kappa} = k + b \times \beta_{ols}^{xw} - \widehat{\beta}\beta_{ols}^{\widetilde{x}w}. \tag{58}$$

From equations (53), (56) and (58), we have the following analog of Propositions 2 and 4, establishing the impossibility of getting something from nothing, with:

$$b = 0 \Rightarrow \left(\widehat{\alpha}, \widehat{\beta}, \widehat{\kappa}\right) = (a, b, k) = (a, 0, k). \tag{59}$$

It thus follows that if $b = 0$, then $(a, 0, k)$ represents both a Goodhart estimate and a fixed point, consistent with Proposition 6.

We also have the following result demonstrating the analog of Propositions 3 and 5, the necessity of at least some downward coefficient slope if $b > 0$. In particular,

$$\widehat{\beta} \geq b > 0 \text{ and } \widehat{\kappa} \geq k \Rightarrow \widehat{\alpha} < a. \tag{60}$$

Since a fixed point model is just a special case of the estimator here, it follows that any fixed point model must also feature some downward coefficient shift, consistent with Proposition 6.

Finally, consistent with Proposition 7, it is readily verified that $b > 0$ implies a fixed point model cannot feature $\widehat{\beta} = 0$. After all, if the posted model features a coefficient of zero on the manipulable covariate, there will be no manipulation, in which case the MSPE estimator would be (a, b, k), a contradiction.

7 Conclusion

This paper contributes to a growing literature on econometric responses to data manipulation, focusing on default prediction models. We suggest a number of natural directions for future work. First, it would be useful to consider settings in which multiple covariates can be manipulated, although it is likely that analytical results would be much more difficult to obtain. Second, it would be useful to consider whether and how standard machine-learning tools could be adapted in light of data manipulation, again in the context of logit and probit-type credit risk prediction. Finally, as the stock of such models grows, it would be useful to evaluate the performance of alternative models empirically.

Appendix

LEMMA 1. *Let* $\Omega(z) \equiv [F(z)]^{-1}$ *where* $F(z) \equiv e^z(1+e^z)^{-1}$ *or* $F(z) \equiv \mathcal{N}(z)$. *Then* Ω *is strictly decreasing and strictly convex on* $\Re$. *If* $F(z) \equiv \min\{1, \max\{0, z\}\}$, *then* Ω *is strictly decreasing and strictly convex on* $(0,1)$.

Proof.

To begin, note that, assuming differentiability, we have

$$\begin{aligned} \Omega'(z) &= -[F(z)]^{-2} f(z) \leq 0 \\ \Omega''(z) &= [F(z)]^{-2} \left[\frac{2[f(z)]^2}{F(z)} - F''(z) \right] \end{aligned}$$

Notice, the first inequality is strict for Logit and Probit models. Consider next the linear probability model for $z \in (0,1)$. We have

$$\begin{aligned} \Omega'(z) &= -\frac{1}{z^2} < 0 \\ \Omega''(z) &= 2z^{-3} > 0 \end{aligned}$$

Consider next Logit. We have:

$$\begin{aligned} F(z) &\equiv \frac{e^z}{1+e^z} \\ F'(z) &= \frac{(1+e^z)e^z - e^{2z}}{(1+e^z)^2} = \frac{e^z}{(1+e^z)^2} \\ F''(z) &= \frac{(1+e^z)^2 e^z - 2e^{2z}(1+e^z)}{(1+e^z)^4} = \frac{e^z(1-e^z)}{(1+e^z)^3} \end{aligned}$$

Thus,

$$\begin{aligned}
\Omega''(z) &= [F(z)]^{-2}\left[\frac{2[f(z)]^2}{F(z)} - F''(z)\right] \\
&= [F(z)]^{-2}\left[\frac{2e^{2z}}{(1+e^z)^4}\frac{1+e^z}{e^z} - \frac{e^z(1-e^z)}{(1+e^z)^3}\right] \\
&= [F(z)]^{-2}\left[\frac{2e^{z}}{(1+e^z)^3} - \frac{e^z(1-e^z)}{(1+e^z)^3}\right] \\
&= [F(z)]^{-2}\left[\frac{2e^{z} - e^z + e^{2z}}{(1+e^z)^3}\right] \\
&= [F(z)]^{-2}\left[\frac{e^z(1+e^z)}{(1+e^z)^3}\right] \\
&= [F(z)]^{-2}\left[\frac{e^z}{(1+e^z)^2}\right] \\
&= \frac{(1+e^z)^2}{e^{2z}}\frac{e^z}{(1+e^z)^2} \\
&= \frac{1}{e^z} > 0
\end{aligned}$$

Thus we have established that Ω is strictly decreasing and convex in the case of Logit.

Finally, let us establish convexity when we consider the Normal CDF. We have:

$$\begin{aligned}
F(z) &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-\frac{1}{2}t^2}dt \\
F'(z) &= \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2} \\
F''(z) &= -\frac{z}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2} = -zF'(z)
\end{aligned}$$

Thus we have:

$$\begin{aligned}
\Omega''(z) &= [F(z)]^{-2}\left[\frac{2[f(z)]^2}{F(z)} - F''(z)\right] \\
&= [F(z)]^{-2}\left[\frac{2[f(z)]^2}{F(z)} + zf(z)\right] \\
&= [F(z)]^{-2}f(z)\left[\frac{2f(z)}{F(z)} + z\right] \\
&= [F(z)]^{-2}f(z)\left[\frac{2f(-z)}{1-F(-z)} + z\right] \\
&= [F(z)]^{-2}f(z)\left[2h(-z) + z\right] \\
&= [F(z)]^{-2}f(z)\left[h(-z) + h(-z) + z\right] > 0
\end{aligned}$$

Where the last line follows from Baricz (2008), who shows that for a standard normal random variable $h(-s) + s > 0$.■

References

[1] Altman, Stewart, 1968, Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy, *Journal of Finance*, 23, 589-619.

[2] Barboza, Flavio, Herbert Kimura and Edward Altman, 2017, Machine learning models and bankruptcy prediction, *Expert Systems with Applications* (83), 405-417.

[3] Björkegren, Daniel, Joshua E. Blumenstock and Samsun Knight, 2020, Manipulation-proof machine learning, arXiv working paper.

[4] Brückner, Michael and Tobias Scheffer, 2011. Stackelberg games for adversarial prediction problems. *Journal of Machine Learning Research*, 13, 2617-2654.

[5] Caton, Gary L., Chiraphol N. Chiyachantana, Choong-Tze Chua and Jeremy Goh, 2011, Earnings management surrounding seasoned bond offerings: Do managers mislead ratings agencies and the bond market?, *Journal of Financial and Quantitative Analysis* (46), 687-708.

[6] Chan-Lau, Jorge A., 2017, Lasso regressions and forecasting models in applied stress testing, IMF Working Paper WP/17/108.

[7] Chen, Yiling, Chara Podimata, Ariel D. Procaccia and Nisarg Shah, 2018. Strategyproof linear regression in high dimensions. Working paper, Harvard University.

[8] Chen, Hongmei and Yaoxin Xiang, 2017, The study of credit scoring models based upon group Lasso, *Information Technology and Quantitative Management*, 677-684.

[9] Conniffe, Denis, 1987, Expected Log Likelihood Estimation. *Journal of the Royal Statistical Society* 36 (4), 317-329.

[10] O. Dekel, F. Fischer, and A. D. Procaccia, 2010. Incentive compatible regression learning. *Journal of Computing System Science*, 76 (8), 759–777.

[11] J. Dong, A. Roth, Z. Schutzman, B. Waggoner, and Z. S. Wu, 2017. Strategic classification from revealed preferences. In *Proceedings of the 2018 ACM Conference.*

[12] Eliaz, Kfir and Ran Spiegler, 2019. The model selection curse. *American Economic Review: Insights*, 1 (2), 127-140.

[13] Frankel, Alex and Navin Kartik, 2019. Muddled information. *Journal of Political Economy*, 129, 1739-1776.

[14] Frankel, Alex and Navin Kartik, 2023. Improving information from manipulable data. *International Economic Review.*

[15] Goodhart, Charles A., 1975. Problems of monetary management: The U.K. experience. *Papers in Monetary Economics* I, Reserve Bank of Australia.

[16] Greene, William H., 1997, *Econometric Analysis*, Third Edition. Prentice Hall, Saddle River, N.J.

[17] Hardt, Moritz, Nimrod Megiddo, Christos Papadimitriou, and Mary Wooters, 2016. Strategic Classification. *Proceedings of the 7th Innovations in Theoretical Computer Science Conference*, 111–122.

[18] Hastie, Trevor, Robert Tibshirani and Jerome Friedman, 2017. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer.

[19] Hennessy, Christopher A., and Charles A.E. Goodhart, 2020. Goodhart's Law and Machine Learning: A Structural Perspective. *International Economic Review.*

[20] Lucas, Robert, 1976. Econometric policy evaluation: A critique, *Carnegie-Rochester Conference Series on Public Policy*, 19-46.

[21] Merton, Robert, 1974, On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance* 29, 449-470.

[22] Rajan, Uday, Amit Seru and Vikrant Vig, 2010, Statistical default models and incentives, *American Economic Review Papers and Proceedings* (100), 506-510.

[23] Stein, Jeremy C., 1989, Efficient capital markets, inefficient firms: A model of myopic corporate behavior, *Quarterly Journal of Economics.*

[24] Tian, Shaonan and Yan Yu, 2017, Financial rations and bankruptcy predictions: International evidence, *International Review of Economics and Finance* (51), 510-526.

[25] Tian, Shaonan, Yan Yu and Hui Guo, 2015, Variable selection and corporate bankruptcy forecasts, *Journal of Banking and Finance* (52), 89-100.

[26] Zhou, Ying, Mohammad Shamsu Uddin, Tabassum Habib, Guotai Chi and Kunpeng Yuan, 2021, Feature selection in credit risk modeling: International evidence, *Economic Research.*

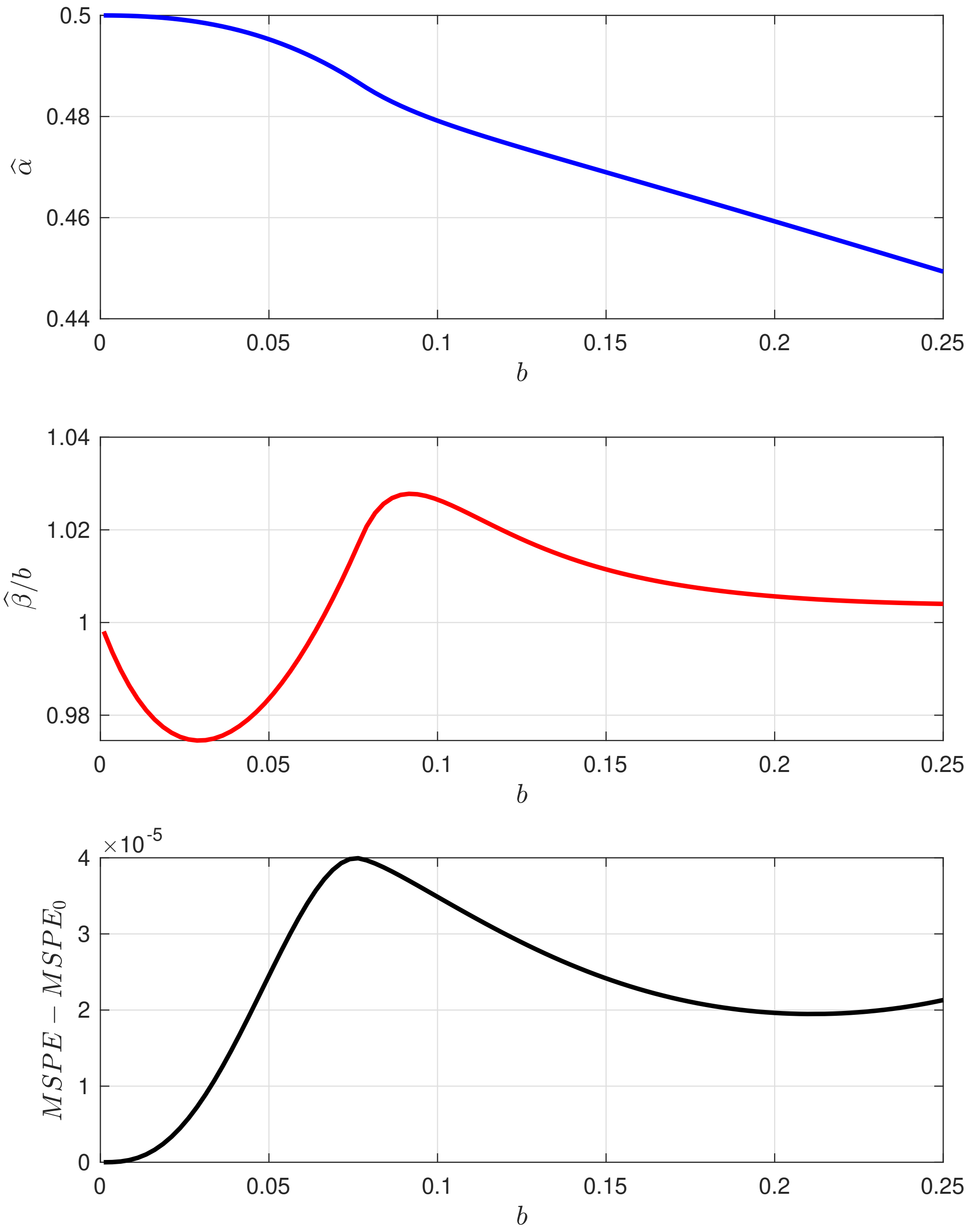

Figure 1: **Goodhart's Law - Linear prediction model - Binary manipulation**. We plot OLS Goodhart estimates against $\widetilde{\beta}$ of the posted credit prediction model, with $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$, and for $b \in]0, \overline{b}]$. In the first panel we plot $\widehat{\alpha}$ and in the second panel $\widehat{\beta}/b$. In the third panel we plot the difference between the optimal mean square prediction error, $MSPE(\widehat{\alpha}, \widehat{\beta}; \widetilde{\alpha}, \widetilde{\beta}; a, b)$, with manipulation and using $(\widehat{\alpha}, \widehat{\beta})$, and the optimal mean square prediction error with no manipulation and therefore based on estimates (a, b). The figure assumes a binary manipulation decision, with $m = \delta = 0.2$, probability distribution of manipulation cost, G, uniform in $[0, c_{\max}]$, and a linear prediction model of credit risk, $\widetilde{\alpha} + \widetilde{\beta}x$, where x has uniform distribution on $[x_{\min}, x_{\max}]$, with $x_{\min} = 0$, and $x_{\max} = \frac{1-a}{\overline{b}} - \delta = 1.8$ to ensure a positive probability for all x and all $\widetilde{\beta}$. We set $a = 0.5$, and $\overline{b} = 0.25$ to ensure consistency across the different credit models, that is the survival probabilities (i.e., the interest rates) are similar, for the same x, for the linear and the logit model. $c_{\max} = 0.03$ is set to make sure at $x = x_{\max}$ manipulation probability is lower than 1. The remaining parameters are $r = 0$ and $l = 0.5$.

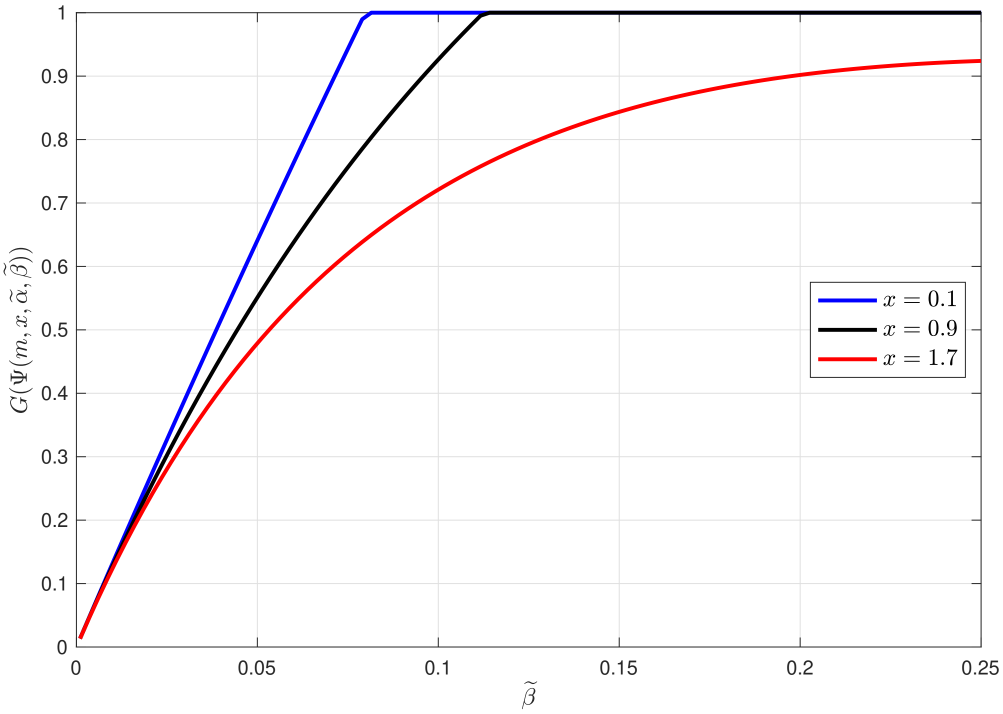

Figure 2: **Incentive compatible manipulation - Linear prediction model - Binary manipulation**. We plot the probability of manipulation, $G(\Psi(m, x, \widetilde{\alpha}, \widetilde{\beta}))$, for three values of x, against $\widetilde{\beta}$ of the posted credit prediction model, with $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$, and for $b \in]0, \overline{b}]$. The model has binary manipulation decision, with $m = \delta = 0.2$ and probability distribution of manipulation cost, G, uniform in $[0, c_{\max}]$. The figure is based on a linear prediction model of credit risk, $\widetilde{\alpha} + \widetilde{\beta}x$, where x has uniform distribution on $[x_{\min}, x_{\max}]$, with $x_{\min} = 0$, and $x_{\max} = \frac{1-a}{\overline{b}} - \delta = 1.8$ to ensure a positive probability for all x and all $\widetilde{\beta}$. We set $a = 0.5$, and $\overline{b} = 0.25$ to ensure consistency across the different credit models, that is the survival probabilities (i.e., the interest rates) are similar, for the same x, for the linear and the logit model. $c_{\max} = 0.03$ is set to make sure at $x = x_{\max}$ manipulation probability is lower than 1. The remaining parameters are $r = 0$ and $l = 0.5$.

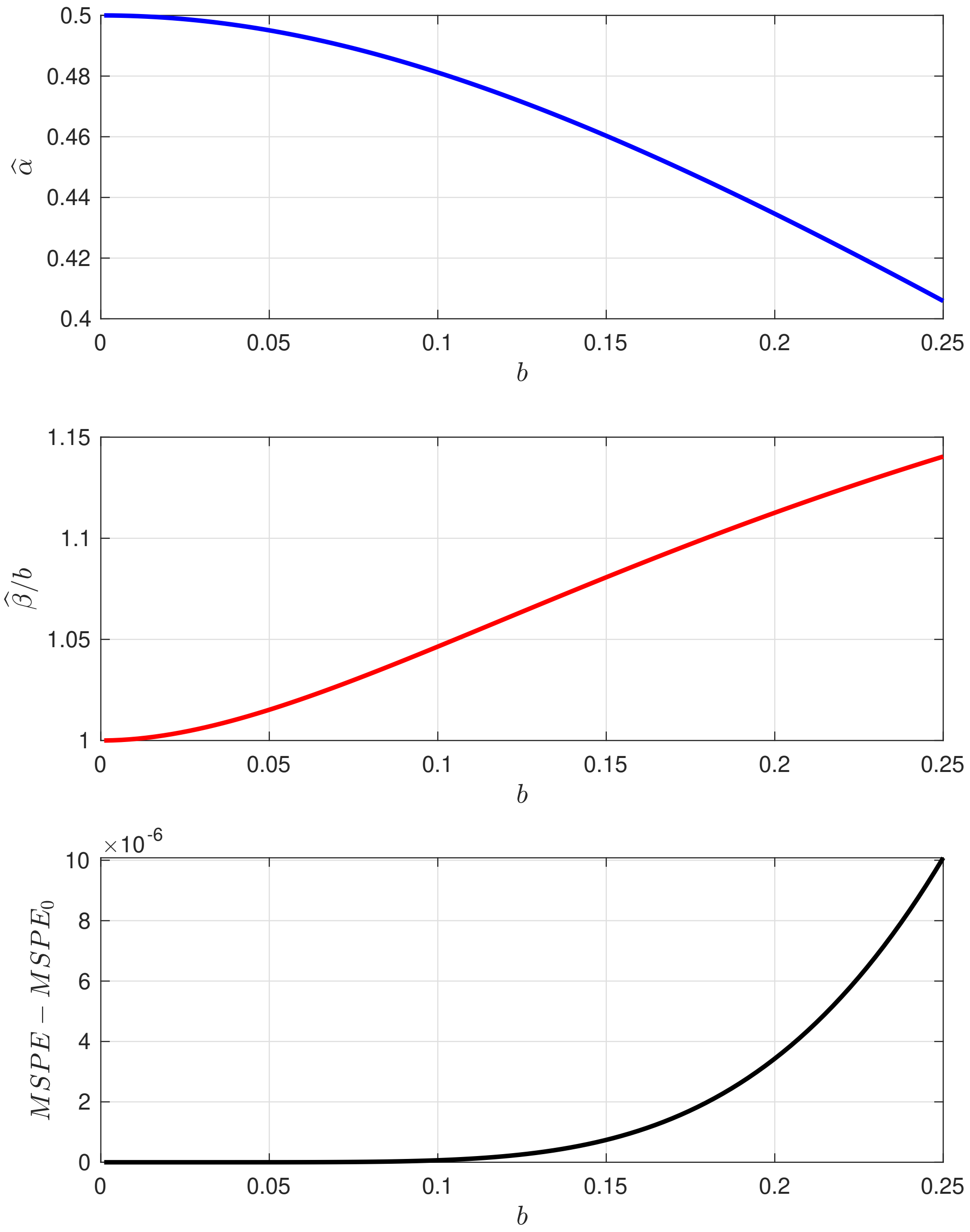

Figure 3: **Goodhart's Law - Linear prediction model - Convex manipulation cost**. We plot OLS Goodhart estimates against $\widetilde{\beta}$ of the posted credit prediction model, with $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$, and for $b \in]0, \overline{b}]$. In the first panel we plot $\widehat{\alpha}$ and in the second panel $\widehat{\beta}/b$. In the third panel we plot the difference between the optimal mean square prediction error, $MSPE(\widehat{\alpha}, \widehat{\beta}; \widetilde{\alpha}, \widetilde{\beta}; a, b)$, with manipulation and using $(\widehat{\alpha}, \widehat{\beta})$, and the optimal mean square prediction error with no manipulation and therefore based on estimates (a, b). The figure assumes convex manipulation cost function, with $c = 1$, and a linear prediction model of credit risk, $\widetilde{\alpha} + \widetilde{\beta}x$, where x has uniform distribution on $[x_{\min}, x_{\max}]$, with $x_{\min} = 0$, and $x_{\max} = \frac{1-a}{\overline{b}} - \delta = 1.8$ to ensure a positive probability for all x and all $\widetilde{\beta}$. We set $a = 0.5$, and $\overline{b} = 0.25$ to ensure consistency across the different credit models, that is the survival probabilities (i.e., the interest rates) are similar, for the same x, for the linear and the logit model. The remaining parameters are $r = 0$ and $l = 0.5$.

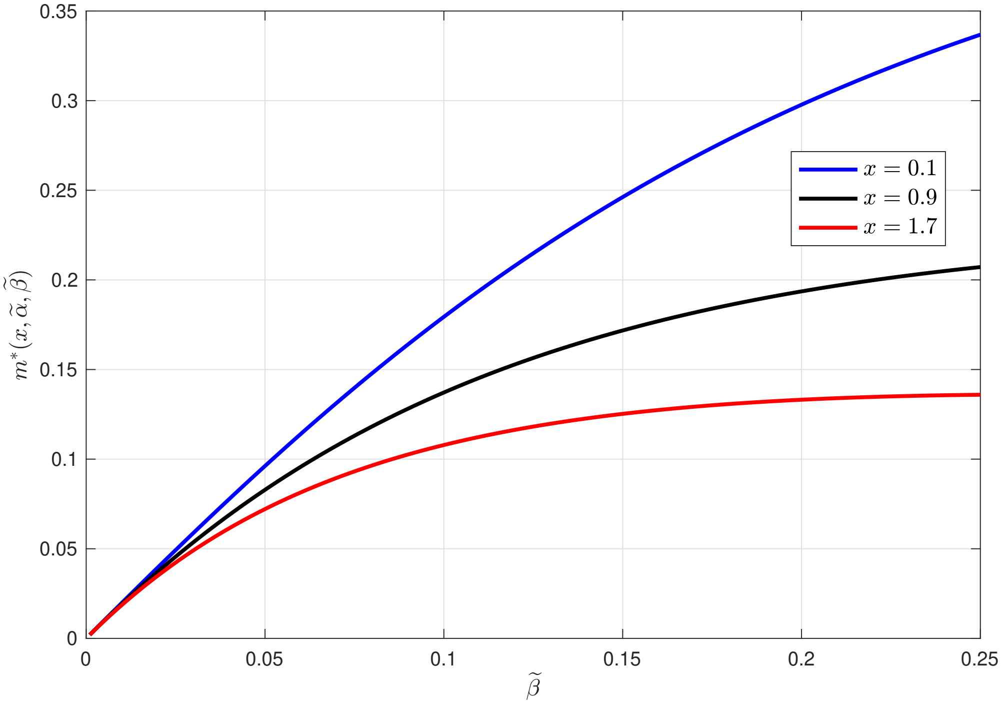

Figure 4: **Incentive compatible manipulation - Linear prediction model - Convex manipulation cost**. We plot optimal manipulation, $m^*(x, \widetilde{\alpha}, \widetilde{\beta})$, for three values of x, against $\widetilde{\beta}$ of the posted credit prediction model, with $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$, and for $b \in]0, \bar{b}]$. The model has convex manipulation cost function, with $c = 1$. The figure is based on a linear prediction model of credit risk, $\widetilde{\alpha} + \widetilde{\beta}x$, where x has uniform distribution on $[x_{\min}, x_{\max}]$, with $x_{\min} = 0$, and $x_{\max} = \frac{1-a}{\bar{b}} - \delta$ to ensure a positive probability for all x and all $\widetilde{\beta}$. We set $a = 0.5$, and $\bar{b} = 0.25$ to ensure consistency across the different credit models, that is the survival probabilities (i.e., the interest rates) are similar, for the same x, for the linear and the logit model. The remaining parameters are $r = 0$ and $l = 0.5$.

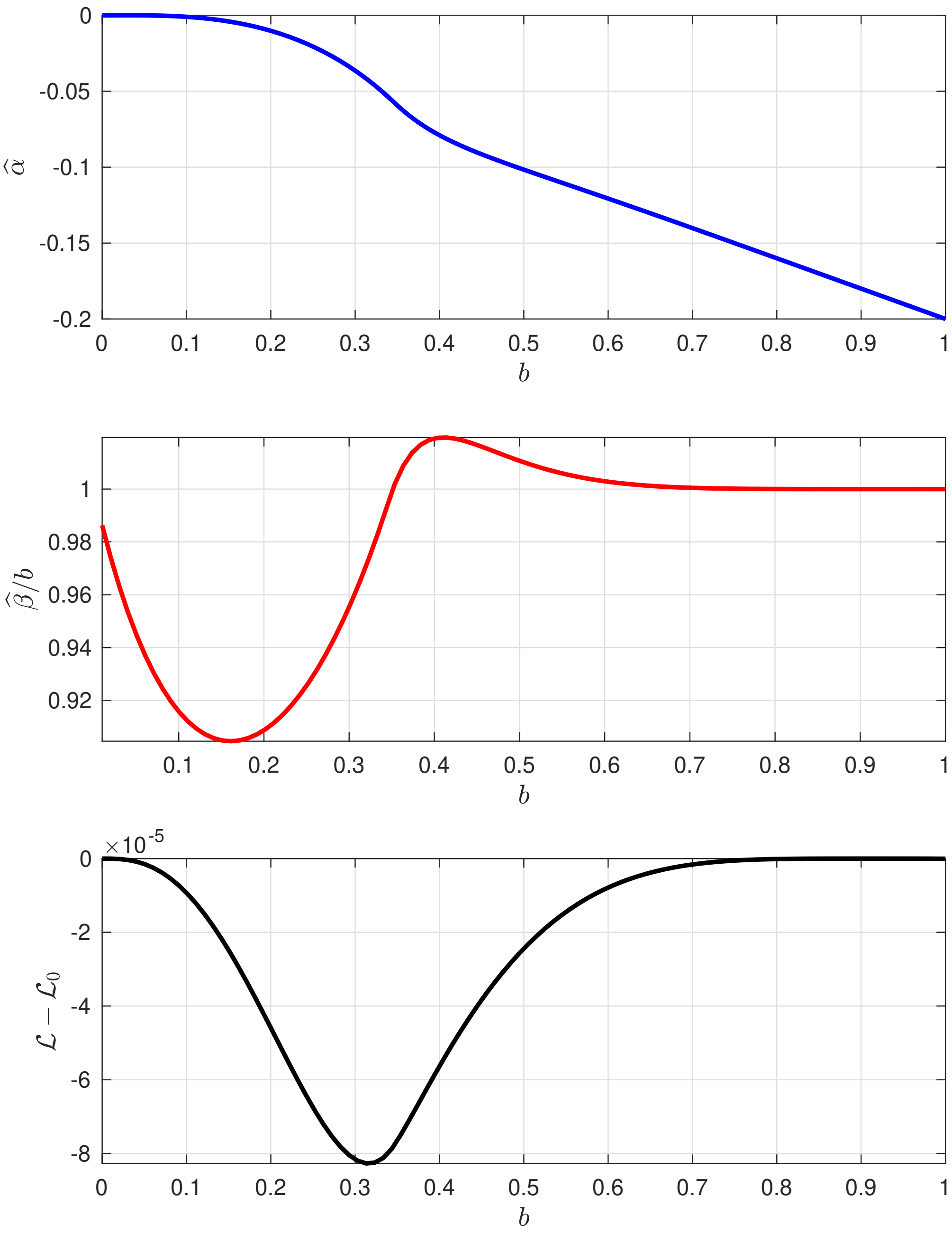

Figure 5: **Goodhart's Law - Logit prediction model - Binary manipulation**. We plot MLE Goodhart estimates against $\widetilde{\beta}$ of the posted credit prediction model, with $(\widetilde{\alpha},\widetilde{\beta}) = (a,b)$, and for $b \in]0,\bar{b}]$. In the first panel we plot $\widehat{\alpha}$ and in the second panel $\widehat{\beta}/b$. In the third panel we plot the difference between the optimal likelihood with manipulation, $\mathcal{L} = \mathcal{L}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta};a,b)$, based on estimates $(\widehat{\alpha},\widehat{\beta})$, and the optimal likelihood with no manipulation, $\mathcal{L}_0$, and therefore based on estimates (a,b). The figure assumes a binary manipulation decision, with $m=\delta=0.2$, probability distribution of manipulation cost, G, uniform in $[0,c_{\max}]$, and a logit prediction model of credit risk, $F(\widetilde{\alpha}+\widetilde{\beta}x)$, with $F(z)=e^z(1+e^z)^{-1}$, where x has uniform distribution on $[x_{\min},x_{\max}]$, with $x_{\min}=0$, and $x_{\max}=1$. The remaining parameters are $r=0$, $l=0.5$, $a=0$, and $\bar{b}=1$. Because $\Psi(m,x_{\max},0,\widetilde{\beta})$ is a concave function of $\widetilde{\beta}$ with $\Psi^* = \max_{\widetilde{\beta}}\Psi(m,x_{\max},0,\widetilde{\beta}) \approx 0.0335$, to achieve that manipulation takes place for all $x \in [x_{\min},x_{\max}]$ for high $\widetilde{\beta}$, we set $c_{\max}=\Psi^*$.

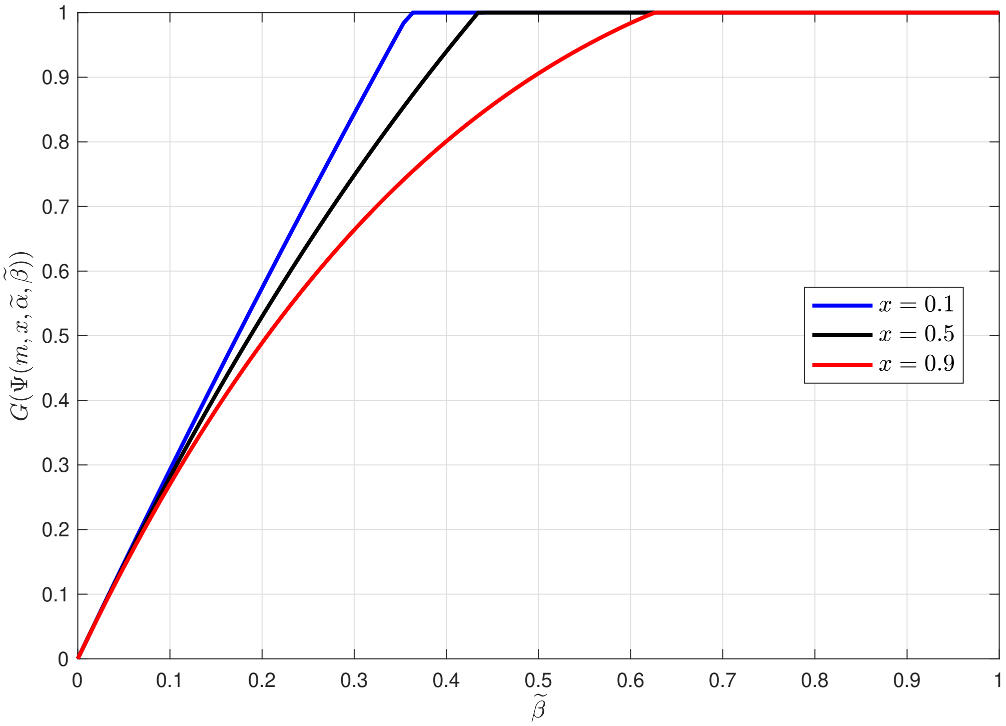

Figure 6: **Incentive compatible manipulation - Logit prediction model - Binary manipulation**. We plot the probability of manipulation, $G(\Psi(m, x, \widetilde{\alpha}, \widetilde{\beta}))$, for three values of x, against $\widetilde{\beta}$ of the posted credit prediction model, with $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$, and for $b \in]0, \bar{b}]$. The figure assumes a binary manipulation decision, with $m = \delta = 0.2$, probability distribution of manipulation cost, G, uniform in $[0, c_{\max}]$, and a logit prediction model of credit risk, $F(\widetilde{\alpha} + \widetilde{\beta}x)$, with $F(z) = e^z(1 + e^z)^{-1}$, where x has uniform distribution on $[x_{\min}, x_{\max}]$, with $x_{\min} = 0$, and $x_{\max} = 1$. The remaining parameters are $r = 0$, $l = 0.5$, $a = 0$, and $\bar{b} = 1$.

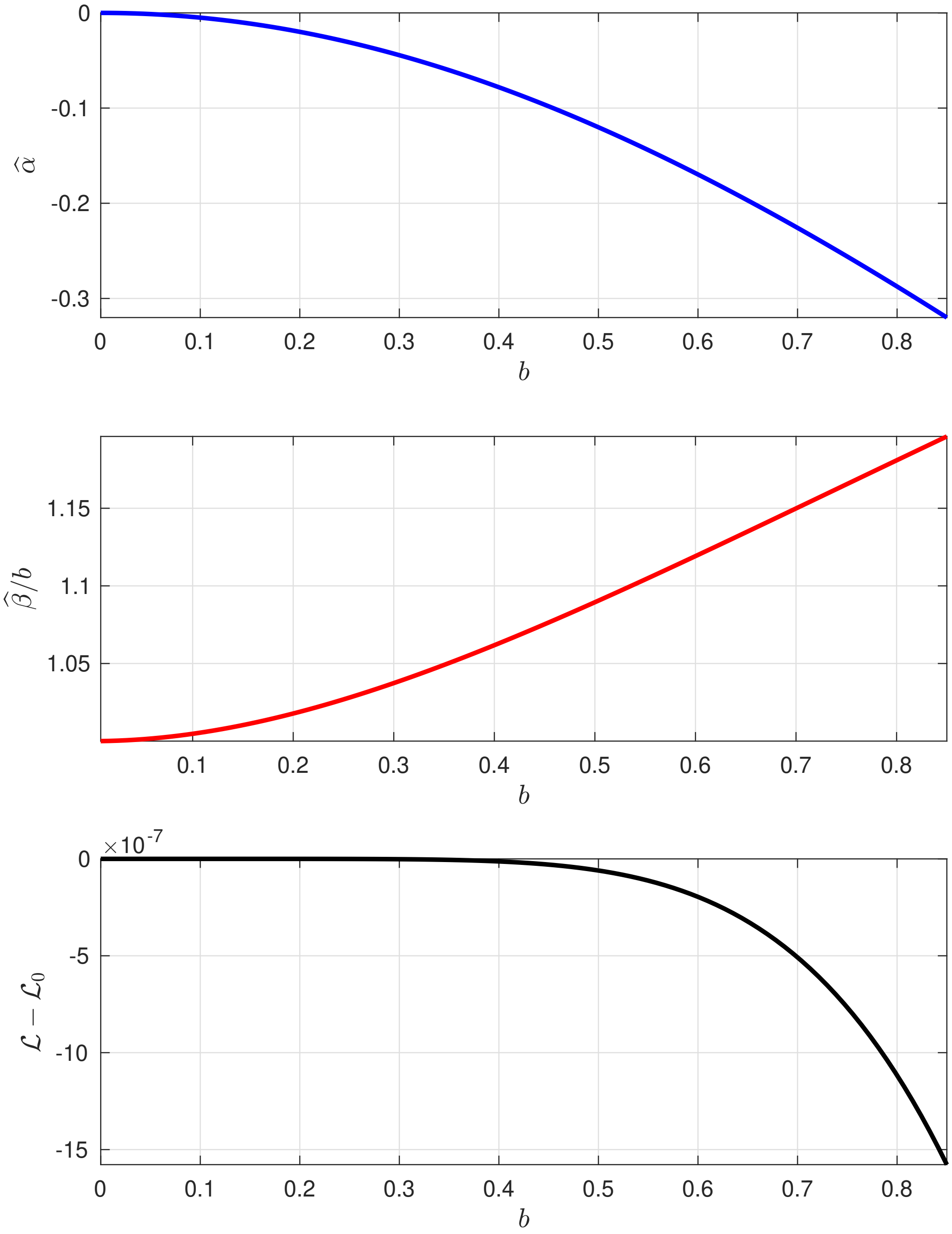

Figure 7: **Goodhart's Law - Logit prediction model - Convex manipulation cost**. We plot MLE Goodhart estimates against $\widetilde{\beta}$ of the posted credit prediction model, with $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$, and for $b \in]0, \overline{b}]$. In the first panel we plot $\widehat{\alpha}$ and in the second panel $\widehat{\beta}/b$. In the third panel we plot the difference between the optimal likelihood with manipulation, $\mathcal{L} = \mathcal{L}(\widehat{\alpha}, \widehat{\beta}; \widetilde{\alpha}, \widetilde{\beta}; a, b)$, based on estimates $(\widehat{\alpha}, \widehat{\beta})$, and the optimal likelihood with no manipulation, $\mathcal{L}_0$, and therefore based on estimates (a, b). The figure assumes a convex manipulation cost function, with $c = 1$, and a logit prediction model of credit risk, $F(\widetilde{\alpha} + \widetilde{\beta}x)$, with $F(z) = e^z(1 + e^z)^{-1}$, where x has uniform distribution on $[x_{\min}, x_{\max}]$, with $x_{\min} = 0$, and $x_{\max} = 1$. The remaining parameters are $r = 0$, $l = 0.5$, $a = 0$, and $\overline{b} = 0.85$.

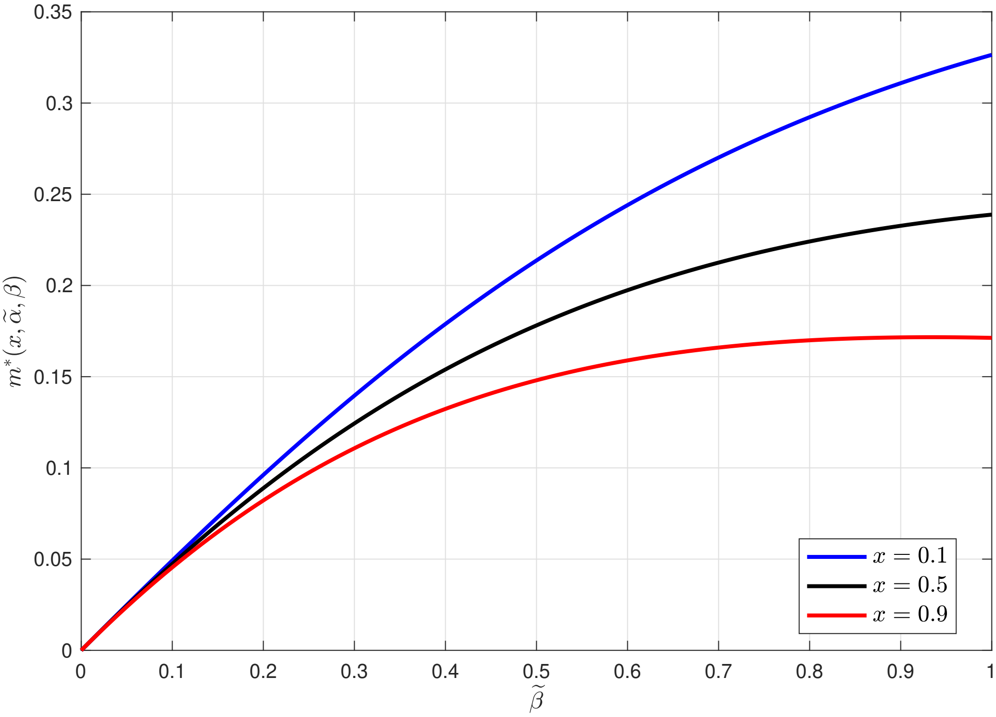

Figure 8: **Incentive compatible manipulation - Logit prediction model - Convex manipulation cost**. We plot optimal manipulation, $m^*(x, \widetilde{\alpha}, \widetilde{\beta})$, for three values of x, against $\widetilde{\beta}$ of the posted credit prediction model, with $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$, and for $b \in]0, \bar{b}]$. The model has convex manipulation cost function, with $c = 1$. The figure is based on a logit prediction model of credit risk, $F(\widetilde{\alpha} + \widetilde{\beta}x)$, with $F(z) = e^z(1 + e^z)^{-1}$, where x has uniform distribution on $[x_{\min}, x_{\max}]$, with $x_{\min} = 0$, and $x_{\max} = 1$. The remaining parameters are $r = 0$, $l = 0.5$, $a = 0$, and $\bar{b} = 1$.

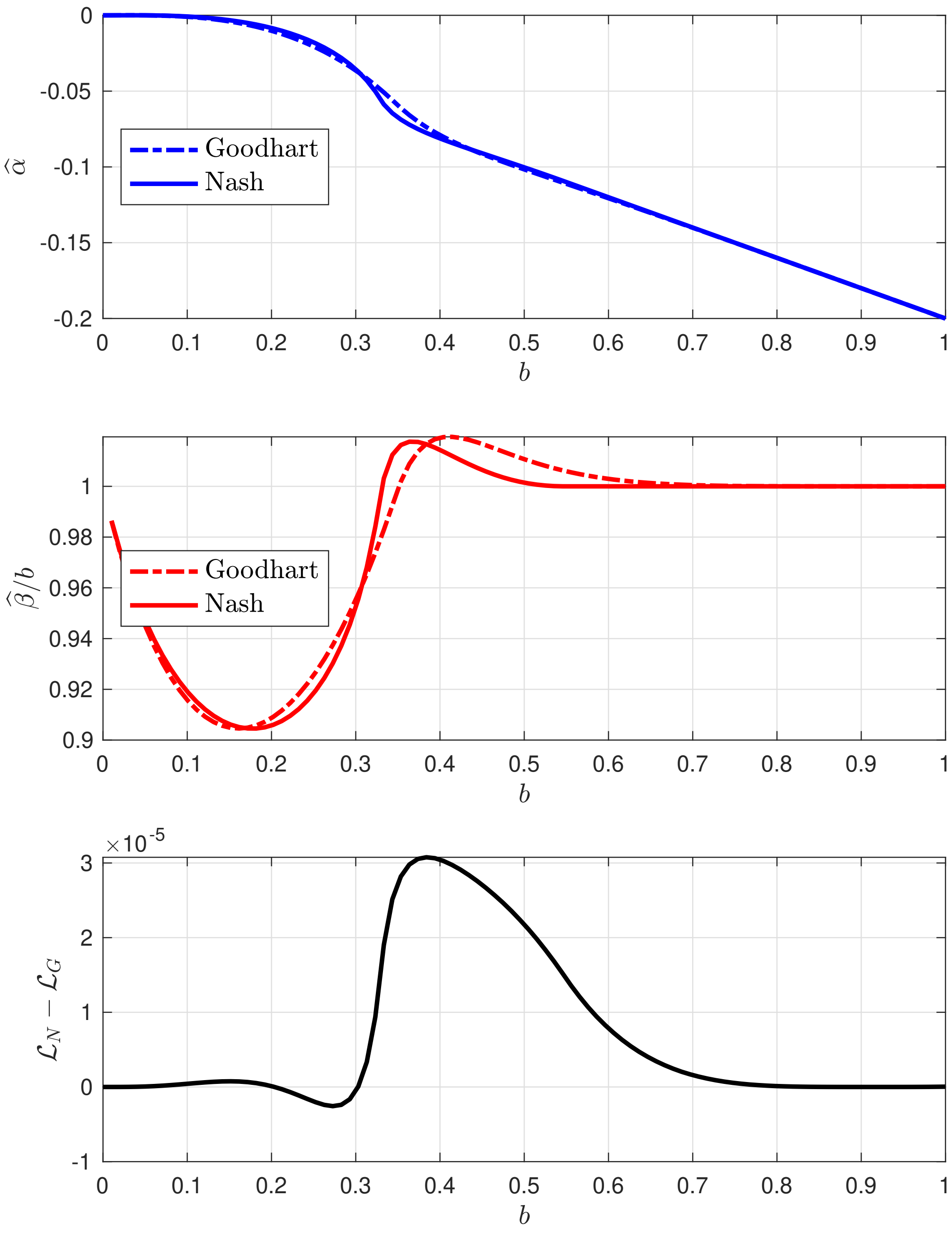

Figure 9: **Nash vs Goodhart - Logit prediction model - Binary manipulation**. We plot MLE estimates against $\widetilde{\beta}$ of the posted credit prediction model, with $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$, and for $b \in]0, \overline{b}]$. In the first panel we plot $\widehat{\alpha}$ and in the second panel $\widehat{\beta}/b$. We consider two alternative programs: the Nash equilibrium, $(\alpha^*, \beta^*) = \arg\max_{(\alpha,\beta)} \mathcal{L}(\alpha, \beta; \alpha^*, \beta^*; a, b)$, and the one used to analyze the effect of Goodhart's Law, $(\widehat{\alpha}, \widehat{\beta}) = \arg\max_{(\alpha,\beta)} \mathcal{L}(\alpha, \beta; a, b; a, b)$. In the third panel we plot the difference between the optimal likelihood in the Nash case, $\mathcal{L}_N$, and the optimal likelihood in the Goodhart's case, $\mathcal{L}_G$. The figure assumes a binary manipulation decision, with $m = \delta = 0.2$, probability distribution of manipulation cost, G, uniform in $[0, c_{\max}]$, and a logit prediction model of credit risk, $F(\widetilde{\alpha} + \widetilde{\beta} x)$, with $F(z) = e^z(1 + e^z)^{-1}$, where x has uniform distribution on $[x_{\min}, x_{\max}]$, with $x_{\min} = 0$, and $x_{\max} = 1$. The remaining parameters are $r = 0$, $l = 0.5$, $a = 0$, and $\overline{b} = 1$. Because $\Psi(m, x_{\max}, 0, \widetilde{\beta})$ is a concave function of $\widetilde{\beta}$ with $\Psi^* = \max_{\widetilde{\beta}} \Psi(m, x_{\max}, 0, \widetilde{\beta}) = 0.0335$, to achieve that manipulation takes place for all $x \in [x_{\min}, x_{\max}]$ for high $\widetilde{\beta}$, we set $c_{\max} = \Psi^*$.

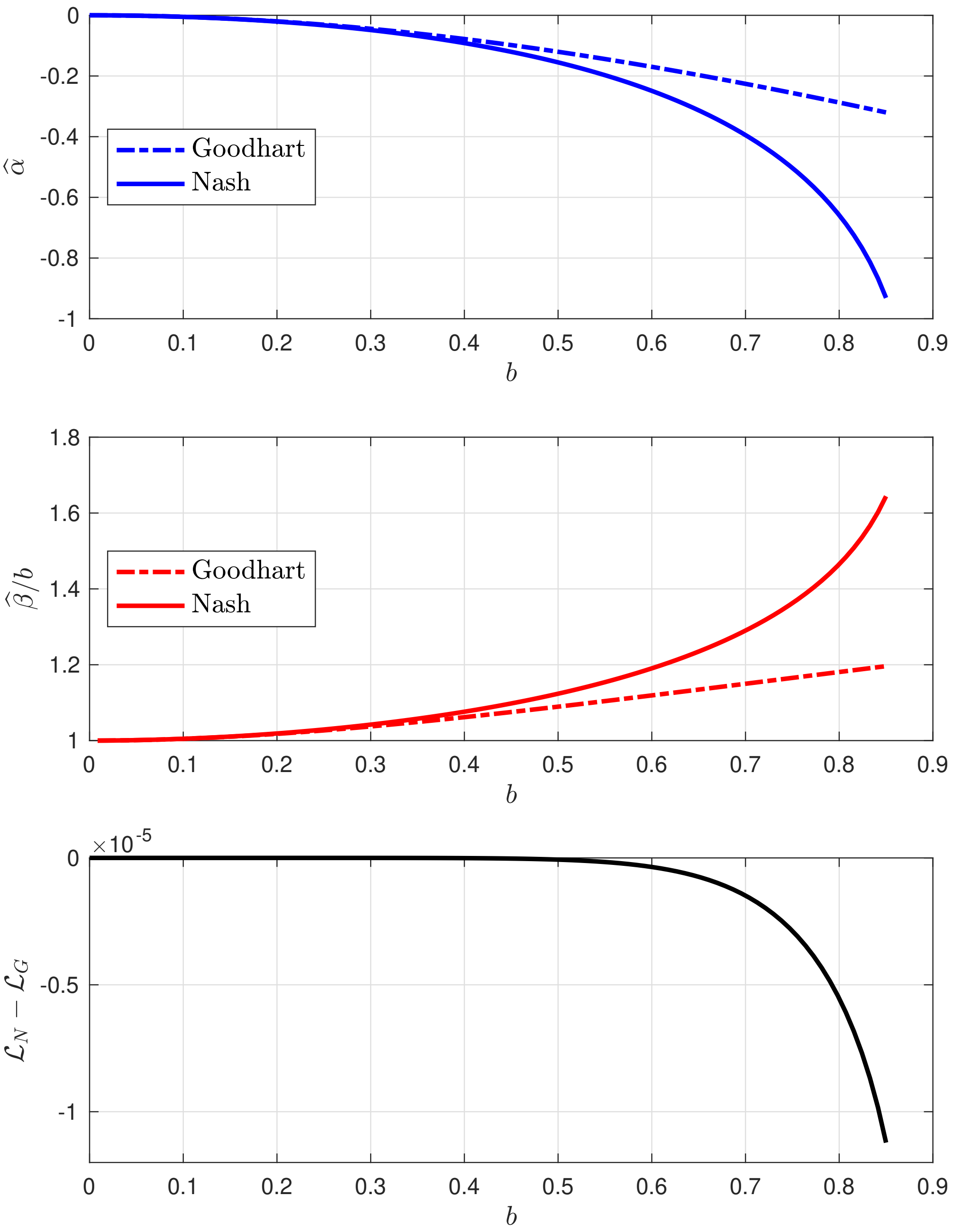

Figure 10: **Nash vs Goodhart - Logit prediction model - Convex manipulation cost**. We plot against $\widetilde{\beta}$ of the posted credit prediction model, with $(\widetilde{\alpha},\widetilde{\beta})=(a,b)$, and for $b\in]0,\overline{b}]$, in the first panel the MLE estimate $\widehat{\alpha}$ and in the second panel $\widehat{\beta}/b$. We consider two alternative programs: the Nash equilibrium, $(\alpha^*,\beta^*)=\arg\max_{(\alpha,\beta)}\mathcal{L}(\alpha,\beta;\alpha^*,\beta^*;a,b)$, and the one used to analyze the effect of Goodhart's Law, $(\widehat{\alpha},\widehat{\beta})=\arg\max_{(\alpha,\beta)}\mathcal{L}(\alpha,\beta;a,b;a,b)$. In the third panel we plot the difference between the optimal likelihood in the Nash case, $\mathcal{L}_N$, and the optimal likelihood in the Goodhart's case, $\mathcal{L}_G$. The figure assumes a convex manipulation cost function, with $c=1$, and a logit prediction model of credit risk, $F(\widetilde{\alpha}+\widetilde{\beta}x)$, with $F(z)=e^z(1+e^z)^{-1}$, where x has uniform distribution on $[x_{\min},x_{\max}]$, with $x_{\min}=0$, and $x_{\max}=1$. The remaining parameters are $r=0$, $l=0.5$, $a=0$, and $\overline{b}=0.85$. Because $\Psi(m,x_{\max},0,\widetilde{\beta})$ is a concave function of $\widetilde{\beta}$ with $\Psi^*=\max_{\widetilde{\beta}}\Psi(m,x_{\max},0,\widetilde{\beta})=0.0335$, to achieve that manipulation takes place for all $x\in[x_{\min},x_{\max}]$ for high $\widetilde{\beta}$, we set $c_{\max}=\Psi^*$.

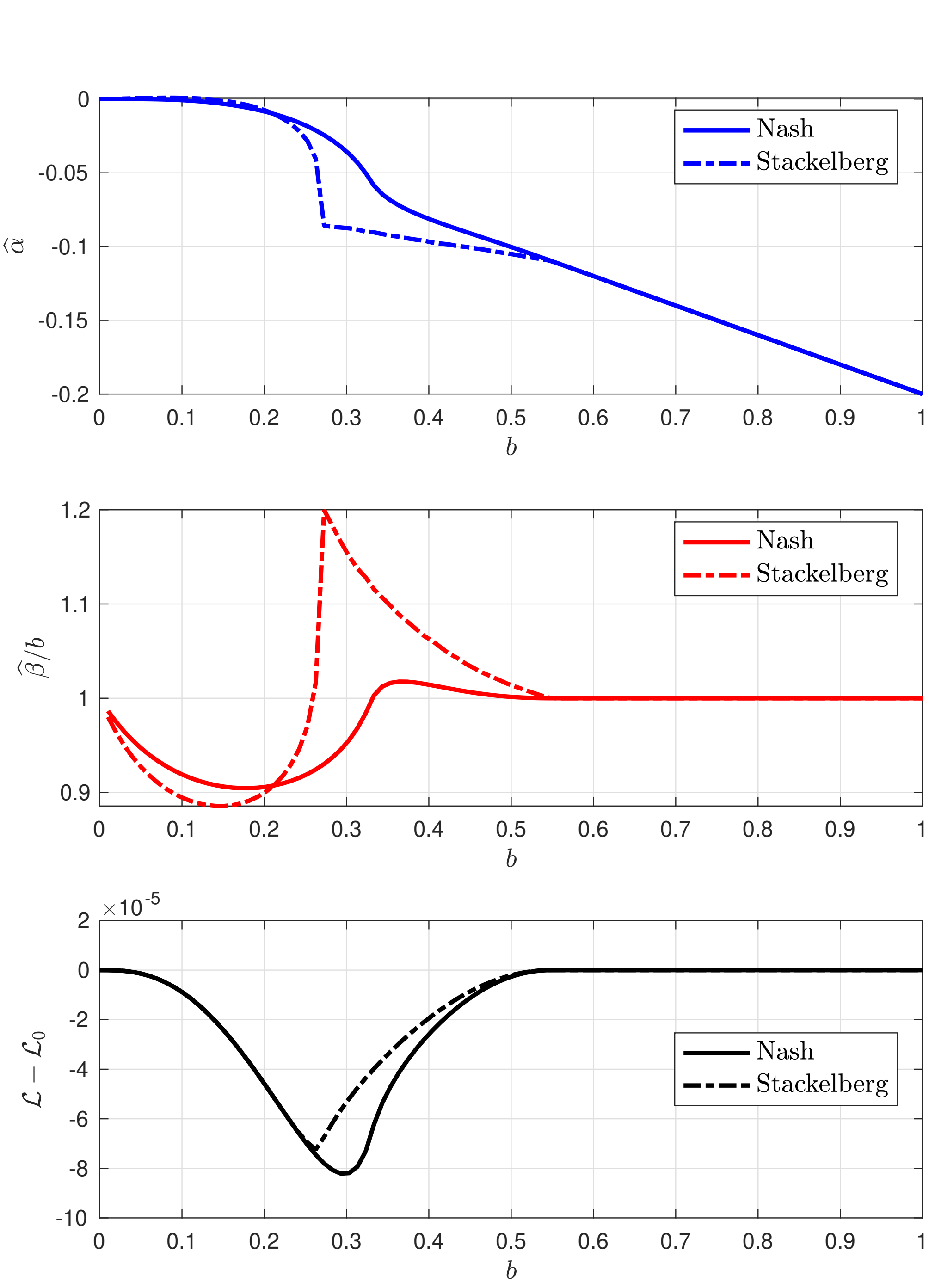

Figure 11: **Nash and Stackelberg equilibria - Logit prediction model - Binary manipulation**. We plot against $\widetilde{\beta}$ of the posted credit prediction model, with $(\widetilde{\alpha},\widetilde{\beta})=(a,b)$, and for $b\in]0,\overline{b}]$, in the first panel the MLE estimate $\widehat{\alpha}$ and in the second panel $\widehat{\beta}/b$. In the third panel we plot the difference between the optimal likelihood with manipulation, $\mathcal{L}=\mathcal{L}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta};a,b)$, based on estimates $(\widehat{\alpha},\widehat{\beta})$, and the optimal likelihood with no manipulation, $\mathcal{L}_0$, and therefore based on estimates (a,b). We consider two alternative programs: the Nash equilibrium, $(\alpha^*,\beta^*)=\arg\max_{(\alpha,\beta)}\mathcal{L}(\alpha,\beta;\alpha^*,\beta^*;a,b)$, and the Stackelberg equilibrium, $(\alpha^{**},\beta^{**})=\arg\max_{(\alpha,\beta)}\mathcal{L}(\alpha,\beta;\alpha,\beta;a,b)$. The figure assumes a binary manipulation decision, with $m=\delta=0.2$, probability distribution of manipulation cost, G, uniform in $[0,c_{\max}]$, and a logit prediction model of credit risk, $F(\widetilde{\alpha}+\widetilde{\beta}x)$, with $F(z)=e^z(1+e^z)^{-1}$, where x has uniform distribution on $[x_{\min},x_{\max}]$, with $x_{\min}=0$, and $x_{\max}=1$. The remaining parameters are $r=0$, $l=0.5$, $a=0$, and $\overline{b}=1$. Because $\Psi(m,x_{\max},0,\widetilde{\beta})$ is a concave function of $\widetilde{\beta}$ with $\Psi^*=\max_{\widetilde{\beta}}\Psi(m,x_{\max},0,\widetilde{\beta})=0.0335$, to achieve that manipulation takes place for all $x\in[x_{\min},x_{\max}]$ for high $\widetilde{\beta}$, we set $c_{\max}=\Psi^*$.

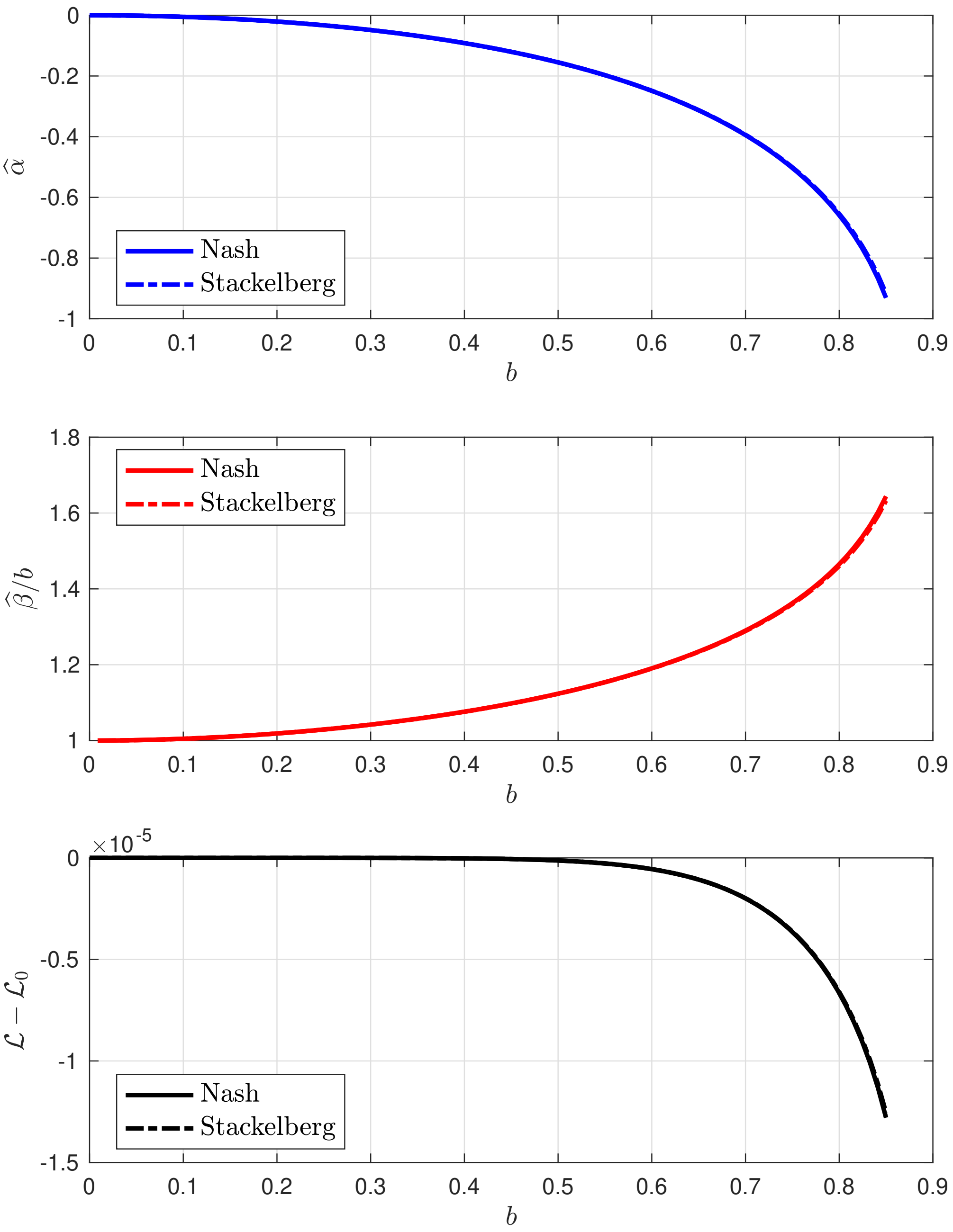

Figure 12: **Nash and Stackelberg equilibria - Logit prediction model - Convex manipulation cost**. We plot against $\widetilde{\beta}$ of the posted credit prediction model, with $(\widetilde{\alpha},\widetilde{\beta})=(a,b)$, and for $b\in]0,\bar{b}]$, in the first panel the MLE estimate $\widehat{\alpha}$ and in the second panel $\widehat{\beta}/b$. In the third panel we plot the difference between the optimal likelihood with manipulation, $\mathcal{L}=\mathcal{L}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta};a,b)$, based on estimates $(\widehat{\alpha},\widehat{\beta})$, and the optimal likelihood with no manipulation, $\mathcal{L}_0$, and therefore based on estimates (a,b). We consider two alternative programs: the Nash equilibrium, $(\alpha^*,\beta^*)=\arg\max_{(\alpha,\beta)}\mathcal{L}(\alpha,\beta;\alpha^*,\beta^*;a,b)$, and the Stackelberg equilibrium, $(\alpha^{**},\beta^{**})=\arg\max_{(\alpha,\beta)}\mathcal{L}(\alpha,\beta;\alpha,\beta;a,b)$. The figure assumes a convex manipulation cost function, with $c=1$, and a logit prediction model of credit risk, $F(\widetilde{\alpha}+\widetilde{\beta}x)$, with $F(z)=e^z(1+e^z)^{-1}$, where x has uniform distribution on $[x_{\min},x_{\max}]$, with $x_{\min}=0$, and $x_{\max}=1$. The remaining parameters are $r=0$, $l=0.5$, $a=0$, and $\bar{b}=0.85$.