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

Andrea Gamba
WBS-University of Warwick

Christopher A. Hennessy
LBS, CEPR, ECGI

Abstract

We analyse default risk coefficient estimation when borrowers can inflate, at quadratic cost, a covariate used in credit scoring, with potential heterogeneity in latent sensitivities to interest rates. A qualified version of Goodhart's law obtains: When the posted model utilizes coefficients from clean historical data, coefficients shift subsequently, provided the coefficient on the true covariate is not zero. As shown, measurement error resulting from manipulation is negatively correlated with the true covariate. This correlation shifts the slope coefficient upward unless noise resulting from heterogeneous interest-rate sensitivities is sufficiently high. We next evaluate internally consistent fixed point (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. Practically, adaptive estimation converges to a fixed point if manipulation costs are sufficiently high. Finally, an econometrician with commitment power optimally discourages manipulation with marginal increases (decreases) in the posted model intercept (slope).

1 Introduction

An extant literature, beginning with Altman (1968), attempts to use historical statistical relationships to estimate default probabilities. Problematically, the use of econometric models to estimate default probabilities brings them into direct conflict with *Goodhart's law*, which states that:[1]

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

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

*We thank Charles Goodhart, discussants Anton Tsoy and Adam Winegar, and other participants at the Aarhus Workshop on Strategic Interaction in Corporate Finance 2023, ASU Sonora Winter Conference 2024, and Lake District CF Workshop 2024 for their comments and suggestions.

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

traceable (transparent), clearly explained and palatable (understandable and acceptable) to lenders, regulators and consumers." Borrowers facing such models have an obvious incentive to game them. Indeed, echoing Goodhart (1975), Mark Zandi, chief economist at Moody's Analytics, has expressed 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." Supporting this concern, Liu, et al. (2010) and Caton, et al. (2011) document that corporations engage in earnings management prior to bond flotations.

Although the issue of gaming of credit risk scores is oft-noted, it is not clear how to model it formally, to say nothing of the 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 canonical econometric models of credit risk. How do we incorporate data manipulation into a formal statement of the default prediction problem? What kind of coefficient shift should we expect to see? What tools are available for addressing Goodhart's law here? Can one expect econometric tâtonnement to equilibrium, with coefficient estimates being stable within manipulated data?

The setting considered is as follows. Each borrower's latent repayment probability is $F(a+bx)$, with $x \geq 0$, $b \geq 0$, and F falling into one of three standard functional forms: uniform, logistic, or normal. In the spirit of Goodhart (1975), the econometrician has access to clean data on outcomes y and covariates x collected from a cohort of borrowers who did not face a default prediction model.[2] The clean data allows the econometrician to correctly estimate the intercept and slope coefficients (a, b).

The main task of the econometrician is to post a default prediction model with intercept and slope parameters $(\widetilde{\alpha}, \widetilde{\beta})$. Borrowers in a future cohort report their manipulated covariate $\widetilde{x} \geq x$, understanding that their repayment probability will be computed according to $F(\widetilde{\alpha} + \widetilde{\beta}\widetilde{x})$. Borrowers incur quadratic costs to manipulate their covariate upward, with heterogeneous cost parameters c associated with heterogeneity in borrower interest-rate sensitivities. Importantly, the cost parameters c are assumed to be independent from the true covariate x. By construction, this independence assumption rules out mechanical correlation between manipulation and the true covariate.

In order to understand how Goodhart's law would manifest itself in default prediction, we first consider a naive econometrician who posts coefficients $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$. That is, the naive econometrician uses the clean data parameters (a, b), and computes repayment probabilities for the future cohort according to $F(a+b\widetilde{x})$, despite $\widetilde{x} \geq x$. Ex post, the econometrician tests for coefficient instability by estimating new coefficients $(\widehat{\alpha}, \widehat{\beta})$ using the manipulated data $(y, \widetilde{x})$ collected from the cohort of strategic borrowers. Consistent with Goodhart's law, estimated coefficients are indeed unstable across the historical and strategic cohort, with $(\widehat{\alpha}, \widehat{\beta}) \neq (a, b)$ unless $b = 0$.

[2] Access to clean data is only assumed in our Goodhart-type thought experiments.

A better understanding of coefficient shift is gained by noting that covariate manipulation represents an endogenous source of measurement error differing in kind from classical white noise error. In particular, conditional upon c, manipulation is negatively correlated with the true covariate x. As we show, this implies that, absent additional noise generated by heterogeneity in manipulation costs (or interest-rate sensitivity), the coefficient on the manipulated covariate will *exceed* the original coefficient on the unmanipulated covariate. That is, here we have a case in which manipulation of a covariate actually serves to increase the apparent sensitivity of outcomes to that covariate.

A simple argument conveys the intuition. As we show, the incentive to manipulate decreases with a borrower's baseline credit score, since the interest rate is decreasing and convex in the repayment probability. Therefore, if one were to consider, say, the best linear fit between two data points x and $x+\Delta x$, the mean change in the numerator (the repayment probability) would be $b\Delta x$. But in manipulated data, the change in the denominator is $\Delta\widetilde{x} < \Delta x$ since manipulation decreases with x. Thus, the slope increases to $b\Delta x/\Delta\widetilde{x} > b$.

Figure 1 provides illustrative econometric output, with manipulated covariates $\widetilde{x}$ on the horizontal axis and true repayment probabilities on the vertical axis.[3] The top panel depicts a setting with substantial cross-sectional heterogeneity in manipulation costs. Such cost heterogeneity scatters observations away from the line of best fit, with the estimated slope falling below the clean data slope b. The bottom panel captures a setting with lower cross-sectional variation in manipulation costs. In this case, observations fall closer to the estimated line of best fit, with the slope coefficient in manipulated data actually shifting above the clean data coefficient b. That is, here we have a case in which manipulation of a covariate serves to increase the estimated sensitivity of outcomes to that covariate. As we show, this type of upward shift in estimated slope tends to occur in settings with low variation in manipulation costs.

After characterizing coefficient shift, we examine alternative econometric responses. We consider first fixed point (Nash) coefficients (α^*, β^*) which are optimal given the induced covariates $\widetilde{x}$, and vice-versa. Notice, fixed points will be immune to the sort of parameter instability that troubled Goodhart. Moreover, fixed points satisfy potential institutional demands for model coefficients to be justifiable by the data. As we show, historical clean data coefficients (a, b) are fixed points if and only if $b = 0$. If $b > 0$, a fixed point model must feature a lower intercept and/or slope than (a, b), and the slope $\beta^* \neq 0$.

We next consider fixed point convergence. Practically, we show that so long as estimated coefficients $(\widehat{\alpha}, \widehat{\beta})$ are not too sensitive to posted model coefficients $(\widetilde{\alpha}, \widetilde{\beta})$, econometricians will converge to a fixed point if they simply iterate, utilizing the prior period's coefficient estimates as the next period's posted model. In turn, this low sensitivity scenario occurs if manipulation costs are sufficiently high and/or borrowers have sufficiently high latent quality (x).

[3] For ease of inspection, we replace (0,1) values of y with mean y.

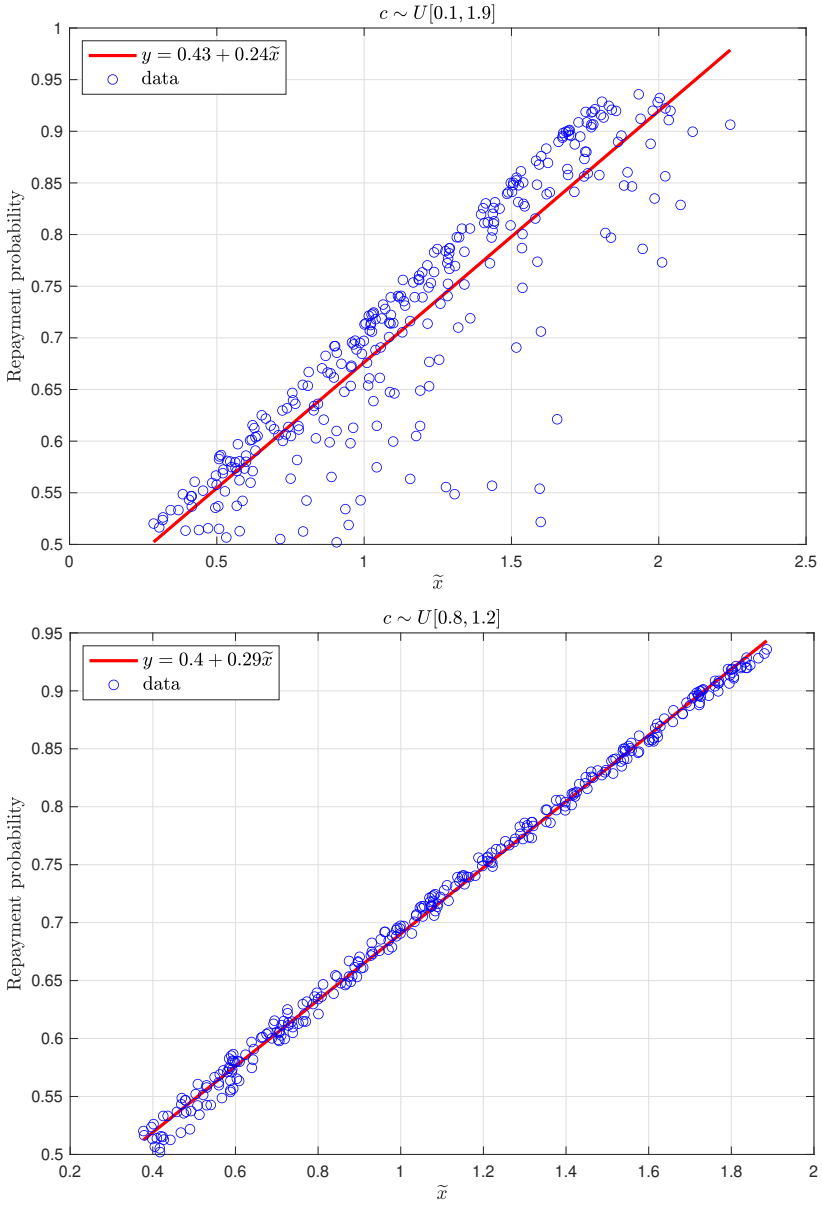

Figure 1: **Goodhart's Law**. OLS estimation based on manipulated data $\widetilde{x}$, with posted model coefficients $(a, b) = (0.5, 0.25)$. The figure assumes uniformly distributed cost $c \sim U[c_d, c_u]$, and uniformly distributed on $x \sim U[x_{\min}, x_{\max}]$, with $x_{\min} = 0$, and $x_{\max} = \max\{\frac{1-a}{b} - \delta, 1\}$ to ensure repayment probability in $[0, 1]$ for all x, where δ equals the maximum manipulation for average x. The variance of manipulation cost is high in the upper panel and low in the lower panel. The remaining parameters are $r = 0$ and $l = 0.5$.

Finally, we analyze optimal coefficients $(\alpha^{**}, \beta^{**})$ with full commitment. In contrast to fixed point (Nash) models, where the econometrician treats the distribution of covariates as given, a Stackelberg parameterization accounts for the effect of the posted model on borrower incentives. Under commitment, the econometrician discourages borrower manipulation with marginal upward (downward) adjustments in the posted model intercept (slope). This leads to an increase in predictive power relative to fixed points. However, the increase in predictive power comes at the cost of the model being difficult to rationalize since Stackelberg coefficients are inconsistent with the data they generate, with $(\alpha^{**}, \beta^{**}) \neq (\widehat{\alpha}, \widehat{\beta})$.

Our paper draws much inspiration from recent work of Frankel and Kartik (2022), and to a lesser extent Hennessy and Goodhart (2023).[4] Frankel and Kartik consider a more abstract setting, linear prediction with idiosyncratic manipulation gains modelled as random variables potentially correlated with the latent covariate. They show a Stackelberg principal finds it optimal to make allocations less sensitive to the covariate than in Nash equilibrium. In contrast, we consider a specific non-linear setting, default prediction via MLE, generating novel predictions regarding incentives for data manipulation and resulting econometric outcomes. For example, in the setting we consider, there is endogenous negative correlation between manipulation and the latent covariate which leads to slope coefficient overshooting absent noise from cross-sectional heterogeneity in manipulation costs. In addition, in the setting we consider, incentives for manipulation are shaped by the posted model intercept, not just the posted slope, complicating the search for fixed points and optimal commitment models. We also show that outcomes vary in a non-linear way with the true causal parameter b, while Frankel and Kartik consider the special case of $b = 1$. Finally, we consider the conditions under which adaptive estimation will converge to a fixed point.

Another closely related paper is that of Rajan, Seru and Vig (2010). They demonstrate a complementary variation of 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 our analysis in an explicit econometric framework, inspired by Frankel and Kartik (2022) and Hennessy and Goodhart (2023).

Eliaz and Spiegler (2019) consider an economy in which each agent is part of the training data. In their setting, 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 likely to be acute under Lasso estimation with sparsity: An agent may have an incentive to misreport given that the 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

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

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 induce truthful reporting in 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 considers fixed points and convergence. Section 6 considers optimal models under commitment. Section 7 extends some key results to a multivariate setup.

2 Data Manipulation Incentives

This section begins with a description of the assumed institutional setting which features strategic interaction between an econometrician and borrowers. We then consider the incentive of borrowers to manipulate their covariate report. As shown, 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. The loan amount is 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 borrower recovers zero and 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 representing an independent draw from an atomless cumulative distribution function H with probability density h. The variance of x is $\sigma_x^2 > 0$. The conditional expectation of y given x, or the probability of debt repayment conditional upon x, is:

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

where $b \geq 0$. The argument, call it z, to which the c.d.f. F is applied is labelled the *credit score.* Each borrower's true latent credit score is $z = a + bx$.

Equation (1) can be motivated by canonical structural models of default. For example, in the log-normal economy considered by Merton (1974), the probability of repayment by a borrower holding an asset with value V facing a zero coupon debt face value D due at time τ is given by:

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

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.[5] The question pondered by Goodhart is the extent to which statistical regularities gleaned from clean 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 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 each loan must satisfy:

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

For future borrowers, the lender only observes an endogenously manipulated covariate $\widetilde{x} \geq x$. A borrower's cost of manipulation is $cm^2/2$, with $m \equiv \widetilde{x} - x$ and $c > 0$ being a private independent draw from cumulative distribution G with density g. The variance of c is $\sigma_c^2 \geq 0$. Manipulation costs are incurred at the end of the period, at the same time debt comes due. For example, one may think of manipulation costs as capturing pecuniary penalties assessed once manipulation is eventually detected. Alternatively, one may think of these costs as representing loss of going-concern value, or ongoing costs associated with papering over the prior period's accounting misstatement.

In summary, each prospective borrower has latent two-dimensional type (x, c), with type drawn from joint distribution $H \times G$. The maintained assumption that $x \perp c$ precludes mechanistic correlation between manipulation and the latent covariate. Any correlation between manipulation and the true covariate is endogenous, arising from the implicit incentives generated by model-based loan pricing.

[5] Alternatively, the first posted model can feature zero loading on the covariate, eliminating manipulation in that cohort.

The econometrician's *default prediction model* (DPM) states how reported covariates ($\widetilde{x}$) will be mapped to an assessed repayment probability for future borrowers. In particular, repayment probabilities will be computed according to

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

The task of the econometrician is to specify the coefficients $(\widetilde{\alpha}, \widetilde{\beta})$ of the posted econometric model, perhaps relying on the estimates derived from the clean historical data (y, x). A borrower responds to the posted DPM by reporting a covariate $\widetilde{x}$.[6] Since $b \geq 0$, attention is confined to posted models that load positively on the covariate ($\widetilde{\beta} \geq 0$).

Substituting equation (4) into equation (3), the model-implied interest rate is:

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

Recall, the borrower has limited liability and receives a payoff of zero if $y = 0$, as would be the case in the event of a corporate default. Therefore, each borrower will minimize costs they will incur in the event of solvency ($y = 1$): debt service plus manipulation costs.[7] Optimal manipulation m^* is pinned down by:

$$m^*(x, \widetilde{\alpha}, \widetilde{\beta}, c) \in \arg\min_m \ \frac{1}{2}cm^2 + \iota(m, x, \widetilde{\alpha}, \widetilde{\beta}). \tag{6}$$

Before closing this subsection, it is useful to discuss empirically relevant settings that give rise to the same program as (6), and identical econometric outcomes. To this end, assume manipulation costs are non-pecuniary costs of the form $\xi m^2/2$, with ξ being a privately observed random variable. Assume also that borrowers have heterogenous shadow values for end-of-period funds $\lambda \geq 1$, with λ also being a privately observed random variable. Finally, assume the econometrician only knows the distribution G of the random variable $c \equiv \xi/\lambda$. Then we are in an isomorphic environment where a borrower minimizes:

$$\frac{1}{2}\xi m^2 + \lambda\iota(m, x, \widetilde{\alpha}, \widetilde{\beta}) = \lambda\left[\frac{1}{2}cm^2 + \iota(m, x, \widetilde{\alpha}, \widetilde{\beta})\right]. \tag{7}$$

Alternatively, the preceding program also arises if, say, credit card applicants can achieve total clandestine borrowing λ using infinitesimal loans from atomistic banks. In either environment, borrowers with high λ, and low $c \equiv \xi/\lambda$, are more interest rate sensitive in the sense of placing greater weight on getting a good interest rate ι.

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

[7] Recall, manipulation costs are end-of-period costs.

With the preceding discussion in mind, we can return back to our baseline technology, noting that variance in c can be understood as proxying for heterogeneity in interest-rate sensitivity arising from differences in shadow values of internal funds or differences in total indebtedness.

2.2 Incentive for Data Manipulation

For brevity, let

$$\Omega(z) \equiv [F(z)]^{-1} \Rightarrow \iota(m, x, \widetilde{\alpha}, \widetilde{\beta}) = l + (1 + r - l)\Omega[\widetilde{\alpha} + \widetilde{\beta}(x + m)] - 1. \tag{8}$$

The following lemma, relegated to the appendix, establishes some useful properties of Ω for the three standard classes of default prediction models that we consider: 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 $(0, 1)$.*

We have the following first-order condition (FOC below) pinning down optimal manipulation:

$$cm^* + \iota_m(m^*, x, \widetilde{\alpha}, \widetilde{\beta}) = 0 \Rightarrow cm^* + (1 + r - l)\widetilde{\beta}\Omega'[\widetilde{\alpha} + \widetilde{\beta}(x + m^*)] = 0. \tag{9}$$

Intuitively, borrowers equate marginal manipulation costs with the marginal reduction in interest rate that results from upward manipulation.

Importantly, the marginal gain to manipulation depends upon the specific properties of the distribution F utilized by the econometrician, as well as the credit score z at which it is evaluated. To see this, note that the marginal rate reduction generated by upward manipulation can be decomposed into three parts as follows:

$$\begin{aligned} \frac{\partial \iota}{\partial m} &= \frac{\partial z}{\partial m} \times \frac{\partial \widehat{F}}{\partial z} \times \frac{\partial \iota}{\partial \widehat{F}} \\ &= \widetilde{\beta} \times f(z) \times \left(-\frac{1 + r - l}{[F(z)]^2} \right) \leq 0. \end{aligned} \tag{10}$$

That is, the marginal rate reduction is the product of: $\widetilde{\beta}$ capturing the effect of manipulation m on the credit score z; $f(z)$ capturing the effect of z on imputed repayment probability $\widehat{F}$; and the effect of $\widehat{F}$ on the interest rate.

As discussed below, for standard distributions F, the dominant force in equation (10) is the final term which informs us that the interest rate is most sensitive to the imputed repayment probability at low credit scores. The middle term in the marginal gain equation, $F' = f$, captures

a subtle secondary force: For logit and probit models, manipulation gains are attenuated in the tails of the distribution, where the imputed repayment probability has low sensitivity to the credit score z. The more general point is that, while the choice of F is often based upon tractability, or dictated by assumptions about underlying real technologies, as in the log-normal economy of Merton (1974), the choice of distribution has subtle incentive implications, magnifying or attenuating local manipulation incentives.

For standard distributions F (Lemma 1), the marginal manipulation gain is decreasing in the true covariate, with:

$$\frac{\partial}{\partial x}\frac{\partial \iota}{\partial m} = (1+r-l)\widetilde{\beta}^2 \underbrace{[F(z)]^{-2}\left[\frac{2[f(z)]^2}{F(z)} - F''(z)\right]}_{\equiv \Omega''(z)} (\geq 0)\,. \tag{11}$$

Here too, a decomposition reveals the presence of potentially opposing forces. Applying the chain rule we have:

$$\begin{aligned} \frac{\partial}{\partial x}\frac{\partial \iota}{\partial m} &= \frac{\partial}{\partial x}\left(\frac{\partial \iota}{\partial \widehat{F}}\frac{\partial \widehat{F}}{\partial m}\right) \\ &= \frac{\partial \widehat{F}}{\partial m}\frac{\partial^2 \iota}{\partial x \partial \widehat{F}} + \frac{\partial \iota}{\partial \widehat{F}}\frac{\partial^2 \widehat{F}}{\partial x \partial m} \\ &= \widetilde{\beta}^2 f^2 \underbrace{\frac{\partial^2 \iota}{\partial \widehat{F}^2}}_{>0} + \widetilde{\beta}^2 \underbrace{\frac{\partial \iota}{\partial \widehat{F}}}_{<0} F''. \end{aligned} \tag{12}$$

Notice, the sign of the second term is negative if F is locally convex. Intuitively, as x increases, the imputed repayment probability becomes more sensitive to manipulation m if f is increasing. It follows that if f is increasing, as is the case in the left tail of the distribution for logit and probit, there is one channel causing marginal manipulation gains to increase with x. Nevertheless, equation (11) informs us that for linear, logit and probit models, the dominant force is convexity of the interest rate in $\widehat{F}$ which ensures diminishing marginal manipulation gains ($\iota_{xm} \geq 0$) even where F is locally convex.

For convex Ω (Lemma 1), the second-order condition (SOC below) for a local minimum is necessarily satisfied, with:

$$c + \iota_{mm}(m^*, x, \widetilde{\alpha}, \widetilde{\beta}) = c + (1+r-l)\widetilde{\beta}^2 \Omega''[\widetilde{\alpha} + \widetilde{\beta}(x+m)] > 0. \tag{13}$$

Applying the implicit function theorem to the borrower's FOC, we obtain the following comparative statics:

$$\begin{aligned}
\frac{\partial m^*}{\partial x} &= -\frac{(1+r-l)\widetilde{\beta}^2\Omega''[\widetilde{\alpha}+\widetilde{\beta}(x+m^*)]}{c+\iota_{mm}(m^*,x,\widetilde{\alpha},\widetilde{\beta})} \\
\frac{\partial m^*}{\partial \widetilde{\alpha}} &= -\frac{(1+r-l)\widetilde{\beta}\Omega''[\widetilde{\alpha}+\widetilde{\beta}(x+m^*)]}{c+\iota_{mm}(m^*,x,\widetilde{\alpha},\widetilde{\beta})} \\
\frac{\partial m^*}{\partial \widetilde{\beta}} &= -\frac{(1+r-l)\left[\widetilde{\beta}(x+m^*)\Omega''(\widetilde{\alpha}+\widetilde{\beta}(x+m^*))+\Omega'(\widetilde{\alpha}+\widetilde{\beta}(x+m^*))\right]}{c+\iota_{mm}(m^*,x,\widetilde{\alpha},\widetilde{\beta})} \\
\frac{\partial m^*}{\partial c} &= -\frac{m^*}{c+\iota_{mm}(m^*,x,\widetilde{\alpha},\widetilde{\beta})}.
\end{aligned} \tag{14}$$

If Ω is decreasing and convex, the first two comparative statics immediately above are negative: Manipulation is decreasing in both $\widetilde{\alpha}$ and x. Intuitively, the incentive to manipulate decreases with a borrower's *baseline credit score* $z = \widetilde{\alpha}+\widetilde{\beta}x$. Again, 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 higher initial $\widehat{F}$, an incremental increase in $\widehat{F}$ through data manipulation has a smaller effect on the interest rate.

An increase in the posted slope coefficient $\widetilde{\beta}$ generates competing effects. On one hand, with higher $\widetilde{\beta}$, each manipulation increment has a larger effect on the credit score z, since $\partial z/\partial m = \widetilde{\beta}$. This substitution effect stimulates data manipulation. However, starting at a given $x > 0$, an increase in $\widetilde{\beta}$ raises the baseline credit score. This income effect makes the interest rate (equation (5)) less sensitive to increases in the imputed repayment probability $\widehat{F}$, discouraging manipulation. It is readily verified that the substitution effect dominates for $\widetilde{\beta}$ sufficiently small. Conversely, the income effect potentially dominates for $\widetilde{\beta}$ sufficiently high. For example, in the case of linear probability models, it is readily verified that the income effect dominates if $\widetilde{\beta} > \widetilde{\alpha}/(2x)$.

The following proposition summarizes results from this section.

Proposition 1. *If the posted (linear, logit, or probit) model features slope coefficient $\widetilde{\beta} > 0$, manipulation is: decreasing in the true covariate x; decreasing in the posted model intercept $\widetilde{\alpha}$; and increasing in the posted model slope for $\widetilde{\beta}$ sufficiently small.*

3 Goodhart's Law in Linear Probability Models

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

$$\mathbb{E}[y|x] = \Pr[y=1|x] = a + bx. \tag{15}$$

Despite its practical limitations, the LPM is an attractive starting point since many arguments can be expressed in terms of intuitive objects, such as covariances. Moreover, as we show, many results for the LPM carry over to ML estimation.

To begin, it is useful to decompose the mean squared prediction error (MSPE) generated when a univariate function $\upsilon(\cdot)$ is applied to the measured covariate $\widetilde{x}$.[8] We have:[9]

$$\begin{aligned}
\mathbb{E}\left[(y-\upsilon(\widetilde{x}))^2\right] &= \mathbb{E}\left\{[(y-\mathbb{E}(y|\widetilde{x}))+(\mathbb{E}(y|\widetilde{x})-\upsilon(\widetilde{x}))]^2\right\} \\
&= \mathbb{E}\{(y-\mathbb{E}(y|\widetilde{x}))^2+(\mathbb{E}(y|\widetilde{x})-\upsilon(\widetilde{x}))^2+2(y-\mathbb{E}(y|\widetilde{x}))(\mathbb{E}(y|\widetilde{x})-\upsilon(\widetilde{x}))]\} \\
&= \mathbb{E}\{(y-\mathbb{E}(y|\widetilde{x}))^2\}+\mathbb{E}\left[(\mathbb{E}(y|\widetilde{x})-\upsilon(\widetilde{x}))^2\right]+2\mathbb{E}\left[(y-\mathbb{E}(y|\widetilde{x}))\mathbb{E}(y|\widetilde{x})\right]-2\mathbb{E}\left[(y-\mathbb{E}(y|\widetilde{x}))\upsilon(\widetilde{x})\right] \\
&= \mathbb{E}\left[\left[\mathbb{E}(y|\widetilde{x})-y\right]^2\right]+\mathbb{E}\left[\left[\upsilon(\widetilde{x})-\mathbb{E}(y|\widetilde{x})\right]^2\right].
\end{aligned} \tag{16}$$

Equation (16) 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 inherent loss arising from using the specific covariate $\widetilde{x}$ as a basis for prediction. The second component is the distance between the chosen function υ and the conditional expectation function. Notice, if an econometrician were to treat the distribution of $\widetilde{x}$ as predetermined, as in Nash equilibrium, then the conditional expectation function is the optimal υ. However, a Stackelberg leader would account for the effect of υ on the distribution of $\widetilde{x}$, as stressed by Frankel and Kartik (2022).

Suppose for the moment that data manipulation is impossible, with the econometrician choosing υ to be the affine function $\upsilon(x)=\alpha+\beta x$. We then have:

$$\mathbb{E}\left[(y-\alpha-\beta x)^2\right]=\mathbb{E}\left[\left[\mathbb{E}(y|x)-y\right]^2\right]+\mathbb{E}\left[\left[(\alpha+\beta x)-(a+bx)\right]^2\right]. \tag{17}$$

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

Remark 1. *In an economy without data manipulation, the econometric procedure (OLS/MLE) recovers the deep structural parameters* (a,b) *and each loan is correctly priced if interest rates are set according to equation* (5)*, with posted model coefficients* $(\widetilde{\alpha},\widetilde{\beta})=(a,b)$.

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

[8] Frankel and Kartik (2022) present analogous expressions for their setup.

[9] The penultimate line is zero due to orthogonality of the prediction error to any univariate function of $\widetilde{x}$.

$(y, \widetilde{x})$ 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 MSPE can be expressed as:

$$\mathbb{E}\left[(y - \alpha - \beta\widetilde{x})^2\right] = \alpha^2 + (1 - 2\alpha)\mathbb{E}[y] + 2\alpha\beta\mathbb{E}\left[\widetilde{x}\right] + \beta^2\mathbb{E}[\widetilde{x}^2] - 2\beta\mathbb{E}\{\mathbb{E}\left[y|x\right]\mathbb{E}\left[\widetilde{x}|x\right]\}. \tag{18}$$

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

$$\begin{aligned} \mathbb{E}\{\mathbb{E}\left[y|x\right]\mathbb{E}\left[\widetilde{x}|x\right]\} &= \mathbb{E}\left\{(a + bx)\mathbb{E}\left[\widetilde{x}|x\right]\right\} \\ &= a\mathbb{E}\left[\widetilde{x}\right] + b\mathbb{E}\left[x\widetilde{x}\right]. \end{aligned} \tag{19}$$

Substituting the preceding expression into equation (18), we find that the MSPE can be expressed parametrically as follows:

$$MSPE(\underbrace{\alpha, \beta}_{\text{Candidate}}\ ;\ \underbrace{\widetilde{\alpha}, \widetilde{\beta}}_{\text{Posted}}, \underbrace{a, b}_{\text{DGP}}) = \alpha^2 + (1-2\alpha)[a + b\mathbb{E}(x)] + 2(\alpha - a)\beta\mathbb{E}\left[\widetilde{x}\right] + \beta^2\mathbb{E}[\widetilde{x}^2] - 2\beta b\mathbb{E}\left[x\widetilde{x}\right]. \tag{20}$$

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

$$(\widehat{\alpha}_{ols}, \widehat{\beta}_{ols}) \in \arg\min_{\alpha,\beta}\ MSPE(\alpha, \beta; \widetilde{\alpha}, \widetilde{\beta}, a, b). \tag{21}$$

The preceding objective function is strictly concave, and the FOCs are:

$$\begin{aligned} 0 &= 2\widehat{\alpha} - 2[a + b\mathbb{E}(x)] + 2\widehat{\beta}\mathbb{E}[\widetilde{x}] \\ 0 &= 2(\widehat{\alpha} - a)\mathbb{E}[\widetilde{x}] + 2\widehat{\beta}\mathbb{E}\left[\widetilde{x}^2\right] - 2b\mathbb{E}[x\widetilde{x}]. \end{aligned} \tag{22}$$

Throughout, 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:

$$\begin{aligned} \widehat{\alpha}_{ols} &= a + b\mathbb{E}[x] - \widehat{\beta}\mathbb{E}[\widetilde{x}]. \\ \widehat{\beta}_{ols} &= b \times \underbrace{\frac{\mathbb{E}[x\widetilde{x}] - \mathbb{E}[x]\mathbb{E}[\widetilde{x}]}{\mathbb{E}\left[\widetilde{x}^2\right] - \left(\mathbb{E}[\widetilde{x}]\right)^2}}_{\equiv\widehat{\beta}^{x\widetilde{x}}_{ols}}. \end{aligned} \tag{23}$$

That is, the OLS slope coefficient here is the product of the clean-data slope coefficient and the slope coefficient from a regression of x on $\widetilde{x}$. To develop intuition, note that the slope coefficient in equation (23) can be thought of as a regression chain rule, in the sense that:

$$\underbrace{\widehat{\beta}_{ols}}_{\frac{d\widehat{y}}{d\widetilde{x}}} \approx \underbrace{b}_{\frac{\partial\widehat{y}}{\partial x}} \times \underbrace{\widehat{\beta}^{x\widetilde{x}}_{ols}}_{\frac{\partial x}{\partial\widetilde{x}}}. \tag{24}$$

It is also useful to observe that the OLS slope coefficient can be rewritten as:[10]

$$\begin{aligned}\widehat{\beta}_{ols} &= b \times \rho_{x\widetilde{x}} \times \frac{\sigma_x^2}{\sigma_{\widetilde{x}}^2} \\ &= b \times \frac{\sigma_x^2 + \sigma_{xm}}{\sigma_x^2 + \sigma_m^2 + 2\sigma_{xm}} \\ &= b \times \left[1 - \widehat{\beta}_{ols}^{m\widetilde{x}}\right] \\ &= b \times \left[\frac{1}{1 + \widehat{\beta}_{ols}^{mx}}\right].\end{aligned} \tag{25}$$

The next lemma, demonstrated in the appendix, will prove useful in the analysis that follows.

Lemma 2. *Suppose $b > 0$ and consider a posted model featuring $\widetilde{\beta} > 0$. If $\sigma_c^2 > 0$, then in the limit as σ_x^2* tends to 0, *$\widehat{\beta}_{ols} < b$. If $\sigma_c^2 = 0$, then $\widehat{\beta}_{ols} > b$.*

The first part of Lemma 2, which follows from inspection of equation (25), is analogous to classical econometric results. In particular, as σ_x^2 tends to zero, most of the variation in manipulation is caused by cross-sectional variation in manipulation costs. This type of orthogonal (to x) measurement error is akin to classical white noise measurement error, albeit with positive support and non-spherical distribution.

The second part of Lemma 2 is more novel: The OLS slope coefficient is greater than its clean data counterpart if $\sigma_c^2 = 0$. A simple graphical argument complementary to that in the introduction conveys the intuition here. If $\sigma_c^2 = 0$, manipulation is a univariate decreasing function of x. With this in mind, suppose, say, only two points in the support of x, say x_1 and x_2 where $x_2 > x_1$. Suppose also that only the low types manipulate by more than a trivial amount. Noting that the outcome variable y is unaffected by manipulation, it is apparent that the slope of the line of best fit must rotate upwards. Finally, inspection of the final lines of equation (25) reveals that the estimated slope is greater than the clean data slope b if manipulation decreases at rate sufficient to ensure the best linear fit to m is decreasing in x, as well as $\widetilde{x}$, which is apparently the case if $\sigma_c^2 = 0$ (Lemma 2).

Inspection of equation (23) also leads directly 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 posted model, 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 $b = 0$,

[10] See Jorn-Steffen Pischke's lecture notes on measurment error for similar expressions.

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 can actually remain robust over time, even if borrowers are induced to manipulate by a posted model featuring $\widetilde{\beta} > 0$.

Let us now formally evaluate 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 ML in later sections) using the resulting manipulated data drawn from this strategic cohort. We label the resulting estimates as *Goodhart estimates*, since this is the type of econometric practice Goodhart contemplated, although it is certainly not a practice he advocated. In the present context we have

$$\text{Goodhart Estimates: } (\widehat{\alpha}_{ols}, \widehat{\beta}_{ols}) \in \arg\min_{\alpha,\beta} MSPE(\underbrace{\alpha, \beta}_{\text{Candidate}} ; \underbrace{a, b}_{\text{Posted}}, \underbrace{a, b}_{\text{Historical}}). \tag{26}$$

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:

$$\widetilde{\beta} = b > 0 \Rightarrow \mathbb{E}[\widetilde{x}] > \mathbb{E}[x]. \tag{27}$$

Combining the preceding equation with equation (23), 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 values (a, b) within 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 $\widehat{\beta}_{ols} < b$. Further, if $b > 0$ and $\sigma_c^2 = 0$, then $\widehat{\beta}_{ols} > b$ and $\widehat{\alpha}_{ols} < a$.*

Figure 2 illustrates Goodhart's law in linear default prediction models, as detailed in Proposition 3. The left (right) panels consider relatively high (low) cross-sectional variation in manipulation costs c. In each figure, the clean data causal parameter b is varied along the horizontal axis, allowing us to illustrate how coefficient shift varies with the deep structural parameter b. Notice, as b is varied, the causal effect of x increases. The econometrician relies on the clean data coefficient and posts a model with $\widetilde{\beta} = b$. Thus, as one moves along the horizontal axis incentives for manipulation are changing along with changes in the true causal relationship between y and the latent variable x.

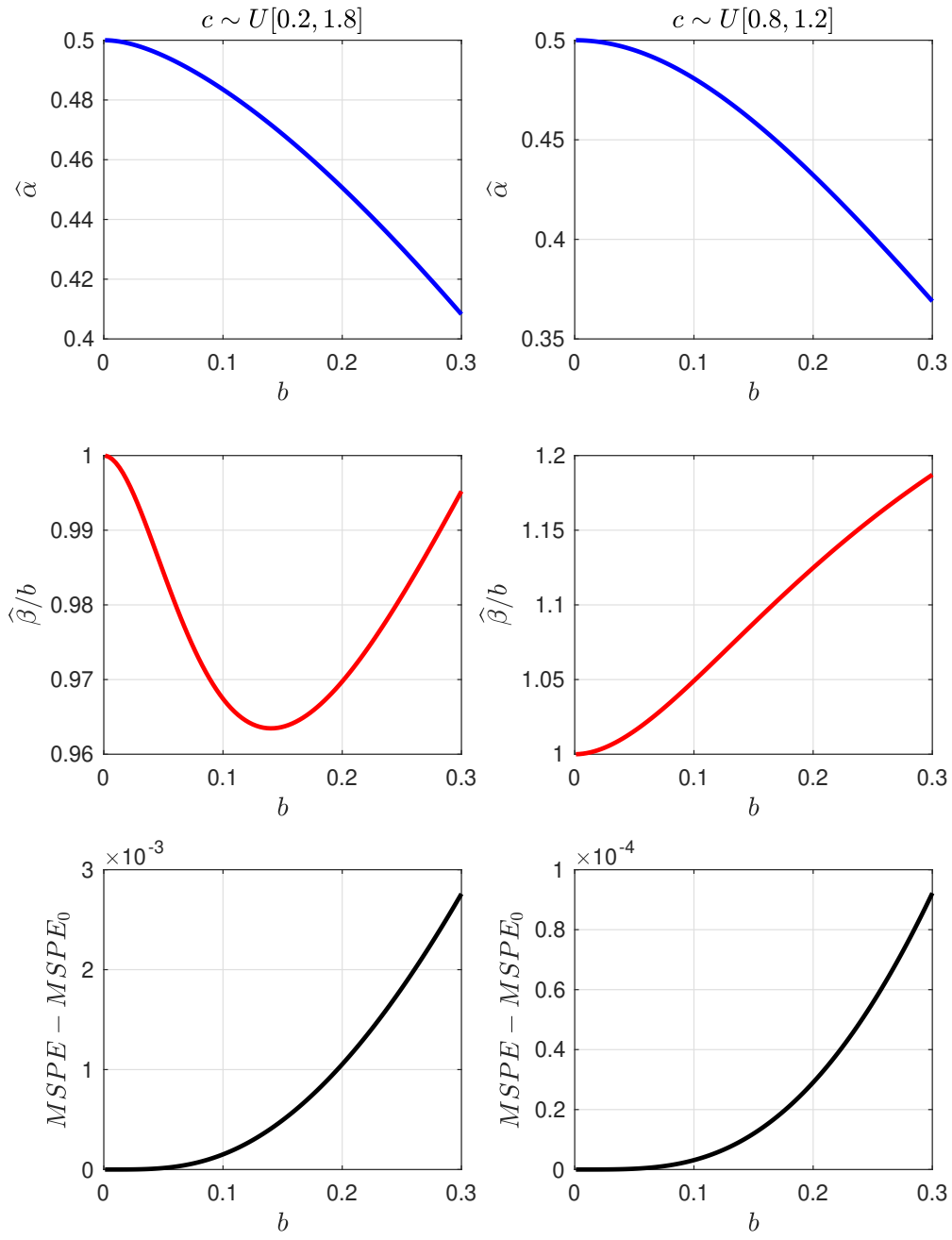

Figure 2: **Goodhart's Law - Linear prediction model**. We plot OLS Goodhart estimates for $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$, with $b \in]0, \overline{b}]$. In the first column the variance of manipulation cost is high, in the second it is low. In the third row we plot the difference between the MSPE with and without manipulation. The figure assumes $c \sim U[c_d, c_u]$ and $x \sim U[x_{\min}, x_{\max}]$, with $x_{\min} = 0$, and $x_{\max} = \max\{\frac{1-a}{\overline{b}} - \delta, 1\}$ to ensure repayment probability is in $[0, 1]$ for all x and all $\widetilde{\beta}$, where δ equals the maximum manipulation for average x. We set $a = 0.5$, and $\overline{b} = 0.3$ to ensure consistency across the different models, that is manipulation incentives are similar for the linear and the logit model. The remaining parameters are $r = 0$ and $l = 0.5$.

To begin, we note that Figure 2 assumes the clean data intercept term is $a = 0.5$. It is apparent from the top two panels that the estimated intercept $\widehat{\alpha}$ undershoots a, as a crude means of countering upward manipulation of the covariate. As the posted model slope $\widetilde{\beta} = b$ is increased, the estimated intercept falls further in order to counter higher manipulation.

The middle panel of Figure 2 examines the behavior of slope coefficients in manipulated data. As the posted model slope b increases, manipulation increases, shifting the estimated coefficient in manipulated data ($\widehat{\beta}$) away from b. Comparing across the two middle panels, we see that the nature of slope coefficient shift depends upon the degree of cross-sectional variation in manipulation costs c. There is upward coefficient shift ($\widehat{\beta} > b$) in the right panel and attenuation in the left panel ($\widehat{\beta} < b$), consistent with Lemma 2.

Consider finally the bottom panels of Figure 2 which measure the difference between the MSPE that emerges in the manipulated data versus the MSPE in clean data. When the lender posts a model with $\widetilde{\beta} = b$, and b is large, manipulation levels are high. This manipulation leads to lower predictive power compared to the clean data benchmark.

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:

$$L \equiv \prod_{i=1}^{I} (F(\alpha + \beta x_i))^{y_i} (1 - F(\alpha + \beta x_i))^{1-y_i} \tag{28}$$

$$\mathcal{L} \equiv \sum_{i=1}^{I} y_i \ln(F(\alpha + \beta x_i)) + (1 - y_i) \ln(1 - F(\alpha + \beta x_i)). \tag{29}$$

Using the law of iterated expectations, an expected log likelihood function can be computed as:

$$\begin{aligned}
\frac{1}{I}\mathbb{E}\left[\mathbb{E}(\mathcal{L}|\mathbf{x})\right] &= \frac{1}{I}\int_X \left[\mathbb{E}\left(\sum_{i=1}^{I} y_i \ln F(\alpha+\beta x_i) + (1-y_i)\ln(1-F(\alpha+\beta x_i))|x_i\right)\right] h(x)dx \quad (30)\\
&= \frac{1}{I}\int_X \left[\sum_{i=1}^{I} \mathbb{E}\left\{y_i \ln F(\alpha+\beta x_i) + (1-y_i)\ln(1-F(\alpha+\beta x_i))|x_i\right\}\right] h(x)dx\\
&= \int_X \left\{\ln F(\alpha+\beta x)\mathbb{E}(y|x) + \ln(1-F(\alpha+\beta x))[1-\mathbb{E}(y|x)]\right\} h(x)dx.
\end{aligned}$$

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

$$\mathcal{L} = \int_X \left\{ \begin{array}{c} F(a+bx)\ln F(\alpha+\beta x) \\ +[1-F(a+bx)]\ln[1-F(\alpha+\beta x)] \end{array} \right\} h(x)dx. \quad (31)$$

Considering the integrand in $\mathcal{L}$, we note that, as shown by Pratt (1981), $\ln F$ and $\ln(1-F)$ are concave for F taken to be either standard normal or logistic. Moreover, the objective is concave since the composition of a concave function with a linear function is concave.

In the absence of data manipulation, the MLE intercept and slope $(\widehat{\alpha}, \widehat{\beta})$ satisfy the following FOCs:

$$\begin{aligned}
\int_X \left[\frac{F(a+bx)}{F(\widehat{\alpha}+\widehat{\beta}x)} - \frac{1-F(a+bx)}{1-F(\widehat{\alpha}+\widehat{\beta}x)}\right] f(\widehat{\alpha}+\widehat{\beta}x)h(x)dx &= 0 \quad (32)\\
\int_X \left[\frac{F(a+bx)}{F(\widehat{\alpha}+\widehat{\beta}x)} - \frac{1-F(a+bx)}{1-F(\widehat{\alpha}+\widehat{\beta}x)}\right] x f(\widehat{\alpha}+\widehat{\beta}x)h(x)dx &= 0.
\end{aligned}$$

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 when coefficient estimates are set equal to the true parameters (a, b). Conniffe (1987) argues this property represents an argument in favor of MLE estimation.

Moving away from the classical MLE environment, 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, and using the conditional expectation function, we have:[11]

$$\begin{aligned}
\frac{1}{I}\mathbb{E}\left[\mathbb{E}(\mathcal{L}|\mathbf{x},\mathbf{c})\right] &= \frac{1}{I}\int_C\int_X\left[\sum_{i=1}^{I}\mathbb{E}\left[y_i\ln(F(\alpha+\beta\widetilde{x}_i))+(1-y_i)\ln(1-F(\alpha+\beta\widetilde{x}_i))|x_i,c_i\right]\right]h(x)g(c)dxdc \quad (33)\\
&= \int_C\int_X\{\mathbb{E}\left[y\ln(F(\alpha+\beta\widetilde{x}))+(1-y)\ln(1-F(\alpha+\beta\widetilde{x}))|x,c\right]\}h(x)g(c)dxdc\\
&= \int_C\int_X\{\mathbb{E}(y|x)\mathbb{E}[\ln(F(\alpha+\beta\widetilde{x}))|x,c]+(1-\mathbb{E}(y|x))\mathbb{E}[\ln(1-F(\alpha+\beta\widetilde{x}))|x,c]\}h(x)g(c)dxdc\\
&= \int_C\int_X\left\{\begin{array}{c}F(a+bx)\mathbb{E}[\ln(F(\alpha+\beta\widetilde{x}))|x,c]\\+(1-F(a+bx))\mathbb{E}[\ln(1-F(\alpha+\beta\widetilde{x}))|x,c]\end{array}\right\}h(x)g(c)dxdc.
\end{aligned}$$

Substituting the function m this into equation (33), we obtain the following expected log likelihood function:

$$\mathcal{L}(\alpha,\beta;\widetilde{\alpha},\widetilde{\beta},a,b)=\int_C\int_X\left[\begin{array}{c}F(a+bx)\ln[F(\alpha+\beta x+\beta m(x,\widetilde{\alpha},\widetilde{\beta},c))]\\+[1-F(a+bx)]\ln[1-F(\alpha+\beta x+\beta m(x,\widetilde{\alpha},\widetilde{\beta},c))]\end{array}\right]h(x)g(c)dxdc. \quad (34)$$

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:

$$MLE:(\widehat{\alpha},\widehat{\beta})\in\arg\max_{\alpha,\beta}\ \mathcal{L}(\underbrace{\alpha,\beta}_{\text{Candidate}};\underbrace{\widetilde{\alpha},\widetilde{\beta}}_{\text{Posted}},\underbrace{a,b}_{\text{DGP}}). \quad (35)$$

Differentiating equation (34) we have the following FOCs for intercept and slope coefficients, respectively:

$$\begin{aligned}
\mathcal{L}_1(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) &= \int_C\int_X\left[\begin{array}{c}\frac{F(a+bx)}{F[\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}m(x,\widetilde{\alpha},\widetilde{\beta},c)]}\\-\frac{1-F(a+bx)}{1-F[\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}m(x,\widetilde{\alpha},\widetilde{\beta},c)]}\end{array}\right]f[\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}m(x,\widetilde{\alpha},\widetilde{\beta},c)]h(x)g(c)dxdc=0 \quad (36)\\
\mathcal{L}_2(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) &= \int_C\int_X\left[\begin{array}{c}\frac{F(a+bx)}{F[\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}m(x,\widetilde{\alpha},\widetilde{\beta},c)]}\\-\frac{1-F(a+bx)}{1-F[\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}m(x,\widetilde{\alpha},\widetilde{\beta},c)]}\end{array}\right](x+m)f[\widehat{\alpha}+\widehat{\beta}x+\widehat{\beta}m(x,\widetilde{\alpha},\widetilde{\beta},c)]h(x)g(c)dxdc=0.
\end{aligned}$$

The SOCs are:

$$\begin{aligned}
\mathcal{L}_{11}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) &< 0\\
\mathcal{L}_{22}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) &< 0
\end{aligned}$$

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

and

$$\left| \begin{array}{cc} \mathcal{L}_{11}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) & \mathcal{L}_{12}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) \\ \mathcal{L}_{21}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) & \mathcal{L}_{22}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) \end{array} \right| > 0.$$

Notice, the FOCs with data manipulation are identical in form to those arising when manipulation is impossible (equation (32)), but now the manipulated covariate $\widetilde{x}$ takes the place of the true covariate 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

$$\begin{aligned} \mathcal{L}_1(a, 0; \widetilde{\alpha}, \widetilde{\beta}, a, 0) &= 0 \\ \mathcal{L}_2(a, 0; \widetilde{\alpha}, \widetilde{\beta}, a, 0) &= 0. \end{aligned} \tag{37}$$

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

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 Comparative Statics: Posted Model

Related to Goodhart's law is the more general question of how posting a model changes incentives, observables, and coefficient estimates. In order to examine this question, this subsection presents comparative static results for the effect of changes in $(\widetilde{\alpha}, \widetilde{\beta})$ on $(\widehat{\alpha}, \widehat{\beta})$ holding fixed the deep causal parameters (a, b). That is, we examine the effect of changes in the posted model on estimated coefficients arising from agents responding to the posted model. Anticipating, this analysis is of independent technical interest since it is related to the question of fixed point convergence, as discussed in detail below.

Consider first the effect of changing the posted model slope, holding fixed the posted model intercept. Computing the total differential of the FOCs given in equation (36) we obtain:

$$\begin{aligned} \mathcal{L}_{11} d\widehat{\alpha} + \mathcal{L}_{12} d\widehat{\beta} + \mathcal{L}_{14} d\widetilde{\beta} &= 0 \\ \mathcal{L}_{21} d\widehat{\alpha} + \mathcal{L}_{22} d\widehat{\beta} + \mathcal{L}_{24} d\widetilde{\beta} &= 0. \end{aligned} \tag{38}$$

Rearranging terms, we obtain the following comparative statics:

$$\begin{aligned}\begin{bmatrix} d\widehat{\alpha}/d\widetilde{\beta} \\ d\widehat{\beta}/d\widetilde{\beta} \end{bmatrix} &= -\begin{bmatrix} \mathcal{L}_{11}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) & \mathcal{L}_{12}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) \\ \mathcal{L}_{21}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) & \mathcal{L}_{22}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) \end{bmatrix}^{-1} \begin{bmatrix} \mathcal{L}_{14}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) \\ \mathcal{L}_{24}(\widehat{\alpha},\widehat{\beta};\widetilde{\alpha},\widetilde{\beta},a,b) \end{bmatrix} \\ &= \underbrace{-\left(\frac{1}{\mathcal{L}_{11}\mathcal{L}_{22}-\mathcal{L}_{12}\mathcal{L}_{21}}\right)}_{<0} \begin{bmatrix} \mathcal{L}_{22}\mathcal{L}_{14}-\mathcal{L}_{12}\mathcal{L}_{24} \\ -\mathcal{L}_{21}\mathcal{L}_{14}+\mathcal{L}_{11}\mathcal{L}_{24} \end{bmatrix}. \end{aligned} \tag{39}$$

Following analogous steps, we obtain the following comparative statics regarding the effect of changes in the posted model intercept, holding fixed the posted model slope:

$$\begin{bmatrix} d\widehat{\alpha}/d\widetilde{\alpha} \\ d\widehat{\beta}/d\widetilde{\alpha} \end{bmatrix} = \underbrace{-\left(\frac{1}{\mathcal{L}_{11}\mathcal{L}_{22}-\mathcal{L}_{12}\mathcal{L}_{21}}\right)}_{<0} \begin{bmatrix} \mathcal{L}_{22}\mathcal{L}_{13}-\mathcal{L}_{12}\mathcal{L}_{23} \\ -\mathcal{L}_{21}\mathcal{L}_{13}+\mathcal{L}_{11}\mathcal{L}_{23} \end{bmatrix}. \tag{40}$$

Figure 3 presents numerical comparative statics results in the context of logit estimation.[12] Consider first the left panels in the figure which present comparative statics for $(\widehat{\alpha},\widehat{\beta})$ resulting from alternative values of the posted model intercept $\widetilde{\alpha}$. As the posted model intercept increases, each borrower has a higher baseline credit score and diminished incentive to manipulate since, we recall, the interest rate is a decreasing convex function of the credit score. Since borrowers manipulate less, the estimated intercept $\widehat{\alpha}$ shifts upward, with the estimated slope falling by a small amount. As shown in the bottom left panel, the difference between the expected likelihood ratio attained in manipulated data ($\mathcal{L}$) and its clean data counterpart ($\mathcal{L}_0$) also shrinks as the posted intercept $\widetilde{\alpha}$ is increased, consistent with less manipulation.

The right panels in Figure 3 present comparative statics for alternative values of the posted model slope $\widetilde{\beta}$. As the posted model slope increases, borrowers tend to have a stronger incentive to manipulate. With more manipulation, the estimated intercept $\widehat{\alpha}$ shifts downward to counter manipulation. However, the estimated slope increases, consistent with strong negative correlation between manipulation (measurement error) and the regressor. Equation (25) hints that upward slope shift may well be expected. Finally, as shown in the bottom right panel, the difference between the likelihood ratio attained in manipulated data ($\mathcal{L}$) and its clean data counterpart ($\mathcal{L}_0$) increases as the posted slope $\widetilde{\beta}$ is increased, consistent with the idea that more manipulation leads to diminished predictive power.

Finally, it is important to note the magnitude of the various comparative statics. In particular, the estimated coefficients vary less than one-for-one with changes in the posted coefficients, at least at the assumed parameter values. This hints at the possibility that MLE estimation performed on manipulated data can represent a contraction mapping, with concomitant implications for convergence, a topic discussed below.

[12] Probit results are similar and available upon request.

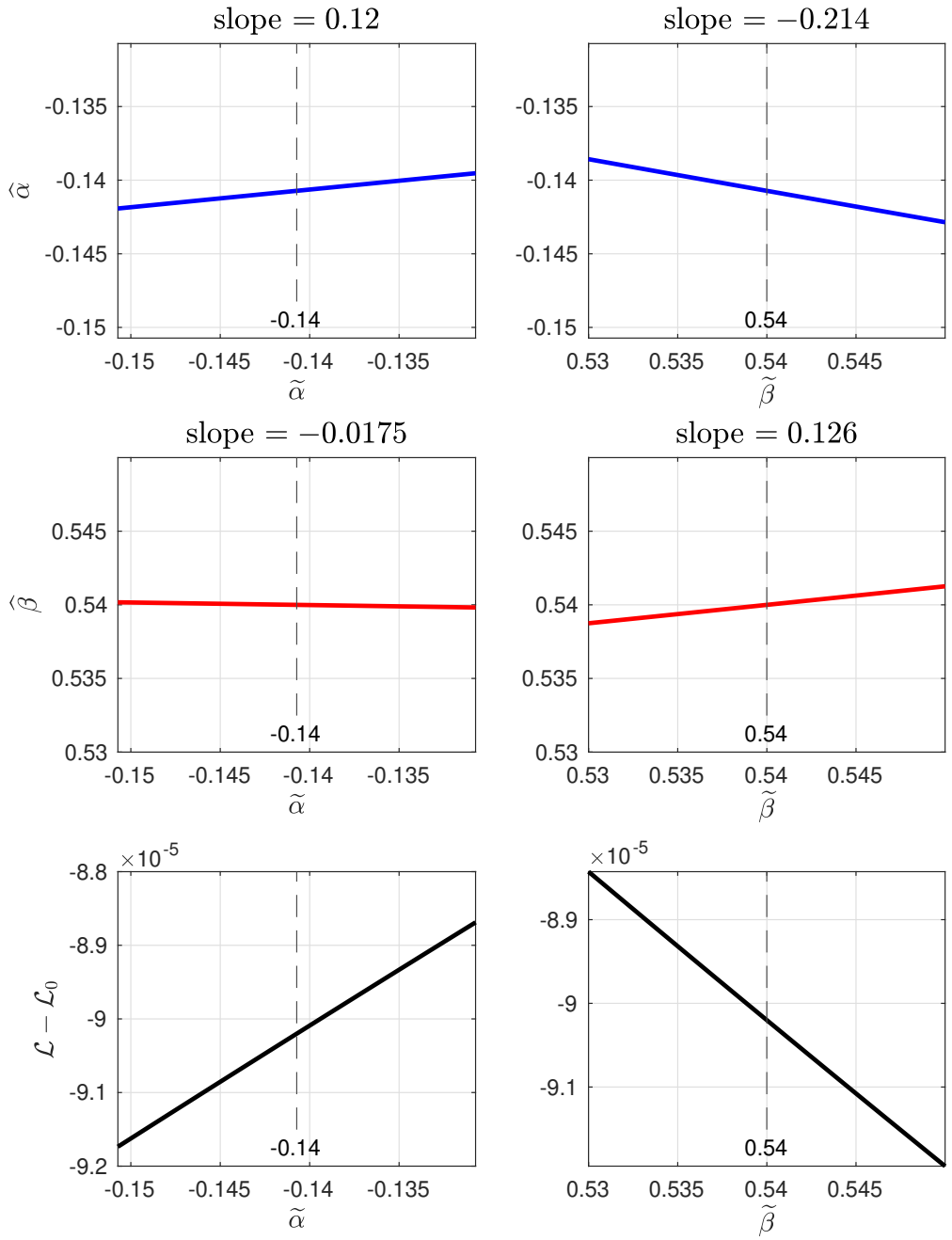

Figure 3: **Comparative Statics - Logit prediction model**. MLE estimates, $(\widehat{\alpha}, \widehat{\beta})$, against the posted credit prediction model, $(\widetilde{\alpha}, \widetilde{\beta})$, given the data generating process $(a, b) = (0, 0.5)$. We plot $(\widehat{\alpha}, \widehat{\beta})$ in a range for $(\widetilde{\alpha}, \widetilde{\beta})$ around the fixed point model $(\alpha^*, \beta^*) = (-0.14, 0.54)$. In the third line we plot the difference between the optimal likelihood with manipulation and without manipulation. We assume $c \sim U[c_d, c_u]$ with $c_d = 0.6$ and $c_u = 1.4$, and a logit prediction model, $F(z) = e^z(1 + e^z)^{-1}$, with $z = \widetilde{\alpha} + \widetilde{\beta}x$), where $x \sim U[x_{\min}, x_{\max}]$, with $x_{\min} = 0$, and $x_{\max} = 1$. The remaining parameters are $r = 0$, and $l = 0.5$.

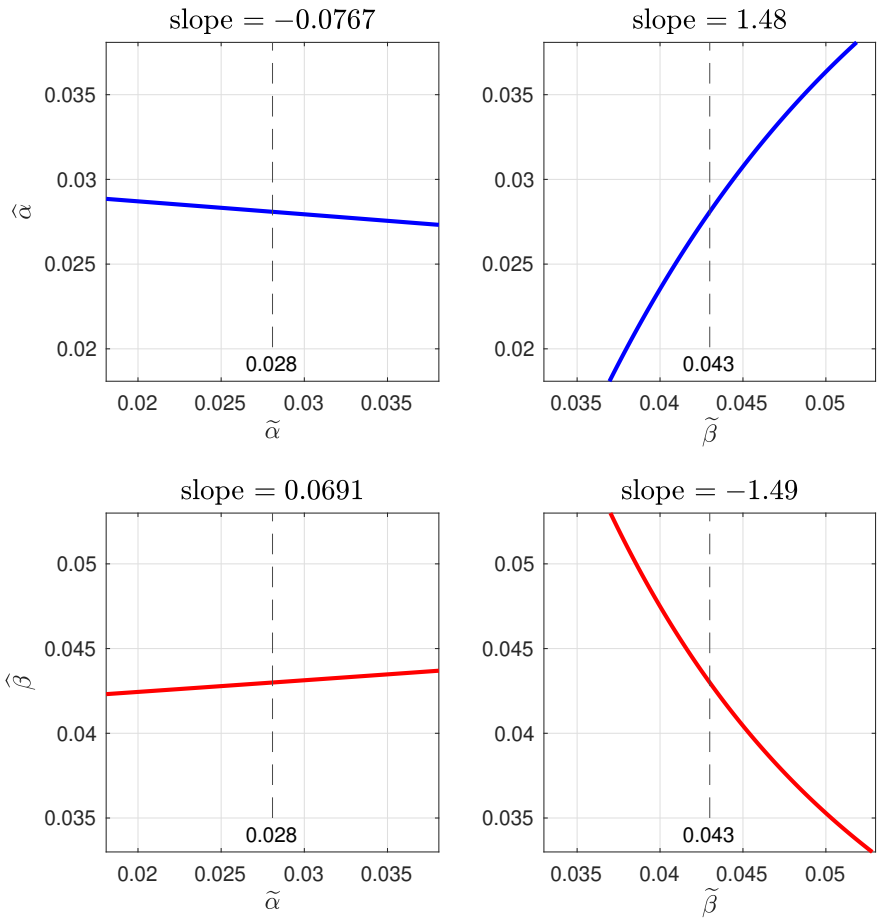

Figure 4: **Comparative Statics - Logit prediction model**. This is the same as Figure 3, under the logit prediction model (MLE) with DGP parameters $(a, b) = (0, 0.5)$, except $x \sim U[0, 0.5]$ and $c \sim U[0.0060, 0.0140]$. With these parameters the fixed point model is $(\alpha^*, \beta^*) = (0.028, 0.043)$.

In Figure 4 we perform the same comparative statics exercise as in Figure 3, but now we shift down the manipulation cost support, which increases manipulation for each x, and reduce the upper bound for x, which increases the proportion of lower quality borrowers who have stronger manipulation incentives. As shown, in this environment, estimated coefficients are much more sensitive to the coefficients of the posted model. Intuitively, borrowers are much more responsive to the posted model if they face low manipulation costs and/or they have low latent credit quality.

4.3 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:

$$\text{Goodhart Estimates} = (\widehat{\alpha}, \widehat{\beta}) \in \arg\max_{\alpha,\beta} \quad \mathcal{L}(\alpha, \beta; \underbrace{a, b}_{\text{Posted}}, \underbrace{a, b}_{\text{Historical}}). \tag{41}$$

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 $\beta \geq b > 0$. Notice, the FOC for the intercept would be violated at such a candidate coefficient vector, with:

$$\mathcal{L}_1(\alpha, \beta; a, b, a, b) = \int_C \int_X \begin{bmatrix} \frac{F(a+bx)}{F[\alpha+\beta x+\beta m(x,a,b,c)]} \\ -\frac{1-F(a+bx)}{1-F[\alpha+\beta x+\beta m(x,a,b,c)]} \end{bmatrix} f[\alpha + \beta x + \widehat{\beta} m(x, a, b, c)] h(x) g(c) dx dc < 0.$$

The preceding inequality, along with Proposition 4, establishes the following proposition, the MLE analog of Proposition 3.[13]

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 $\widehat{\beta} < b$.*

Before proceeding, it is useful to gain a better sense of the consequences of Goodhart's law in MLE settings by way of numerical analysis. To this end, Figure 5 represents the MLE logit counterpart to Figure 2. Once again, we plot the coefficients that arise in data generated by

[13] Proposition 3 offers greater clarity regarding slope overshooting.

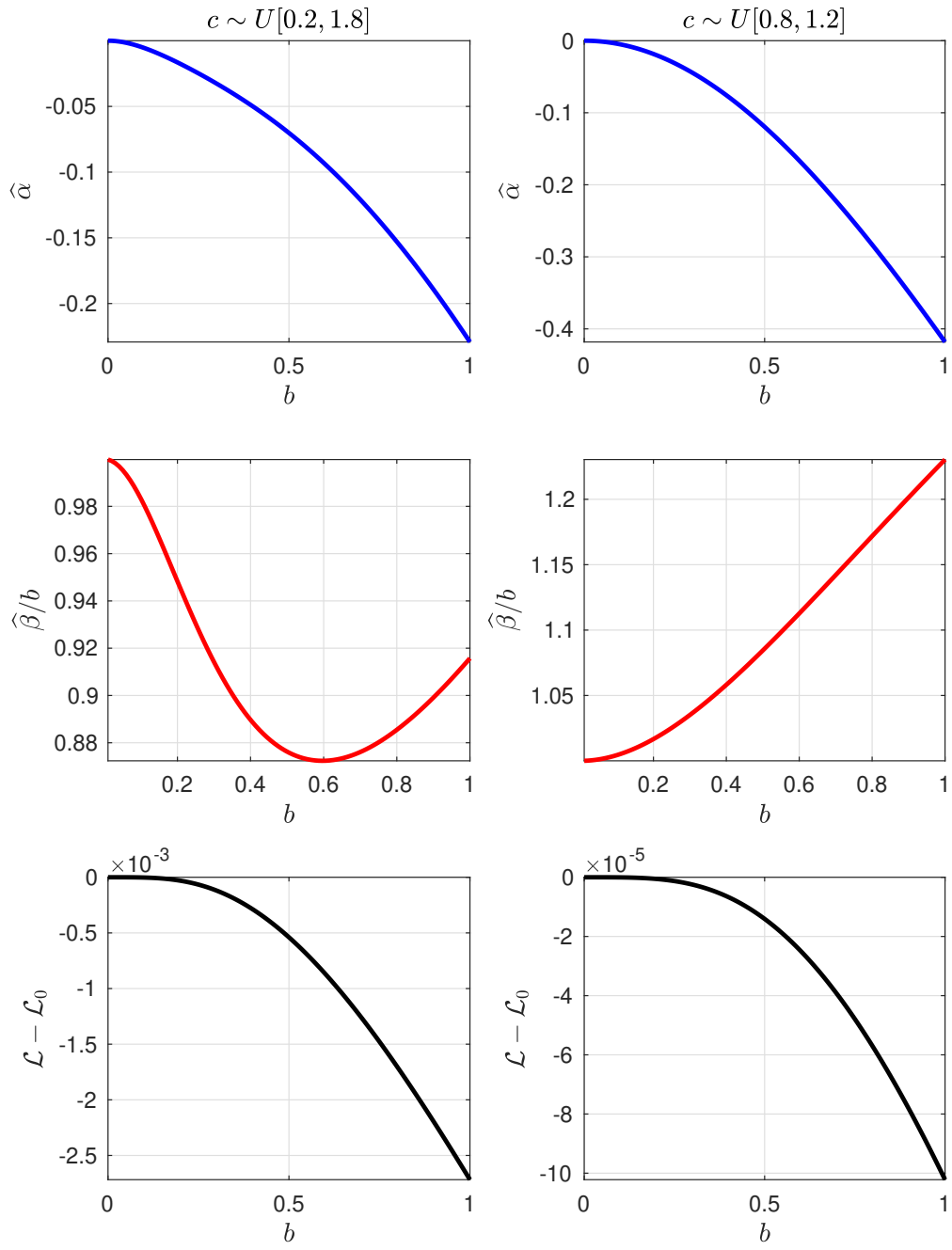

Figure 5: **Goodhart's Law - Logit prediction model**. We plot MLE Goodhart estimates for $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$, with $b \in]0, \overline{b}]$, where $a = 0$ and $\overline{b} = 1$. In the first column the variance of manipulation cost is high, in the second colums it is low. In the third row we plot the difference between the optimal likelihood with and without manipulation. We use the same assumptions as in Figure 3 for the default prediction model, manipulation cost and the remaining parameters, except for the assumption on $c \sim U[c_d, c_u]$.

strategic borrowers manipulate data in response to the posting of a model utilizing the clean data parameters (a, b). This naive econometric procedure is captured in equation (41). The left (right) panels consider relatively high (low) cross-sectional variation in manipulation costs. In each figure, the clean data slope parameter b is varied along the horizontal axis, allowing us to illustrate how the Goodhart effect varies with the true causal parameter b.

To begin, we note that Figure 5 assumes the clean data intercept term is $a = 0$. It is apparent from the top two panels that the estimated intercept $\widehat{\alpha}$ shifts below a, as a crude means of countering upward manipulation of the covariate. More specifically, as the posted model slope $\widetilde{\beta} = b$ is increased, the estimated intercept falls in order to counter ever-increasing manipulation.

The middle panel of Figure 5 illustrates the behavior of slope coefficients. Once again, we see that the direction of slope coefficient shift depends upon the volatility of manipulation cost parameters c. There is upward shift ($\widehat{\beta} > b$) in the right panel (low cost variability) and attenuation ($\widehat{\beta} < b$) in the left panel (high cost variability), as suggested by Lemma 2.

Consider finally the bottom panels of Figure 5 which measure the difference between the expected log likelihood ratio that emerges in the manipulated data versus its clean data analog. As shown, if the lender posts a model with $\widetilde{\beta} = b$, with b large, there will be strong incentives for manipulation. This manipulation leads to lower predictive power compared to the clean data benchmark.

5 Fixed Point Models

Proposition 5 shows that if the true covariate has explanatory power ($b > 0$), there will be inconsistency between a posted econometric model featuring clean data coefficients (a, b) and the MLE estimates $(\widehat{\alpha}, \widehat{\beta})$ that would be obtained if ex post estimation was then performed on the resulting data. This is the sort of inconsistency that troubled Goodhart (1975): Statistical regularities that break down if one and uses them for control purposes. To avoid Goodhart's critique, one might hope to find a model that remains robust even when it is used for control purposes. Such models are the focus of the present section.

5.1 Fixed Points Defined

A *fixed point model* is a pair of coefficients (α^*, β^*) such that:

$$(\alpha^*, \beta^*) \in \arg\max_{\alpha,\beta} \quad \mathcal{L}(\underbrace{\alpha, \beta}_{\text{Candidate}} \; ; \underbrace{\alpha^*, \beta^*}_{\text{Posted}}, \underbrace{a, b}_{\text{DGP}}). \tag{42}$$

In words, a fixed point model maximizes the expected log likelihood function given the distribution of covariates that arises from posting it. In the language Goodhart (1975), a fixed point model constitutes a statistical regularity that remains robust after being used for control purposes.

We note that a fixed point constitutes the Nash equilibrium model of a game in which borrowers submit covariates and the econometrician simultaneously posts coefficients. It is also time-consistent in that the econometrician would have no incentive to change the posted model after collecting data and estimating coefficients using, say, a subsample of data generated by borrowers responding to that posted model. Thus, borrowers could trust a fixed point model even if the econometrician lacked the power to commit to the original posted coefficients. Finally, a fixed point model can be viewed as satisfying an institutional constraint that models be rationalizable in a particular sense: A fixed point model can be shown to be optimal ex post within the data it generates.

5.2 Properties of Fixed Point Models

Consider first existence of fixed points. To begin, we note that a fixed point satisfies the following FOCs:

$$\begin{aligned} \mathcal{L}_1(\alpha^*, \beta^*, \alpha^*, \beta^*, a, b) &= 0 \\ \mathcal{L}_2(\alpha^*, \beta^*, \alpha^*, \beta^*, a, b) &= 0. \end{aligned} \tag{43}$$

Applying the implicit function theorem to the preceding system of two equations, we have the following useful lemma.

Lemma 3. *Suppose* (α_0^*, β_0^*) represents a fixed point for given clean data parameters (a, b). Suppose also that when evaluated at $(\alpha_0^*, \beta_0^*, \alpha_0^*, \beta_0^*, a, b)$,

$$\begin{vmatrix} \mathcal{L}_{11} + \mathcal{L}_{13} & \mathcal{L}_{12} + \mathcal{L}_{14} \\ \mathcal{L}_{21} + \mathcal{L}_{23} & \mathcal{L}_{22} + \mathcal{L}_{24} \end{vmatrix} \neq 0.$$

Then there exists a neighborhood of (a, b) *and a function* Ψ *defined on this neighborhood such that* $(\alpha^*, \beta^*) = \Psi(a, b)$ *uniquely solves:*

$$\begin{aligned} \mathcal{L}_1[\Psi(a,b), \Psi(a,b), a, b] &= 0 \\ \mathcal{L}_2[\Psi(a,b), \Psi(a,b), a, b] &= 0. \end{aligned}$$

That is, if one can pin down a fixed point at a particular point in (a, b) space, a local continuum of unique fixed points can be shown to exist provided the relevant Jacobian condition is satisfied. With this in mind, it is convenient to note that Proposition 5 guarantees existence of a unique fixed point when $b = 0$, as discussed in the following remark.

Remark 2. *Suppose $b = 0$. Then regardless of the intercept a, there is a unique fixed point model $(\alpha_0^*, \beta_0^*) = (a, b)$.*

For further insight, we expand the terms in the FOC for a fixed point:

$$\begin{aligned} 0 &= \mathcal{L}_1(\alpha^*, \beta^*; \alpha^*, \beta^*, a, b) \\ &= \int_C \int_X \left[\begin{array}{c} \frac{F(a+bx)}{F[\alpha^*+\beta^* x+\beta^* m(x,\alpha^*,\beta^*,c)]} \\ -\frac{1-F(a+bx)}{1-F[\alpha^*+\beta^* x+\beta^* m(x,\alpha^*,\beta^*,c)]} \end{array} \right] f[\alpha^* + \beta^* x + \beta^* m(x, \alpha^*, \beta^*, c)] h(x) g(c) dx dc \\ 0 &= \mathcal{L}_2(\alpha^*, \beta^*; \alpha^*, \beta^*, a, b) \\ &= \int_C \int_X \left[\begin{array}{c} \frac{F(a+bx)}{F[\alpha^*+\beta^* x+\beta^* m(x,\alpha^*,\beta^*,c)]} \\ -\frac{1-F(a+bx)}{1-F[\alpha^*+\beta^* x+\beta^* m(x,\alpha^*,\beta^*,c)]} \end{array} \right] [x + m(x, \alpha^*, \beta^*, c)] f[\alpha^* + \beta^* x + \beta^* m] h(x) g(c) dx dc. \end{aligned} \tag{44}$$

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

$$\widetilde{x} = x \Rightarrow (\widehat{\alpha}, \widehat{\beta}) = (a, b) = (a, 0) \equiv (\widetilde{\alpha}, \widetilde{\beta}). \tag{45}$$

Since the necessity component of the first stated claim follows from the second stated claim, we need only establish that claim. To this end, suppose $b > 0$, 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:

$$\begin{aligned} \beta^* &\geq b > 0 \Rightarrow \alpha^* < a \\ \alpha^* &\geq a \Rightarrow \beta^* < b. \end{aligned} \tag{46}$$

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.

Figure 6 contrasts fixed point coefficients (42) with Goodhart estimates (41), evaluated at alternative values of the clean data causal parameter b. The figure allows one to assess how close the econometrician would get to an internally consistent fixed point model after just one round of estimation on the manipulated data that emerges in response to naively posting a model with parameters (a, b). As shown, the gap between Goodhart estimates and fixed points grows larger for higher values of b. Nevertheless, it is worth noting the bottom panel which shows that although Goodhart estimates are not internally consistent, they may well achieve the same predictive power as fixed point models. Indeed, insisting upon a fixed point model is properly viewed as a constraint on the econometrician.

5.3 Fixed Point Iteration

In theory, a social planner who knew the entire structure of the economy, including the relevant manipulation technologies and parameters, could solve for fixed points by finding roots of equation (43). However, this model-based approach is likely to be infeasible given its informational requirements. This then leaves open the question of whether and how econometricians could work their way to a fixed point model.

With this question in mind, recall Proposition 6 informs us that, if $b > 0$, posting the clean data coefficients (a, b) will result in coefficients $(\widehat{\alpha}, \widehat{\beta}) \neq (a, b)$. Nevertheless, it is natural to ask whether econometricians could converge to a fixed point simply by iterating on this procedure, using this period's estimated coefficients as the next periods's posted coefficients.

To this end, let $(\alpha_{n+1}, \beta_{n+1})$ denote the model that will be posted to borrower cohorts in iteration round $n + 1$. Then consider the following mapping T, which corresponds to an adaptive iteration strategy:

$$(\alpha_{n+1}, \beta_{n+1}) \equiv T(\alpha_n, \beta_n) = \arg\max_{\alpha,\beta} \quad \mathcal{L}(\alpha, \beta, \alpha_n, \beta_n, a, b). \tag{47}$$

That is, let T correspond to a simple adaptive econometric strategy which uses the current period's MLE estimate as the next period's posted model. Notice, in practice, an econometrician using an adaptive strategy need only use the $(y, \widetilde{x})$ data generated by the prior cohort.

Applied here, the contraction mapping theorem offers sufficient conditions for the adaptive iteration strategy to converge to a fixed point. In particular, we have the following remark.

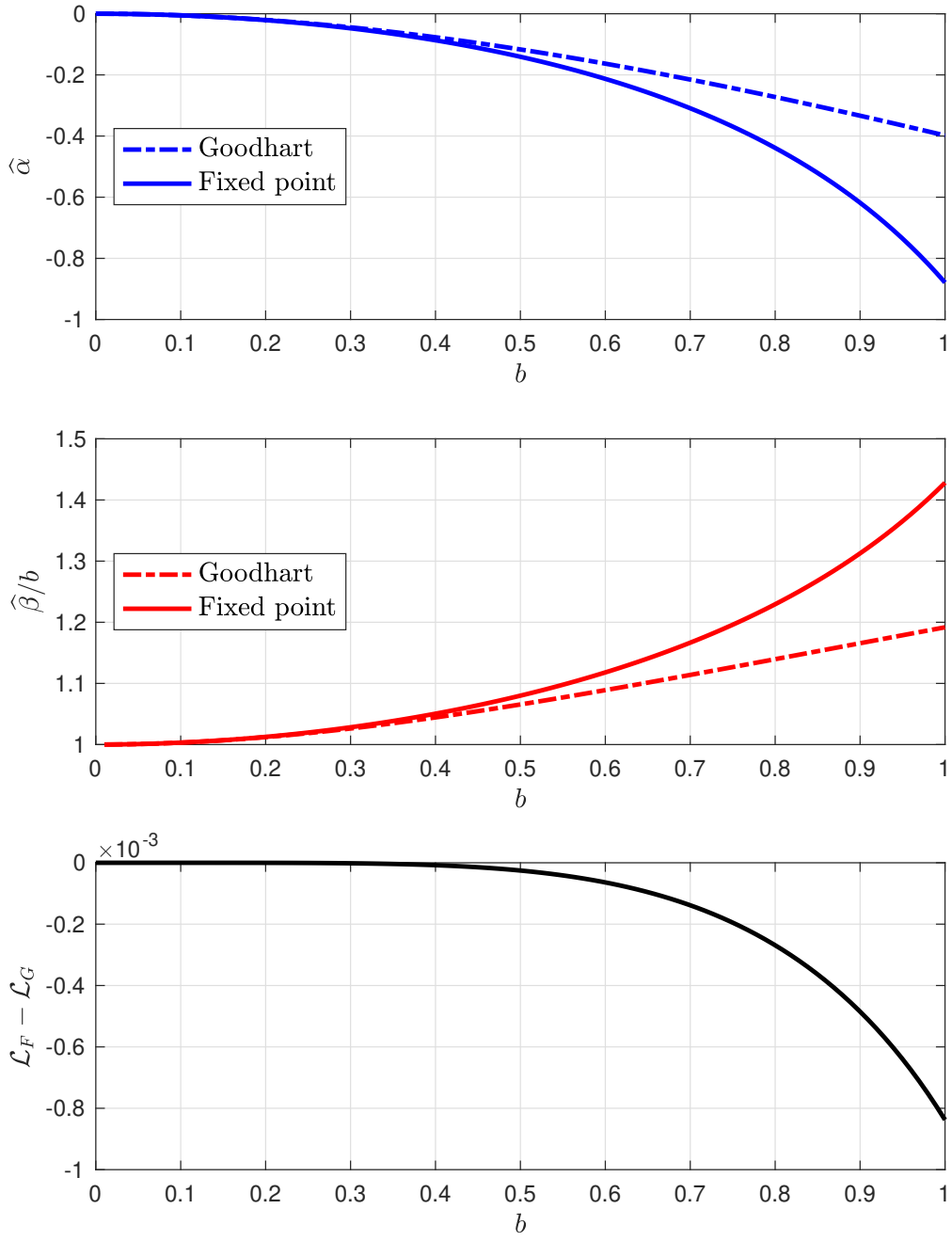

Figure 6: **Fixed point model vs Goodhart - Logit prediction model**. For $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$, with $b \in]0, \overline{b}]$, we plot against b in the first panel the MLE estimate $\widehat{\alpha}$ and in the second panel $\widehat{\beta}/b$. We consider the fixed point program, $(\alpha^*, \beta^*) = \arg\max_{(\alpha,\beta)} \mathcal{L}(\alpha, \beta; \alpha^*, \beta^*; a, b)$, and the program $(\widehat{\alpha}, \widehat{\beta}) = \arg\max_{(\alpha,\beta)} \mathcal{L}(\alpha, \beta; a, b; a, b)$. In the third panel we plot the difference between the optimal likelihood under fixed point, $\mathcal{L}_F$, and the optimal likelihood in the Goodhart's case, $\mathcal{L}_G$. We use the same assumptions as in Figure 5 for the default prediction model, manipulation cost and the remaining parameters.

Remark 3. *Let $\mathcal{D}$ be a closed convex domain in $\Re^2$ and let $T : \mathcal{D} \to \Re^2$ be continuously differentiable. Suppose that*

$$(\alpha, \beta) \in \mathcal{D} \Rightarrow T(\alpha, \beta) \in \mathcal{D}.$$

Suppose also that there exists $q < 1$ such that at all points in $\mathcal{D}$:

$$\left\| \begin{array}{cc} \frac{\partial T_1}{\partial \alpha} & \frac{\partial T_1}{\partial \beta} \\ \frac{\partial T_2}{\partial \alpha} & \frac{\partial T_2}{\partial \beta} \end{array} \right\| \leq q.$$

Then T is a contraction mapping on $\mathcal{D}$. Moreover, for any initial posted model $(\alpha_0, \beta_0) \in \mathcal{D}$, the sequence $(\alpha_{n+1}, \beta_{n+1}) \equiv T(\alpha_n, \beta_n)$ converges to a unique point (α^*, β^*) satisfying $T(\alpha^*, \beta^*) = (\alpha^*, \beta^*)$.

In order to clarify the preceding remark, we note that in the present application it has been shown above that:

$$\begin{aligned} \left[\begin{array}{cc} \frac{\partial T_1}{\partial \alpha} & \frac{\partial T_1}{\partial \beta} \\ \frac{\partial T_2}{\partial \alpha} & \frac{\partial T_2}{\partial \beta} \end{array} \right] &= \left[\begin{array}{cc} \frac{\partial \widehat{\alpha}}{\partial \widetilde{\alpha}} & \frac{\partial \widehat{\alpha}}{\partial \widetilde{\beta}} \\ \frac{\partial \widehat{\beta}}{\partial \widetilde{\alpha}} & \frac{\partial \widehat{\beta}}{\partial \widetilde{\beta}} \end{array} \right] \\ &= -\left(\frac{1}{\mathcal{L}_{11}\mathcal{L}_{22} - \mathcal{L}_{12}\mathcal{L}_{21}} \right) \left[\begin{array}{cc} \mathcal{L}_{22}\mathcal{L}_{13} - \mathcal{L}_{12}\mathcal{L}_{23} & \mathcal{L}_{22}\mathcal{L}_{14} - \mathcal{L}_{12}\mathcal{L}_{24} \\ -\mathcal{L}_{21}\mathcal{L}_{13} + \mathcal{L}_{11}\mathcal{L}_{23} & -\mathcal{L}_{21}\mathcal{L}_{14} + \mathcal{L}_{11}\mathcal{L}_{24} \end{array} \right]. \end{aligned}$$

Thus, convergence of the adaptive econometric procedure to a unique fixed point can be understood as hinging upon estimated coefficients not being too sensitive to posted coefficients, consistent with T being a contraction mapping. To take a trivial example, if data manipulation were impossible, MLE coefficient estimates would be completely insensitive to posted coefficients, and all elements of the relevant Jacobian would be equal to zero.

Two other technical points are worthy of note here. First, in order for the Jacobian condition in Remark 3 to be satisfied on some closed convex domain $\mathcal{D}$ about a fixed point (α^*, β^*), it must be satisfied at the fixed point itself. Thus, evaluating the Jacobian at (α^*, β^*) provides a useful first check for potential convergence. Second, and more practically, the converse of Remark 3 provides a useful diagnostic: If adaptive estimation fails to converge, borrower behavior must be responsive to posted model coefficients.

To further illustrate these arguments, we provide two numerical examples, one with convergence and the other without convergence. To begin, consider Figure 7, which utilizes the same parameter values and distributions as Figure 3, with the first posted model naively using the clean data coefficients (a, b). As shown in Figure 7, in this setting adaptive estimation (equation (47)) converges to a fixed point. Notice, there is a large gap between the first round estimates and the fixed point, consistent with Goodhart's law. However, the gap narrows dramatically after the second round of estimation. Apparently, here the opportunity to estimate coefficients on just one round of manipulated data helps the econometrician greatly in converging to a fixed point.

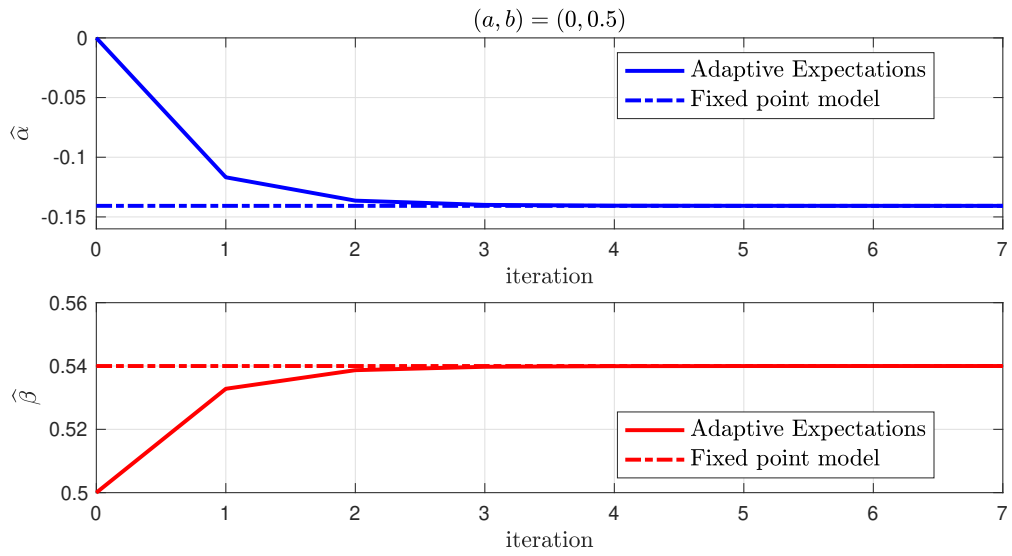

Figure 7: **Convergence of adaptive expectations to fixed point model**. We compare $\widehat{\alpha}$ and $\widehat{\beta}$ obtained by iterating the Goodhart estimation under adaptive expectations to the solution of fixed point model, under the logit prediction model (MLE) with DGP parameters $(a, b) = (0, 0.5)$. The fixed point model is $(\alpha^*, \beta^*) = (-0.14, 0.54)$. In each iteration, the Goodhart estimate is $(\widehat{\alpha}_n, \widehat{\beta}_n) = \arg\max_{\alpha,\beta} \mathcal{L}(\alpha, \beta; \widehat{\alpha}_{n-1}, \widehat{\beta}_{n-1}, a, b)$, for $n = 1, 2, \ldots$, where $(\widehat{\alpha}_0, \widehat{\beta}_0) = (a, b)$. We use the same assumptions as in Figure 3 for the default prediction model, manipulation cost and the remaining parameters.

A better understanding of Figure 7 can be gained by revisiting Figure 3. We note that Figure 3 captures the elements of the relevant Jacobian (Remark 3) at the fixed point. Of particular interest is the fact that, at least for these parameter values, the elements of the Jacobian are less than 1 in absolute value, consistent with the preceding sufficient conditions for fixed point convergence.

Notwithstanding the example in Figure 7, fixed point convergence is by no means guaranteed. After all, convergence requires that estimated coefficients be sufficiently insensitive to the coefficients of the posted model. However, we recall from Figure 4 that estimated coefficients will have high sensitivity to the posted model if borrowers are of low quality and/or have low manipulation costs. Indeed, Figure 8 shows that for the parameter values assumed in Figure 4, convergence fails. In fact, estimated coefficients jump sharply from round to round, with the oscillations becoming wider with more rounds of estimation.

Failure to converge might well be expected in Figure 8 if one accounts for the fact that the norm of the relevant Jacobian is:[14]

$$\|J\| = \left\| \begin{array}{cc} \frac{\partial T_1}{\partial \alpha} & \frac{\partial T_1}{\partial \beta} \\ \frac{\partial T_2}{\partial \alpha} & \frac{\partial T_2}{\partial \beta} \end{array} \right\| = \sqrt{\Lambda_{\max}\left(J^\top J\right)} = 2.0977,$$

[14]While there are several possible definitions of the norm, these definitions generate the same topology in the space of square matrices. For convenience, we use the spectral norm, $\|A\| = \sup\left\{\frac{\|Ax\|}{\|x\|}, x \neq 0\right\}$, which is the norm induced on the space of square matrices by the vector norm $\|\cdot\|$ on $\mathbb{R}^2$.

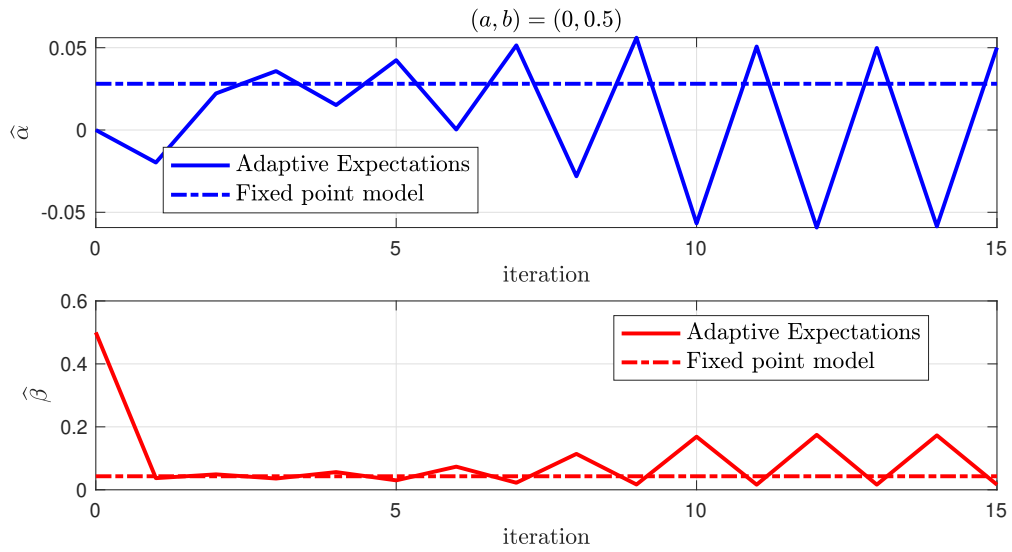

Figure 8: **A case of non-convergence to fixed point model**. We compare $\widehat{\alpha}$ and $\widehat{\beta}$ obtained by iterating the Goodhart estimation under adaptive expectations to the solution of fixed point model. We use the same parameters and model assumptions as in Figure 4.

where $\Lambda_{\max}(\cdot)$ is the largest eigenvalue. By way of contrast, the corresponding matrix norm for the example in Figure 7, with convergence, is only 0.2730.

6 Default Prediction with Commitment Power

The preceding section considered fixed point models as a potential response to Goodhart's law. Indeed, in contrast to clean data models ($a, b > 0$), fixed point models satisfy an attractive internal consistency standard, being ex post optimal responses to the distribution of data they generate. Nevertheless, the broader admonition of Goodhart is not an insistence upon internal consistency. Rather, the broader admonition is that the data will change if incentives change. But notice, the Nash econometrician in program (42) fails to take this argument fully on board. After all, the Nash econometrician effectively treats the distribution of reported covariates as given, rather than accounting for endogeneity of the data.

By way of contrast, a *commitment model* $(\alpha^{**}, \beta^{**})$ satisfies

$$(\alpha^{**}, \beta^{**}) \in \arg\max_{\alpha,\beta} \mathcal{L}(\underbrace{\alpha, \beta}_{\text{Estimated}} ; \underbrace{\alpha, \beta}_{\text{Posted}}, \underbrace{a, b}_{\text{DGP}}). \tag{48}$$

At this point it is instructive to contrast the Nash econometric program in equation (42) with the preceding commitment model program. Notice, in the Nash program, the econometrician takes as predetermined the posted model and resulting distribution of covariates. In stark contrast, in the

commitment program, the econometrician chooses 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, a commitment model satisfies the following FOCs:

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

That is, an econometrician with commitment power sacrifices 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 covariate reports, and these effects are captured by the partial derivatives $\mathcal{L}_3$ and $\mathcal{L}_4$. Phrased differently, with commitment power, the econometrician achieves a higher likelihood ratio than the econometrician posting a Nash model, but does so by way of adopting an econometric model that is ex post inefficient, violating $\mathcal{L}_1 = \mathcal{L}_2 = 0$. Turning next to signs, we recall that an increase in the posted model intercept (slope) tends to decrease (increase) manipulation. This would suggest that $\mathcal{L}_3 > 0$ and $\mathcal{L}_4 < 0$. In turn, one anticipates that the commitment model features a higher intercept ($\mathcal{L}_1 < 0$) and lower slope than a fixed point model ($\mathcal{L}_2 > 0$).

Denoting $m(x, \alpha^{**}, \beta^{**}, c)$ by m^{**}, the FOC for the commitment model intercept is:

$$\begin{aligned}&\int_C \int_X \begin{bmatrix} \frac{F(a+bx)}{F[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]} \\ -\frac{1-F(a+bx)}{1-F[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]} \end{bmatrix} f[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]h(x)g(c)dxdc \qquad (49) \\ = \; &-\int_C \int_X \begin{bmatrix} \frac{F(a+bx)}{F[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]} \\ -\frac{1-F(a+bx)}{1-F[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]} \end{bmatrix} f[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]\beta^{**}m_2(x, \alpha^{**}, \beta^{**}, c)h(x)g(c)dxdc.\end{aligned}$$

Or, more compactly:

$$0 = \int_C \int_X \begin{bmatrix} \left(\frac{F(a+bx)}{F[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]} - \frac{1-F(a+bx)}{1-F[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]}\right) \\ f[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}][1+\beta^{**}m_2(x, \alpha^{**}, \beta^{**}, c)]h(x)g(c) \end{bmatrix} dxdc. \qquad (50)$$

The FOC for the commitment model slope is:

$$\begin{aligned}&\int_C \int_X \begin{bmatrix} \frac{F(a+bx)}{F[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]} - \\ \frac{1-F(a+bx)}{1-F[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]} \end{bmatrix} f[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}][x+m(x, \alpha^{**}, \beta^{**}, c)]h(x)g(c)dxdc \\ = \; &-\int_C \int_X \begin{bmatrix} \frac{F(a+bx)}{F[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]} - \\ \frac{1-F(a+bx)}{1-F[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]} \end{bmatrix} f[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]\beta^{**}m_3(x, \alpha^{**}, \beta^{**}, c)h(x)g(c)dxdc\end{aligned}$$

Or more compactly:

$$0 = \int_C \int_X \begin{bmatrix} \left(\frac{F(a+bx)}{F[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]} - \frac{1-F(a+bx)}{1-F[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}]} \right) \\ f[\alpha^{**}+\beta^{**}x+\beta^{**}m^{**}][x+m^{**}+\beta^{**}m_3(x,\alpha^{**},\beta^{**},c)]h(x)g(c)dxdc \end{bmatrix}. \tag{51}$$

The following proposition follows directly upon inspecting the preceding FOCs.

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

Figure 9 contrasts fixed point coefficients with optimal coefficients under commitment, for alternative values of the clean data causal parameter b. As shown, the commitment model features a higher intercept than the fixed point model. Intuitively, the commitment model nudges borrowers away from manipulation by using a higher intercept. On the other hand, the commitment model features a lower slope, which induces less manipulation. It is also interesting to note that the gap between the commitment model and fixed point model is non-monotonic in b. We conjecture that as b gets larger, the paramount concern is capturing the predictive power of the true covariate, so that ex post efficiency of the fixed model takes precedence over nudging.

7 Multivariate Extension

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 are readily 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:

$$\Pr[y = 1|x, w] = \mathbb{E}[y|x, w] = a + bx + kw. \tag{52}$$

Assume that in contrast to x, the covariate $w \geq 0$ cannot be manipulated.

The objective is to find coefficients:

$$(\widehat{\alpha}, \widehat{\beta}, \widehat{\kappa}) \in \arg \min_{(\alpha,\beta,\kappa)} \mathbb{E}\left[(y - \alpha - \beta\widetilde{x} - \kappa w)^2 \right]. \tag{53}$$

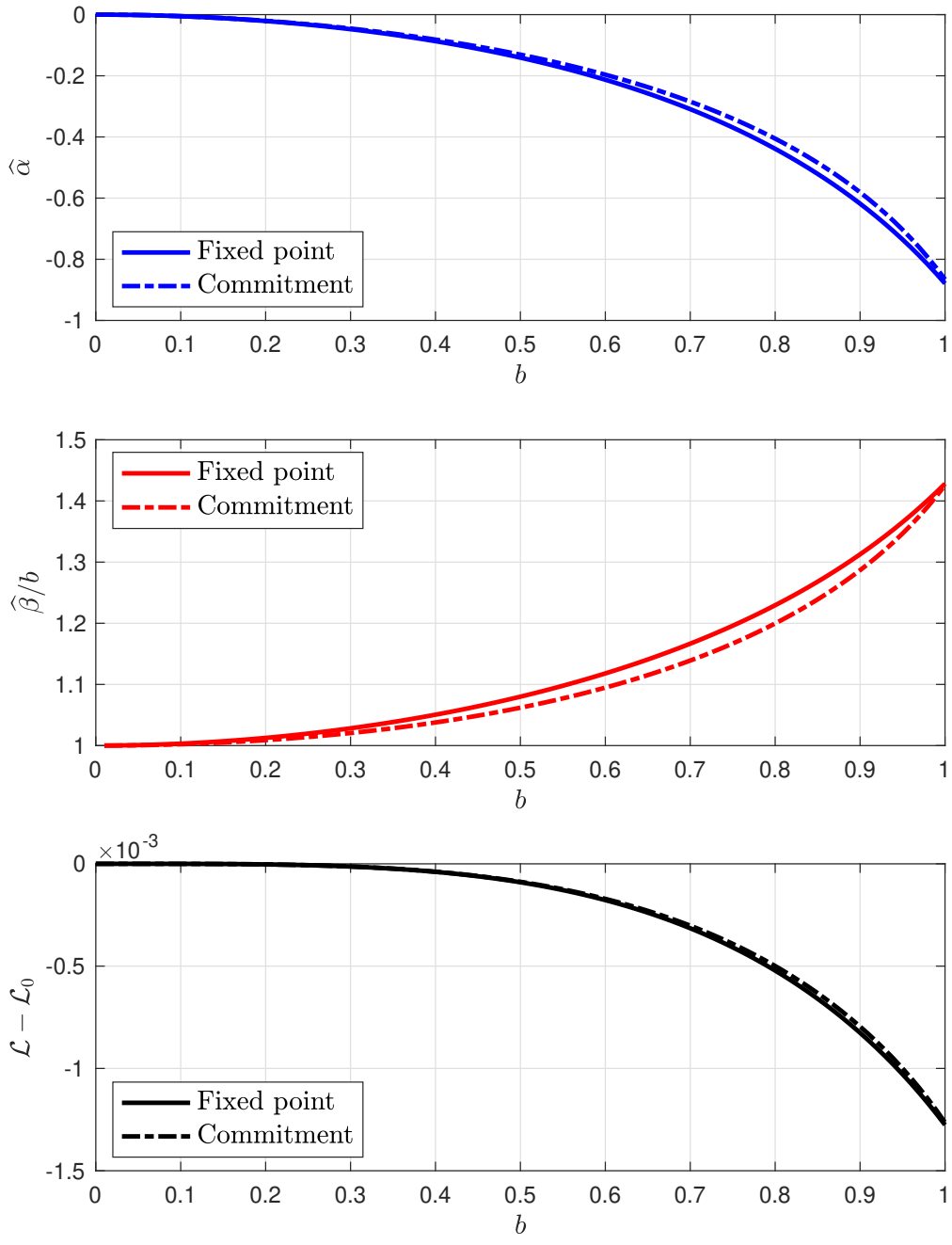

Figure 9: **Fixed point model and commitment model - Logit prediction model**. For $(\widetilde{\alpha}, \widetilde{\beta}) = (a, b)$, with $b \in]0, \overline{b}]$, we plot against b in the first panel the MLE estimate $\widehat{\alpha}$ and in the second panel $\widehat{\beta}/b$. We consider two alternative programs: the fixed point program, $(\alpha^*, \beta^*) = \arg\max_{(\alpha,\beta)} \mathcal{L}(\alpha, \beta; \alpha^*, \beta^*; a, b)$, and the program under commitment, $(\alpha^{**}, \beta^{**}) = \arg\max_{(\alpha,\beta)} \mathcal{L}(\alpha, \beta; \alpha, \beta; a, b)$. In the third panel we plot the difference between the optimal likelihood with manipulation (under commitment or fixed point), $\mathcal{L}$, and the optimal likelihood with no manipulation, $\mathcal{L}_0$. We use the same assumptions as in Figure 5 for the default prediction model, manipulation cost and the remaining parameters.

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

$$\begin{aligned} MSPE = \alpha^2 + (1-\alpha)\mathbb{E}[y] + 2\alpha\beta\mathbb{E}[\widetilde{x}] + 2\alpha\kappa\mathbb{E}[w] + 2\kappa\beta\mathbb{E}[w\widetilde{x}] \\ +\beta^2\mathbb{E}[\widetilde{x}^2] + \kappa^2\mathbb{E}[w^2] - 2\kappa\mathbb{E}[wy] - 2\beta\mathbb{E}[\widetilde{x}y]. \end{aligned} \tag{54}$$

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

$$\begin{aligned} \mathbb{E}\left\{\mathbb{E}[\widetilde{x}y|x,w]\right\} &= \mathbb{E}\left\{\mathbb{E}[\widetilde{x}|x,w]\mathbb{E}[y|x,w]\right\} \\ &= \mathbb{E}\left\{\mathbb{E}[\widetilde{x}|x,w](a+bx+kw)\right\} \\ &= a\mathbb{E}[\widetilde{x}] + b\mathbb{E}[\widetilde{x}x] + k\mathbb{E}[\widetilde{x}w]. \end{aligned} \tag{55}$$

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 (54) allows us to rewrite the MSPE as follows:

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

From the FOC for the intercept, we find:

$$\widehat{\alpha} = a + b\mathbb{E}[x] + k\mathbb{E}[w] - \widehat{\beta}\mathbb{E}[\widetilde{x}] - \widehat{\kappa}\mathbb{E}[w]. \tag{56}$$

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

$$0 = \widehat{\alpha}\mathbb{E}[\widetilde{x}] + \widehat{\kappa}\mathbb{E}[w\widetilde{x}] + \widehat{\beta}\mathbb{E}[\widetilde{x}^2] - \{a\mathbb{E}[\widetilde{x}] + b\mathbb{E}[\widetilde{x}x] + k\mathbb{E}[\widetilde{x}w]\}. \tag{57}$$

Substituting the expression for the intercept (56) into the FOC for $\widehat{\beta}$ we obtain:

$$\begin{aligned} 0 &= a\mathbb{E}[\widetilde{x}] + b\mathbb{E}[x]\mathbb{E}[\widetilde{x}] + k\mathbb{E}[w]\mathbb{E}[\widetilde{x}] - \widehat{\beta}\left(\mathbb{E}[\widetilde{x}]\right)^2 - \widehat{\kappa}\mathbb{E}[w]\mathbb{E}[\widetilde{x}] \\ &\quad +\widehat{\kappa}\mathbb{E}[w\widetilde{x}] + \widehat{\beta}\mathbb{E}[\widetilde{x}^2] - \{a\mathbb{E}[\widetilde{x}] + b\mathbb{E}[\widetilde{x}x] + k\mathbb{E}[\widetilde{x}w]\}. \end{aligned} \tag{58}$$

Rearranging terms we obtain:

$$\widehat{\beta} = b \times \beta_{ols}^{x\widetilde{x}} + (k - \widehat{\kappa})\beta_{ols}^{w\widetilde{x}}. \tag{59}$$

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

$$0 = \widehat{\alpha}\mathbb{E}[w] + \widehat{\beta}\mathbb{E}[w\widetilde{x}] + \widehat{\kappa}\mathbb{E}[w^2] - \{aE[w] + bE[xw] + kE[w^2]\}. \tag{60}$$

Substituting the expression for the intercept (56) into the FOC for $\widehat{\kappa}$ we obtain:

$$\begin{aligned} \widehat{\kappa} &= k + b \times \beta_{ols}^{xw} - \widehat{\beta} \times \beta_{ols}^{\widetilde{x}w} \\ &= k + (b - \widehat{\beta}) \times \beta_{ols}^{xw} - \widehat{\beta} \times \beta_{ols}^{mw}. \end{aligned} \tag{61}$$

From equations (56), (59) and (61), we have the following analog of Propositions 2 and 4, establishing the impossibility of getting something from nothing, with:

$$b = 0 \Rightarrow \left(\widehat{\alpha}, \widehat{\beta}, \widehat{\kappa}\right) = (a, b, k) = (a, 0, k). \tag{62}$$

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,

$$\widehat{\beta} \geq b > 0 \text{ and } \widehat{\kappa} \geq k \Rightarrow \widehat{\alpha} < a. \tag{63}$$

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.

8 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 $h(-s) + s > 0$.■

Lemma 2. *Suppose $b > 0$ and consider any posted model featuring $\widetilde{\beta} > 0$. If $\sigma_c^2 = 0$, then $\widehat{\beta}_{ols} > b$. If $\sigma_c^2 > 0$, then in the limit as σ_x^2* tends to 0, *$\widehat{\beta}_{ols} < b$.*

Proof.

The second result in the lemma follows from the fact that the slope coefficient can be rewritten as:

$$\widehat{\beta}_{ols} = b \times \rho_{x\widetilde{x}} \times \frac{\sigma_x^2}{\sigma_{\widetilde{x}}^2}.$$

For the first result, note that with c constant, the manipulated covariate is a univariate function of x. Moreover, an arbitrary point x_0:

$$\frac{d}{dx}\ \widetilde{x}(x_0) = 1 + m'(x_0) \in (0, 1).$$

Note that, in the preceding equation, the fact that $m' > -1$ follows from the fact that an agent with a lower true type cannot find it optimal to choose the same or higher report than a higher type, since the lower type faces the same marginal benefit to data manipulation point-wise, but faces higher marginal costs. Since the reported covariate is strictly monotone in x, we may apply the implicit function theorem and note that:

$$\frac{d}{d\widetilde{x}}\ x(\widetilde{x}_0) = \frac{1}{1 + m'[x(\widetilde{x}_0)]} > 1.$$

With this in mind, consider that $\widehat{\beta}^{x\widetilde{x}}$ represents the slope of the best linear approximation of x using $\widetilde{x}$. Suppose then to the contrary of the maintained assertion that $\widehat{\beta}^{x\widetilde{x}} < 1$. Then the assumed line of best fit would cross the function x at most one time, from above. This cannot be optimal since a positive rotation of the line originally posited would result in lower MSPE.■

References

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

[2] Baricz, Arpad, 2008, Mills' ratio: Monotonicity patterns and functional inequalities, *Journal of Mathematical Analysis and Applications*.

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

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

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

[9] 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.*

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

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

[12] Frankel, Alex and Navin Kartik, 2022. Improving information from manipulable data. Journal of the European Economic Association 20 (1), 79-115.

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

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

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

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

[17] Hennessy, Christopher A., and Charles A.E. Goodhart, 2023. Goodhart's law and machine learning: A Structural Perspective. *International Economic Review*.

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

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

[20] Pratt, John W., 1981, Concavity of the log likelihood, *Journal of the American Statistical Association* 76, 103-106.

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