BLS WORKING PAPERS

U.S. Department of Labor
U.S. Bureau of Labor Statistics
Office of Prices and Living Conditions

Another Look at the Linear Probability Model and Nonlinear Index Models

Kaicheng Chen,
Michigan State University
Robert S. Martin,
U.S. Bureau of Labor Statistics
Jeffrey M. Wooldridge,
Michigan State University

Working Paper 569
February 21, 2025

Another Look at the Linear Probability Model and Nonlinear Index Models

Kaicheng Chen* Robert S. Martin† Jeffrey M. Wooldridge‡

February 21, 2025

Abstract

We reassess the use of linear models for binary responses, focusing on average partial effects (APEs). We confirm that under certain conditions, linear projection parameters correspond to APEs even when the true model is nonlinear. Simulations demonstrate a large fraction of fitted values in [0, 1] is neither necessary nor sufficient for OLS to approximate the APEs. To reduce bias, excluding observations with fitted values outside [0, 1] has been proposed. We show that iteratively trimming the sample is equivalent to nonlinear least squares estimation of a piece-wise linear (ramp) model, for which we establish consistency and asymptotic normality results.

Keywords: Binary response; linear probability model; average partial effect; nonlinear least square; probit model.

JEL Classification Code: C25

*Department of Economics, Michigan State University. Email: chenka19@msu.edu
†Division of Price and Index Number Research, Bureau of Labor Statistics. Email: martin.robert@bls.gov
‡Department of Economics, Michigan State University. Email: wooldri1@msu.edu

1 Introduction

When an outcome variable, y, is binary, empirical researchers usually choose between two general strategies given a vector of (exogenous) explanatory variables, $\mathbf{x}$: (i) approximate the response probability, $P(y = 1|\mathbf{x})$, using a model linear in parameters or (ii) use a non-linear model, such as logit or probit. The first strategy is commonly known as using a *linear probability model* (LPM). The benefits of the LPM are well-known and include ease of interpretation and simple estimation. The shortcomings of the LPM are also well known and discussed in most introductory econometrics texts; see, for example, Wooldridge, 2019, Section 7.5. More advanced discussions of the LPM recognize that one should not take the linear model for $P(y = 1|\mathbf{x})$ literally but only as an approximation. The approximation can be exact in special cases—such as when $\mathbf{x}$ consists of binary indicators that are exhaustive and mutually exclusive—and it may be poor in other cases. However, for the most part, prediction is not the primary use of LPMs specifically or binary response models generally. Rather, researchers are largely interested in using binary response models to measure ceteris paribus or causal effects, and it is from this perspective that the LPM approximation should be evaluated. Angrist and Pischke, 2009, Section 3.4.1 and Wooldridge, 2010, Section 15.2 take this perspective. Wooldridge, 2010, Section 15.6, p. 579 shows how the results of Stoker (1986) can be applied to OLS estimation of the parameters in a LPM. Remarkably, there are situations where the linear projection exactly recovers the average partial effects (APEs) across a broad range of binary response models.[1]

Even though it is natural to study the LPM from the linear projection perspective, this opinion is not universally held. In an influential paper, Horrace and Oaxaca (2006) study both the bias and inconsistency of the OLS estimator for the parameters of an underlying piecewise linear model for the response probability that ensures the probabilities are in the unit interval.[2] The Horrace and Oaxaca paper is regularly cited in empirical research,[3] sometimes as a cautionary tale in using the LPM and sometimes as support for using the LPM when relatively few fitted values lie outside the unit interval. While Horrace and Oaxaca take the piecewise linear model seriously, much if not most of the citing literature seeks to use their results to choose between the LPM and an alternative like probit or logit.[4]

[1]Note, extensions of the LPM do not necessarily recover the APE. For instance, see Li et al. (2022) for the case of the LPM with endogenous $\mathbf{x}$ and two-stage least squares estimation.

[2]Horrace and Oaxaca (2006) defines the LPM as the piecewise linear ramp model. However, in this paper, we differentiate between the "ramp model" and the "LPM" (which is linear everywhere).

[3]In recent years (2020-2024), Horrace and Oaxaca (2006) has more than 300 Google Scholar citations.

[4]See, for example, Footnote 20 of van den Berg and Siflinger (2022).

In the current paper, we revisit the Horrace and Oaxaca framework but, rather than focus on parameters, we focus on APEs. We show that Horrace and Oaxaca set up the problem so that, in general, the response probability is nonlinear in the underlying linear index, $\mathbf{x}\boldsymbol{\beta} = \beta_1 + \beta_2 x_2 + \cdots + \beta_K x_K$:

$$P(y = 1|\mathbf{x}) = R(\mathbf{x}\boldsymbol{\beta}) = \begin{cases} 0, & x\beta \leq 0 \\ x\beta, & x\beta \in (0, 1). \\ 1, & x\beta \geq 1 \end{cases} \tag{1.1}$$

The nonlinear function $R(\cdot)$—sometimes called the ramp function—is piecewise linear and continuous, but it is not strictly increasing, and it is nondifferentiable at two inflection points. Nevertheless, under fairly weak assumptions, one can define the APEs. For continuous variables, the associated APEs are necessarily smaller in magnitude than the index slope coefficients in the underlying nonlinear model. Consequently, Horrace and Oaxaca's focus on index parameters rather than APEs is essentially the same as focusing on parameters in smooth response probabilities such as the logit and probit functions. Therefore, any conclusions about the usefulness of the LPM should be reexamined from the perspective of identifying APEs rather than coefficients.

It is important to understand that we are not advocating the ramp function as an especially sensible model of the response probability. Rather, we primarily study that specification from the perspective of average partial effects to determine how the Horrace and Oaxaca conclusions hold up. Briefly, in some cases, the linear projection parameters do a very good job of approximating the APEs even when a large percentage of the fitted values are outside the unit interval. Conversely, in other cases, the linear projection parameters do a very poor job of approximating the APEs even when a high percentage of the fitted values are within the unit interval. A practical implication is that there is little justification for how the Horrace and Oaxaca paper is cited in empirical research.

We compare the OLS estimation of the LPM to a few nonlinear competitors, including probit and logit quasi-maximum likelihood estimation (QMLE), as natural benchmarks. Horrace and Oaxaca cite a few theoretical rationalizations for the ramp model, so it also makes sense to see if a consistent estimator exists that takes it seriously. Horrace and Oaxaca suggest trimming the sample of fitted values outside the unit interval and re-estimating using OLS, but do not present any theoretical or simulation results. In unreported simulations, we found that trimming the sample once did not necessarily improve performance over OLS

for estimating the APEs. Interestingly, by iteratively trimming the sample and performing OLS estimation (referred to as the ITO procedure, hereafter), we show it produces results equivalent to those from numerically minimizing the nonlinear least square (NLS) objective function with the ramp model. In Section 3, we show that the NLS estimator of the ramp model is consistent and asymptotically normal under mild assumptions, which in turn justifies trimming procedures in practice. For estimating the APEs, we find that NLS estimation of the ramp function performs comparably to quasi-MLE estimation of the logit and probit models and has good finite sample properties even when OLS estimation of the LPM does not.

Section 2 delivers our main theoretical arguments. Starting with a linear index model as the response probability of a binary outcome, we define and contrast parameters of interest, which are the index slope coefficients, average partial effects, and linear projection parameters. By leveraging results from Stoker (1986), we describe scenarios where the linear projection parameters recover APEs. In particular, we extend the discussion in Wooldridge, 2010, Section 15.6 and show that, when the covariates have a multivariate normal distribution, the linear projection identifies the APEs. Section 4 continues with the mission by conducting simulations under scenarios where theory has made predictions and where theory suggests, but does not fully uncover the relations. Related to our main theoretical arguments, we show that a large fraction of fitted values in $[0, 1]$ is neither sufficient nor necessary condition for the LPM to well-approximate the APEs. We revisit an empirical study of mortgage lending decisions in Section 5. The LPM estimated by OLS, with a full set of interactions between the variable of interest and the control variables, delivers a notably smaller and marginally statistically significant estimate of the effect of being white on the approval probability. The NLS estimator of the ramp function, probit QMLE, and logit QMLE are very similar and all statistically significant at the 0.2% level—both because the estimated effects are larger but also because the (robust) standard errors are notably smaller. In Section 6, we conclude with some implications for empirical research.

2 Relevant Parameters of Binary Response Models

Let y be the binary outcome variable and $\mathbf{x}$ the $1 \times K$ vector of explanatory variables, where $x_1 \equiv 1$ allows for an intercept in the index. Consider a linear index model of the response probability for y:

$$P\left(y=1|\mathbf{x}\right)\equiv p(\mathbf{x})=G\left(\mathbf{x}\boldsymbol{\beta}\right)=G\left(\beta_1+\beta_2x_2+\cdots+\beta_Kx_K\right) \tag{2.1}$$

where $G:\mathbb{R}\to[0,1]$. This embeds the probit/logit model by setting $G(\cdot)$ as the standard normal CDF/standard logistic function, and it includes the ramp model in Horrace and Oaxaca (2006) by setting $G(\cdot)=R(\cdot)$ as in 1.1.

The following subsection compares different parameters of interest for the linear index model generally and for the ramp model specifically. The ramp model for the response probability was suggested by Horowitz and Savin (2001) as being suitable when one starts with a linear model for $p\left(\mathbf{x}\right)$ but wants to ensure that the probabilities are within the unit interval. While not necessarily advocating this view, our purpose is to show that Horrace and Oaxaca's conclusions about one set of parameters ($\boldsymbol{\beta}$) do not necessarily apply to the most interesting set of parameters (the APEs).

2.1 APEs, Index Slopes, and Linear Projection Parameters

We will first consider partial effects for continuous variables. Let x_j be a continuously distributed explanatory variable. For simplicity, the discussion here assumes that x_j appears only by itself. If the model includes quadratics, interactions, and so on then the details become more complicated but the conclusions do not change substantively. Assume $G(\cdot)$ is differentiable almost everywhere, with its derivative denoted by $g(\cdot)$. Then, the partial effects and average partial effects of x_j on the response probability of y can be defined as:

$$\mathrm{PE}_j(\mathbf{x})\equiv\beta_jg\left(\mathbf{x}\boldsymbol{\beta}\right),\quad \mathrm{APE}_j\equiv\beta_j\mathrm{E}\left[g\left(\mathbf{x}\boldsymbol{\beta}\right)\right]$$

In the case of the ramp function, even though $R\left(z\right)$ is non-differentiable at $z=0$ and $z=1$, it is still differentiable with probability one as long as $\mathbf{x}\boldsymbol{\beta}$ is continuous, and so $P\left(\mathbf{x}\boldsymbol{\beta}=0\right)=P\left(\mathbf{x}\boldsymbol{\beta}=1\right)=0$. This holds true provided that at least one element of $\mathbf{x}$ is continuous, and that element has a nonzero coefficient, which is a very common assumption imposed in the semiparametric literature on binary response models. In what follows, we maintain that $\mathbf{x}\boldsymbol{\beta}$ is continuous so that partial effects are well-defined with probability one. As a result, we can define a partial effect function as the derivative of $R\left(\mathbf{x}\boldsymbol{\beta}\right)$ and ignore points where the derivative does not exist:

$$\mathrm{PE}_j\left(\mathbf{x}\right)=\frac{\partial p}{\partial x_j}\left(\mathbf{x}\right)=\beta_j1\left[0\leq\mathbf{x}\boldsymbol{\beta}\leq1\right],$$

where $1[\cdot]$ is the indicator function. Therefore, under the ramp model, the APE is

$$\text{APE}_j \equiv \text{E}[\text{PE}_j(\mathbf{x})] = \beta_j P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1). \tag{2.2}$$

There are some simple but useful observations about (2.2). First, similar to the probit or logit cases, APE_j always has the same sign as β_j. Second, because $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) \leq 1$, $|\text{APE}_j| \leq |\beta_j|$; with wide support for $\mathbf{x}\boldsymbol{\beta}$, APE_j can be much smaller in magnitude than β_j. Moreover, $\text{APE}_j = \beta_j$ if and only if $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) = 1$, which means the support of $\mathbf{x}\boldsymbol{\beta}$ is inside the unit interval. This is essentially the condition used by Horrace and Oaxaca (2006) to conclude that the OLS estimator in linear regression is unbiased and consistent for $\boldsymbol{\beta}$. Our goal here is to compare the OLS estimators with the APEs in the general case where $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) < 1$; the Horrace and Oaxaca condition is then a special case where the index coefficient, β_j, is identical to APE_j.

In order to understand the behavior of the OLS estimator under a linear index model, it is important to introduce a third set of parameters: the linear projection parameters, denoted as $\boldsymbol{\gamma}$. Assume that the x_j have finite second moments and that the $K \times K$ matrix $\text{E}(\mathbf{x}'\mathbf{x})$ is nonsingular. Then we can always define the $K \times 1$ vector $\boldsymbol{\gamma}$ as

$$\boldsymbol{\gamma} = [\text{E}(\mathbf{x}'\mathbf{x})]^{-1} \text{E}(\mathbf{x}'y).$$

We then write the linear projection of y on $(1, x_2, ..., x_K)$ as.

$$L(y|\mathbf{x}) = L(y|1, x_2, ..., x_K) = \gamma_1 + \gamma_2 x_2 + \cdots + \gamma_K x_K = \mathbf{x}\boldsymbol{\gamma}.$$

In understanding the findings in Horrace and Oaxaca, and their limitations, it is important to know that, given the model of 2.1, APE_j, β_j, and γ_j are all well-defined parameters and, in general, they are all different. Defining $\boldsymbol{\beta}$ and the APEs require an underlying model for the response probability whereas defining $\boldsymbol{\gamma}$ does not.

As is well known, under random sampling the OLS estimator consistently estimates the parameters of the linear projection; see, for example, Wooldridge (2010, Chapter 4.2). In other words, if we run the OLS regression underlying LPM estimation,

$$y_i \text{ on } 1,\ x_{i2},\ ...,\ x_{iK},\ i = 1, ..., N,$$

and obtain the $\hat{\gamma}_j$, then $\hat{\gamma}_j \xrightarrow{p} \gamma_j$. Again, this result holds free of any kind of underlying

model.

Under the ramp model, Horrace and Oaxaca study the consistency of the $\hat{\gamma}_j$ when considered as estimators of β_j—the coefficients in the index. In other words, their asymptotic analysis is the same as comparing the linear projection parameters γ_j to the index parameters β_j. Our view is that this does usually not make much sense—for the same reason, we do not study the consistency of the OLS estimator for the index parameters in, say, probit or logit. If one explicitly models the response probability as a nonlinear function of $\mathbf{x}\boldsymbol{\beta}$ then one must recognize that nonlinearity when defining the parameters of interest. When interest is in the effects of the explanatory variables on the response probability—which describes almost all modern usages of the LPM—it only makes sense to compare the linear projection parameters to the APEs. In other words, we should ask: When is γ_j "close" to APE_j? This is not the same as studying when γ_j is "close" to β_j (except in the special case where $P\left(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1\right)$ holds).

Under the ramp model we can write

$$\mathrm{E}\left(y|\mathbf{x}\right) = p\left(\mathbf{x}\right) = 1\left[0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1\right]\mathbf{x}\boldsymbol{\beta} + 1\left[\mathbf{x}\boldsymbol{\beta} > 1\right].$$

If $P\left(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1\right)$ holds then, with probability one,

$$\mathrm{E}\left(y|\mathbf{x}\right) = \mathbf{x}\boldsymbol{\beta} = L\left(y|\mathbf{x}\right),$$

in which case $\text{APE}_j = \beta_j = \gamma_j$ and so the OLS estimators, $\hat{\gamma}_j$ are consistent for β_j and APE_j. If for a random sample of size N, $\mathbf{x}_i\boldsymbol{\beta} \in [0,1]$ for all i, then

$$\mathrm{E}\left(y_i|\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\right) = \mathbf{x}_i\boldsymbol{\beta},$$

and it follows that the OLS estimators are conditionally unbiased for the β_j – the conclusion reached in Horrace and Oaxaca.

If $P\left(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1\right) < 1$, then the β_j measure the partial effects when $0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1$, but this restriction depends on the unknown vector and the β_j need not be very useful as summary measures of the partial effects. In the next subsection, we discuss more generally when the linear projection parameters identify the APEs.

2.2 When Linear Projection Recovers the APEs

In addition to being easy to interpret, empirically, the OLS estimates of the LPM are often similar to the corresponding APEs from nonlinear index models—particularly logit or probit. Wooldridge, 2010, Section 15.6 provides a discussion based on a results of Stoker (1986) that helps one understand these empirical findings. Here we expand that discussion to allow for an extension to the ramp model.

As argued in Wooldridge, 2010, Section 15.6, the results of Stoker (1986) imply that, if $(x_2, ..., x_K)$ has a multivariate normal distribution and $G(\cdot)$ is differentiable almost everywhere on $\mathbb{R}$ (with respect to Lebesgue measure), then

$$\gamma_j = \beta_j \mathrm{E}\left[g\left(\mathbf{x}\boldsymbol{\beta}\right)\right] = \mathrm{APE}_j,\ j = 2, ..., K,$$

where γ_j is the slope coefficients on x_j in $L\left(y|\mathbf{x}\right) = \mathbf{x}\boldsymbol{\gamma}$, $g\left(\cdot\right)$ is the almost everywhere derivative of $G\left(\cdot\right)$. The ramp function $R\left(\cdot\right)$ is differentiable everywhere except at zero and one, and so it satisfies Stoker's (1986) assumptions. The result is that OLS consistently estimates APE_j, even though the APE_j are attenuated versions of the β_j. This equality holds even when $P\left(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1\right)$ is very close to zero. Horrace and Oaxaca (2006), and many papers citing their findings, focus on the inconsistency of OLS for β_j, failing to recognize that the OLS estimators from the linear model could be consistent for the more interesting quantities, the APE_j. This point is key to our argument: If the model of the response probability is nonlinear so that $0 \leq p\left(\mathbf{x}\right) \leq 1$ is ensured, one should study estimation of APEs, not underlying index parameters.

Other than the case of multivariate normality of $(x_2, ..., x_K)$, there is another case where the linear projection parameters, γ_j, $j = 2, ..., K$, equal the APEs: x_2, ..., x_K are mutually exclusive binary indicators that, along with a base group given by $x_2 = x_3 = \cdots = x_K = 0$, are exhaustive. See Angrist and Pischke, 2009, Section 3.1.4 and Wooldridge, 2010, Section 15.2. If $x_1 = 1$ denotes the base group then the APEs are simply

$$\mathrm{APE}_j = \mathrm{E}\left(y|x_j = 1\right) - \mathrm{E}\left(y|x_1 = 1\right),\ j = 2, ..., K,$$

and these are identical to the corresponding LPM coefficients.

2.3 More General Cases

Clearly the assumption of multivariate normality of $\mathbf{x}$ is too restrictive to be widely applicable. Nevertheless, the results of Stoker (1986) are suggestive, especially when combined with Ruud (1983). Ruud studies smooth nonlinear function forms that never hit the endpoints of the unit interval, like probit and logit. In these cases, quasi-MLE identifies the index coefficients up to scale.[5] If $\mathbf{x}$ has a centrally symmetric distribution—of which the multivariate normal is a special case—then Ruud's (1983) conditions hold. In Section 4, we will find several cases where the covariates are symmetrically distributed (but not multivariate normal) and the APEs are still approximated well by the linear projection parameters.

Beyond the extreme cases described here, there appears to be no general theory to determine when the linear projection coefficients will be the same or "close" to the APEs. Many empirical applications include a combination of continuous, discrete, and even mixed explanatory variables. Rarely do these all have marginal symmetric distributions, let alone a symmetric joint distribution. Plus, such explanatory variables often appear as quadratics, interactions, and other functional forms—which also do not have symmetric distributions. In Section 4, we use simulations to shed light on when the LPM coefficients closely approximate the APEs—and when they do not. When evaluating the performance of the LPM as an approximation to Horrace and Oaxaca's ramp model, it makes sense to consider an estimator which takes such a model seriously. To that end, the next section describes such an estimator.

3 Asymptotically Valid Estimators of the Ramp Model

3.1 Nonlinear Least Square Estimation

We have already seen how if $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) = 1$, then OLS is consistent for the β_j, which are equal to APE_j in the case of a continuous covariate x_j under model (1.1). If the probability that $\mathbf{x}\boldsymbol{\beta}$ lies outside the unit interval is nonzero, then OLS is no longer consistent for the β_j, and it may or may not approximate the APE_j depending on the distribution of $\mathbf{x}$. In addition to probit and logit quasi-MLE, it makes sense to consider an estimator which is consistent if the ramp model is true. Of course, Bernoulli MLE in the usual fashion using the ramp model as the conditional response probability is not feasible because the log-likelihood

[5] Li et al. (2022) discuss this further and show that in the case of a single normal covariate, logit quasi-MLE identifies the APE, but probit quasi-MLE does not.

is not necessarily defined for $\mathbf{x}\boldsymbol{\beta} \notin (0,1)$. Instead, we consider nonlinear least squares (NLS) using the piecewise ramp function $R(\mathbf{x}\boldsymbol{\beta})$ from (1.1) as the conditional mean. In addition, since there may not be much justification to think the ramp function is the true response probability, we allow for general misspecification. Therefore, we define $\boldsymbol{\beta}_o$ as the pseudo-true value in the sense that $\boldsymbol{\beta}_o$ is the unique solution to

$$\min_{\boldsymbol{\beta}} \mathrm{E}\left[(y - R(\mathbf{x}\boldsymbol{\beta}))^2\right] \equiv \min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta}). \tag{4.1}$$

We say that the model is misspecified if there is no such $\boldsymbol{\beta}$ such that $E[y|\mathbf{x}] = R(\mathbf{x}\boldsymbol{\beta})$. By construction, $\boldsymbol{\beta}_o$ is the true coefficient when the model is correctly specified and otherwise we view $R(\mathbf{x}\boldsymbol{\beta}_o)$ as the best mean squared error approximation to $E[y|\mathbf{x}]$ over all ramp functions $R(\mathbf{x}\boldsymbol{\beta})$.

Assume a random sample indexed by $i = 1, ..., N$. As a sample analogue of (4.1), we define the objective function $Q_N(\boldsymbol{\beta})$ as

$$\begin{aligned} Q_N(\boldsymbol{\beta}) \equiv & \frac{1}{N}\sum_{i=1}^{N} (y_i - R(\mathbf{x}_i\boldsymbol{\beta}))^2 \\ = & \frac{1}{N}\sum_{i=1}^{N} \left(y_i^2 1\left\{\mathbf{x}_i\boldsymbol{\beta} \leq 0\right\} + (y_i - \mathbf{x}_i\boldsymbol{\beta})^2 1\left\{\mathbf{x}_i\boldsymbol{\beta} \in (0,1)\right\} + (y_i - 1)^2 1\left\{\mathbf{x}_i\boldsymbol{\beta} \geq 1\right\}\right), \end{aligned}$$

where N is the sample size. We define the NLS estimator $\hat{\boldsymbol{\beta}}$ as

$$\hat{\boldsymbol{\beta}} \equiv \operatorname*{argmin}_{\boldsymbol{\beta}} \ Q_N(\boldsymbol{\beta}).$$

The following theorem gives the consistency of the NLS estimator for the pseudo-true value, allowing for misspecification of the conditional mean model.

Theorem 1. *Let $\{y_i, \mathbf{x}_i\}_{i=1}^{\infty}$ be an i.i.d. sequence with y only taking on values zero and one, and let $R : \mathbb{R} \to [0,1]$ be the ramp function defined in (1.1). Suppose $\boldsymbol{\beta} \in \mathcal{B}$ such that $\mathcal{B} \subset \mathbb{R}^K$ is compact, and $\boldsymbol{\beta}_o$ is identified in the sense that $\forall\, \boldsymbol{\beta} \in \mathcal{B}, \boldsymbol{\beta} \neq \boldsymbol{\beta}_o$,*

$$\mathrm{E}\left[(y_i - R(\mathbf{x}_i\boldsymbol{\beta}_o))^2\right] < \mathrm{E}\left[(y_i - R(\mathbf{x}_i\boldsymbol{\beta}))^2\right]$$

Then, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_o$ as $N \to \infty$.

The consistency result of Theorem 1 follows directly from Theorem 12.2 of Wooldridge

(2010).

If $\mathbf{x}$ contains a continuously distributed x_j and β_{jo} is nonzero, then the probability of $\mathbf{x}_i\boldsymbol{\beta}_o$ being equal to 0 or 1 is zero. Then, with suitable moment conditions on $\mathbf{x}$ (so the Leibniz integral rule applies), the FOC of the (4.1) is well defined with probability 1 as follows:

$$\mathrm{E}\left[\mathbf{x}_i' u_i 1\{\mathbf{x}_i\boldsymbol{\beta}_o \in (0,1)\}\right] = 0, \tag{4.2}$$

where $u_i(\boldsymbol{\beta}) = y_i - R(\mathbf{x}_i\boldsymbol{\beta})$ and $u_i \equiv u_i(\boldsymbol{\beta}_0)$. Define the score function for random draw i:

$$\mathbf{s}_i(\boldsymbol{\beta}) = -\mathbf{x}_i' u_i(\boldsymbol{\beta}) 1\{\mathbf{x}_i\boldsymbol{\beta} \in (0,1)\}.$$

Then, $\boldsymbol{\beta}_o$ solves $E[\mathbf{s}_i(\boldsymbol{\beta}_o)] = 0$. The variance-covariance matrix of $\mathbf{s}_i(\boldsymbol{\beta})$ is

$$\boldsymbol{\Omega}(\boldsymbol{\beta}) = \mathrm{E}\left[\mathbf{x}_i'\mathbf{x}_i u_i(\boldsymbol{\beta})^2 1\left\{\mathbf{x}_i\boldsymbol{\beta} \in (0,1)\right\}\right]. \tag{4.3}$$

The natural definition of the Jacobian of $\mathbf{s}_i(\boldsymbol{\beta})$ is

$$\mathbf{A}_i(\boldsymbol{\beta}) = \mathbf{x}_i'\mathbf{x}_i 1\left\{\mathbf{x}_i\boldsymbol{\beta} \in (0,1)\right\}.$$

For the similar reason as (4.2), the Hessian of $Q(\boldsymbol{\beta})$ is well-defined with probability 1 at $\boldsymbol{\beta}_o$ as follows

$$\mathbf{A}(\boldsymbol{\beta}_o) = \mathrm{E}\left[\mathbf{x}_i'\mathbf{x}_i 1\left\{\mathbf{x}_i\boldsymbol{\beta}_o \in (0,1)\right\}\right]. \tag{4.4}$$

Note that (4.3) and (4.4) are the same whether the conditional mean model is correctly specified or not. Therefore, the following asymptotic distribution result allows for misspecification of the model.

Theorem 2. *Suppose that the assumptions from Theorem 1 hold, and (i) $\boldsymbol{\beta}_o$ is an interior point of $\mathcal{B}$; (ii) $\mathbf{x}_i$ contains a continuously distributed random variable with a nonzero coefficient; (iii) $\mathrm{E}\left\|\mathbf{x}_i\right\|^2 < \infty$ and $\mathrm{E}\left[\mathbf{x}_i'\mathbf{x}_i 1\left\{\mathbf{x}_i\boldsymbol{\beta}_o \in (0,1)\right\}\right] > 0$, where $\|.\|$ denotes the $l^2 - norm$. Then, as $N \to \infty$,*

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o\right) \xrightarrow{d} \mathbb{N}(0, \mathbf{A}(\boldsymbol{\beta}_o)^{-1}\boldsymbol{\Omega}(\boldsymbol{\beta}_o)\mathbf{A}(\boldsymbol{\beta}_o)^{-1}).$$

The proof of Theorem 2 is given in the Appendix. The asymptotic normality results

does not follow directly from the M-estimator due to the non-smoothness of the objective function. We therefore leverage an asymptotic normality result for estimators with non-smooth objective function from Newey and McFadden (1994).

Taking the sample analogue of the asymptotic variance from Theorem 2, we define a variance estimator of $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o)$ as

$$\hat{\mathbf{V}} = \mathbf{A}_N(\hat{\boldsymbol{\beta}})^{-1}\boldsymbol{\Omega}_N(\hat{\boldsymbol{\beta}})\mathbf{A}_N(\hat{\boldsymbol{\beta}})^{-1},$$

where $\mathbf{A}_N(\hat{\boldsymbol{\beta}}) = N^{-1}\sum_{i=1}^N \mathbf{x}_i'\mathbf{x}_i 1\{\mathbf{x}_i\hat{\boldsymbol{\beta}} \in (0,1)\}$, $\boldsymbol{\Omega}_N(\hat{\boldsymbol{\beta}}) = N^{-1}\sum_{i=1}^N \mathbf{x}_i'\mathbf{x}_i\hat{u}_i^2 1\{\mathbf{x}_i\hat{\boldsymbol{\beta}} \in (0,1)\}$, and $\hat{u}_i = y_i - R(\mathbf{x}_i\hat{\boldsymbol{\beta}})$. Standard errors are obtained the usual way from $\hat{\mathbf{V}}/N$. The next theorem gives the consistency result of the variance estimator.

Theorem 3. *Under the same assumption of Theorem 2 and* $\mathrm{E}\|x\|^4 < \infty$*, as* $N \to \infty$*,* $\hat{\mathbf{V}} \xrightarrow{p} \mathbf{A}(\boldsymbol{\beta}_o)^{-1}\boldsymbol{\Omega}(\boldsymbol{\beta}_o)\mathbf{A}(\boldsymbol{\beta}_o)^{-1}$.

The proof of Theorem 3 is given in the Appendix. As before, we are interested in the APE. Consider the best ramp approximation in (4.1), the APE of a continuous random variable x_k is defined as

$$\mathrm{APE}_k = \mathrm{E}\left[\frac{\partial R(\mathbf{x}_i\boldsymbol{\beta}_o)}{\partial x_k}\right] = \beta_{ko} P\left(\mathbf{x}_i\boldsymbol{\beta}_o \in (0,1)\right).$$

A sample-analogue estimator of the APE is then given by

$$\widehat{\mathrm{APE}}_k = \hat{\beta}_k \frac{1}{N}\sum_{i=1}^N 1\left\{\mathbf{x}_i\hat{\boldsymbol{\beta}} \in (0,1)\right\}$$

Define $g(\mathbf{x}_i, \boldsymbol{\beta}) = \beta_k 1\{\mathbf{x}_i\boldsymbol{\beta} \in (0,1)\}$, $\delta_o = E[g(\mathbf{x}_i, \boldsymbol{\beta}_o)]$, and $\mathbf{G}_o = \nabla_{\boldsymbol{\beta}} g(\mathbf{x}_i, \boldsymbol{\beta}_o)$. Following problem 12.17 of Wooldridge (2010), the asymptotic variance of the estimated APE is given by

$$\mathrm{AVar}\left(\sqrt{N}\left(\widehat{\mathrm{APE}}_k - \mathrm{APE}_k\right)\right) = \mathrm{Var}\left(g(\mathbf{x}_i, \boldsymbol{\beta}_o) - \delta_o - \mathbf{G}_o\mathbf{A}(\boldsymbol{\beta}_o)^{-1}\mathbf{s}_i(\boldsymbol{\beta}_o)\right),$$

where $\mathbf{G}_o$ is a $1 \times K$ vector with the k^{th} element being $p_o \equiv P\left(\mathbf{x}_i\boldsymbol{\beta}_o \in (0,1)\right)$ and all else 0. The asymptotic variance can be estimated by the sample variance of $g(\mathbf{x}_i, \hat{\boldsymbol{\beta}}) - \hat{\delta} - \widehat{\mathbf{G}}\mathbf{A}_N(\hat{\boldsymbol{\beta}})^{-1}\mathbf{s}_i(\hat{\boldsymbol{\beta}})$, where $\hat{\delta} = \frac{1}{N}\sum_{i=1}^N g(\mathbf{x}_i, \hat{\boldsymbol{\beta}})$, $\widehat{\mathbf{G}}$ is a $1 \times K$ vector with the k^{th} element being $\hat{p} = \frac{1}{N}\sum_{i=1}^N 1\left\{\mathbf{x}_i\hat{\boldsymbol{\beta}} \in (0,1)\right\}$.

The APE for a discrete random variable x_k can be defined as

$$\text{APE}_k = \text{E}\left[R(\mathbf{x}_{i,-k}\boldsymbol{\beta}_{-ko} + \beta_{ko}) - R(\mathbf{x}_{i,-k}\boldsymbol{\beta}_{-ko})\right].$$

A sample analogue estimator of APE_k is given by

$$\widehat{\text{APE}}_k = \frac{1}{N}\sum_{i=1}^{N} R(\mathbf{x}_{i,-k}\hat{\boldsymbol{\beta}}_{-k} + \hat{\beta}_k) - R(\mathbf{x}_{i,-k}\hat{\boldsymbol{\beta}}_{-k}).$$

The asymptotic variance can be found and estimated in a similar manner as the continuous case.

3.2 Iterative Trimming OLS Estimation

To estimate $\boldsymbol{\beta}_o$, Horrace and Oaxaca suggest running OLS on a trimmed sample (i.e., those observations for which initial OLS fitted values are inside the unit interval) to reduce bias. We find in practice that a single round of trimming may not reduce the bias for the APEs in the cases where OLS is not consistent for them. However, we find an iterative trimming OLS procedure (ITO) does reduce the bias for estimating APEs, as well as $\boldsymbol{\beta}_o$. The procedure goes: 1) estimate the LPM by OLS. 2) Compute fitted values. 3) Drop observations with fitted values outside the unit interval, and 4) Repeat starting at 1) until no further observations are dropped. In fact, we find in simulations that the NLS estimates are numerically the same as the ITO estimates up to machine precision.[6] It turns out that ITO is implicitly minimizing the NLS sample objective function using the OLS estimates as starting values and following the Newton-Raphson numerical method, which is iterative (see Wooldridge, 2010, Section

[6]With some DGP, it was occasionally necessary to specify OLS starting values for the NLS function evaluator for the NLS and ITO estimates to match to machine precision. The two were still otherwise very close.

12.7.1). Given an estimate $\boldsymbol{\beta}^{\{g\}}$, the next iteration is given (using our notation) by

$$\begin{aligned}
\boldsymbol{\beta}^{\{g+1\}} &= \boldsymbol{\beta}^{\{g\}} - \left[N^{-1} \sum_{i=1}^{N} \mathbf{A}_i(\boldsymbol{\beta}^{\{g\}}) \right]^{-1} N^{-1} \sum_{i=1}^{N} \mathbf{s}_i(\boldsymbol{\beta}^{\{g\}}) \\
&= \boldsymbol{\beta}^{\{g\}} + \left[N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i 1\left\{ \mathbf{x}_i \boldsymbol{\beta}^{\{g\}} \in (0,1) \right\} \right]^{-1} N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \left(y_i - \mathbf{x}_i \boldsymbol{\beta}^{\{g\}} \right) 1\left\{ \mathbf{x}_i \boldsymbol{\beta}^{\{g\}} \in (0,1) \right\} \\
&= \left[N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' \mathbf{x}_i 1\left\{ \mathbf{x}_i \boldsymbol{\beta}^{\{g\}} \in (0,1) \right\} \right]^{-1} N^{-1} \sum_{i=1}^{N} \mathbf{x}_i' y_i 1\left\{ \mathbf{x}_i \boldsymbol{\beta}^{\{g\}} \in (0,1) \right\}.
\end{aligned}$$

The second equality above substitutes our expressions for $\mathbf{s}_i(\boldsymbol{\beta})$ and $\mathbf{A}_i(\boldsymbol{\beta})$ and uses the fact that $R(\mathbf{x}_i\boldsymbol{\beta}) = \mathbf{x}_i\boldsymbol{\beta}$ for $\mathbf{x}_i\boldsymbol{\beta} \in (0,1)$. This shows that $\boldsymbol{\beta}^{\{g+1\}}$ is simply the OLS estimator on the sample with $\mathbf{x}_i\boldsymbol{\beta}^{\{g\}} \in (0,1)$.

As a consequence, the preceding consistency and asymptotic normality results for the NLS estimator justify using the ITO procedure to reduce the OLS bias. However, it is worth mentioning that, at least in Stata, the pre-loaded NLS solver (the"nl" command) may have a performance advantage over ITO in practice. We find in simulations that ITO can result in a dead loop when only a very small portion of observations are left for estimation after iterative trimming. The pre-loaded NLS algorithm continues to work well in those cases.

4 Simulations

In this section, we present several Monte Carlo simulations that provide insights into the behavior of different modeling/estimation approaches. The LPM is estimated by OLS and the ramp function is estimated by NLS. For the LPM, the APE estimates come directly from the linear projection (e.g., the estimated slope coefficient for a non-interacted variable). For the ramp model, the APEs are estimated using averages of derivatives and differences of the ramp function, as discussed in Section 2. These resemble the familiar formulas for the linear model, though the individual unit partial effects need to be scaled by $1\left[0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{NLS} \leq 1\right]$ before averaging, where $\widehat{\boldsymbol{\beta}}_{NLS}$ corresponds to the NLS slope estimate. The logit and probit parameters are estimated by the (quasi-) maximum likelihood estimator, and then the APEs are estimated using the usual APE formulas. We used Stata®17 for simulation. [7].

[7]The Stata code is available via the repository `https://kaichengchen.github.io/lpm_simulation_post.rar`

To better evaluate the findings from Horrace and Oaxaca (2006), we generate the responses to follow the ramp model for their true conditional probabilities. We also show that our main arguments hold when the true responses are probit. It is useful to observe that the response probability can be derived from a latent variable formulation:

$$y^* = \mathbf{x}\boldsymbol{\beta} - u, \tag{4.1}$$

$$y = 1\left[y^* > 0\right]. \tag{4.2}$$

For the ramp model in (1.1), suppose that

$$u|\mathbf{x} \sim \text{Uniform}\left(0,1\right). \tag{4.3}$$

Under 4.3, the CDF of u is identical to the ramp function $R\left(\cdot\right)$, it follows immediately that (4.1), (4.2), (4.3) lead to the response probability in (1.1). In the Appendix, we show an extension of the above model where u has variable support, which is another way to represent the role of the unit interval bounds for response probabilities. For the probit model, suppose that

$$u|\mathbf{x} \sim \text{N}(0,1). \tag{4.4}$$

Initially, the true models take the form (we are dropping o on beta here)

$$y = 1\left[\beta_0 + \beta_1 x_1 + \beta_2 x_2 - u > 0\right],$$

For a given choice of $(\beta_0, \beta_1, \beta_2) = (b_0, b_1, b_2)$, we can scale (β_1, β_2) by a positive constant c,

$$(\beta_0, \beta_1, \beta_2) = (b_0, cb_1, cb_2),$$

to govern how close to linear the response probability is. When $u \sim \text{Uniform}(0,1)$, the ramp model is correctly specified, but the LPM is misspecified to varying degrees. For given initial values (b_0, b_1, b_2), a larger scaling factor c makes the kinks in the ramp function more likely to be binding and the LPM can provide a poor approximation to the response probability. Naturally, the logit and probit models are always misspecified in this case. As stated before, here we focus on the APEs rather than the underlying parameters or how well the models approximate the true response probability.

The sample size is $N = 1,000$ and $10,000$ replications are used. The population (or true)

APEs are not available in closed form, so we simulate these along with the estimators. In the tables to follow, the columns labeled "Simulated Truth" include the empirical means and standard deviations of the sample APEs at the true parameter values. We also simulate the probabilities $P(y = 1)$ and $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1)$ where the first quantity is the (Monte Carlo) population response probability and the second one tells us how binding are the ramp function inflection points. We also simulate the fraction of OLS fitted values in the unit interval, $P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{OLS} \leq 1)$. This is practically relevant because researchers often check the fraction of fitted values outside the unit interval to determine the adequacy of the LPM.

4.1 Symmetrically Distributed Explanatory Variables

In the first design, (x_1, x_2) are generated as

$$x_1 = v/2\sqrt{2} + e/2\sqrt{2}$$
$$x_2 = 1\left[v + r > 0\right],$$

where v, e, and r are independent standard normals. The initial choice of parameters is set to

$$(b_0, b_1, b_2) = (1/2, 1/4, 1/4).$$

Table 1 reports the findings when $c = 1$. There is a small probability that $\mathbf{x}\boldsymbol{\beta} \notin [0, 1]$ – roughly, about 0.021. Moreover, across all simulations, about 2.0% of the OLS fitted values are outside the unit interval. The pattern is clear: All the estimators of the APEs show very little bias and have the same precision. This is true for the continuous variable, x_1, and the binary variable, x_2. Note that this is not predicted by application of the Stoker results because x_2 is a discrete variable.[8] Nevertheless, this table illustrates what is often observed in practice: the LPM coefficients estimated by OLS are often close to the probit and logit APEs.

The story does not change when the constraints of the ramp function are strongly binding. In Table 2, we scale the initial choice coefficients by $c = 2$ and we see $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1)$ is only about 0.67, and about 28% of the OLS fitted values are outside $[0, 1]$. And yet, for estimating the APEs, the LPM does essentially as well as probit and logit, with the bias being slightly larger for APE_2. This delivers the first argument: having a large fraction of

[8] Admittedly, when the LPM is used for approximation, the bias for APE_2 is slightly larger compared to APE_1, but the bias is still reasonably small and comparable to the ones by probit and logit approximation.

Table 1. $u \sim \text{Uniform}(0,1)$, x_1 normal, x_2 binary; $c = 1$

$N = 1000$		Simulated Truth*	LPM (OLS)	Ramp (NLS)	Probit (QMLE)	Logit. (QMLE)
APE_1	mean	0.2448	0.2444	0.2450	0.2483	0.2452
	sd	0.0011	0.0288	0.0292	0.0287	0.0290
APE_2	mean	0.2489	0.2506	0.2493	0.2454	0.2449
	sd	0.0003	0.0325	0.0326	0.0323	0.0324
$P(y=1) = 0.6238$, $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) = 0.9792$						
$P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{OLS} \leq 1) = 0.9806$, $P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{NLS} \leq 1) = 0.9774$						

*This column contains the empirical means and standard deviations of the sample APEs at the true parameter values.

observations with fitted values within $0, 1$ is not a *necessary* condition for the OLS estimator to produce a good estimate of the APE.

Table 2. $u \sim \text{Uniform}(0,1)$, x_1 normal, x_2 binary; $c = 2$

$N = 1000$		Simulated Truth*	LPM (OLS)	Ramp (NLS)	Probit (QMLE)	Logit. (QMLE)
APE_1	mean	0.3219	0.3200	0.3221	0.3242	0.3220
	sd	0.0075	0.0237	0.0237	0.0226	0.0236
APE_2	mean	0.4003	0.4186	0.4006	0.4051	0.4036
	sd	0.0044	0.0270	0.0274	0.0263	0.0270
$P(y=1) = 0.6769$, $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) = 0.6738$						
$P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{OLS} \leq 1) = 0.8155$, $P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{NLS} \leq 1) = 0.6406$						

*This column contains the empirical means and standard deviations of the sample APEs at the true parameter values.

Table 3 shows the case where $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1)$ is very close to one (the consistency result of OLS estimator for the index coefficients in Horrace and Oaxaca (2006) applies when $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1)$ is exactly one). We would expect the LPM to work very well in this case, and it does. What is, perhaps, more surprising is that probit and logit work just as well, even though the true response probability is largely linear over the support of $\mathbf{x}\boldsymbol{\beta}$. These findings are a good reminder of why statements such as "the linear probability model is preferred to probit because the latter assumes normality" are not just misleading: they are wrong. In the end, what we care about is how well each approach approximates the partial effects on $P(y = 1|\mathbf{x})$. When we consider the APEs, all methods do well even when the response probability has the peculiar ramp shape.

We next consider the response probability resulting from 4.1, 4.2, and 4.4, under which

Table 3. $u \sim \text{Uniform}(0,1)$, x_1 normal, x_2 binary; $c = 0.75$

$N = 1000$		Simulated Truth*	LPM (OLS)	Ramp (NLS)	Probit (QMLE)	Logit. (QMLE)
APE_1	mean	0.1874	0.1873	0.1875	0.1886	0.1877
	sd	0.0001	0.0312	0.0314	0.0313	0.0313
APE_2	mean	0.1875	0.1881	0.1880	0.1859	0.1856
	sd	0.0000	0.0334	0.0334	0.0333	0.0334
$P(y=1) = 0.5937$, $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) = 0.9996$						
$P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{OLS} \leq 1) = 0.9991$, $P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{NLS} \leq 1) = 0.9990$						

*This column contains the empirical means and standard deviations of the sample APEs at the true parameter values.

only the probit model is correctly specified. Tables 4 and 5 maintain the same true index slopes as Tables 1 and 2, but due to scaling from the standard normal PDF, the true APEs are lower. Nevertheless, a similar pattern is revealed. Table 4 can be compared to Table 1 where $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1)$ is close to one. In this case, OLS even does a better job in fitting the response probability into the unit interval and, not surprisingly, LPM estimated by OLS performs just as well as the the correctly specified probit model in producing the estimated APEs. As c increases from 1 to 2 in Tables 5, $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1)$ drops to 0.64. Due to the better-behaved Gaussian error, we still observe a large fraction of OLS fitted values are within $[0,1]$ and there is not much difference across different methods. To better compare with the true APEs of Table 1, we increase c even further in Table 6. In this case, support of the linear index becomes really wide, and $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1)$ is as small as 0.37. Correspondingly, only 86% of observations have OLS fitted values within $[0,1]$. However, OLS still produces estimates of APEs as good as those produced by nonlinear methods. Not surprisingly, probit and logit QMLE have low bias, while NLS of the ramp model has slightly higher bias in the cases (e.g., Table 6) where the true APEs are larger.

We also generated the outcome y using an interaction between x_1 and x_2, with u having a uniform distribution. Specifically,

$$y = 1\left[\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \left(x_1 \cdot x_2\right) + u > 0\right]$$

The initial choice of parameters and the scaled parameters are

$$(b_0, b_1, b_2, b_3) = (1/2, 1/4, 1/4, 1/8)$$
$$(\beta_0, \beta_1, \beta_2, \beta_3) = (b_0, cb_1, cb_2, cb_3)$$

Table 4. $u \sim \mathrm{N}(0,1)$, x_1 normal, x_2 binary; $c = 1$

$N = 1000$		Simulated Truth*	LPM (OLS)	Ramp (NLS)	Probit (QMLE)	Logit. (QMLE)
APE_1	mean	0.0810	0.0814	0.0814	0.0814	0.0814
	sd	0.0003	0.0302	0.0303	0.0302	0.0302
APE_2	mean	0.0815	0.0817	0.0817	0.0815	0.0816
	sd	0.0002	0.0306	0.0306	0.0306	0.0306
$P(y = 1) = 0.7296$, $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) = 0.9793$						
$P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{OLS} \leq 1) = 0.9999$, $P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{NLS} \leq 1) = 0.9999$						

*This column contains the empirical means and standard deviations of the sample APEs at the true parameter values.

Table 5. $u \sim \mathrm{N}(0,1)$, x_1 normal, x_2 binary; $c = 2$

$N = 1000$		Simulated Truth*	LPM (OLS)	Ramp (NLS)	Probit (QMLE)	Logit. (QMLE)
APE_1	mean	0.1448	0.1450	0.1484	0.1451	0.1450
	sd	0.0013	0.0281	0.0305	0.0280	0.0281
APE_2	mean	0.1478	0.1488	0.1484	0.1480	0.1483
	sd	0.0008	0.0288	0.0288	0.0287	0.0288
$P(y = 1) = 0.7550$, $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) = 0.6438$						
$P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{OLS} \leq 1) = 0.9890$, $P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{NLS} \leq 1) = 0.9845$						

*This column contains the empirical means and standard deviations of the sample APEs at the true parameter values.

Tables 7 and 8 display simulation results under uniformly distributed u and normally distributed u, respectively. The scaling factor c is set as 2 to focus on the scenarios with small $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1)$ and potentially small $P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{OLS} \leq 1)$. Remember, both x_1 and x_2 have symmetric distributions, but this functional form falls outside Stoker's results because x_2 is discrete and so is $x_1 \cdot x_2$: it has a mass point at zero and is otherwise continuous. However, the four approaches—where the interaction term is included in the estimation—delivered similar estimated APEs that were close to the sample "true" APEs (as previously, probit, logit, and LPM approaches use a misspecified response probability).

Table 6. $u \sim \mathrm{N}(0,1)$, x_1 normal, x_2 binary; $c = 4$

$N = 1000$		Simulated Truth*	LPM (OLS)	Ramp (NLS)	Probit (QMLE)	Logit. (QMLE)
APE_1	mean	0.2296	0.2295	0.2368	0.2298	0.2296
	sd	0.0043	0.0253	0.0257	0.0247	0.0249
APE_2	mean	0.2375	0.2396	0.2420	0.2375	0.2393
	sd	0.0029	0.0257	0.0279	0.0255	0.0256
$P(y=1) = 0.7733$, $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) = 0.3725$						
$P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{OLS} \leq 1) = 0.8605$, $P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{NLS} \leq 1) = 0.7028$						

*This column contains the empirical means and standard deviations of the sample APEs at the true parameter values.

Table 7. $u \sim \mathrm{Uniform}(0,1)$, x_1 normal, x_2 binary; $c = 2$; with interaction

$N = 1000$		Simulated Truth*	LPM (OLS)	Ramp (NLS)	Probit (QMLE)	Logit. (QMLE)
APE_1	mean	0.3634	0.3606	0.3638	0.3664	0.3641
	sd	0.0089	0.0245	0.0249	0.0241	0.0246
APE_2	mean	0.3509	0.3777	0.3512	0.3471	0.3456
	sd	0.0040	0.0281	0.0289	0.0275	0.0278
$P(y=1) = 0.6645$, $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) = 0.6436$						
$P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{OLS} \leq 1) = 0.8554$, $P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{NLS} \leq 1) = 0.6403$						

*This column contains the empirical means and standard deviations of the sample APEs at the true parameter values.

4.2 Asymmetrically Distributed Explanatory Variables

The story changes markedly when the distributions of x_1 and x_2 are asymmetric. With v, e, and r generated as before, x_1 and x_2 are now generated as

$$\begin{aligned} x_1 &= \exp\left(-1/4 + v/2\sqrt{2} + e/2\sqrt{2}\right) \\ x_2 &= 1\left[-1/4 + v + e > 0\right], \end{aligned}$$

so that x_1 has a lognormal distribution. The variable x_2 is still binary but the response probability is below 0.5. The unscaled parameter values are, again, $(b_0, b_1, b_2) = (1/2, 1/4, 1/4)$.

Table 9 repeats the same experiment as Table 1, with the scaling factor $c = 1$ except that the explanatory variables are not asymmetrically distributed. We observe that the OLS estimated APE for $x1$ under the LPM is severely biased. The misspecified probit model and logit model estimated by QMLE appear to be slightly biased too. The Ramp model

Table 8. $u \sim \mathrm{N}(0,1)$, x_1 normal, x_2 binary; $c = 2$; with interaction

$N = 1000$		Simulated Truth*	LPM (OLS)	Ramp (NLS)	Probit (QMLE)	Logit. (QMLE)
APE_1	mean	0.1685	0.1684	0.1717	0.1689	0.1689
	sd	0.0013	0.0280	0.0295	0.0279	0.0280
APE_2	mean	0.1393	0.1427	0.1418	0.1392	0.1384
	sd	0.0005	0.0290	0.0292	0.0292	0.0293
$P(y = 1) = 0.7566$, $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) = 0.6437$						
$P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{OLS} \leq 1) = 0.9842$, $P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{NLS} \leq 1) = 0.9728$						

*This column contains the empirical means and standard deviations of the sample APEs at the true parameter values.

Table 9. $u \sim \mathrm{Uniform}(0,1)$, x_1 lognormal, x_2 asym. binary; $c = 1$

$N = 1000$		Simulated Truth*	LPM (OLS)	Ramp (NLS)	Probit (QMLE)	Logit. (QMLE)
APE_1	mean	0.1975	0.1299	0.1988	0.2225	0.2203
	sd	0.0032	0.0220	0.0350	0.0361	0.0383
APE_2	mean	0.2203	0.2354	0.2211	0.2291	0.2298
	sd	0.0020	0.0226	0.0233	0.0226	0.0230
$P(y = 1) = 0.8024$, $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) = 0.7900$						
$P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{OLS} \leq 1) = 0.9011$, $P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{NLS} \leq 1) = 0.7857$						

*This column contains the empirical means and standard deviations of the sample APEs at the true parameter values.

is correctly specified and, as predicted by the asymptotic properties given in Section 3, the NLS estimator continues to perform well. The relative bias of probit and logit are higher as well compared to the previous tables, but not to as high a degree as the LPM.

The findings in Table 10 are striking. Even though $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1)$ is high—around 0.95—and the OLS fitted values are very rarely outside the unit interval (only about 3.3 percent of the time), LPM/OLS is badly biased for the APEs and notably worse than other methods. This would seem to go against the conventional wisdom of checking the proportion of fitted values within $[0, 1]$, and this confirms the second argument: having a large fraction of OLS fitted values within the unit interval is not sufficient. Among the other estimators, Ramp/NLS has a smaller bias in terms of both APEs, while probit and logit appear to have small bias for the discrete APE, but higher bias for the continuous APE. With respect to the performance of the LPM, the results with a normally distributed u and with an interaction term are similar and so are skipped for brevity.

Table 10. $u \sim \text{Uniform}(0,1)$, x_1 lognormal, x_2 asym. binary; $c = 0.75$

$N = 1000$		Simulated Truth*	LPM (OLS)	Ramp (NLS)	Probit (QMLE)	Logit. (QMLE)
APE_1	mean	0.1776	0.1486	0.1796	0.2110	0.2111
	sd	0.0013	0.0248	0.0345	0.0358	0.0378
APE_2	mean	0.1828	0.1910	0.1829	0.1835	0.1835
	sd	0.0007	0.0278	0.0282	0.0281	0.0285
$P(y = 1) = 0.7413$, $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1) = 0.9471$						
$P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{OLS} \leq 1) = 0.9671$, $P(0 \leq \mathbf{x}\widehat{\boldsymbol{\beta}}_{NLS} \leq 1) = 0.9446$						

*This column contains the empirical means and standard deviations of the sample APEs at the true parameter values.

5 Mortgage Approval Probabilities

As an illustration of linear and nonlinear estimators for binary response models, we revisit the analysis of mortgage lending decisions from Hunter and Walker (1996).[9] We compare linear and nonlinear estimates of the average effect of being white on the probability of loan approval, holding constant a number of loan, property, and borrower characteristics. Table 9 presents basic summary statistics for the dependent variable "approve" and 23 covariates.

For our index model, we include interactions between "white" and all other explanatory variables to allow for the factors like loan amount and credit history to have a differential impact on approval probability by group. Let w denote "white" and $\mathbf{z}$ be a vector including the 22 other covariates, so that $\mathbf{x} = \{1, \mathbf{z}, w, w\mathbf{z}\}$ and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_0, \boldsymbol{\beta}_z, \beta_w, \boldsymbol{\beta}_{wz}\}$, where $\boldsymbol{\beta}_0$ is the intercept, $\boldsymbol{\beta}_z$ and β_w are the coefficients on $\mathbf{z}$ and w, respectively, while $\boldsymbol{\beta}_{wz}$ is the coefficient on $w\mathbf{z}$. Then the partial effects we average are formed by evaluating the difference in the probabilities evaluated at $w = 1$ and $w = 0$, respectively, as given below.

$$APE_w = \mathrm{E}\left[G(\beta_0 + \beta_w + \mathbf{z}(\boldsymbol{\beta}_z + \boldsymbol{\beta}_{wz})) - G(\beta_0 + \mathbf{z}\boldsymbol{\beta}_z)\right],$$

where $G()$ is either the identity function (for the LPM), the probit CDF, the logit CDF, or the ramp function.

Table 10 presents the results. Using the LPM estimated by OLS, about 18% of observations have predicted probabilities outside the unit interval.[10] The Horrace and Oaxaca results then clearly imply OLS is inconsistent for the slope parameters if the ramp model is

[9] We use a version of the loan applications dataset provided by Mary Beth Walker for Wooldridge (2019).
[10] Within this 18%, 98% of observations had predicted values greater than 1.

Table 9: Loan Approval Summary Statistics ($N = 1989$)

Variable	Description	Mean	SD	Skew.	Kurt.
approve	=1 if loan approved	0.88	0.33	-2.30	6.29
white	=1 if white	0.85	0.36	-1.91	4.64
loanamt	Loan amount $1000s	143.25	80.52	3.13	20.36
suffolk	=1 if in Suffolk County	0.15	0.36	1.91	4.66
appinc	Applicant income $1000s	84.68	87.06	5.26	36.70
unit	Number of units in property	1.12	0.44	4.01	19.89
married	=1 if applicant married	0.66	0.47	-0.67	1.45
dep	Number of dependents	0.77	1.10	1.47	5.33
emp	Years employed in line of work	0.21	1.00	6.69	50.57
yjob	Years at this job	0.45	1.12	5.32	36.18
atotinc	Total monthly income	5195.55	5269.06	6.36	65.34
self	=1 if self employed	0.13	0.34	2.21	5.89
other	Other financing $1000s	2.37	28.23	26.80	886.84
rep	Number of credit reports	1.50	0.99	1.45	7.37
pubrec	=1 if filed bankruptcy	0.07	0.25	3.40	12.59
hrat	Housing expense % of total inc.	24.79	7.12	0.25	6.74
obrat	Other obligations % of total inc.	32.39	8.26	0.44	7.40
cosign	=1 if there is a cosigner	0.03	0.17	5.65	32.92
sch	=1 if $>$ 12 years schooling	0.77	0.42	-1.29	2.68
mortno	=1 if no mortgage history	0.33	0.47	0.71	1.51
mortlat1	=1 if one or two late payments	0.02	0.14	7.03	50.36
mortlat2	=1 if more than two late payments	0.01	0.10	9.58	92.72
chist	=0 if accounts are delinq. $\geq$60 days	0.84	0.37	-1.83	4.35
loanprc	Loan amount / purchase price	0.77	0.19	0.44	14.39

correct. There is little reason to expect the LPM will approximate this APE, either based on the theoretical results of Stoker (1986) or our simulation study. Many of the explanatory variables are binary, and the continuous variables (e.g., income) tend to be skewed. For each variable, normality is strongly rejected by a Jarque-Bera test (a joint test of the skewness and kurtosis) with p-values well below 1%. The model also includes interactions between the continuous variables and a binary variable. Using the LPM estimates, the APE for *white* is 5.3 percentage points and it is only marginally significant. Using the nonlinear estimators, the APEs are each a bit larger at about 7.0 percentage points, and they are all significant at the 1% level.

Interestingly, OLS predicts only 18% of observations with indexes outside the unit interval, whereas NLS predicts nearly 40%, which follows the pattern of many of our simulations from the previous section and suggests trimming the sample once is not sufficient to consis-

Table 10: Estimates of the APE of "White" on Loan Approval

	LPM (OLS)	Ramp (NLS)	Probit (QMLE)	Logit (QMLE)
Estimate	0.0532	0.0706	0.0695	0.0712
Robust SE	0.0278	0.0227	0.0220	0.0219
Mean Squared Error	0.1171	0.0839	0.0840	0.0837

Notes: There were only 1976 complete cases out of 1989 observations total. All robust standard errors were computed using the sandwich forms and the delta method. The fraction of predicted linear indexes within the unit interval is 0.8173 by OLS and 0.6027 by NLS.

tently estimate the parameters or APEs under the piecewise linear model.[11] Of this 40%, 99% had NLS predicted linear indexes greater than 1 and most had high predicted probabilities of approval regardless of the model or counterfactual race.[12] Model selection by the minimum mean squared error favors logit, though the other nonlinear models are very similar.

6 Implications for Empirical Research

We have revisited the conclusions reached by Horrace and Oaxaca (2006) concerning the ability of the linear projection parameters—consistently estimated by OLS—to recover interesting parameters. We argue that Horrace and Oaxaca's focus on the parameters in the underlying index model is misguided; instead, one should focus on the APEs. Focusing on the APEs is hardly controversial, as almost every modern study that employs any model nonlinear in the explanatory variables reports estimated APEs.

Once the focus is on the APEs, a few useful conclusions emerge. First, having a high of estimated response probabilities in $[0, 1]$ is neither necessary nor sufficient for good performance of the LPM. Notably, when the explanatory variables have a multivariate normal distribution, the linear projection parameters are identical to the population APEs under a

[11]The reason we report the fraction of NLS predicted linear index outside 0 and 1 here is to illustrate what proportion of observations would have been trimmed by the iterated trimming OLS procedure. We should note that this quantity is not of essential interest just as the linear indexes in probit and logit models are not.

[12]NLS drops these observations because they have predicted indexes outside the unit interval, not necessarily because they have high leverage. In fact, under the logit model, the average Pregibon (1981) leverage statistic for the 40% ("predict lev, hat" in Stata following logit estimation) was lower (0.008) than the average for the included observations (0.033).

general index model, and this is true even when the flat parts of the ramp function occur with high probability, i.e. $P(0 \leq \mathbf{x}\boldsymbol{\beta} \leq 1)$ is small. In this case, the linear projection parameters, γ_j, will be greatly attenuated toward zero compared with the index parameters, β_j. We find that OLS estimation of the LPM continues to have good finite sample properties for the APE in many cases when the covariates are symmetrically distributed. When the explanatory variables have asymmetric distributions, however, the conclusions for the LPM are not as sanguine—unless the support of $\mathbf{x}\boldsymbol{\beta}$ is contained entirely in the interval $[0, 1]$. Some simulations show that even if the probability of $\mathbf{x}\boldsymbol{\beta}$ being in the unit interval is high (e.g. 97% in Table 10), the linear projection parameters are not very close to the true APEs.

For the DGPs we study, we also find the logit and probit models, estimated by quasi-MLE (because the response probabilities are misspecified), tend to approximate the APEs very well. Especially when the support of $\mathbf{x}\boldsymbol{\beta}$ is wide relative to $[0, 1]$, the logit and probit approximations to the APEs may be notably better than those for the LPM when the covariates have asymmetric distributions, though this is not guaranteed. Although the ramp function may not be particularly realistic as a model for the response probability, we have shown that NLS estimation based on it is consistent (for the best MSE approximation to the true response probability) and asymptotically normal. A nonlinear model, of course, offers other advantages over the LPM, such as more realistic response probabilities and nonconstant partial effects. Especially given the ease of modern computation, an implication of our simulation findings is that researchers should generally try a nonlinear estimator, as they may be more robust to covariate asymmetry and variance than OLS estimation of the LPM.

To summarize, in evaluating different strategies, one needs to make sure we have carefully defined the population quantities of interest, and then we make proper comparisons across different approaches. We find probit, logit, and the ramp models have the best finite sample properties for estimating the APEs across the model DGPs we study. However, when the APEs are of interest, we also find the LPM is more widely applicable than a simple reading of Horrace and Oaxaca might suggest. The conclusions drawn here are easily extended to the case where y is a fractional response, where the limit values zero and one can occur with positive probability. In particular, the results of Stoker (1986) apply to $\mathrm{E}(y|\mathbf{x})$. If this conditional mean follows the same ramp function, the qualitative conclusions obtained in the binary case will remain.

References

Angrist, J. D. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Horowitz, J. L. and Savin, N. (2001). Binary response models: Logits, probits and semiparametrics. *Journal of Economic Perspectives*, 15(4):43–56.

Horrace, W. C. and Oaxaca, R. L. (2006). Results on the bias and inconsistency of ordinary least squares for the linear probability model. *Economics Letters*, 90(3):321–327.

Hunter, W. C. and Walker, M. B. (1996). The cultural affinity hypothesis and mortgage lending decisions. *Journal of Real Estate Finance and Economics*, 13:57–70.

Li, C., Poskitt, D. S., Windmeijer, F., and Zhao, X. (2022). Binary outcomes, OLS, 2SLS and IV probit. *Econometric Reviews*, 41(8):859–876.

Newey, W. K. and McFadden, D. (1994). Large sample estimation testing. *Handbook of Econometrics*, 4:2113–2245.

Pregibon, D. (1981). Logistic regression diagnostics. *The Annals of Statistics*, 9(4):705–724.

Ruud, P. A. (1983). Sufficient Conditions for the Consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution in Multinomial Discrete Choice Models. *Econometrica*, 51(1):225–228.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica*, 54(6):1461–1481.

van den Berg, G. J. and Siflinger, B. M. (2022). The effects of a daycare reform on health in childhood–Evidence from Sweden. *Journal of Health Economics*, 81:102577.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Wooldridge, J. M. (2019). *Introductory Econometrics: A Modern Approach*. Cengage Learning.

Appendix

Proof of Theorem 2

Proof. We will obtain the asymptotic normality of the NLS estimator by applying Theorem 7.1 of Newey and McFadden (1994). Condition (i) and (ii) of Theorem 7.1 follows from our assumptions. As we discussed in the main context, condition (iii) is satisfied as long as $\mathbf{x}$ contains a continuous variables x_j with nonzero β_{jo} so that $P(\mathbf{x}_i\boldsymbol{\beta}_o = 0 \text{ or } \mathbf{x}_i\boldsymbol{\beta}_o = 1) = 0$.

For condition (iv), notice that the first derivative of the object function is well defined at $\boldsymbol{\beta}_o$ with probability 1:

$$D_N(\boldsymbol{\beta}_o) = \nabla_\beta Q_N(\boldsymbol{\beta}_o) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i'(y_i - \mathbf{x}_i\boldsymbol{\beta}_o)1\{\mathbf{x}_i\boldsymbol{\beta}_o \in (0,1)\} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i' u_i 1\{\mathbf{x}_i\boldsymbol{\beta}_o \in (0,1)\},$$

where $u_i = y_i - R(\mathbf{x}_i\boldsymbol{\beta}_o)$. Since $\mathrm{E}\|\mathbf{x}_i' u_i\|1\{\mathbf{x}_i\boldsymbol{\beta}_o \in (0,1)\} < \infty$ under the assumption $\mathrm{E}\|\mathbf{x}_i\|^2 < \infty$, the vector Lindberg-Levy CLT applies:

$$\sqrt{N} D_N(\boldsymbol{\beta}_o) \xrightarrow{d} \mathbb{N}\left(0, \boldsymbol{\Omega}(\boldsymbol{\beta}_o)\right),$$

giving condition (iv). Lastly, for condition (v), following Newey and McFadden (1994), we can rewrite

$$\begin{aligned}
&\sqrt{N}[Q_N(\boldsymbol{\beta}) - Q_N(\boldsymbol{\beta}_o)] \\
=&\sqrt{N}\left[D_N(\boldsymbol{\beta}_o)(\boldsymbol{\beta} - \boldsymbol{\beta}_o) + Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}_o)\right] + \|\boldsymbol{\beta} - \boldsymbol{\beta}_o\| M_N(\boldsymbol{\beta}),
\end{aligned}$$

where $M_N(\boldsymbol{\beta})$ is the remainder term, defined as:

$$M_N(\boldsymbol{\beta}) = \frac{\sqrt{N}\left[Q_N(\boldsymbol{\beta}) - Q_N(\boldsymbol{\beta}_o) - D_N'(\boldsymbol{\beta}_o)(\boldsymbol{\beta} - \boldsymbol{\beta}_o) - (Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}_o))\right]}{\|\boldsymbol{\beta} - \boldsymbol{\beta}_o\|}.$$

Let U_N be an neighborhood of $\boldsymbol{\beta}_o$: $U_N = \{\boldsymbol{\beta} \in \mathcal{B} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_o\| < \varepsilon_N\}$ where $\varepsilon_N \to 0$. Consider any $\boldsymbol{\beta} \in U_N$. Since $D_N(\boldsymbol{\beta}_o)$ is the gradient of $Q_N(\boldsymbol{\beta})$ at $\boldsymbol{\beta}_o$, $Q_N(\boldsymbol{\beta}) - Q_N(\boldsymbol{\beta}_o) - D_N(\boldsymbol{\beta}_o)(\boldsymbol{\beta} - \boldsymbol{\beta}_o)$ goes to zero faster than $\|\boldsymbol{\beta} - \boldsymbol{\beta}_o\|$ as $\boldsymbol{\beta}$ goes to $\boldsymbol{\beta}_o$, by the definition of the gradient. Similarly, due to $\nabla_\beta Q(\boldsymbol{\beta}_o) = \mathrm{E}\left(s_i(\boldsymbol{\beta}_o)\right) = 0$, $Q(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}_o)$ goes to 0 faster than $\|\boldsymbol{\beta} - \boldsymbol{\beta}_o\|$ as $\boldsymbol{\beta}$ goes to $\boldsymbol{\beta}_o$. Under the moment conditions, we can easily show $Q_N(\boldsymbol{\beta}) - Q(\boldsymbol{\beta}) \to 0$ in probability for each $\boldsymbol{\beta}$ and so $\sqrt{N}[Q_N(\boldsymbol{\beta}) - Q(\boldsymbol{\beta})]$ is bounded in probability for each $\boldsymbol{\beta}$. Also note that $\sqrt{N} D_N(\boldsymbol{\beta}_o)$ is bounded in probability due to the

asymptotic normality. Since the numerator is bounded in probability and converges to zero faster than the denominator, we conclude that $\lim_{N\to\infty} \sup_{\boldsymbol{\beta}\in U_N} M_N(\boldsymbol{\beta}) \to 0$ in probability, which implies condition (v). □

Proof of Theorem 3

Proof. Consider $\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}})$:

$$\begin{aligned}\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}}) =& \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i'\mathbf{x}_i \left(y_i - R(\mathbf{x}_i\hat{\boldsymbol{\beta}})\right)^2 1\{\mathbf{x}_i\hat{\boldsymbol{\beta}} \in (0,1)\} \\ \equiv& \frac{1}{N}\sum_{i=1}^{N} a(\mathbf{x_i}, \hat{\boldsymbol{\beta}})\end{aligned}$$

Note that $E|y_i - R(\mathbf{x}_i\boldsymbol{\beta})|^4 \leq 1$ for any $\boldsymbol{\beta} \in \mathcal{B}$ since both y_i and $R(.)$ are naturally bounded in $[0,1]$ with probability 1. Then, we have

$$\mathrm{E} \sup_{\boldsymbol{\beta}\in\mathcal{B}} \|a(x,\boldsymbol{\beta})\| \leq (\mathrm{E}\,\|\mathbf{x}_i\|^4\, E|y_i - R(\mathbf{x}_i\boldsymbol{\beta})|^4)^{1/2} < \infty,$$

where the first inequality follows from Hölder's inequality. Also note that $a(\mathbf{x_i}, \boldsymbol{\beta})$ is continuous at $\boldsymbol{\beta}_o$ with probability one given that $P(\mathbf{x}_i\boldsymbol{\beta}_o = 0) = P(\mathbf{x}_i\boldsymbol{\beta}_o = 1) = 0$. Then, we can apply Lemma 4.3 of Newey and McFadden (1994):

$$\boldsymbol{\Omega}(\hat{\boldsymbol{\beta}}) = \frac{1}{N}\sum_{i=1}^{N} a(\mathbf{x_i}, \hat{\boldsymbol{\beta}}) \xrightarrow{p} E(a(\mathbf{x_i}, \boldsymbol{\beta}_o)) = \boldsymbol{\Omega}(\boldsymbol{\beta}_o).$$

Similarly, Lemma 4.3 also applies to $\mathbf{A}_N(\hat{\boldsymbol{\beta}}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i'\mathbf{x}_i 1\{\mathbf{x}_i\hat{\boldsymbol{\beta}} \in (0,1)\}$:

$$\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} 1\{\mathbf{x}_i\hat{\boldsymbol{\beta}} \in (0,1)\}\mathbf{x}_i'\mathbf{x}_i = \mathbf{A}(\boldsymbol{\beta}_o).$$

So, we conclude that

$$\hat{\mathbf{V}} = \mathbf{A}_N(\hat{\boldsymbol{\beta}})^{-1}\boldsymbol{\Omega}_N(\hat{\boldsymbol{\beta}})\mathbf{A}_N(\hat{\boldsymbol{\beta}})^{-1} \xrightarrow{p} \mathbf{A}(\boldsymbol{\beta}_o)^{-1}\boldsymbol{\Omega}(\boldsymbol{\beta}_o)\mathbf{A}(\boldsymbol{\beta}_o)^{-1}. \square$$

The Ramp Model with Variable Support

In this appendix, we modify and extend the Horrace and Oaxaca setup in order to show how the constraint on the linear index through a ramp function can be interpreted as relating to the support of the latent model error term. In particular, write

$$\begin{aligned} y^* &= \mathbf{x}\boldsymbol{\beta} - u \\ u|\mathbf{x} &\sim \text{Uniform}\left(-a, a\right) \\ y &= 1\left[y^* > 0\right] \end{aligned} \tag{6.1}$$

for some $a > 0$. Compared with Horrace and Oaxaca, we have shifted the intercept so that u has a symmetric distribution about its mean of zero. Also, we allow u to have narrow or wide support, depending on a. The CDF for the Uniform $\left(-\sqrt{3}, \sqrt{3}\right)$ distribution, which has unit variance, is graphed in Figure 1.

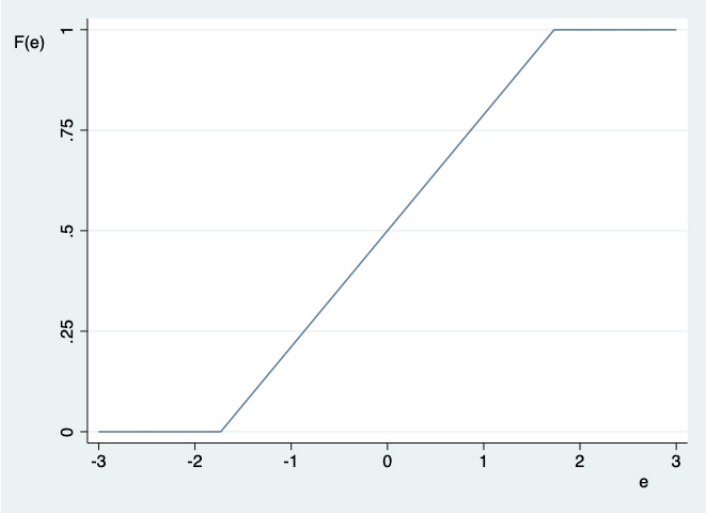

Figure 1: The CDF of y with $u|x \sim \text{U}\left(-\sqrt{3}, \sqrt{3}\right)$.

Given the latent variable model in (6.1), we can derive the response probability:

$$\begin{aligned} p\left(\mathbf{x}\right) \equiv P\left(y = 1|\mathbf{x}\right) &= P\left(y^* \geq 0|\mathbf{x}\right) = P\left(u \geq -\mathbf{x}\boldsymbol{\beta}|\mathbf{x}\right) = P\left(u \leq \mathbf{x}\boldsymbol{\beta}|\mathbf{x}\right) = F_u\left(\mathbf{x}\boldsymbol{\beta}\right) \\ &= \begin{cases} 0, & \mathbf{x}\boldsymbol{\beta} < -a \\ \frac{\mathbf{x}\boldsymbol{\beta}+a}{2a}, & -a \leq \mathbf{x}\boldsymbol{\beta} \leq a \\ 1, & \mathbf{x}\boldsymbol{\beta} > a \end{cases} \end{aligned}$$

We write this function as $F_u(\mathbf{x}\boldsymbol{\beta}) \equiv R_a(\mathbf{x}\boldsymbol{\beta})$, which is a ramp function that is nondifferentiable at $-a$ and a. For an x_j with a positive coefficient, the response probability has the same shape as in Figure 1.

As a increases relative to $\boldsymbol{\beta}$, the response probability is linear over more of the support of $\mathbf{x}$. If

$$P(-a \leq \mathbf{x}\boldsymbol{\beta} \leq a) = 1 \tag{6.2}$$

then, with probability one, $R_a(\mathbf{x}\boldsymbol{\beta}) = (\mathbf{x}\boldsymbol{\beta} + a)/2a$, a linear function of $\mathbf{x}$. In this case, the partial effects are constant and equal to $\beta_j/2a$, $j = 2, ..., K$. These are also the linear projection parameters γ_j and so OLS consistently estimates the APEs under (6.2).

If x_j is a continuous variable, we are interested in the APE defined as a derivative, which exists with probability one when $\mathbf{x}\boldsymbol{\beta}$ is continuous. At $\mathbf{x}\boldsymbol{\beta} \in \{-a, a\}$ the definition of the partial effect is immaterial. To be concrete, take

$$PE_j(\mathbf{x}) = \frac{\beta_j}{2a} \cdot 1[-a \leq \mathbf{x}\boldsymbol{\beta} \leq a].$$

Notice that $PE_j(\mathbf{x}) = 0$ if $\mathbf{x}\boldsymbol{\beta} < -a$ or $\mathbf{x}\boldsymbol{\beta} > a$ because we are on one of the flat parts of the ramp. This feature of $PE_j(\mathbf{x})$ is taken into account in computing the APE:

$$\text{APE}_j = \text{E}[PE_j(\mathbf{x})] = \frac{\beta_j}{2a} \cdot P(-a \leq \mathbf{x}\boldsymbol{\beta} \leq a)$$

The case that aligns with Horrace and Oaxaca is $a = 1/2$—so that the $Uniform(0, 1)$ distributed has just been shifted to have zero mean—in which case $|\text{APE}_j| \leq |\beta_j|$, and the difference between APE_j and β_j can be large. It is easily seen that $|\text{APE}_j| < |\beta_j|$ for any $a \geq 1/2$. In the extended model (6.1), depending on the values of a and $P(-a \leq \mathbf{x}\boldsymbol{\beta} \leq a)$, $|\text{APE}_j|$ need not be smaller than $|\beta_j|$.

While the latent error support parameter a is not separately identified from $\boldsymbol{\beta}$, this model can be a convenient device for generating data where the unit interval for probabilities is binding to varying degrees.