NONLINEAR AND BINARY REGRESSION, PREDICTIVE EFFECTS, AND M-ESTIMATION

ABSTRACT. We begin by formulating nonlinear regression models, where nonlinearity may arise via parameters or via variables. We discuss the key estimands for nonlinear regression – the predictive effects, average predictive effects, and sorted predictive effects. We then focus on regression models for binary outcomes. Binary outcomes naturally lead to nonlinear regression functions. A natural way to estimate nonlinear regression models is through nonlinear least squares or through (quasi)-maximum likelihood methods. The latter methods are special cases of the M-estimation framework, which could in turn be viewed as a special case of GMM. We provide two applications to racial-based discrimination in mortgage lending and gender-based discrimination in wages. Here we find heterogeneous, sometimes quite large, predictive effects of race on the probability of mortgage denial and of gender on average wages.

1. Nonlinear Regression, Predictive Effects, Average and Sorted PE

In this lecture we will be concerned with nonlinear predictive models. What do we mean by nonlinear models? We consider two kinds of nonlinearities, which are not mutually exclusive:

- nonlinearity in key variables of interest;
- nonlinearity in parameters.

Nonlinearities in variables arise from the fact that often we use transformations of variables in formulating predictive and causal models. For instance, consider the case where Y is an outcome variable of interest, and $X = (D, W)$ is a vector of covariates, where D is a binary treatment indicator and W is the set of controls. Then a natural interactive model for the expectation of Y given X is

$$p(X) := \mathrm{E}[Y \mid X] = B(W)'\alpha_0 + DB(W)'\delta_0,$$

where $B(W)$ is dictionary of transformations of W, e.g. polynomials and interactions. The fact that D is interacted with functions of W makes the model nonlinear with respect to D and W. The model is still linear in parameters, which can be estimated by least squares.

Nonlinearity in parameters arises for example from considering models of the sort

$$p(X) = p(X, \beta_0),$$

where β_0 is a parameter value and p is a nonlinear function with respect to β. Such models are natural when we consider binary, nonnegative, count, and other types of outcome variables. While linear in parameters approximations may still perform well, there is value in considering both.

For instance, a linear in parameters model may be natural when modeling log wages or log durations, but an exponential model might be preferred to model directly expected values of wages or durations given covariates, i.e.

$$p(X) = \exp(B(X)'\beta_0),$$

which respects the fact that wages and durations are nonnegative.

1.1. **What estimands are of interest? Introducing Average and Sorted Effects.** The parameter β often has a useful structural interpretation, as in the binary response models of Section 2, and so it is a good practice to estimate and report it. However, nonlinearities often make the interpretation of this parameter difficult, because it seems to play only a technical role in the model.

What should we target instead? Here we discuss other estimands that we can identify as functionals of β and estimate them using the plug-in principle. They correspond to predictive effects of D on Y holding W fixed. Let $(d, w) \mapsto p(d, w)$ be some predictive function of Y given D and W, e.g.

$$p(d, w) = \mathrm{E}[Y \mid D = d, W = w].$$

When the variable of interest D is binary, we can consider the following estimands:

(a) predictive effect (PE) of changing $D = 0$ to $D = 1$ given $W = w$:

$$\theta_w = p(1, w) - p(0, w),$$

(b) average PE (APE) over a subpopulation with the distribution of covariates M:

$$\theta = \int \theta_w dM(w) = \mathrm{E}\theta_{\tilde{W}}, \quad \tilde{W} \sim M,$$

(c) sorted PE (SPE) by percentiles,

$$\theta_\alpha = \alpha\text{-quantile of } \theta_{\tilde{W}}, \quad \tilde{W} \sim M.$$

We could also use the name "treatment effect" (TE) instead of "predictive effect" (PE), when there is a sense in which the estimated effects have a casual interpretation.

All of the above are interesting estimands:

- The PE characterizes the impact of changing D on the prediction at a given control value w. The problem with PE alone is that there are many possible values w at which we can compute the effect. We can further summarize the PE θ_w in many ways.

- For example, we can use PEs θ_w for *classification* of individuals into groups that are "most affected" (all i's such that $\theta_{W_i} > t_1$ for some threshold t_1) and the "least affected" (all i's such that $\theta_{W_i} < t_2$ for some threshold t_2). We do so in our empirical applications and then present the average characteristics of those most affected and least affected. For example, in the mortgage application, the group that will be most strongly affected are those who are either single or black or low income, or all of the above.

- We can use the PEs to compute APEs by averaging θ_w with respect to different distributions M of the controls. For example, in the mortgage example, the APE for all applicants and for black applicants are different.

- We can use the PEs to compute the SPEs, which give a more complete summary of all PE's in the population with controls distributed according to M. Indeed, for example, APE could be small, but there could be groups of people who are much more strongly affected.

As for the choice of the distribution M, in the empirical applications we shall report the results for

- $M = F_W$, the distribution of controls in the entire population, and

- $M = F_{W|D=1}$, the distribution of controls in the "treated" population.

In the context of treatment effect analysis, using $M = F_W$ corresponds to computing the so called "average treatment effect," and using $M = F_{W|D=1}$ to the "average treatment effect on the treated."

When D is continuous we can also look at estimands defined in terms of partial derivatives:

(a)′ predictive partial effect (PE) of D when $W = w$ at $D = d$:

$$\theta_x = (\partial/\partial d)p(d, w),$$

(b)′ average PE (APE) over a subpopulation with distribution of covariates M,

$$\theta_x dM(x) = \mathrm{E}\theta_{\tilde{X}}, \quad \tilde{X} \sim M,$$

(c)′ sorted PE (SPE) by percentiles,

$$\theta_\alpha = \alpha\text{-quantile of } \theta_{\tilde{X}}, \quad \tilde{X} \sim M.$$

Figure 1 shows estimated APE and SPE of the gender wage gap on the treated female population conditional on worker characteristics. Figure 2 shows estimated APE and SPE of race on the probability of mortgage denial conditional on applicant characteristics for the entire population. These two applications are discussed in more detail in Section 4.

1.2. **Estimation and Inference on Technical Parameters.** The models described above can be estimated by nonlinear least squares

$$\hat{\beta} \in \arg\min_{\beta \in \mathcal{B}} \mathbb{E}_n m(Z, \beta),$$

where m is the square loss function $m(Z, \beta) = (Y - p(X, \beta))^2$, and $\mathcal{B} \subseteq \mathbb{R}^{\dim \beta}$ is the parameter space for β. This approach is motivated by the analogy principle, since the true value of the parameter solves the population problem:

$$\beta_0 = \arg\min_{\beta \in \mathcal{B}} \mathrm{E} m(Z, \beta).$$

This formulation emphasizes that the nonlinear least squares estimator is an M-estimator with loss function m.

The consistency of $\hat{\beta}$ is immediate from the Extremum Consistency Lemma. Root-n consistency and approximate normality follow from the fact that the nonlinear least squares estimator is a GMM estimator with the score function

$$g(Z, \beta) := \frac{\partial}{\partial \beta} m(Z, \beta),$$

since $\hat{\beta}$ solves $\mathbb{E}_n g(Z, \hat{\beta}) = 0$ in the sample, and β_0 solves $\mathrm{E} g(Z, \beta_0) = 0$ in the population, provided that the solutions are interior to the parameter space $\mathcal{B}$. Thus the approximate normality and bootstrap results of L4 and L5 apply here. Even though nonlinear least squares can be reformulated as GMM for figuring out its theoretical properties, it is often not convenient to treat nonlinear least squares as GMM for computational purposes.

1.3. **Estimation and Inference on Target Parameters.** Having obtained estimators $\hat{\beta}$ of the technical parameters, we can obtain the estimators of various estimands – (a), (b), (c), (a)′,

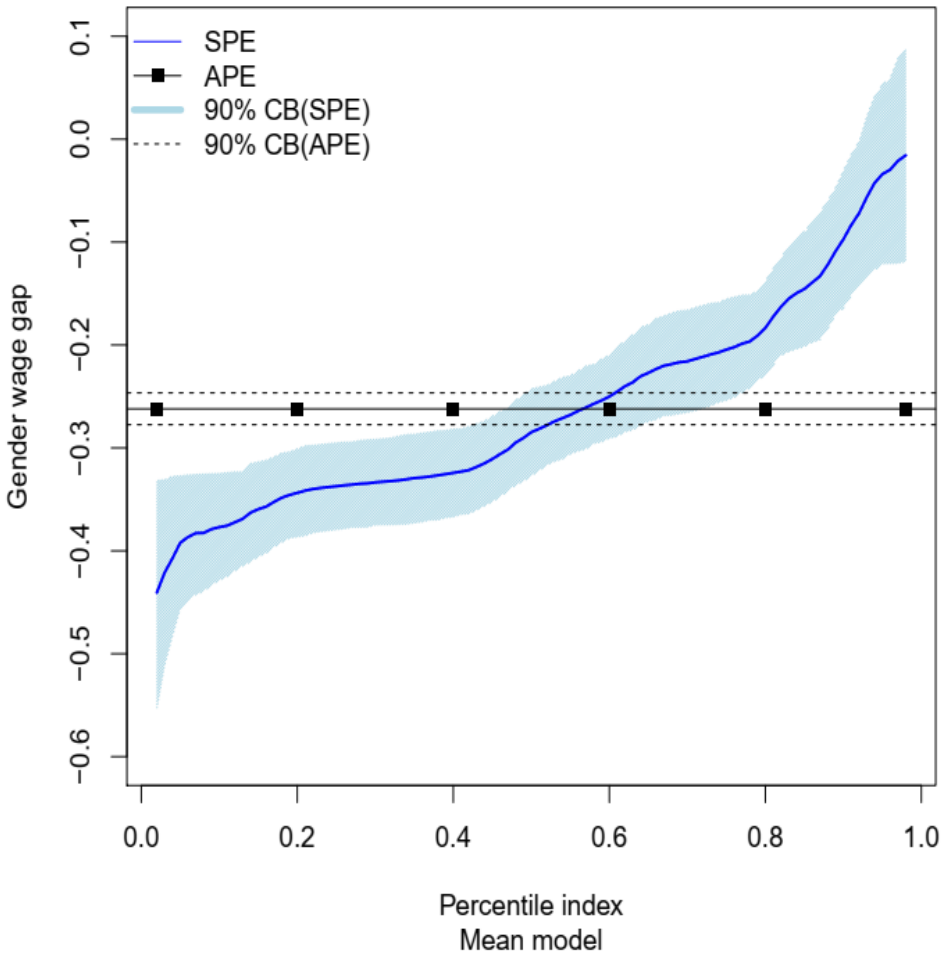

FIGURE 1. The APE and SPE of gender wage gap on the treated female population conditional on job market characters. The estimates are obtained via the plug-in principle and the 90% confidence sets are obtained by empirical bootstrap with 500 repetitions.

(b)′, and (c)′ using the *plug-in* principle. For example, the estimator of the predictive effect (PE) of changing $D = 0$ to $D = 1$ given $W = w$ is given by

$$\hat{\theta}_w = \hat{p}(1, w) - \hat{p}(0, w) = p(1, w, \hat{\beta}) - p(0, w, \hat{\beta}).$$

The asymptotic properties of this estimator follow from the delta method. We can also use the bootstrap to approximate the distribution of this estimator since it is approximately linear with respect to $\hat{\beta}$, which is a GMM estimator.

A more complicated quantity is the average PE. Consider, for example, the case where the distribution M over which we integrate is the overall population, so that $M = F_W$.

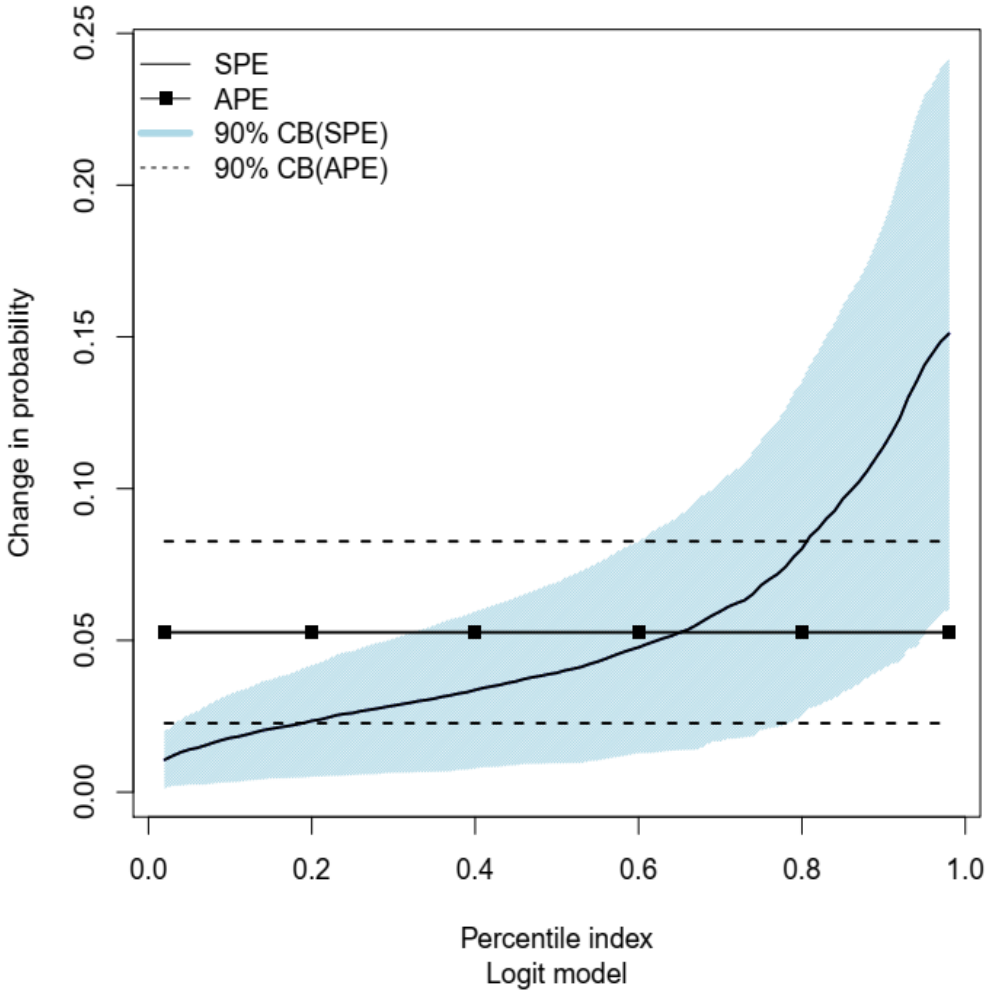

FIGURE 2. The APE and SPE of being black on the probability of mortgage denial conditional on applicant characteristics for the entire population. The estimates are obtained via the plug-in principle and the 90% confidence sets are obtained by empirical bootstrap with 500 repetitions.

Then the natural estimator is

$$\hat{\theta} = \int \hat{\theta}_w d\hat{F}_W(w) = \frac{1}{n}\sum_{i=1}^{n} \hat{\theta}_{W_i},$$

where $\hat{F}_W$ is the empirical distribution function of W_i's, an estimator of F_W. The estimator $\hat{\theta}$ has two sources of uncertainty: one is created by estimation of $\hat{\theta}_w$ and one is created by estimation of F_W, so the situation is potentially more complicated. However, we can approximate the distribution of this estimator using the bootstrap. To see why bootstrap

works we can represent the estimator as a GMM estimator with the score function:

$$\tilde{g}(Z_i, \gamma) = \begin{pmatrix} \theta - p(1, W_i, \beta) + p(0, W_i, \beta), \\ \frac{\partial}{\partial \beta} m(Z_i, \beta) \end{pmatrix}, \quad \gamma := (\theta, \beta')'.$$

Here we "stack" the score functions for the two estimation problems together to form the joint score function. This allows us to invoke the GMM machinery for the analysis of the theoretical properties of this estimator – we can write down the large sample variance and distribution of $\hat{\gamma} = (\hat{\theta}, \hat{\beta}')'$. Moreover, we can use the bootstrap for calculating the large sample variance and distribution. We shall tend to use the bootstrap as a more convenient and practical approach.

Here we provide an explicit algorithm for the bootstrap construction of the confidence band for the APE θ:

(1) Obtain many bootstrap draws $\hat{\theta}^{*(j)}$, $j = 1, \ldots, B$, of the estimator $\hat{\theta}$, where the index j enumerates the bootstrap draws.
(2) Compute the bootstrap variance estimator

$$\hat{s}^2 = B^{-1} \sum_{j=1}^{B} (\hat{\theta}^{*(j)} - \hat{\theta})^2,$$

(or use the estimate based on the interquartile range).
(3) Compute the critical value

$$c(1-a) = (1-a)\text{-quantile of } \left\{ |\hat{\theta}^{*(j)} - \hat{\theta}|/\hat{s}, \quad j = 1, \ldots, B \right\}.$$

(4) Report the confidence region for θ with confidence level $1 - a$ as $[\hat{\theta} \pm c(1-a)\hat{s}]$.

Figures 1 and 2 present confidence intervals for the APE obtained using this algorithm. There are other versions of the confidence intervals we can report. For example, we can report the confidence interval as simply the region between the $a/2$ and $1 - a/2$ quantiles of the sample of the bootstrap draws $\hat{\theta}^{*(j)}$ for $j = 1, \ldots, B$.

The reasoning for other parameters is very similar. For example, the estimators of the sorted PEs for the population $M = F_W$ are obtained by sorting the values of PEs:

$$\mathsf{PE} = (\hat{\theta}_{W_i}, \quad i = 1, \ldots, n)$$

in the increasing order. Given a grid of percentile indices $\mathcal{A} \in [0, 1]$ we then obtain

$$\hat{\theta}_\alpha = \alpha - \text{quantile of } \mathsf{PE}, \quad \alpha \in \mathcal{A}.$$

The sorted PEs θ_α could be represented as GMM estimands and the same logic as for APE applies to them – however, we now do skip the details, because they are immaterial to the

discussion that follows below (see [4]). Moreover, the sorted PEs carry the percentile index $\alpha \in [0,1]$ and we can construct joint confidence band for θ_α for α's on a grid $\mathcal{A} \in [0,1]$ along the lines of L1, where we used the normal approximations to do so, but we can also use the bootstrap instead of the normal approximations.

Here we provide an explicit algorithm for the bootstrap construction of the joint confidence bands for SPEs $(\theta_\alpha)_{\alpha\in\mathcal{A}}$:

(1) Obtain many bootstrap draws
$$(\hat\theta^{*(j)}_\alpha)_{\alpha\in\mathcal{A}}, \quad j = 1,\dots,B$$
of the estimator $(\hat\theta_\alpha)_{\alpha\in\mathcal{A}}$, where index j enumerates the bootstrap draws.

(2) For each α in $\mathcal{A}$ compute the bootstrap variance estimate
$$\hat s^2(\alpha) = B^{-1}\sum_{j=1}^{B}(\hat\theta^{*(j)}_\alpha - \hat\theta_\alpha)^2,$$
(or use the estimates based on the interquartile ranges).

(3) Compute the critical value
$$c(1-a) = (1-a)\text{-quantile of } \left\{\max_{\alpha\in\mathcal{A}} |\hat\theta^{*(j)}_\alpha - \hat\theta_\alpha|/\hat s(\alpha), \quad j = 1,\dots,B\right\}.$$

(4) Report the joint confidence region for $(\theta_\alpha)_{\alpha\in\mathcal{A}}$ of level $1-a$ as
$$[\hat\theta_\alpha \pm c(1-a)\hat s(\alpha)], \quad \alpha\in\mathcal{A}.$$

The confidence region for $(\theta_\alpha)_{\alpha\in\mathcal{A}}$ might contain nonincreasing functions if the lower and upper-end functions, $\alpha \mapsto \hat\theta_\alpha - c(1-a)\hat s(\alpha)$ and $\alpha \mapsto \hat\theta_\alpha + c(1-a)\hat s(\alpha)$, are not increasing. Since the target function $\alpha \mapsto \theta_\alpha$ is nonincreasing, we can improve the finite sample properties of the confidence region by monotonizing the lower and upper functions using the monotone rearrangement of [3]. This method is described in L7.

Figures 1 and 2 present confidence bands obtained using this algorithm.

2. The Case of Binary Outcomes: An In-Depth Look

2.1. **Modeling.** Consider the problem where the outcome variable Y is binary, taking values in $\{0,1\}$, D is a variable of interest, for example a treatment indicator, and W is the set of controls. We are interested in the predictive effect of D on Y controlling for W. We shall use an example of racial discrimination in lending, where Y is the indicator of mortgage denial, D is an indicator for the applicant being black, and W is a set of controls including financial variables and other characteristics of the applicant.

We can always begin the analysis by building predictive linear models that project outcomes on the main variable D and the controls W in the sample. Such predictive models are linear and we might wonder if we can do better with nonlinear models.

The best predictor of Y in the mean squared error sense is the conditional expectation:

$$\mathrm{E}[Y \mid X] = \mathrm{P}[Y = 1 \mid X] =: p(X), \quad X := (D, W),$$

which is nonlinear in general. This observation suggests a possibility that we can do better than linear models. Basic binary outcome models postulate the nonlinear functional forms for $p(X)$:

$$p(X) = F(B(X)'\beta),$$

where $B(X)$ is dictionary of transformations of X (such as polynomials, cosine series, linear splines, and their interactions) and F is a known link function. Such F could be either of the following:

$F(t) = 0 \vee t \wedge 1$	uniform cdf	uniform
$F(t) = \Lambda(t) = e^t/(1+e^t)$	logistic cdf	logit
$F(t) = \Phi(t)$	normal cdf	probit
$F(t) = \mathrm{C}(t) = 1/2 + \arctan(t)/\pi$	cauchy cdf	cauchit
$F(t) = \mathrm{T}_\nu(t)$	cdf of rv $t(\nu)$	student

The functional forms given above have structural interpretation in specific contexts. For example, suppose that the expected loss of the bank from denying a loan is given by

$$\underbrace{Y^*}_{\text{utility}} = \underbrace{B(X)'\gamma}_{\text{systematic part}} - \underbrace{\sigma\epsilon}_{\text{noise}},$$

where $\sigma\epsilon$ is the component that is not observed by econometrician. Then $Y = 1(Y^* > 0)$. If ϵ conditional on X is distributed according to the link F, then

$$\mathrm{P}(Y = 1 \mid X) = \mathrm{P}(\epsilon \leq B(X)'\beta \mid X) = F(B(X)'\beta), \quad \beta = \gamma/\sigma.$$

Thus, in the structural interpretation, we can think of $B(X)'\beta$ as describing the systematic or the mean part of the decision-maker's utility or value function.

Note that β identifies the structural parameter γ only up to a scale.

The models may not hold exactly, but we can expect that

$$p(X) \approx F(B(X)'\beta),$$

if the dictionary $B(X)$ is rich enough. Thus, if $B(X)$ is rich enough, the choice of F is not important theoretically, but could still matter for finite-sample performance.

Formally we could state this as follows: if $B(X)$ is a dictionary of r terms that can approximate any function $x \mapsto f(x)$ such that $\mathrm{E}f^2(X) < \infty$ in the mean square sense, namely

$$\min_{b \in \mathbb{R}^r} \mathrm{E}(f(X) - B(X)'b)^2 \to 0, \quad r \to \infty, \tag{2.1}$$

then if $f(X) = F^{-1}(p(X))$ has finite second moment,

$$\begin{aligned} &\min_{b \in \mathbb{R}^r} \mathrm{E}(p(X) - F(B(X)'b))^2 = \min_{b \in \mathbb{R}^r} \mathrm{E}(F(f(X)) - F(B(X)'b))^2 \\ &\leq \min_{b \in \mathbb{R}^r} (\bar{F}')^2 \mathrm{E}(f(X) - B(X)'b)^2 \to 0, \quad r \to \infty, \quad \bar{F}' := \sup_{t \in \mathbb{R}} \partial F(t)/\partial t. \end{aligned}$$

From this reasoning we could conclude that the choice of the link F is not theoretically important for approximating $p(X)$ if the dictionary of technical regressors is sufficiently rich. On the other hand the same theoretical observation suggests that the choice of F may be important if the dictionary is not rich enough, so in practice the choice of F could matter. In the empirical example below we observe little difference between logit and probit links, which seems to be generally the case, and observe larger differences between the predicted probability implied by the logit, cauchit, and linear models. When we observe large differences in predicted probabilities, we have to think about choosing a good link function F.

So how to choose F? A simple device we could use for choosing the functional form F as well as the number of terms in the dictionary is sample splitting. We designate a randomly selected part of the sample as a training sample and the other part as a validation sample:

$$\underbrace{(Y_1, X_1), \ldots, (Y_m, X_m)}_{\text{training sample}}, \underbrace{(Y_{m+1}, X_{m+1}), \ldots, (Y_n, X_n)}_{\text{validation sample}}.$$

We estimate β in the training sample using the maximum (quasi) likelihood estimator $\hat{\beta}$ of Section 2.2.[1] Then we form the predicted probabilities $\{F(B(X_i)'\hat{\beta})\}$, and compute the mean squared error (MSE) for predicting Y in the validation sample:

$$\frac{1}{n-m} \sum_{i=m+1}^{n} (Y_i - F(B(X_i)'\hat{\beta}))^2.$$

We choose the link that exhibits the smallest MSE in the validation sample. Unlike the in-sample MSE, this measure does not suffer from over-fitting and is a good measure of quality of various prediction procedures.

In the empirical example below we used the 2 to 1 splitting of the sample, and we conclude that logit and probit slightly outperform the linear and the cauchit model. Another approach we can pursue is sensitivity analysis, where we compute the empirical results

[1] The word quasi designates the fact that the model may not be perfect and may be misspecified in the sense that $F(B(X)'\beta) \neq p(X)$ with positive probability.

using, say, logit, and then also report additional empirical results using other links as a robustness check. In the empirical example below, for example, using the cauchit link leads to qualitatively and quantitatively similar empirical results as the logit.

2.2. **Estimation and Inference on Structural Parameters β.** Given the postulated models, we can write conditional log-likelihood of Y_i given X_i as

$$\ln f(Y_i \mid X_i, \beta_0) = Y_i \ln F(B(X_i)'\beta_0) + (1 - Y_i) \ln(1 - F(B(X_i)'\beta_0)),$$

where β_0 will designate the true value. The maximum (conditional) likelihood (ML) estimator based on the entire sample is

$$\hat{\beta} \in \arg\min_{\beta \in \mathcal{B}} -\mathbb{E}_n \ln f(Y_i \mid X_i, \beta).$$

If the model is not correctly specified, we can call the estimator the maximum quasi-likelihood estimator (QML).

By the Extremum Consitency Lemma this estimator is consistent for

$$\beta_0 = \arg\min_{\beta \in \mathcal{B}} -\mathrm{E} \ln f(Y \mid X, \beta),$$

provided that β_0 is unique. Since $\beta \mapsto \ln f(Y \mid X, \beta)$ is concave in the case of logit and probit models (and some others), β_0 is unique if the Hessian of the population objective function, the information matrix, is positive definite:

$$G = -\frac{\partial\partial}{\partial\beta\partial\beta'} \mathrm{E} \ln f(Y \mid X, \beta_0) > 0.$$

This holds under weak assumptions provided that $\mathrm{E}B(X)B(X)'$ is of full rank. The logit and probit estimators are computationally efficient because of the smoothness and convexity of the sample objective functions.

Given the postulated model, we can also use nonlinear least squares (NLS) estimators:

$$\tilde{\beta} \in \arg\min_{\beta \in \mathcal{B}} \mathbb{E}_n (Y_i - F(B(X_i)'\beta))^2.$$

By the Extremum Consistency Lemma, this will be consistent for

$$\beta_0^* = \arg\min_{\beta \in \mathcal{B}} \mathrm{E}(Y - F(B(X)'\beta))^2,$$

provided that β_0^* is unique. Note that under correct specification, NLS is less efficient than the ML and is also less computationally convenient, because the objective function is no longer convex. To have NLS as efficient as ML we need to do additional weighting by the inverse of the conditional variance of Y_i given X_i,

$$p(X_i)(1 - p(X_i)),$$

which needs to be pre-estimated by NLS. Under misspecification of the model, the NLS and ML will be consistent for different quantities β_0 and β_0^*. Note that both ML and NLS will have some good interpretability under misspecification.

Given all of the above considerations, a popular choice is to use the ML estimators even under misspecification.

We can treat both NLS and MLE as a special case of the so called M-estimators.

3. M-Estimation and Inference: General Principles

A generic M-estimator takes the form

$$\hat{\beta} \in \arg\min_{\beta \in \mathcal{B}} \mathbb{E}_n m(Z_i, \beta),$$

where $(z, \beta) \mapsto m(z, \beta)$ is a scalar valued loss function, Z_i is a random vector containing the data for the observational unit i, and β is a parameter vector defined over a parameter space $\mathcal{B} \subset \mathbb{R}^d$.

Many estimators are special cases: ordinary least squares, nonlinear least squares, maximum (quasi) likelihood, least absolute deviation and quantile regression, just to name a few.

Results such as the extremum consistency lemma suggest that $\hat{\beta}$ will be generally consistent for the solution of the population analog of the sample problem above:

$$\beta_0 = \arg\min_{\beta \in \mathcal{B}} \mathrm{E} m(Z, \beta),$$

provided that β_0 is a unique solution of the population problem.

We can also recognize the M-estimators as GMM estimators for the purpose of stating their approximate distributions. If the following FOC hold for the M-estimator

$$\mathbb{E}_n \frac{\partial}{\partial \beta} m(Z_i, \hat{\beta}) = 0,$$

then this estimator is a GMM estimator with the score function:

$$g(Z_i, \beta) = \frac{\partial}{\partial \beta} m(Z_i, \beta).$$

This reasoning suggests the following general result.

Assertion 2 (General Properties of M-estimators) *There exist mild regularity conditions under which*

$$\sqrt{n}(\hat{\beta} - \beta_0) \overset{a}{\sim} -G^{-1}\sqrt{n}\hat{g}(\beta_0) \overset{a}{\sim} G^{-1}N(0,\Omega) = N(0, G^{-1}\Omega G^{-1}),$$

where

$$\hat{g}(\beta_0) = \mathbb{E}_n g(Z_i, \beta_0), \quad G = \frac{\partial\partial}{\partial\beta\partial\beta'}\mathrm{E}m(Z_i, \beta_0), \quad \Omega = \mathrm{Var}(\sqrt{n}\hat{g}(\beta_0)).$$

This follows from the simplification of the more general result we have stated for the GMM estimator. Primitive rigorous conditions for this result could be stated along the lines of the result stated for the GMM estimator.

Note that for convex M-problems, when the loss function $\beta \mapsto m(Z_i, \beta)$ is convex and the parameter space $\mathcal{B}$ is convex, the conditions for consistency follow under very weak conditions. We state the corresponding result as a technical tool in the Appendix.

A special class of M-estimators are the ML estimators. They correspond to using the loss function

$$m(Z_i, \beta) = -\ln f(Z_i, \beta),$$

where $z \mapsto f(z, \beta)$ is the parametric density function such that $z \mapsto f(z, \beta_0)$ is the true density function of Z_i. Note that β_0 minimizes $-\mathrm{E}\ln f(Z_i, \beta)$ as long as $f(Z_i, \beta_0) \neq f(Z_i, \beta)$ with positive probability for $\beta \neq \beta_0$. Indeed, by strict Jensen's inequality,

$$\begin{aligned} \mathrm{E}\ln f(Z_i, \beta) - \mathrm{E}\ln f(Z_i, \beta_0) &= \mathrm{E}\ln f(Z_i, \beta)/f(Z_i, \beta_0) \\ &< \ln \mathrm{E} f(Z_i, \beta)/f(Z_i, \beta_0) \\ &= \ln \int \frac{f(z, \beta)}{f(z, \beta_0)} f(z, \beta_0) dz = 0, \end{aligned}$$

provided $\mathrm{E}\ln f(Z_i, \beta_0)$ is finite. This is known as the information or Kullback-Leibler inequality, and the quantity

$$\mathrm{E}\ln f(Z_i, \beta_0) - \mathrm{E}\ln f(Z_i, \beta)$$

is known as Kullback-Leibler divergence criterion. Minimizing $-\mathrm{E}\ln f(Z_i, \beta)$ is the same as minimizing the divergence criterion.

In the case of probit or logit we can think of Z_i as (Y_i, X_i) and of the density having the product form:

$$f(Z_i, \beta) = f(Y_i \mid X_i, \beta) g(X_i),$$

where g is the density of X, which is not functionally related to β. This density function drops out from the ML estimation of β, since

$$\ln f(Z_i, \beta) = \ln f(Y_i \mid X_i, \beta) + \ln g(X_i).$$

We can think of the probit or logit ML estimators as either conditional or unconditional ML estimators. Thus, effectively, the density of X drops out from the picture.

The ML is a GMM estimator with the score function

$$g(Z_i, \beta) = \frac{\partial}{\partial \beta} \ln f(Z_i, \beta).$$

Under mild smoothness conditions and under correct specification, the following relation, called the *information matrix equality*, holds:

$$G = -\frac{\partial \partial}{\partial \beta \partial \beta'} \mathrm{E} \ln f(Z_i, \beta_0) = \mathrm{Var}(g(Z_i, \beta_0)).$$

Assertion 3 (General Properties of ML-estimators) *Under correct specification, there exist mild regularity conditions under which the information matrix equality holds and the ML estimator obeys:*

$$\sqrt{n}(\hat{\beta}_{ML} - \beta_0) \quad \overset{a}{\sim} \quad -G^{-1}\sqrt{n}\hat{g}(\beta_0) \overset{a}{\sim} G^{-1}N(0, G) = N(0, G^{-1}),$$

where $\hat{g}(\beta_0) = \mathbb{E}_n g(Z_i, \beta_0)$. *Moreover, asymptotic variance matrix* G^{-1} *of the maximum likelihood estimator is smaller than asymptotic variance* V *of any other consistent and asymptotically normal GMM estimator for* β_0, *i.e.*

$$G^{-1} \leq V$$

in the matrix sense.

Note that we are better off using the robust variance formula $G^{-1}\Omega G^{-1}$ from the previous assertion for inference, because it applies both under correct and incorrect specification of the model. By contrast, we don't advise to use the nonrobust variance formula G^{-1} for inference, since the model could be misspecified, causing the information matrix equality to break, making G^{-1} an incorrect variance formula.

Note that even though we can reformulate the M-estimators as GMM estimators for theoretical purposes, we typically don't use this reformulation for computation. Indeed, some M-estimation problems, for example, logit and probit ML estimators, are computationally efficient because they solve convex minimization problems, whereas their GMM reformulation does not lead to convex optimization problems and hence is not computationally efficient.

4. Empirical Applications

4.1. **Gender Wage Gap in 2012.** We consider the gender wage gap using data from the U.S. March Supplement of the Current Population Survey (CPS) in 2012. We select white,

nonhispanic individuals who are aged 25 to 64 years and work more than 35 hours per week during at least 50 weeks of the year. We exclude self-employed workers; individuals living in group quarters; individuals in the military, agricultural or private household sectors; individuals with inconsistent reports on earnings and employment status; and individuals with allocated or missing information in any of the variables used in the analysis. The resulting sample consists of $29,217$ workers including $16,690$ men and $12,527$ of women.

We estimate interactive linear models by least squares. The outcome variable Y is the logarithm of the hourly wage rate constructed as the ratio of the annual earnings to the total number of hours worked, which is constructed in turn as the product of number of weeks worked and the usual number of hours worked per week. The key covariate D is an indicator for female worker, and the control variables W include 5 marital status indicators (widowed, divorced, separated, never married, and married); 6 educational attainment indicators (0-8 years of schooling completed, high school dropouts, high school graduates, some college, college graduate, and advanced degree); 4 region indicators (midwest, south, west, and northeast); and a quartic in potential experience constructed as the maximum of age minus years of schooling minus 7 and zero, i.e., $experience = \max(age - education - 7, 0)$, interacted with the educational attainment indicators.[2] All calculations use the CPS sampling weights to account for nonrandom sampling in the March CPS.

Table 1 reports sample means for the variables used in the analysis. Working women are more highly educated than working men, have slightly less potential experience, and are less likely to be married and more likely to be divorced or widowed. The unconditional gender wage gap is 25%.

Figure 1 of Section 1 plots point estimates and 90% confidence bands for the APE and SPEs on the treated of the conditional gender wage gap. The PEs are obtained using the interactive specification $P(T, W) = (TW, (1 - T)W)$. The distribution $M = F_{W|D=1}$ is estimated by the empirical distribution of W for women. The confidence bands are constructed by empirical bootstrap with $B = 500$ repetitions, and are uniform for the SPEs over the grid $\mathcal{A} = \{.01, .02, \ldots, .98\}$. We monotonize the bands using the rearrangement method of [3]. After controlling for worker characteristics, the gender wage gap for women remains on average around 26%. More importantly, we uncover a striking amount of heterogeneity, with the PE ranging between 0 and 43%.

Table 2 shows the results of a classification analysis, exhibiting characteristics of women that are most and least affected by gender discrimination. According to this table the 5% of the women *most affected* by gender discrimination earn higher wages, are much more likely to be married, have either very low or very high education, and possess much more potential experience than the 5% least affected women.

[2] The sample selection criteria and the variable construction follow [7].

TABLE 1. Descriptive Statistics

	All	Men	Women
Log wage	2.79	2.90	2.65
Female	0.43	0.00	1.00
Married	0.66	0.69	0.63
Widowed	0.01	0.00	0.02
Divorced	0.12	0.10	0.15
Separated	0.02	0.02	0.02
Never married	0.19	0.19	0.18
0-8 years completed	0.00	0.01	0.00
High school dropout	0.02	0.03	0.02
High school graduate	0.25	0.27	0.23
Some college	0.28	0.27	0.30
College graduate	0.28	0.28	0.29
Advanced degree	0.15	0.14	0.17
Northeast	0.20	0.20	0.19
Midwest	0.27	0.27	0.28
South	0.35	0.35	0.35
West	0.18	0.19	0.18
Potential experience	18.96	19.01	18.90
Number of observations	29,217	16,690	12,527

Source: March Supplement CPS 2012

We further explore these findings by analyzing the APE and SPE on the treated *conditional* on marital status and potential experience. Here we show estimates and 90% confidence bands of the APE and SPEs of the gender wage gap for 3 subpopulations defined by marital status (never married, married and divorced women) and 3 subpopulations defined by experience (low, medium and high experienced women, where the experience cutoffs are 11 and 26, the first and third sample quartiles of potential experience for women). The confidence bands are constructed as in fig. 1. We find significant heterogeneity in the gender gap within each subpopulation, and also between subpopulations defined by marital status and experience. The SPEs are much more heterogeneous for women with low experience and women that never married. Married and high experienced women suffer from the highest gender wage gaps. This pattern is consistent with preferences that make single young women be more career-oriented.

TABLE 2. Classification Analysis – Averages of Characteristics of the Women Least and Most Affected by Gender Discrimination

Characteristics of the Group	**5% Least Affected** PE $> -.03$	**5% Most Affected** PE $< -.39$
Log Wage	2.61	2.87
Female	1.00	1.00
Married	0.03	0.94
Widowed	0.00	0.01
Divorced	0.01	0.03
Separated	0.00	0.01
Never married	0.96	0.01
0-8 years completed	0.01	0.03
High school dropout	0.04	0.15
High school graduate	0.00	0.01
Some College	0.09	0.00
College graduate	0.61	0.12
Advanced Degree	0.25	0.69
Notheast	0.35	0.23
Midwest	0.26	0.26
South	0.23	0.29
West	0.16	0.22
Potential experience	4.31	25.70

4.2. **Analyzing Predictive Effect of Race on Mortgage Denials.** To illustrate the methods for binary outcomes, we consider an empirical application to racial discrimination in the bank decisions of mortgage denials. We use data of mortgage applications in Boston for the year 1990 collected by the Federal Reserve Bank of Boston in relation to the Home Mortgage Disclosure Act (HMDA), see [8]. The HMDA was passed to monitor minority access to the mortgage market. Providing better access to credit markets can arguably help the disadvantaged groups escape poverty traps. Following [9], we focus on white and black applicants for single-family residences. The sample comprises $2,380$ observations corresponding to $2,041$ white applicants and 339 black applicants.

The outcome variable Y is an indicator of mortgage denial, the key variable D is an indicator for the applicant being black, and the controls W contain financial and other characteristics of the applicant that the banks take into account in the mortgage decisions. These include the monthly debt to income ratio; monthly housing expenses to income ratio; a categorial variable for "bad" consumer credit score with 6 categories (1 if no slow

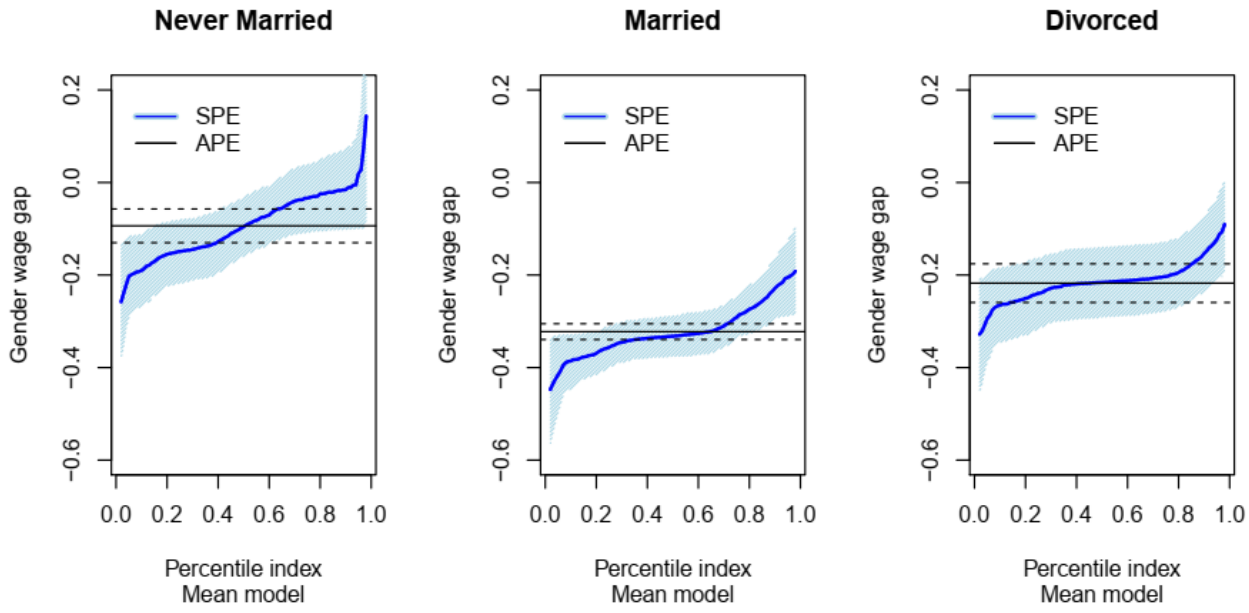

FIGURE 3. APE and SPE of the gender wage gap for women by marital status. Estimates and 90% bootstrap uniform confidence bands based on a linear model with interactions for the conditional expectation function are shown.

payments or delinquencies, 2 if one or two slow payments or delinquencies, 3 if more than two slow payments or delinquencies, 4 if insufficient credit history for determination, 5 if delinquent credit history with payments 60 days overdue, and 6 if delinquent credit history with payments 90 days overdue); a categorical variable for "bad" mortgage credit score with 4 categories (1 if no late mortgage payments, 2 if no mortgage payment history, 3 if one or two late mortgage payments, and 4 if more than two late mortgage payments); an indicator for public record of credit problems including bankruptcy, charge-offs, and collective actions; an indicator for denial of application for mortgage insurance; two indicators for medium and high loan to property value ratio, where medium is between .80 and .95 and high is above .95; and three indicators for self-employed, single, and high school graduate.

Table 3 reports the sample means of the variables used in the analysis. The probability of having the mortgage denied is 19% higher for black applicants than for white applicants. However, black applicants are more likely to have socio-economic characteristics linked to a denial of the mortgage, as Table 3 shows. Table 4 compares the unconditional effect of race with a conditional effect that controls for the variables in W using linear projection. After controlling for characteristics, black applicants are still 8% more likely to have the mortgage denied than white applicants with the same characteristics. We can interpret

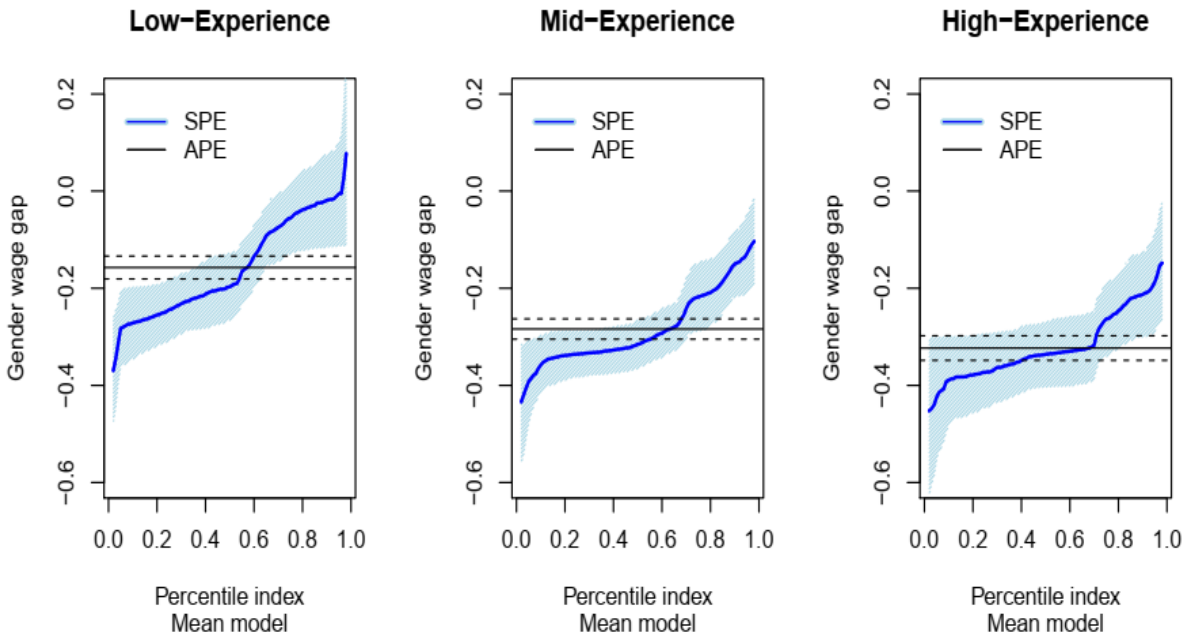

FIGURE 4. APE and SPE of the gender wage gap for women by experience level. Estimates and 90% bootstrap uniform confidence bands based on a linear model with interactions for the conditional expectation function are shown.

TABLE 3. Descriptive Statistics

	All	Black	White
deny	0.12	0.28	0.09
black	0.14	1.00	0.00
payment-to-income ratio	0.33	0.35	0.33
expenses-to-income ratio	0.26	0.27	0.25
bad consumer credit	2.12	3.02	1.97
bad mortgage credit	1.72	1.88	1.69
credit problems	0.07	0.18	0.06
denied mortgage insurance	0.02	0.05	0.02
medium loan-to-value ratio	0.37	0.56	0.34
high loan-to-value ratio	0.03	0.07	0.03
self-employed	0.12	0.07	0.12
single	0.39	0.52	0.37
high school graduated	0.98	0.97	0.99
number of observations	2,380	339	2,041

this as race having an economically and statistically significant predictive impact on being denied a mortgage.

TABLE 4. Basic OLS Results

A. The predictive effect of black on the mortgage denial rate

	Estimate	Std. Error	t value	Pr(>|t|)
base rate	0.0926	0.0070	13.16	0.0000
black effect	0.1906	0.0186	10.22	0.0000

B. The predictive effect of black on the mortgage denial rate, controlling linearly for other characteristics

	Estimate	Std. Error	t value	Pr(>|t|)
black effect	0.0771	0.0172	4.48	0.0000

We next consider nonlinear specifications to model the conditional probabilities of being denied a mortgage. We begin by examining the sensitivity to the choice of link function. Figure 5 compares the linear, cauchit and probit models with the logit model in terms of predicted probabilities. All the models use a linear index $B(X)'\beta = \beta_D D + W'\beta_W$. For this parsimonious specification of the index, the predicted probabilities are sensitive to the choice of link function. Linear and cauchit links produce substantially different probabilities from the logit link, while logit and probit probabilities are very similar. To select the link function, we apply the procedure described in Section 2 by randomly splitting the data into a training sample and a validation sample, which contain 2/3 and 1/3 of the original sample, respectively. We estimate the models in the training sample and evaluate their goodness of fit in terms of mean squared prediction error in the validation sample. The results in Table 5 show that logit and probit outperform the linear and cauchit links, but the difference is small in this application.

TABLE 5. Out-of-Sample Mean Squared Prediction Error

Logit	Probit	Cauchit	Linear
0.2703	0.2706	0.2729	0.2753

Figures 2 and 6 plot estimates and 90% confidence sets for the APE and SPE for all the applicants and the black applicants, respectively. The PEs are obtained estimating a logit model in the entire sample. The confidence sets are obtained by empirical bootstrap with

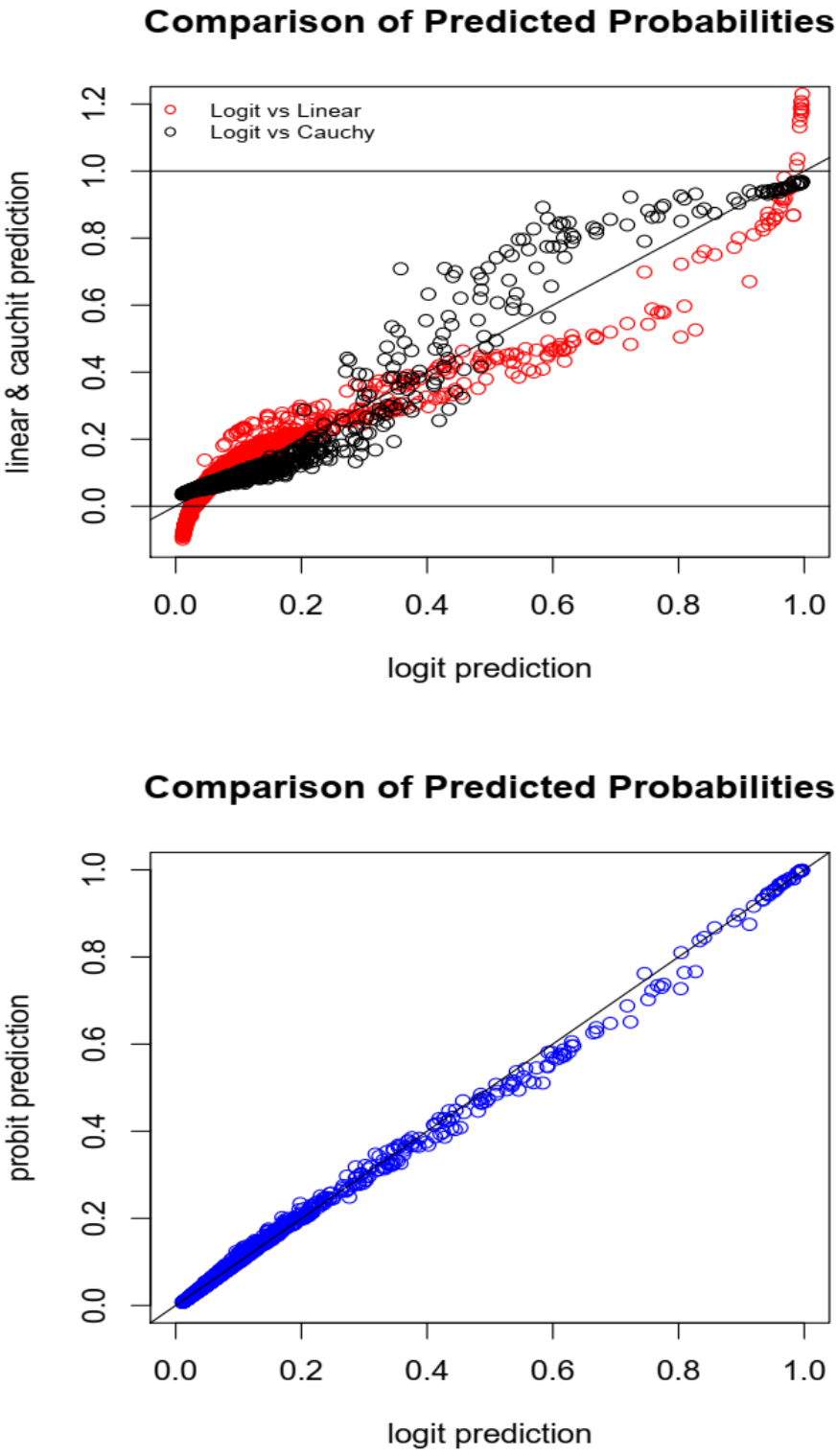

FIGURE 5. Comparison of predicted conditional probabilities of mortgage denial. The top panel compares cauchit and linear against the logit model. The bottom panel compares the probit vs logit models.

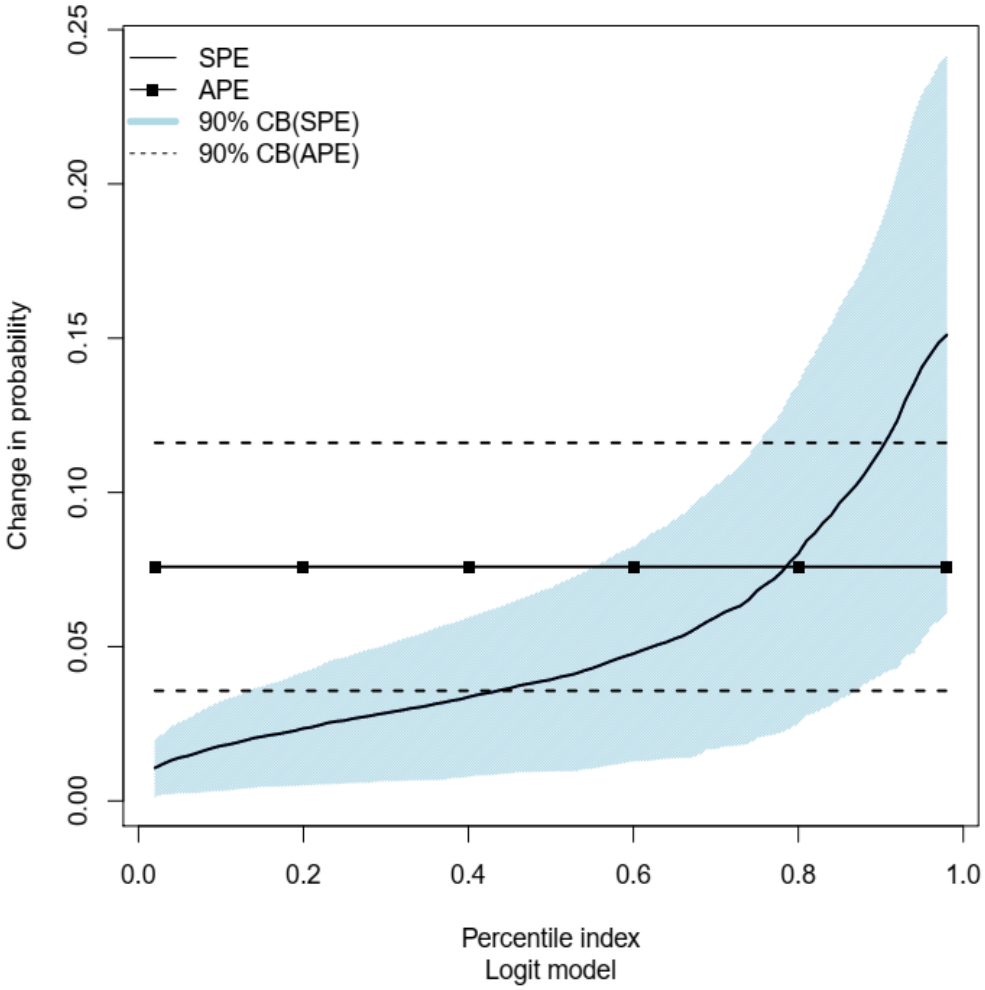

FIGURE 6. The APE and SPE of race on the predicted conditional probabilities for black applicants. 90% confidence sets obtained by empirical bootstrap with 500 repetitions.

500 repetitions and are uniform for the SPE in that they cover the entire SPE with probability 90% in large samples. Interestingly, the APE for black applicants is 7.6%, higher than the APE of 5.3% for all the applicants. The SPEs show significant heterogeneity in the effect of race, with the PE ranging between 0 and 15%. Table 6 shows that the applicants most affected by race discrimination are more likely to have either of the following characteristics: black, self employed, single and not graduated from high school, with high debt to income, expense to income and loan to value ratios, bad consumer and credit scores, credit problems, and the mortgage insurance application not denied.

TABLE 6. Classification Analysis: Averages of Characteristics of the Least and Most Affected Groups

Characteristics of the Groups	**5% Most Affected** Predictive Effect $> .14$	**5% Least Affected** Predictive Effect $< .01$
deny	0.54	0.15
black	0.41	0.09
debt-to-income ratio	0.40	0.24
expense-to-income ratio	0.29	0.20
consumer credit score	4.85	1.49
mortgage credit score	1.99	1.33
credit problem	0.64	0.10
denied mortgage insurance	0.00	0.10
medium loan-to-house ratio	0.60	0.08
high loan-to- house value	0.10	0.03
self employed	0.18	0.08
single	0.56	0.13
high school graduate	0.92	0.99

NOTES

Probit and logit binary regressions were introduced by Chester Bliss [1], Ronald Fisher [2, Appendix], and David Cox [5]. Peter Huber developed the theory for M-estimators in [6]. Reporting sorted partial effects in nonlinear regression models was proposed in [4].

APPENDIX A. TOOL: EXTREMUM CONSISTENCY LEMMA FOR CONVEX PROBLEMS

It is possible to relax the assumption of compactness in extremum consistency lemma, if something is done to keep the objective function $\hat{Q}(\theta)$ from turning back down towards its minimum, i.e. prevent *ill-posedness*. One condition that has this effect, and has many applications, is convexity (for minimization, concavity for maximization). Convexity also facilitates efficient computation of estimators.

Lemma 1. *(Consistency of Argmins with Convexity): If $\theta \mapsto \hat{Q}(\theta)$ is convex and i) $\theta \mapsto Q(\theta)$ is continuous and is uniquely minimized at θ_0; ii) Θ is a convex subset of $\mathbb{R}^k$; iii) $\hat{Q}(\theta) \to_P Q(\theta)$ for each $\theta \in \Theta$; then $\hat{\theta} \to_P \theta_0$.*

The proof of this lemma partly relies on the following result, which is a stochastic version of a well-known result from convex analysis.

Lemma 2 (Uniform pointwise convergence under convexity). *If $\theta \mapsto \hat{Q}(\theta)$ is convex and $\hat{Q}(\theta) \to_P Q(\theta) < \infty$ on an open convex set A, then for any compact subset K of A, $\sup_{\theta \in K} |\hat{Q}(\theta) - Q(\theta)| \to_P 0$, and the limit criterion function $\theta \mapsto Q(\theta)$ is continuous on A.*

Appendix B. Problems

(1) State the score function $g(\cdot, \cdot)$ and moment Jacobian matrix G for the probit and logit quasi-maximum likelihood estimators.

(2) Argue that large sample properties of the estimator of the average predictive effect (b) could be obtained via the GMM approach. Argue that you can apply bootstrap to approximate the distribution of this estimator. A challenge problem for an extra credit: provide a similar argument for the sorted predictive effect.

(3) Estimate average predictive effects and sorted predictive effects (at various percentile indices) of race on the probability of mortgage denial using the mortgage data. Explain the choice of the link function you have made and provide results for two different link functions. Report confidence intervals based on the bootstrap. You can present the results in a table or as in the figures above.

References

[1] C. I. Bliss. The method of probits. *Science*, 79(2037):38–39, 1934.
[2] C. I. Bliss. The calculation of the dosage-mortality curve. *Annals of Applied Biology*, 22(1):134–167, 1935.

[3] V. Chernozhukov, I. Fernández-Val, and A. Galichon. Improving point and interval estimators of monotone functions by rearrangement. *Biometrika*, 96(3):559–575, 2009.
[4] V. Chernozhukov, I. Fernandez-Val, and Y. Luo. The Sorted Effects Method: Discovering Heterogeneous Effects Beyond Their Averages. *ArXiv e-prints*, December 2015.
[5] D. R. Cox. The regression analysis of binary sequences. *J. Roy. Statist. Soc. Ser. B*, 20:215–242, 1958.
[6] Peter J. Huber. *Robust statistics*. John Wiley & Sons, Inc., New York, 1981. Wiley Series in Probability and Mathematical Statistics.
[7] Casey B. Mulligan and Yona Rubinstein. Selection, investment, and women's relative wages over time. *The Quarterly Journal of Economics*, 123(3):1061–1110, 2008.
[8] Alicia H. Munnell, Geoffrey M. B. Tootell, Lynn E. Browne, and James McEneaney. Mortgage lending in boston: Interpreting hmda data. *The American Economic Review*, 86(1):pp. 25–53, 1996.
[9] James Stock and Mark Watson. *Introduction to Econometrics (3rd edition)*. Addison Wesley Longman, 2011.