14.385 Fall, 2007

Nonlinear Econometrics

Lecture 2.

Theory: Consistency for Extremum Estimators

Modeling: Probit, Logit, and Other Links.

Example: Binary Choice Models. The latent outcome is defined by the equation

$$y_i^* = x_i'\beta - \varepsilon_i, \quad \varepsilon_i \sim F(\cdot).$$

We observe

$$y_i = 1(y_i^* \geq 0).$$

The cdf F is completely known. Then

$$\begin{aligned} P(y_i = 1|x_i) &= P(\varepsilon_i \leq x_i'\beta|x_i) \\ &= F(x_i'\beta). \end{aligned}$$

We can then estimate β using the log-likelihood function

$$\widehat{Q}(\beta) = E_n[y_i \ln F(x_i'\beta) + (1 - y_i) \ln(1 - F(x_i'\beta))].$$

The resulting MLE are CAN and efficient, under regularity conditions.

The story is that a consumer may have two choices, the utility from one choice is $y_i^* = x_i'\beta - \varepsilon_i$ and the utility from the other is normalized to be 0. We need to estimate the

parameters of the latent utility based on the observed choice frequencies.

Estimands: The key parameters to estimate are $P[y_i = 1|x_i]$ and the partial effects of the kind

$$\frac{\partial P[y_i = 1|x_i]}{\partial x_{ij}} = f(x_i'\beta)\beta_j,$$

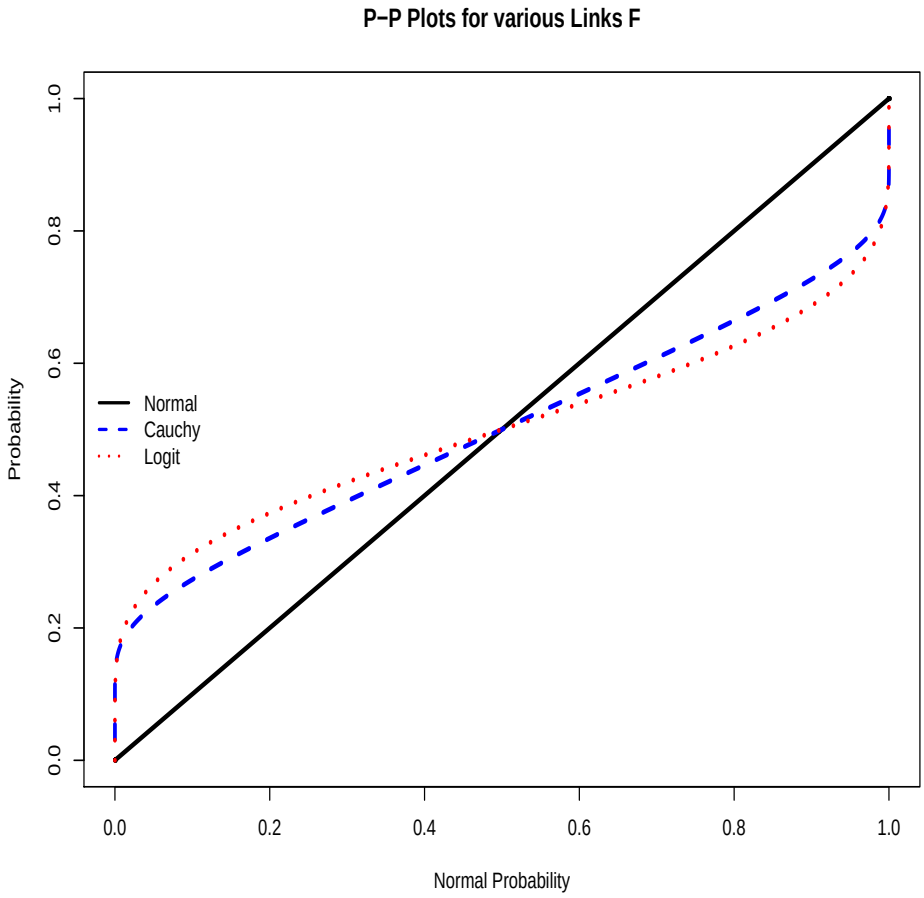
where $f = F'$. These parameters are functionals of parameter β and the link F.

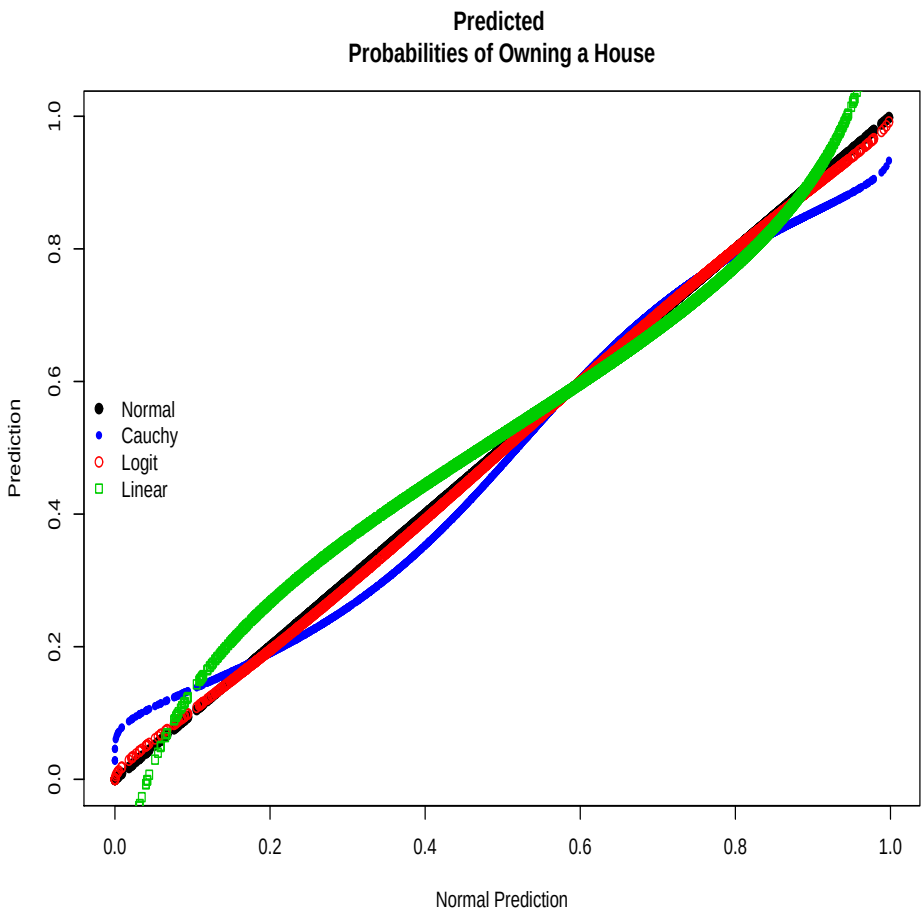
Choices of F:

- Logit: $F(t) = \Lambda(t) = \frac{\exp(t)}{1+\exp(t)}$.
- Probit: $F(t) = \Phi(t)$, standard normal cdf.
- Cauchy: $F(t) = C(t) = \frac{1}{2} + \frac{1}{\pi}\arctan(t)$, the Cauchy cdf.
- Gosset: $F(t) = T(t, v)$, the cdf of t-variable with v degrees of freedom.

Choice of $F(\cdot)$ can be important especially in the tails. The prediction of small and large

probabilities by different models may differ substantially. For example, Probit and Cauchit links, $\Phi(t)$ and $C(t)$, have drastically different tail behavior and give different predictions for the same value of the index t. See Figure 1 for a theoretical example and Figure 2 for an empirical example. In the housing example, y_i records whether a person owns a house or not, and x_i consists of an intercept and person's income.



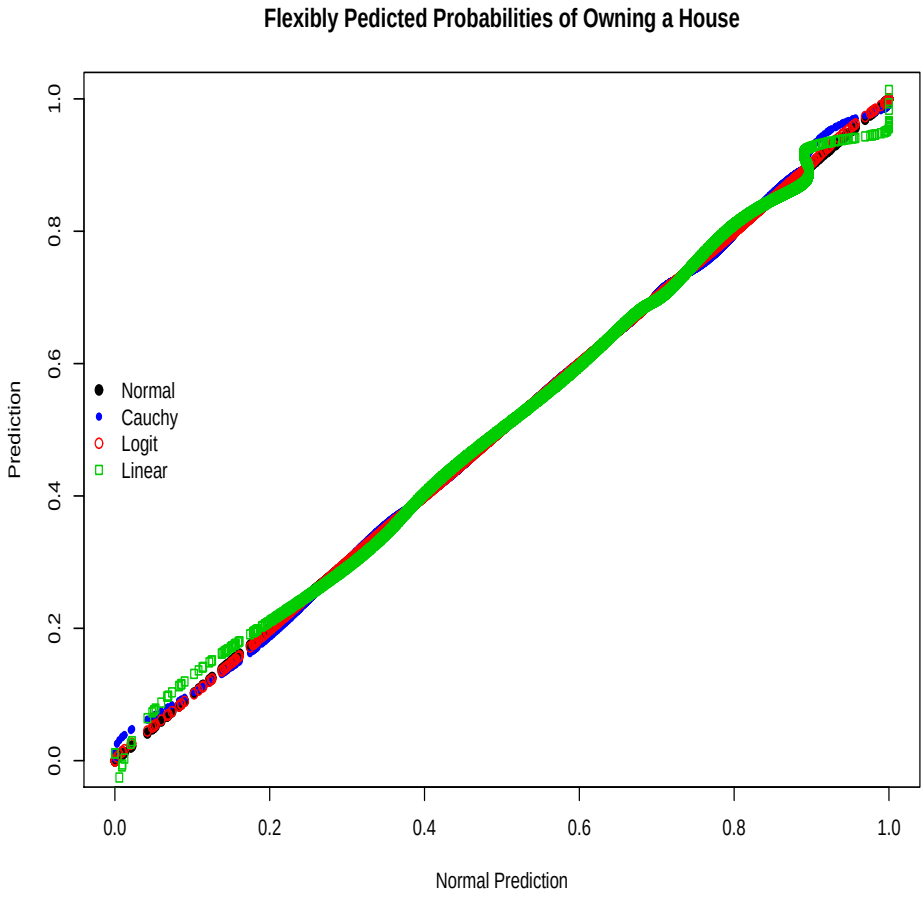


Choice of $F(\cdot)$ can be less important when using flexible functional forms. Indeed, for any F we can approximate

$$P[y_i = 1|x] \approx F[P(x)'\beta],$$

where $P(x)$ is a collection of approximating functions, for example, splines, powers, or other

series, as we know from the basic **approximation theory**. This point is illustrated in the following Figure, which deals with an earlier housing example, but uses flexible functional form with $P(x)$ generated as a cubic spline with ten degrees of freedom. Flexibility is great for this reason, but of course has its own price: additional parameters lead to increased estimation variance.



Discussion: Choice of the right model is a hard and very important problem is statistical analysis. Using flexible links, e.g. t-link vs. probit link, comes at a cost of additional parameters. Using flexible expansions inside the links also requires additional parameters. Flexibility reduces the approximation error (bias),

but typically increases estimation variance. Thus an optimal choice has to balance these terms. A useful device for choosing best performing models is cross-validation.

Reading: A very nice reference is R. Koenker and J. Yoon (2006) who provide a systematic treatment of the links, beyond logits and probits, with an application to propensity score matching. The estimates plotted in the figures were produced using R language's package glm. The Cauchy, Gosset, and other links for this package were implemented by Koenker and Yoon (2006).

(http://www.econ.uiuc.edu/~roger/research/links/links.html)

References:

Koenker, R. and J. Yoon (2006), "Parametric Links for Binary Response Models."

1. Extremum Consistency

Extremum estimator

$$\widehat{\theta} = \arg \min_{\theta \in \Theta} \widehat{Q}(\theta).$$

As we have seen in Lecture 1, extremum estimator encompasses nearly all estimators (MLE, M, GMM, MD). Thus, consistency of all these estimators will follow from consistency of extremum estimators.

Theorem 1 *(Extremum Consistency Theorem) If i) (Identification) $Q(\theta)$ is uniquely minimized at the true parameter value θ_0; ii) (Compactness) Θ is compact; iii) (Continuity) $Q(\cdot)$ is continuous; iv) (uniform convergence)* $\sup_{\theta \in \Theta} \left| \widehat{Q}(\theta) - Q(\theta) \right| \xrightarrow{p} 0$; *then* $\widehat{\theta} \xrightarrow{p} \theta_0$.

Intuition: Just draw a picture that describes the theorem.

Proof: The proof has two steps: 1) we need to show $Q(\hat{\theta}) \to_p Q(\theta_0)$ using the assumed uniform convergence of $\hat{Q}$ to Q, and 2) we need to show that this implies that $\hat{\theta}$ must be close to θ_0 using continuity of Q and the fact that θ_0 is the unique minimizer.

Step 1. By uniform convergence,

$$\hat{Q}(\hat{\theta}) - Q(\hat{\theta}) \xrightarrow{p} 0 \text{ and } \hat{Q}(\theta_0) - Q(\theta_0) \xrightarrow{p} 0.$$

Also, by $\hat{Q}(\hat{\theta})$ and $Q(\theta_0)$ being minima,

$$Q(\theta_0) \leq Q(\hat{\theta}) \quad \text{and} \quad \hat{Q}(\hat{\theta}) \leq \hat{Q}(\theta_0).$$

Therefore

$$\begin{aligned} Q(\theta_0) &\leq Q(\hat{\theta}) = \hat{Q}(\hat{\theta}) + [Q(\hat{\theta}) - \hat{Q}(\hat{\theta})] \\ &\leq \hat{Q}(\theta_0) + [Q(\hat{\theta}) - \hat{Q}(\hat{\theta})] \\ &= Q(\theta_0) + \underbrace{[\hat{Q}(\theta_0) - Q(\theta_0) + Q(\hat{\theta}) - \hat{Q}(\hat{\theta})]}_{o_p(1)}, \end{aligned}$$

implying that

$$Q(\theta_0) \leq Q(\hat{\theta}) \leq Q(\theta_0) + o_p(1)$$

It follows that $Q(\widehat{\theta}) \xrightarrow{p} Q(\theta_0)$.

Step 2. By compactness of Θ and continuity of $Q(\theta)$, for any open subset $\mathcal{N}$ of Θ containing θ_0, we have that

$$\inf_{\theta \notin \mathcal{N}} Q(\theta) > Q(\theta_0).$$

Indeed, $\inf_{\theta \notin \mathcal{N}} Q(\theta) = Q(\theta^*)$ for some $\theta^* \in \Theta$. By identification, $Q(\theta^*) > Q(\theta_0)$.

But, by $Q(\widehat{\theta}) \xrightarrow{p} Q(\theta_0)$, we have

$$Q(\widehat{\theta}) < \inf_{\theta \notin \mathcal{N}} Q(\theta)$$

with probability approaching one, and hence $\widehat{\theta} \in \mathcal{N}$ with probability approaching one. *Q.E.D.*

Discussion:

1. The first and big step in verifying consistency is that the limit $Q(\theta)$ is minimized at the

parameter value we want to estimate. $Q(\theta)$ is always minimized at something under the stated conditions, but it may not be the value we want. If we have verified this part, we can proceed to verify other conditions. Because $Q(\theta)$ is either an average (in M-estimation) or a transformation of an average (in GMM), we can verify the uniform convergence by an application of one of the uniform laws of large numbers (ULLN).

The starred comments below are technical in nature.

2.* The theorem can be easily generalized in several directions: (a) continuity of the limit objective function can be replaced by the lower-semi-continuity, and (b) uniform convergence can be replaced by, for example, the one-sided uniform convergence: $\sup_{\theta\in\Theta} |\hat{Q}(\theta) - Q(\theta)|_- \to_p 0$ and $|\hat{Q}(\theta_0) - Q(\theta_0)| \to_p 0$.

Van der Vaart (1998) presents several generalizations. Knight (1998) "Epi-convergence and Stochastic Equi-semicontinuity" is another good reference.

3.* Measurability conditions are subsumed here in the general definition of $\to_p$ (convergence in outer probability). Element

$$X_n = \sup_{\theta \in \Theta} \left| \hat{Q}(\theta) - Q(\theta) \right|$$

converges in outer probability to 0, if for any $\epsilon > 0$ there exists a measurable event Ω_ϵ which contains the event $\{X_n > \epsilon\}$, not necessarily measurable, and $Pr(\Omega_\epsilon) \to 0$. We also denote this statement by $Pr^*(X_n > \epsilon) \to 0$. This approach, due to Hoffman-Jorgensen, allows to bypass measurability issues and focus more directly on the problem. Van der Vaart (1998) provides a very good introduction to the Hoffman-Jorgensen approach.

3. Uniform Convergence and Uniform Laws of Large Numbers.

Here we discuss how to verify the uniform convergence of the sample objective function $\widehat{Q}$ to the limit objective function Q. It is important to emphasize here that **pointwise convergence**, that is, that for each θ, $\widehat{Q}(\theta) \to_p Q(\theta)$, does not imply **uniform convergence** $\sup_{\theta \in \Theta} |\widehat{Q}(\theta) - Q(\theta)| \to_p 0$. It is also easy to see that pointwise convergence is not sufficient for consistency of argmins. (Just draw a "moving spike" picture).

There are two approaches:

a. Establish uniform convergence results directly using a ready uniform laws of large numbers that rely on the fact that many objective functions are averages or functions of averages.

b. Convert familiar pointwise convergence in probability and pointwise laws of large numbers to uniform convergence in probability and the uniform laws of large numbers. This conversion is achieved by the means of the **stochastic equi-continuity.**

a. Direct approach

Often $\widehat{Q}(\theta)$ is equal to a sample average (M-estimation) or a quadratic form of a sample average (GMM).

The following result is useful for showing conditions iii) and iv) of Theorem 1, because it gives conditions for continuity of expectations and uniform convergence of sample averages to expectations.

Lemma 1 *(Uniform Law of Large Numbers) Suppose* $\widehat{Q}(\theta) = E_n[q(z_i, \theta)]$. *Assume that data*

$z_1, ..., z_n$ is stationary and strongly mixing and i) $q(z, \theta)$ is continuous at each θ with probability one; ii) Θ is compact, and iii) $E\left[\sup_{\theta \in \Theta} |q(z, \theta)|\right] < \infty$; then $Q(\theta) = E[q(z, \theta)]$ is continuous on Θ and $\sup_{\theta \in \Theta} |\widehat{Q}(\theta) - Q(\theta)| \xrightarrow{p} 0$.

Condition i) allows for $q(z, \theta)$ to be discontinuous (with probability zero). For instance, $q(z, \theta) = 1(x'\theta > 0)$ would satisfy this condition at any θ where $\Pr(x'\theta = 0) = 0$.

Theoretical Exercise. Prove the ULLN lemma. Lemmas that follow below may be helpful for this.

b. Equicontinuity Approach*.

First of all, we require that the sample criterion function converges to the limit criterion function **pointwise**, that is for, each θ,

$\widehat{Q}(\theta) \to_p Q(\theta)$. We also require that $\widehat{Q}$ has equicontinuous behavior. Let

$$\widehat{\omega}(\epsilon) := \sup_{\|\theta - \theta'\| \leq \epsilon} |\widehat{Q}(\theta) - \widehat{Q}(\theta')|$$

be a measure of oscillation of a function over small neigborhoods (balls of radius ϵ). This measure is called the modulus of continuity. We require that these oscillations vanish with the size of the neigborhoods, in large samples. Thus, we rule out "moving spikes" that can be consistent with pointwise convergence but destroy the uniform convergence.

Lemma 2 *For a compact parameter space Θ, suppose that the following conditions hold: i) (pointwise or finite-dimensional convergence)*

$$\widehat{Q}(\theta) \to_p Q(\theta) \text{ for each } \theta \in \Theta \ .$$

ii) (stochastic equicontinuity) For any $\eta > 0$ there is $\epsilon > 0$ and a sufficiently large n' such that for all $n > n'$

$$Pr^*\{\widehat{\omega}(\epsilon) > \eta\} \leq \eta.$$

Then $\sup_{\theta \in \Theta} |\widehat{Q}(\theta) - Q(\theta)| \to_p 0$.

The last condition is a stochastic version of the Arzella-Ascolli's equicontinuity condition. It is also worth noting that this condition is equivalent to the finite-dimensional approximability: the function $\widehat{Q}(\theta)$ can be approximated well by its values $\widehat{Q}(\theta_j), j = 1, ..., K$ on a finite fixed grid in large sample. The uniform convergence then follows from the convergence in probability of $\widehat{Q}(\theta_j), j = 1, ..., K$ to $Q(\theta_j), j = 1, ..., K$.

Theoretical Exercise. Prove Lemma 2, using the comment above as a hint.

Simple sufficient conditions for stochastic equicontinuity and thus uniform convergence include the following.

Lemma 3 (Uniform Holder Continuity) *Suppose* $\widehat{Q}(\theta) \to_p Q(\theta)$ *for each* $\theta \in \Theta$ *compact. Then*

the uniform convergence holds if the uniform Holder condition holds: for some $h > 0$, uniformly for all $\theta, \theta' \in \Theta$

$$|\widehat{Q}(\theta) - \widehat{Q}(\theta')| \leq B_T \|\theta - \theta'\|^h, \quad B_T = O_p(1).$$

Lemma 4 (Convexity Lemma) *Suppose $\widehat{Q}(\theta) \to_p Q(\theta)$ for almost every $\theta \in \Theta_0$, where Θ_0 is an open convex set in $\mathbb{R}^k$ containing compact parameter set Θ, and that $\widehat{Q}(\theta)$ is convex on Θ_0. Then the uniform convergence holds over Θ. The limit $Q(\theta)$ is necessarily convex on Θ_0.*

Theoretical Exercise. Prove Lemma 3 and 4. Proving Lemma 3 is trivial, while proving Lemma 4 is not. However, the result is quite intuitive, since, a convex function cannot oscillate too much over small neighborhood, provided this function is pointwise close to a continuous function. We can draw a picture to convince ourself of this. Lemma 4 is a stochastic version of a result from convex analysis. Pollard (1991) provides a nice proof of this result.

A good technical reference on the uniform convergence is Andrews (1994).

4. Consistency of Extremum Estimators under Convexity. It is possible to relax the assumption of compactness, if something is done to keep $\widehat{Q}(\theta)$ 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.

Theorem 2 *(Consistency with Convexity): If* $\widehat{Q}(\theta)$ *is convex on an open convex set* $\Theta \subseteq R^d$ *and (i)* $Q(\theta)$ *is uniquely minimized at* $\theta_0 \in \Theta$*. ; (ii) (iii)* $\widehat{Q}(\theta) \xrightarrow{p} Q(\theta)$ *for almost every* $\theta \in \Theta$ *; then* $\widehat{\theta} \xrightarrow{p} \theta_0$*.*

Theoretical Exercise. Prove this lemma. Another consequence of convexity is that the uniform

convergence is replaced by pointwise convergence. This is a consequence of the Convexity Lemma.

References:

Pollard, D. Asymptotics for least absolute deviation regression estimators. Econometric Theory 7 (1991), no. 2, 186–199.

Andrews, D. W. K. Empirical process methods in econometrics. Handbook of econometrics, Vol. IV, 2247–2294, Handbooks in Econom., 2, North-Holland, Amsterdam, 1994.