Optimal Rates for Regression Coefficients
Heteroskedastic Binary Response Models

Shakeeb Khan*
Duke University

May 2011

Abstract

In this paper optimal rates of convergence for regression coefficients in distribution free heteroskedastic binary response models is considered. The motivation for deriving these rates is based on the relation between a distribution free model with a conditional median restriction on the disturbance term and a heteroskedastic probit model discussed in Khan(2011). The main result in this note is that the optimal attainable rate for the regression coefficients is the same in the two models.

JEL Classification: C13,C14,C24

Key Words: binary response, heteroskedasticity, probit, optimal rates.

*Corresponding author. Department of Economics, Duke University I am grateful to S. Chen, X. Chen, M. Coppejans, B. Honoré, A. Lewbel, W. Newey, W. Ploberger, J. Powell, and seminar participants at Boston College, Brown, McGill, Harvard/MIT, Rice, and Texas A&M for helpful comments. This research was supported in part by the National Science Foundation through grant SES-0213621.
1 Model and Optimal Rates of Convergence

As discussed in the Khan(2011) the heteroskedastic probit model is equivalent to a distribution free model with a conditional median restriction on the disturbance term. These models impose stronger restrictions than those imposed in Horowitz(1992), and are more similar to the model considered in Chamberlain(1986). Horowitz(1993) derived the optimal rates for estimating $\beta_0$ under the assumptions in Horowitz(1992). Chamberlain(1986) showed the parametric rate of $\sqrt{n}$ was not achievable in the model he considered, but did not derive the fastest achievable rate.

In this note we derive an upper bound for the fastest achievable rate\(^{1}\) for $\beta_0$ in a heteroskedastic probit model where we impose smoothness conditions on the probability function and the distribution of the regressors. Here we strengthen the conditions in Model 2 in the previous section to enable estimation of the probability function, and derive the upper bound on the fastest achievable rate for estimating regression coefficients.

Before detailing the stronger assumptions imposed, we first introduce some notation which will be used in imposing smoothness and compactness conditions. The notation adopted here is identical to that used in Ai and Chen(2003), Chen et al.(2003a). For any $k \times 1$ vector $v = (v_1, v_2, ... v_k)',$ let $|v|$ denote $\sum_{i=1}^{k} v_i.$ Let $h(\cdot)$ denote any function on $\mathcal{X}.$ We denote the $|v|$-th derivative of the function $h(\cdot)$ as:

$$\nabla v h(x) = \frac{\partial^{|v|}}{\partial x_1^{v_1} ... \partial x_k^{v_k}} h(x)$$

Also, for $\gamma > 0$ we let $\Lambda^\gamma(\mathcal{X})$ denote the space of functions which have up to $[\gamma]$ (here $[\cdot]$ denotes the integer operator) continuous derivatives with the highest derivatives that are Holder continuous\(^{2}\) with exponent $(\gamma - [\gamma]).$

Let $\| \cdot \|_E$ denote the Euclidean norm. For a real valued function $h(\cdot) \in \Lambda^\gamma(\mathcal{X})$ we define its Holder norm as

$$\|h\|_{\Lambda^\gamma} = \sup_{x \in \mathcal{X}} |h(x)| + \max_{|v| = [\gamma]} \max_{x \neq \bar{x}} \frac{\|\nabla v h(x) - \nabla v h(\bar{x})\|_E}{\sqrt{(x - \bar{x})'(x - \bar{x})^{-[\gamma]}}}$$

\(^{1}\)For a formal definition of the upper bound on rates of convergence of estimators see Stone(1980) or Horowitz(1993).

\(^{2}\)See Ai and Chen(2003) for a formal definition of Holder continuity and a more detailed discussion on Holder Spaces.
Finally we denote a space of functions that will be used in defining the parameter space:

$$\Lambda^\gamma(\mathcal{X}, w_1) \equiv \{ h \in \Lambda^\gamma(\mathcal{X}) : \left\| h(\cdot)(1 + x'x)^{-w_1/2} \right\|_{\Lambda^\gamma} \leq c < \infty \}$$

where $w_1 > 0$ and $c$ is a known constant.

The weighting function of the regressors $(1 + x'x)^{-w_1/2}$ goes to 0 as $\|x\|_E$ goes to infinity and permits $h(\cdot)$ and its derivatives to be unbounded.\(^3\)

The model to be considered for deriving an upper bound on rates of convergence for the regression coefficients is characterized by the following conditions:

**OR1** the vector $(y_i, x_i)$ is i.i.d. where $y_i = I[x'_i\beta_0 - \sigma_0(x_i)\epsilon_i \geq 0]$, $x_i \in \mathbb{R}^k$ is a random vector, $\beta_0 \in \mathbb{R}^k$ is a constant vector, and $\epsilon_i$ is a random variable distributed standard normal.

**OR2** The support of the distribution of $x_i$ is not contained in any proper linear subspace of $\mathbb{R}^k$. Furthermore, letting $x_i^{[k,1]}$ denote the last component of the vector $x_i$, and letting $\tilde{x}_i$ denote the first $k - 1$ components, the distribution of $x_i^{[k,1]}$ conditional on $\tilde{x}_i$ has everywhere positive density with respect to Lebesgue measure.

**OR3** Let $z_{0i} = x'_i\beta_0$. Denoting the conditional density of $z_{0i}$ given $\tilde{x}_i$ by $f_{Z|\tilde{X}}(\cdot|\cdot)$ we assume $f_{Z|\tilde{X}}(z_0|\tilde{x})$ is continuously differentiable in $z_0$ with bounded derivatives for all $z$ in a neighborhood of 0 and all $\tilde{x}$.

**OR4** Let $g_0(x_i) \equiv 1/\sigma_0(x_i)$. The function $\Phi(x'_i\beta_0 \cdot g_0(x)) \in \Lambda^p_c(\mathcal{X}, w_1)$, where $p > k/2, w_1 > 0$.

**OR5** $\int (1 + \|x\|_E^2)^{w_2} f_X(x) dx < \infty$ where $w_2 \geq w_1 + p$.

The fastest possible rate for estimating $\beta_0$ in the model characterized by the above conditions is stated in the following theorem, whose proof is left to the appendix:

**Theorem 1.1** Under conditions OR1-OR5, an upper bound on the fastest achievable rate for estimating $\beta_0$ is $n^{-p/(2p+1)}$.

We note that this rate is the same as in Horowitz(1993). What the theorem implies is that although the assumptions of the heteroskedastic probit model considered here are much

\(^3\)We note the results in this paper do not require adopting a weighting function of this form. Examples of other weighting functions, such as $\exp(-x'_ix_i)$, that can be used can be found in Gallant and Nychka(1987).
stronger than those in Horowitz(1992,1993), they are not that much stronger in the sense that \( \beta_0 \) cannot be estimated at a faster rate. The global structure imposed in assumption A5, which is stronger than the “local” structure imposed in Horowitz(1993) has no effect on the attainable rates.

**References**


**A Appendix**

**A.1 Proof of Theorem 1.1**

In deriving an upper bound on the rate of convergence, we parallel the arguments in Horowitz(1993), and deliberately keep notation as close as possible to notation used in that paper. We refer the reader to Horowitz(1993) or Stone(1980) for the formal definition of the upper bound on the rate of convergence of an estimator.

From Proposition 1 in Horowitz(1993), to establish that \( n^{-\frac{p}{2k+1}} \) is indeed an upper bound on the rate of convergence, we wish to construct a sequence of scale functions \( g_n(z_i, \tilde{x}_i) \) which satisfies the assumptions of the model, and a local asymptotic normality condition (LAN) holds.

We define this LAN condition as follows. Let \( g_0(x_i) \) denote the true scale function, and reparameterize so \( g_0(x_i) = \tilde{g}_0(z_{0i}, \tilde{x}_i) \), where \( z_{0i} = x_i' \beta_0 \), with \( \beta_0 = (\theta_0', 1)' \) denoting the true value. Let \( \psi(\beta_0) \) be the \((k-1) \times (k-1)\) matrix:

\[
\left[ 2 \phi(0)^2 \int \tilde{g}_0^2(0, \tilde{x}) \tilde{x} \tilde{x}' f_{Z|X}(0|\tilde{x}) dP(\tilde{x}) \right]^{-1/2}
\]
Let \( \theta_n = \theta_0 + n^{-\frac{p}{p+1}} \Psi(\beta_0)'t \) where \( t \in \mathbb{R}^{k-1} \). Let \( \beta_n = (1, \theta_n)' \), \( z_{ni} = x_i'\beta_n \), and \( \tau_n = \theta_n - \theta_0 \). Let \( \tilde{g}_n(z_{ni}, \tilde{x}_i) \) denote a sequence of positive scale functions. The LAN condition involves showing the likelihood ratio

\[
\sum_{i=1}^{n} (1 - y_i) \log \left( \frac{\Phi(-z_{ni}\tilde{g}_n(z_{ni}, \tilde{x}_i))}{\Phi(-z_{0i}\tilde{g}_0(z_{0i}, \tilde{x}_i))} \right) + y_i \log \left( \frac{1 - \Phi(-z_{ni}\tilde{g}_n(z_{ni}, \tilde{x}_i))}{1 - \Phi(-z_{0i}\tilde{g}_0(z_{0i}, \tilde{x}_i))} \right) \tag{A.1}
\]

has the asymptotic representation:

\[
Z_n^* - \frac{1}{2} t't + o_p(1) \tag{A.2}
\]

where \( Z_n \) has a standard multivariate normal distribution.

Thus to prove the theorem, we wish to define a sequence of scale functions such that the assumptions of the model are satisfied, and LAN holds. To simplify exposition, we will assume in the rest of this section that \( \tilde{x}_i \) has bounded support. This will not affect the main results which will continue to hold if sufficiently many moments of \( \tilde{x}_i \) are finite. (Nor will it affect identification of the model, as we still have \( x_i'\beta_0 \) has support on the real line.

First let \( h_n = n^{-\frac{1}{2}} \). Define the sequence as follows:

\[
\tilde{g}_n(z_n, \tilde{x}) = I[|z_n| > h_n]\tilde{g}_0(z_n, \tilde{x}) \frac{z_n}{h_n} + I[|z_n| \leq h_n]\tilde{g}_0(z_n, \tilde{x}) \frac{\text{sgn}(z_0)h_n}{z_n - z_0 + \text{sgn}(z_0)h_n} \tag{A.3}
\]

We comment on the how the above sequence behaves as \( z_n \) get near 0. First note that since \( z_n - z_n = \tilde{x}'(\theta_0 - \theta_n) \), if \( |z_n| > h_n \), for \( n \) large enough, \( |z_n| > h_n/2 \) either since \( \tilde{x}_i \) has been assumed to have bounded support. (Note this event will occur with probability approaching 1 sufficiently rapidly with \( n \) if \( \tilde{x}_i \) has unbounded support but enough moments of \( \tilde{x}_i \) exist.) We also note that for \( n \) sufficiently large each scale function in the sequence is smooth at \( h_n \) and positive for \( z_n \) close to \( z_0 \). The spline constructed is continuous, but stronger smoothness could be imposed by constructing a polynomial spline. This too will not change the main results but change the constant \( (\frac{2}{3}) \) in the definition of \( \Psi(\beta_0) \).

To show LAN we follow the procedure in proving theorem 1 in Horowitz(1993). We first adopt the notation \( \tilde{g}_{ni} \) for \( \tilde{g}_n(z_{ni}, \tilde{x}_i) \) and \( \tilde{g}_0 \) for \( \tilde{g}_0(z_{0i}, \tilde{x}_i) \). In our proof \( z_{ni} = x_i'\beta_n \) where \( \beta_n = (\theta_n', 1)' \) and recall \( \theta_n = \theta_0 + n^{-\frac{p}{p+1}} \Psi(\theta_0)'t \).

Define:

\[
Z_{ni} = \frac{\Phi(-z_{ni}\tilde{g}_{ni}) - \Phi(-z_{0i}\tilde{g}_0)}{\Phi(-z_{0i}\tilde{g}_0)(1 - \Phi(-z_{0i}\tilde{g}_0))} (1 - y_i - \Phi(-z_{0i}\tilde{g}_0)) \tag{A.4}
\]

Let \( L_n \) denote the expression in (A.1). Note we have

\[
L_n = \sum_{i=1}^{n} \log(1 + Z_{ni}) \tag{A.5}
\]
A mean value expansion yields that the right hand side of the above equation can be expressed as:

\[
\sum_{i=1}^{n} Z_{ni} - \frac{1}{2} \sum_{i=1}^{n} (1 + K_{ni})^{-2} Z_{ni}^2
\]  
(A.6)

where \( K_{ni} \) is in between 0 and \( Z_{ni} \). We first turn attention to \( \sum_{i=1}^{n} Z_{ni} \). Note we have

\[
E[Z_{ni}|z_{ni}, z_{0i}, \tilde{x}_i] = 0
\]  
(A.7)

Note (since \( z_{ni} = z_{0i} + \tilde{x}'_i \tau_n \)) the variance can be expressed as

\[
n \int_{|z_0| \leq h_n} \frac{(\Phi((z_0 - \tilde{x}'_n \tau_n)\tilde{g}_n) - \Phi(z_0\tilde{g}_0))^2}{\Phi(z_0\tilde{g}_0)(1 - \Phi(-z_0\tilde{g}_0))} dP(z_0, \tilde{x})
\]  
(A.8)

where here \( \tilde{g}_n \) denotes \( \tilde{g}_0(z_0, \tilde{x}) \frac{sgn(z_0)h_n}{\tilde{x}'_n \tau_n + sgn(z_0)h_n} \), \( \tilde{g}_0 \) denotes \( \tilde{g}_0(z_0, \tilde{x}) \), and \( P(z_0, \tilde{x}) \) denotes the distribution function of \( z_{0i}, \tilde{x}_i \). For notational convenience we drop the \( sgn(z_0) \) here. A mean value expansion around \( \tau_n = 0 \) yields that

\[
\Phi((z_0 - \tilde{x}'_n \tau_n)\tilde{g}_n) - \Phi(z_0\tilde{g}_0) = \\
\phi \left( (z_0 - \tilde{x}'\lambda_{1n})\tilde{g}_0(z_0, \tilde{x}) \frac{h_n}{h_n + \tilde{x}'\lambda_{1n}} \right) \times \\
\tilde{g}_0(z_0, \tilde{x}) \left( \frac{h_n}{h_n + \tilde{x}'\lambda_{1n}} + (z_0 - \tilde{x}'\lambda_{1n}) \frac{h_n}{(h_n + \tilde{x}'\lambda_{1n})^2} \right) \tilde{x}'_n \tau_n
\]  
(A.9)

where \( \lambda_{1n} \) denotes an intermediate value. Note that \( \tau_n = O(n^{-p/(2p+1)}) \), so \( ||\tau_n||^2_F = O(n^{-2p/(2p+1)}) \). As the integral in (A.8) is over \( |z_0| \leq h_n \), we can perform a change of variables, and because \( h_n = n^{-1/(2p+1)} \), the expression in (A.8) converges to \( t't \) as \( n \to \infty \). It follows that (see, e.g. Horowitz(1992(Lemma 6b))

\[
L_{ni} = Z't + o_p(1)
\]  
(A.10)

Next we turn attention to

\[
-\frac{1}{2} \sum_{i=1}^{n} (1 + K_{ni})^{-2} Z_{ni}^2
\]  
(A.11)

Since \( K_{ni} \) converges to 0, the same arguments may be used to conclude that

\[
E \left[ -\frac{1}{2} \sum_{i=1}^{n} (1 + K_n)^{-2} Z_{ni}^2 \right] \to -\frac{1}{2} t't
\]  
(A.12)
turning attention to the variance of (A.11), we note that

\[ n^{-1/2} E \left[ -\frac{1}{2} \sum_{i=1}^{n} (1 + K_n)^{-2} Z^2_{ni} \right] \rightarrow 0 \]  

(A.13)

Also, we note that

\[ \frac{n}{4} E[(1 + K_n)^{-4} Z^4_{ni}] \]

is of the same order as

\[ \frac{n}{4} \left\{ \int_{|z_0| \leq h_n} (\tilde{x}' \tau_n)^4 \phi^4 \left( (-z_0 - \tilde{x}' \lambda_{1n}) \tilde{g}_0(z_0, \tilde{x}) \frac{h_n}{h_n + \tilde{x}' \lambda_{1n}} \right) \times \right. \]

\[ \left. \tilde{g}_0^4(z_0, \tilde{x}) \left( \frac{h_n}{h_n + \tilde{x}' \lambda_{2n}} + (z_0 - \tilde{x}' \lambda_{2n}) \frac{h_n}{(h_n + \tilde{x}' \lambda_{2n})^2} \right)^4 dP(z_0, \tilde{x}) \right\} \]  

(A.14)

where \( \lambda_{2n} \) denotes an intermediate value. This sequence (again after employing a change of variables) is \( O(n \| \tau_n \|_E^4 h_n) \) which converges to 0, implying the variance of (A.11) converges to 0. Thus we have (A.11) converges in probability to \( -\frac{1}{2} t't \). Combining results we have established the LAN condition, proving the theorem.