Discrete Choice: Qualitative Response
FALL 2007

Types of discrete choice models:

1. Univariate Models
   1. Binary: Linear; Probit; Logit; Arctan, etc.
   2. Multinomial: Logit; Nested Logit; GEV; Probit.

2. Multivariate Models

Chapter 9 in Amemiya covers essentially every thing you need to know. Read pp267–278; pp281–
311. pp319–pp348 are also good to look at if you have time. Maddala Ch1-3, Ch5 are also good to
look at for more intuitions.

Binary Choice: $y_i^* = x_i'\beta + \varepsilon_i$. $\varepsilon_i \sim F(\cdot)$. $F(\cdot)$ is completely known, symmetric around 0.
Observe $y_i = 1 (y_i^* \geq 0)$. The $P(y_i = 1) = P(\varepsilon_i > -x_i'\beta) = 1 - F(-x_i'\beta) = F(x_i'\beta)$. The choice
maximizes the latent utility $y_i^*$. Only the choice is observed. Can’t estimate scale parameter in
$F(\cdot)$ so $F(\cdot)$ must be assumed known.

Choice of $F$:

1. Logit: $F(t) = \Lambda(t) = \frac{\exp(t)}{1+\exp(t)}$.
2. Probit: $F(t) = \Phi(t)$ for $\Phi(t)$ the standard normal distribution function.
3. Arctan: $F(t) = \frac{1}{2} + \frac{1}{\pi}\arctan t$. This is for $\varepsilon$ having a cauchy distribution: $f(t) = \frac{1}{\pi(1+t^2)}$.

Is the choice of $F(\cdot)$ important? NO and Yes. If the truth is $P(y_i = 1) = G(x_i'beta)$ for some
$G(\cdot)$ but you specify $P(y = 1) = F(x'\beta)$, then there should be, for some unknown $H(\cdot)$, the truth
is $P(y_i = 1) = F(H(x_i, \beta))$. To the extend that $H(x_i, \beta)$ can be approximated by a linear function
function $z_i'\beta$, for some transformation $z_i$ of $x_i$, then it is just ok to run $P(y_i = 1) = F(z_i'\beta)$. So the
choice of $z_i$, the transformation of $x_i$, is the important factor. Although in practice, you usually
determine your choice of $x_i$ by how you need to interprete your choice and won’t transform it into
$z_i$ simply for getting better fit. For that matter the choice of $F(\cdot)$ is in fact important.

Is the estimate $\hat{\beta}$ important: No. $\hat{\beta}$ depends on how you specify and scale $F(\cdot)$. The important
fact for comparing the estimate of different models is $\frac{\partial F(x, \beta)}{\partial x}$, the marginal propensity of choosing
$y = 1$ given $x$. Usually this is evaluated at the average of $x$, the regressors, and this is the number
to report in estimation result tables. Comparing this number across Probit or Logit or possibly other $F(\cdot)$ usually gives very similar results.

Maximum Likelihood Estimation:

$$L = \prod_{i=1}^{n} F(x_i'\beta)^{y_i} (1 - F(x_i'\beta))^{1-y_i}$$

$$\log L = \sum_{i=1}^{n} y_i \log F(x_i'\beta) + (1-y_i) \log (1 - F(x_i'\beta))$$

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} \left( \frac{y_i}{F(x_i'\beta)} - \frac{1-y_i}{1 - F(x_i'\beta)} \right) f(x_i'\beta)x_i' = \sum_{i=1}^{n} \frac{y_i - F(x_i'\beta)}{F(x_i'\beta)(1 - F(x_i'\beta))} f(x_i'\beta)x_i'$$
\[ \frac{\partial^2 \log L (\beta_0)}{\partial \beta \partial \beta'} = \sum_{i=1}^{n} \frac{y_i - F_i}{F_i (1 - F_i)} f'(x_i' \beta) x_i x'_i + \sum_{i=1}^{n} \frac{(y_i - F_i) f_i^2}{F_i (1 - F_i)} \left( \frac{1}{F_i (1 - F_i)} \right) x_i x'_i - \sum_{i=1}^{n} \frac{f_i^2}{F_i (1 - F_i)} x_i x'_i \]

\[ \overset{LD}{=} - \sum_{i=1}^{n} \frac{f_i^2}{F_i (1 - F_i)} x_i x'_i \quad \text{since} \quad E(y_i - F(x'_i \beta) | x_i) = 0. \]

On the other hand,

\[ \frac{1}{\sqrt{n}} \frac{\partial \log L}{\partial \beta_0} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{y_i - F_i}{F_i (1 - F_i)} f'(x_i' \beta) x_i \overset{A}{\to} N \left( 0, \frac{1}{n} \sum_{i=1}^{n} \frac{f_i^2}{F_i (1 - F_i)} x_i x'_i \right) \]

Since \( \text{Var}(y_i - F_i | x_i) = F_i (1 - F_i) \). Therefore, as usual for MLE:

\[ \sqrt{n} \left( \hat{\beta} - \beta \right) \overset{LD}{=} - \left[ \frac{\partial^2 \log L (\beta_0)}{\partial \beta \partial \beta'} \right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial \log L}{\partial \beta_0} \overset{d}{\to} N \left( 0, \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{f_i^2}{F_i (1 - F_i)} x_i x'_i \right]^{-1} \right) \]

**Probit and Logit Case:** The two distribution functions are similar in both location and scale. Only the tail of Logit is flatter. The variance of logit = \( \frac{\pi^2}{3} \). Scaling the logit estimate of \( \beta \) by \( \frac{\sqrt{3}}{\pi} \) should be close to the probit estimate.

For probit: \( \sqrt{n} \left( \hat{\beta} - \beta \right) \overset{d}{\to} N \left( 0, \left[ \frac{1}{n} \sum_{i=1}^{n} \frac{\phi^2}{F_i (1 - F_i)} x_i x'_i \right]^{-1} \right) \).

For logit: \( f(t) = \Lambda'(t) = \Lambda(t) (1 - \Lambda(t)) \). So

\[ \sqrt{n} \left( \hat{\beta} - \beta \right) \overset{d}{\to} N \left( 0, \left[ \frac{1}{n} \sum_{i=1}^{n} \Lambda(x'_i \beta) (1 - \Lambda(x'_i \beta)) x_i x'_i \right]^{-1} \right) \]

You can also do nonlinear least square by minimizing \( \hat{\beta} = \arg \min \sum_{i=1}^{n} (y_i - F(x'_i \beta))^2 \) by using the relations:

\[ y_i = F(x'_i \beta) + \epsilon_i \quad E(\epsilon_i | x_i) = 0 \quad V(\epsilon_i | x_i) = F(x'_i \beta) (1 - F(x'_i \beta)) \]

\( \hat{\beta} \) is not efficient because of heteroscedasticity so you can do weighted least square:

\[ \hat{\beta} = \arg \min \sum_{i=1}^{n} \frac{1}{F(x'_i \beta) \left( 1 - F(x'_i \beta) \right)} (y_i - F(x'_i \beta))^2 \]

But the usual choice for Probit and Logit is MLE, because the least square problem is also nonlinear and doesn’t seem to have advantage. Plus log \( L \) for probit and logit are both concave, so any local max is global max.

**Discriminant Analysis (DA):** Although we will find similar functional form to the logit model and to linear probability model, this is something conceptual different. Instead of specifying the conditional probability of \( P(y = 1 | x) \) given \( x \) (and leave \( f(x) \) unspecified), the discriminant analysis method specifies the density of \( x \) given \( y = 1 \) or \( y = 0 \), and tries to predict whether \( x \) belongs to group \( y = 1 \) or \( y = 0 \) by looking at whether \( P(y|x) > \frac{1}{2} \) or not. The logit model results when
you specifies the distribution of $x$ given $y$ as normal given equal variance: $(x|y = 1) \sim N(\mu_1, \Sigma)$, $(x|y = 0) \sim N(\mu_0, \Sigma)$, $P(y = 1) = p_1$, $P(y = 0) = p_0$, then just use the Bayes rule:

$$P(y = 1|x) = \frac{q_1 N(\mu_1, \Sigma)}{q_1 N(\mu_1, \Sigma) + q_0 N(\mu_0, \Sigma)} = \frac{1}{1 + \exp \left( x' \Sigma (\mu_1 - \mu_0) + \ln \frac{q_1}{q_0} - \frac{1}{2} \mu_1 \Sigma^{-1} \mu + \frac{1}{2} \mu_0 \Sigma^{-1} \mu_0 \right)} = \frac{1}{1 + \exp \left( x' \beta + C \right)}$$

for some constant $C$ and for $\beta = \Sigma^{-1} (\mu_1 - \mu_0)$.

You can estimate $\hat{p}_0, \hat{q}_1, \hat{\mu}_1, \hat{\mu}_0, \hat{\Sigma}$ from sample moments, then construct $\hat{\beta}$ from the above formula. Read Amemiya for details. If you only care about $P(y = 1|x)$ then you can estimate $\beta$ and $C$ using a logit model. Both are consistent, the sample moments are more efficient.

But if the true model is fact a logit model, and NOT a DA model, then you have to estimate it using the logit likelihood. If you use the sample moments and construct $\beta$ through $\hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}, \Sigma$, etc, as if it is a DA model, then it will give inconsistent estimate of $\beta$. Without specifying $f(x)$ in the logit model, it is not clear what relation each of these sample moments has with respect to $\beta$.

If the true model is DA, then it can be shown that if you run the linear regression: $y = \hat{a} + \hat{b} x$, then $\hat{b} = \lambda \beta = \lambda \Sigma^{-1} (\hat{\mu}_1 - \hat{\mu}_0)$ for some constant proportional $\lambda$. For details see Maddala pp16-21.

**Grouped data, min $\chi^2$:** There is one exception when you can’t use MLE. This is when you have only grouped data but not individual data available. Say, if you have the proportion ($\hat{P}_t$) of people voting for “yes” for some bill, as well as the population for each of the states, but you don’t have the individual record of voting yes or no. Regressor $x_t$ is discrete, finite. Let $P_t = F(x_t \beta)$, then $\hat{P}_t = P_t + \epsilon_t$ for $\epsilon_t = 0$ and $\text{Var}(\epsilon_t) = P_t (1 - P_t)$. Note that $F^{-1}(P_t) = x_t \beta$ since you can invert $F(\cdot)$. If you do a Taylor expansion

$$F^{-1}(\hat{P}_t) \approx F^{-1}(P_t) + \frac{\partial F^{-1}(P_t)}{\partial P_t} (\hat{P}_t - P_t) = x_t \beta + v_t$$

for $E v_t = 0$ and $\text{Var}(v_t) = \frac{P_t (1 - P_t)}{n_t f^2(F^{-1}(P_t))}$. You can then estimate $\text{Var}(v_t)$ by plug in $\hat{P}_t$ into where $P_t$ is and run a WLS of this taylor expansion:

$$\hat{\beta} = \arg\min \sum_{t=1}^T n_t f^2 \frac{\left( F^{-1}(\hat{P}_t) \right)^2}{\hat{P}_t (1 - \hat{P}_t)} \left( F^{-1}(\hat{P}_t) - x_t \beta \right)^2$$

Of course you can also run a nonlinear weighted least without inverting $F^{-1}(P_t)$:

$$\hat{\beta} = \arg\min \sum_{t=1}^T \frac{n_t}{\hat{P}_t (1 - \hat{P}_t)} \left( \hat{P}_t - F(x_t \beta) \right)^2$$

But $\hat{\beta}$ is preferable because it is a linear problem rather a nonlinear problem. For details and asymptotic, read pp275-278 in Amemiya.

**Multinomial Models:** Utility from taking car, train and bus:

1. Bus($y = 1$): $U_{1i} = \mu_{1i} + \epsilon_{1i}$
2. Train\((y = 2)\): \(U_{2i} = \mu_{2i} + \epsilon_{2i}\)

3. Car\((y = 3)\): \(U_{3i} = \mu_{3i} + \epsilon_{3i}\)

Then \(P (y_i = 1) = P (U_{1i} > U_{2i}, U_{1i} > U_{3i})\). \(P (y_i = 2) = P (U_{2i} > U_{1i}, U_{3i} > U_{1i})\), and \(P (y_i = 3) = P (U_{3i} > U_{1i}, U_{3i} > U_{2i})\).

**Multinomial Logit:** \(\epsilon_{1i}, \epsilon_{2i}, \epsilon_{3i}\) iid, each one distributed as Type-I extreme Value: \(F (\epsilon) = e^{-e^{-\epsilon}}\). \(f (\epsilon) = e^{-\epsilon} e^{-e^{-\epsilon}}\). The probability of choice can be calculated:

\[
P (y_i = 2) = P (\epsilon_2 + \mu_2 - \mu_1 > \epsilon_1, \epsilon_2 + \mu_2 - \mu_3 > \epsilon_3)
= \int_{-\infty}^{\infty} f (\epsilon_2) P (\epsilon_3 < \epsilon_2 + \mu_2 - \mu_1) P (\epsilon_1 < \epsilon_2 + \mu_2 - \mu_1) \, d\epsilon_2
= \int_{-\infty}^{\infty} e^{-\epsilon_2} \left[ e^{-\epsilon_2} \left[ 1 + e^{\mu_1 - \mu_2 + \epsilon_2} \right] \right] \, d\epsilon_2
= \int_{0}^{\infty} e^{-[1 + e^{\mu_1 - \mu_2 + \epsilon_2}] x} \, dx
= \frac{1}{1 + e^{\mu_1 - \mu_2 + \mu_2 - \mu_2}} = \frac{e^{\mu_2}}{e^{\mu_1} + e^{\mu_2} + e^{\mu_3}}.
\]

Similarly, \(P (y_i = 1) = \frac{e^{\mu_1}}{e^{\mu_1} + e^{\mu_2} + e^{\mu_3}}\), \(P (y_i = 2) = \frac{e^{\mu_2}}{e^{\mu_1} + e^{\mu_2} + e^{\mu_3}}\), \(P (y_i = 3) = \frac{e^{\mu_3}}{e^{\mu_1} + e^{\mu_2} + e^{\mu_3}}\). Using these numbers it is easy to that

\[
P (y = 1|y = 1 \text{ or } y = 2) = \frac{e^{\mu_1}}{e^{\mu_1} + e^{\mu_2}}

P (y = 3|y = 1 \text{ or } y = 3) = \frac{e^{\mu_3}}{e^{\mu_1} + e^{\mu_3}}
\]

etc.

**IIA property and nested logit:** IIA-independence of irrelevant alternatives. I make some incorrect claims in class by looking at a uninformative pair of conditional probabilities. Here is the correct way of looking at the IIA property and how the nested logit helps in overcoming the problem. In order not to be more confusing, the clarification of the (incorrect)example we look at in class is postponed after we first state the correct examples below.

We will look at the correct statement first: suppose \(y = 2\) is not available, then

\[
P (y = 3 \text{ without choice } 2) = P (U_3 > U_1) = P (\mu_3 + \epsilon_3 > \mu_1 + \epsilon_1) = \int_{-\infty}^{\infty} F (\epsilon_3 + \mu_3 - \mu_1) f (\epsilon_3) \, d\epsilon_3
= \int_{-\infty}^{\infty} e^{-e^{-(\epsilon_3 + \mu_3 - \mu_1)}} e^{-\epsilon_3} e^{-e^{-\epsilon_3}} \, d\epsilon_3
= \int_{-\infty}^{\infty} e^{-\epsilon_3 e^{-[1 + e^{\mu_1 - \epsilon_3} + \mu_3]}} e^{\epsilon_3} \, d\epsilon_3
= \left[ \frac{1}{1 + e^{-\mu_3 + \mu_1}} e^{-[1 + e^{\mu_1 - \epsilon_3}] e^{-\epsilon_3}} \right]_{-\infty}^{\infty} = \frac{1}{1 + e^{-\mu_3 + \mu_1}} = \frac{e^{\mu_3}}{e^{\mu_1} + e^{\mu_3}}.
\]

This defines the IIA(independence of irrelevant alternatives) property: the relative frequency of choosing between 1(bus) and 3(car) does not change whether or not choice 2(train) is present. This appears unreasonable since bus(1) and train(2) are both public transportation and appears correlated with each other. Therefore the introduction of train should affect the relative probability between bus and car.

**Nested Logit:** To allow for correlation between choice 1 and 2, introduce \(F (\epsilon_1, \epsilon_2, \epsilon_3) = G (\epsilon_1, \epsilon_2) F (\epsilon_3) = G (\epsilon_1, \epsilon_2) e^{-e^{-\epsilon_3}}\) and \(G (\epsilon_1, \epsilon_2) = e^{-[e^{-\epsilon_1/r} + e^{-\epsilon_2/r}]^\rho}\), for \(0 < \rho \leq 1\). Contrary to what I tried to claim...
In class, the following relation, although true, is derived from both the independence of \( \epsilon_3 \) and the nested logit specification for \( G(\epsilon_1, \epsilon_2) \):

\[
P(U_1 > U_2) = P(U_1 > U_2|U_1 > U_3 \text{ or } U_2 > U_3) = \frac{P(y = 1)}{P(y = 1) + P(y = 2)}
\]

In fact, both sides are verified to be \( \frac{e^{\mu_1}}{e^{\mu_1} + e^{\mu_2}} \). To verify the right hand side \( P(U_1 > U_2|U_1 > U_3 \text{ or } U_2 > U_3) \), it is easiest to follow Maddala pp70-71 on the GEV (Generalized Extreme Value) models, using (3.26) to arrive at (3.28) to plug in \( P(y = 1) \) and \( P(y = 2) \). The left hand side \( P(U_1 > U_2) \) can be verified directly:

\[
P(U_1 > U_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(\epsilon_1, \epsilon_2) \, d\epsilon_2 \, d\epsilon_1 = \int_{-\infty}^{\infty} \frac{dG(\epsilon_1, \mu_1 - \mu_2)}{d\epsilon_1} \, d\epsilon_1
\]

\[
= \int_{-\infty}^{\infty} e^{-\epsilon_1/\rho} \left[ e^{-\epsilon_1/\rho} + e^{-\mu_1 + \mu_2/\rho} \right]^{\rho - 1} e^{-\mu_1/\rho} \, d\epsilon_1
\]

\[
= \frac{1}{1 + e^{(\mu_2 - \mu_1)/\rho}} = \frac{e^{\mu_1/\rho}}{e^{\mu_1/\rho} + e^{(\mu_2 - \mu_1)/\rho}}
\]

You can also use Maddala (3.28) pp71, or by brute-force integration as in (9.3.53) p301 in Amemiya, to find that:

\[
P(y = 3) = P(U_3 > U_1, U_3 > U_2) = \frac{e^{\mu_3}}{e^{\mu_3} + (e^{\mu_1/\rho} + e^{(\mu_2 - \mu_1)/\rho})^3}
\]

Now the IIA property does not hold since

\[
P(y = 3 \text{ with no choice 2}) = P(U_3 > U_1) \neq P(U_3 > U_1|U_3 > U_2 \text{ or } U_1 > U_3) = P(y = 3|y = 1 \text{ or } y = 3)
\]

**Estimation Method:** A convenient way is to use a 2-step procedure to estimate \( \beta \) and \( \rho \) first, and then use the resulting estimate as the initial value for MLE iteration. This is important especially when the choice set in nest logit is large.

Write the likelihood function:

\[
L = \prod_{i=1}^{n} P(y = 1)^{y_{1i}} \cdot P(y = 2)^{y_{2i}} \cdot P(y = 3)^{y_{3i}} = L_1 \cdot L_2
\]

for

\[
L_1 = \prod_{i=1}^{n} P(y = 1|y = 1\text{ or }2)^{y_{1i}} \cdot P(y = 2|y = 1\text{ or }2)^{y_{2i}}
\]

\[
L_2 = \prod_{i=1}^{n} P(y = 1\text{ or }2)^{y_{1i} + y_{2i}} \cdot P(y = 3)^{y_{3i}}
\]

Note that \( L_1 \) is the logit likelihood for the subsample that choose between 1 and 2, and is a function of only \( \beta/\rho \):

\[
\log L_1 = \sum_{i=1}^{n} y_{1i} \frac{e^{x_{1i}\beta/\rho}}{e^{x_{1i}\beta/\rho} + e^{x_{2i}\beta/\rho}} + y_{2i} \frac{e^{x_{2i}\beta/\rho}}{e^{x_{1i}\beta/\rho} + e^{x_{2i}\beta/\rho}}
\]

The two step estimation goes by first estimate \( \left( \frac{\hat{\beta}}{\hat{\rho}} \right) \) by maximizes \( L_1 \), and then plug this into \( L_2 \) and the maximizes wrt \( \beta \) and \( \rho \):

\[
\log L_2 = \sum_{i=1}^{n} (y_{1i} + y_{2i}) \log \left( 1 - \frac{e^{x_{1i}\beta}}{e^{x_{1i}\beta} + \left( e^{x_{1i}(\hat{\beta}/\hat{\rho})} + e^{x_{1i}(\hat{\beta}/\hat{\rho})} \right)^\rho} \right) + y_{3i} \log \left( \frac{e^{x_{2i}\beta}}{e^{x_{3i}\beta} + \left( e^{x_{1i}(\hat{\beta}/\hat{\rho})} + e^{x_{1i}(\hat{\beta}/\hat{\rho})} \right)^\rho} \right)
\]
After you obtain estimate of $\hat{\beta}$ of $\rho$ from this two step procedure, use them as the initial starting value for maximizing the likelihood function $L$. In fact for this simple example, there is not much point, it is easier to simply maximize the $L$ directly (FIML: full information maximum likelihood), instead of using the sequential procedure first. However, if the choice set is large and there are many branches, the two step procedure is important to find a good starting value. For general setup of nest logit and high level nest, look in Amemiya pp303–305.

**Limitation of Logit and Probit:** Nested Logit is one way to allow for correlation between different choices. However, it suffers from the problem that you have to decide on the independence and correlation structure of the choice set. In particular, you need to decide beforehand which choices are correlated and which ones are independent. Probit is a way to allow for flexible correlation among different choices, read Amemiya pp307–309. But multinomial probit suffers from the computational burden of integrating high dimensional normal density. This comes the big business of simulation estimation. We shall defer discussion of probit until we get into simulation problem later.

**Distribution Free Methods:** This is extra material. Recall the latent utility model is specify as $y^* = x'\beta + \epsilon$ and we observe $y = 1 (y^* \geq 0)$. All parametric models rely on specifying a completely known distribution function $F(\cdot)$ for $\epsilon$. The aim of distributional free (or you may call semiparametric) methods is to relax this parametric assumption, for the reason that it has been shown that an incorrect parametric specification of $F(\cdot)$, as well as the presence of heteroscedasticity, can lead to inconsistent parameter estimates. There are essentially two types of distributional free methods. One is to allow for heteroscedasticity of $\epsilon$ while maintaining minimal identifying moment conditions, this is Manski’s maximum score estimator. The other is to relax all restriction on $F(\cdot)$ while maintaining conditional homoscedasticity, this is Cossett (1983), Klein and Spady (1993), among others. You may read about Cossett (1983) and Manski (1975)(1985) in Amemiya pp339-347, as well as consulting the original papers. Only a few points are highlighted here.

Maximum score estimator of Manski: The identifying assumption is that $P(\epsilon < 0|x) \equiv \frac{1}{2}$ for all $x$. The estimator is to maximize the number of correct predictions (which is considered the “score”), as choosing $\beta$ to maximize (subject to identification conditions on $\beta$):

$$Q_n(\beta) = \sum y_i 1(x_i' \beta \geq 0) + (1 - y_i) 1(x_i' \beta < 0) = \frac{1}{n} \sum_{i=1}^{n} q(y_i, x_i, \beta)$$

Alternatively, you can think of it as minimizing the number of wrong predictions. Since the prediction is wrong if and only if $|y_i - 1(x_i' \beta)| = 1$, it can be written as a LAD (least absolute deviation):

$$\frac{1}{n} \sum_{i=1}^{n} |y_i - 1(x_i' \beta)\geq 0|$$

In fact, when $x_i' \beta_0 \geq 0$, $P(y_i = 1) \geq \frac{1}{2}$, so $med(y_i|x_i) = 1$. On the other hand, when $x_i' \beta_0 < 0$, $P(y_i = 0) \geq \frac{1}{2}$ so $med(y_i|x_i) = 0$. The maximum score estimator is indeed a LAD conditional median estimator.

Kim and Pollard (1990) derives the asymptotic distribution and $N^{\frac{1}{4}}$ rate of convergence of this estimator. The asymptotic distribution is very difficult but there is an intuitive reason for the slow rate $n^{\frac{-1}{2}}$ of convergence:

Consider $\beta$ maximizes:

$$Q_n(\beta) - Q_n(\beta_0) = [Q(\beta) - Q(\beta_0)] + [Q_n(\beta) - Q_n(\beta_0) - Q(\beta) + Q(\beta_0)] = A + B_n$$
with \( A \) and \( B_n \) defined as the first and second brackets, similar to that in lecture note 3. Now \( A \) is as before approximated by a quadratic function: \( A \approx -O \left( (\beta - \beta_0)^2 \right) \), but the convergence rate for \( B_n \) is somewhat different:

As usual, the convergence rate for \( B_n \), which is a mean 0 process, should be \( \sqrt{\text{Var}(B_n)} \), so it is sufficient to look at \( \text{Var}(B_n) \). To make things convenient, let's assume \( x_i' \beta = \beta - z_i \)

\[
\text{Var}(B_n) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(q(y_i, x_i, \beta) - q(y_i, x_i, \beta_0)) \approx \frac{1}{n} E \left( q(y_i, x_i, \beta) - q(y_i, x_i, \beta_0) \right)^2
\]

\[
= \frac{1}{n} E \left[ y_i 1(x_i' \beta \geq 0, x_i' \beta < 0) - (1-y_i) 1(x_i' \beta_0 < 0, x_i' \beta \geq 0) \right]^2
\]

\[
\approx O \left( \frac{1}{n} E \left[ 1(\beta \geq z_i \geq \beta_0) + 1(\beta \leq z_i \leq \beta_0) \right] \right)
\]

\[
= O \left( \frac{1}{n} |\beta - \beta_0| \right)
\]

So \( B_n = O_p \left( \frac{1}{\sqrt{n}} \sqrt{|\beta - \beta_0|} \right) \), this is different from the regular linear regression case, where we have found that \( B_n = O_p \left( \frac{1}{\sqrt{n}} |\beta - \beta_0| \right) \). Then you could expect \( \hat{\beta} \) to maximize an objective function that looks like:

\[
A + B_n = -O \left( (\beta - \beta_0)^2 \right) + O_p \left( \frac{1}{\sqrt{n}} \sqrt{|\beta - \beta_0|} \right)
\]

But since \( A(\beta_0) + B_n(\beta_0) = 0 \), it is necessary that \( A(\hat{\beta}) + B_n(\hat{\beta}) \geq 0 \), so that

\[
-O \left( (\hat{\beta} - \beta_0)^2 \right) + O_p \left( \frac{1}{\sqrt{n}} \sqrt{|\beta - \beta_0|} \right) \geq 0
\]

which implies that

\[
O \left( (\hat{\beta} - \beta_0)^2 \right) \leq O_p \left( \frac{1}{\sqrt{n}} \sqrt{|\beta - \beta_0|} \right) \implies O \left( |\hat{\beta} - \beta_0|^2 \right) \leq O_p \left( \frac{1}{\sqrt{n}} \right)
\]

which in turn finally implies that

\[
O \left( |\hat{\beta} - \beta_0| \right) \leq O_p \left( n^{-\frac{1}{4}} \right)
\]