8. Appendix II. Proof of Theorem 3 in

"Testing Independence And Goodness-of-fit In Linear Regression Models Allowing Non-existence Of The Mean Of The Reponse Variable"

Qiqing Yu [1] **and Ruiqi Liu** [2]

***This version* November 11, 2017**

The proof of Theorem 3 is quite long. We need to establish certain preliminary results for its proof.

Let Q be a probability measure and $\mathcal{F} \subset L_2(Q)$ be a class of functions. We say a measurable function $F_{\mathcal{F}}$ is an envelop of $\mathcal{F}$ if $\sup_{f\in\mathcal{F}} |f| \leq F_{\mathcal{F}}$. Define a ball with radius $\epsilon > 0$ and center $f \in \mathcal{F}$ by $B(f, \epsilon) = \{g \in \mathcal{F} | \int |g - f|^2 dQ \leq \epsilon^2\}$. Define the covering number $N(\epsilon, L_2(Q), \mathcal{F})$ as the smallest integer m such that there exist $f_1, f_2, ..., f_m \in L_2(Q)$ satisfying $\mathcal{F} \subset \cup_{i=1}^m B(f_i, \epsilon)$. Let $h(x, y)$ be a symmetric function in the sense that $h(x, y) = h(y, x)$ and h is said to be a Q-degenerate kernel if $E(h(Z_1, Z_2)|Z_2) = 0$ almost sure, where Z_1, Z_2 are i.i.d. with the distribution associated with the probability measure Q.

Let $(\mathbb{D}, d, \mathcal{B}_d)$ be a metric space with metric d and Borel sigma algebra $\mathcal{B}_d$ generated by all open subsets of $\mathbb{D}$. Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}}, P)$ be a probability space, define outer measure P^* as

$$P^*(B) = \inf_{A\in\mathcal{B}_{\mathcal{X}}, A\supset B} P(A), \text{ for } B \subset \mathcal{X},$$

and inner measure P_* as

$$P_*(B) = 1 - P^*(\mathcal{X} \setminus B), \text{ for } B \subset \mathcal{X}.$$

The outer expectation of D, denoted as E^*D, is defined as the infimum over all EU, where $U : \mathcal{X} \to \mathbb{R}$, U is measurable and $U \geq T$. Analogically, the inner expectation is defined as $E_*D = -E^*(-D)$. Denote $C_b(\mathbb{D})$ the collection of all bounded continuous functions from $\mathbb{D} \to \mathbb{R}$. Let $\{X_n, n \geq 1\}, X$ be maps from $\mathcal{X}$ to $\mathbb{D}$, and X be measurable. $\{X_n, n \geq 1\}$ is said to be asymptotically measurable if

$$E^*(f(X_n)) - E_*(f(X_n)) \to 0, \text{ for all } f \in C_b(\mathbb{D}).$$

[1]Department of Mathematics, State University of New York, Binghamton, NY, 13902
[2]Department of Mathematics, State University of New York, Binghamton, NY, 13902

And $\{X_n, n \geq 1\}$ is asymptotically tight if for every $\epsilon > 0$, there is a compact subset K_ϵ of $\mathbb{D}$ such that

$$\liminf_{n\to\infty} P_*(X_n \in K_\epsilon^\delta) \geq 1 - \epsilon, \text{ for all } \delta > 0,$$

where $K_\epsilon^\delta = \{x \in \mathbb{D} | d(x, y) < \delta, \text{ for some } y \in K_\epsilon\}$.

A measurable map X from $\mathcal{X}$ to $\mathbb{D}$ is said to be tight if for each $\epsilon > 0$, there is a compact set $K_\epsilon \subset \mathbb{D}$ such that

$$P(X \in K_\epsilon) \geq 1 - \epsilon.$$

We denote $\rightsquigarrow$ as weakly convergence, that is, $X_n \rightsquigarrow X$ if $E^* f(X_n) \to E(f(X))$, for each $f \in C_b(\mathbb{D})$, where X is measurable. A common choice of $\mathbb{D}$ is $l^\infty(T)$, which is defined as the collection of all bounded maps $\phi : T \to \mathbb{R}$ and equipped with norm $\|\phi\| = \sup_{t\in T} |\phi(t)|$. By $a_n \lesssim b_n$ we mean that there exists a positive number c such that $a_n \leq c b_n$ for all large n.

Lemma 0.1 (Kosorok (2008) Lemma 7.1)**.** *Let X and Y be tight and measurable with values in metric space $(\mathbb{D}, \mathcal{B}_d, d)$ and $\mathcal{G}$ be a subset of $C_b(\mathbb{D})$ such that*

(i). *$\mathcal{G}$ is a vector space containing constant functionals;*
(ii). *If $g \in \mathcal{G}$, then $\max(g, 0) \in \mathcal{G}$;*
(iii). *For any point $x, y \in \mathbb{D}$, there exists functional $g \in \mathcal{G}$ such that $g(x) \neq g(y)$,*

then X and Y has the same distribution if and only if $E(g(X)) = E(g(Y))$ for all $g \in \mathcal{G}$.

Lemma 0.2 (Kosorok (2008) Lemma 7.12)**.** *Let $\{X_n, n \geq 1\}, X$ be maps from $\mathcal{X}$ to $\mathbb{D}$ and X is measurable. Assume $X_n \rightsquigarrow X$. Then X is tight if and only if X_n is asymptotically tight.*

Lemma 0.3 (Kosorok (2008) Lemma 7.14)**.** *Let $\{X_n, Y_n, n \geq 1\}$ be sequences of maps from $\mathcal{X}$ to $\mathbb{D}$. Then the following are true:*

(i). *$\{X_n, n \geq 1\}$ and $\{Y_n, n \geq 1\}$ are asymptotically tight if and only if $(X_n, Y_n), n \geq 1$ is asymptotically tight.*
(ii). *Asymptotically sequences $\{X_n, n \geq 1\}$ and $\{Y_n, n \geq 1\}$ are asymptotically measurable if and only if $(X_n, Y_n), n \geq 1$ is asymptotically measurable.*

Lemma 0.4 (A version of Prohorove's Theorem in Kosorok (2008) Theorem 7.13)**.** *Let $\{X_n, n \geq 1\}$ be a sequence of maps from $\mathcal{X}$ to $\mathbb{D}$ which is asymptotically measurable and asymptotically tight, then there exists a subsequence $\{X_{n_k}, k \geq 1\}$ converging weakly to a tight random variable X.*

Lemma 0.5. *Let $X(t), Y(t)$ be tight Gaussian processes on $l^\infty(T)$ and T is the index. The sequences of processes $X_n(t) \rightsquigarrow X(t)$ and $Y_n(t) \rightsquigarrow Y(t)$ in $l^\infty(T)$. And for each positive integers k, m and each indexes $t_1, .., t_k, s_1, ..., s_m \in T$, $(X_n(t_1), ..., X_n(t_k), Y_n(s_1), ..., Y_n(s_m))$ converges in distribution to some $(k+m)$-dimensional normal random vector $(Z^{(1)}_{t_1}, ..., Z^{(1)}_{t_k}, Z^{(2)}_{s_1}, ..., Z^{(2)}_{s_m})$. Then there exist a tight 2-dimensional Gaussian process $(D_1(t), D_2(t))$ in $l^\infty(T)$ such that*

(i). $(X_n(t), Y_n(t)) \rightsquigarrow (D_1(t), D_2(t))$ *in* $l^\infty(T)$.
(ii). $(D_1(t_1), D_1(t_2), ..., D_1(t_k), D_2(s_1), D_2(s_2), ..., D_2(s_m))$ *has the same distribution as* $(Z^{(1)}_{t_1}, ..., Z^{(1)}_{t_k}, Z^{(2)}_{s_1}, ..., Z^{(2)}_{s_m})$.
(iii). $Cov(D_1(t), D_1(s)) = Cov(Z^{(1)}_t, Z^{(1)}_s)$, $Cov(D_2(t), D_2(s)) = Cov(Z^{(2)}_t, Z^{(2)}_s)$ *and* $Cov(D_1(t), D_2(s)) = Cov(Z^{(1)}_t, Z^{(2)}_s)$.

Proof of Lemma 0.5. By Lemma 0.2, we can see both $\{X_n, n \geq 1\}$ and $\{Y_n, n \geq 1\}$ are asymptotically measurable and asymptotically tight on $l^\infty(T)$. So $\{(X_n, Y_n), n \geq 1\}$ is also asymptotically measurable and asymptotically tight due to Lemma 0.3. Now, it suffices to show that for any subsequence of $\{(X_n, Y_n), n \geq 1\}$, there exists a further subsequence converging weakly in $l^\infty(T) \times l^\infty(T)$. Abusing notation, denote $\{(X_n, Y_n), n \geq 1\}$ as a subsequence, by Lemma 0.4, there exist a subsequence $\{(X_{n_i}, Y_{n_i}), i \geq 1\}$ and a tight random variable (R_1, R_2) such that (X_{n_i}, Y_{n_i}) converges weakly to (R_1, R_2) on $l^\infty(T) \times l^\infty(T)$.

Next, we will show (R_1, R_2) is (D_1, D_2). Now construct a two dimensional Gaussian process $(\widetilde{D}_1(t), \widetilde{D}_2(t))$ on $l^\infty(T) \times l^\infty(T)$ satisfying (i)-(iii). Let $\mathcal{G}$ be the collection of bounded continuous functionals $g : l^\infty(T) \times l^\infty(T) \to \mathbb{R}$ with the form:

$$g(x, y) = f_g(x(t_1), ..., x(t_k), y(s_1), ..., y(s_m)),$$

for some positive integers k, m, some indexes $t_1, ..., t_k, s_1, ..., s_m \in T$ and some bounded continuous function $f_g : \mathbb{R}^{(k+m)} \to \mathbb{R}$. By definition of weakly convergence, for all $g \in \mathcal{G}$, it follows that

$$\begin{aligned} & \lim_{i\to\infty} E(f_g(X_{n_i}(t_1), ..., X_{n_i}(t_k), Y_{n_i}(s_1), ..., Y_{n_i}(s_k))) \\ = & \lim_{i\to\infty} E(g(X_{n_i}, Y_{n_i})) \\ = & E(g(R_1, R_2)) \\ = & E(f_g(R_1(t_1), ..., R_1(t_k), R_2(s_1), ..., R_2(s_k))). \end{aligned}$$

Since $(X_n(t_1), ..., X_n(t_k), Y_n(s_1), ..., Y_n(s_m))$ converges in distribution to some $(k+m)$-dimensional normal random $(Z^{(1)}_{t_1}, ..., Z^{(1)}_{t_k}, Z^{(2)}_{s_1}, ..., Z^{(2)}_{s_m})$, so we have

$$\begin{aligned} & \lim_{n\to\infty} E(f_g(X_n(t_1), ..., X_n(t_k), Y_n(s_1), ..., Y_n(s_k))) \\ = & E(f_g(Z^{(1)}_{t_1}, ..., Z^{(1)}_{t_k}, Z^{(2)}_{s_1}, ..., Z^{(2)}_{s_m})) \\ = & E(f_g(\widetilde{D}_1(t_1), ..., \widetilde{D}_1(t_k), \widetilde{D}_2(s_1), ..., \widetilde{D}_2(s_k))) \\ = & E(g(\widetilde{D}_1, \widetilde{D}_2)). \end{aligned}$$

These two equations above suggest $E(g(\widetilde{D}_1, \widetilde{D}_2)) = E(g(R_1, R_2))$, for all $g \in \mathcal{G}$ and it is not difficult to verify that $\mathcal{G}$ satisfy the conditions in Lemma 0.1. So it follows that (R_1, R_2) has the same distribution as $(\widetilde{D}_1, \widetilde{D}_2)$ and satisfies (i)-(iii). □

A map $\phi : \mathbb{D} \to \mathbb{E}$ between normed spaces $\mathbb{D}$ and $\mathbb{E}$ is Hadamard differentiable at $\theta \in \mathbb{D}$ if there exists a continuous linear map $\phi'_\theta : \mathbb{D} \to \mathbb{E}$ such that

$$\lim_{n\to\infty} \left\| \frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} - \phi'_\theta - \phi'_\theta(h) \right\| = 0,$$

for every converging sequence $t_n \to 0$ and $h_n \to h \in \mathbb{D}$.

Lemma 0.6 (Kosorok (2008) Theorem 2.8)**.** *For normed spaces $\mathbb{D}$ and $\mathbb{E}$, let $\phi : \mathbb{D} \to \mathbb{E}$ be Hadamard differentiable at θ with derivative ϕ'_θ. Let $\{X_n, n \geq 1\}$ be maps from $\mathcal{X}$ to $\mathbb{D}$ and $r_n(X_n - \theta) \rightsquigarrow X$ for some sequence of constants $r_n \to \infty$ and some tight random variable X with value in $\mathbb{D}$. Then $r_n(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(X)$ on $\mathbb{E}$.*

Lemma 0.7 (Nolan and Pollard (1987) Lemma 16)**.** *Support $\mathcal{F}_\Theta \subset L_2(Q), \mathcal{F}_{i,\Theta} \subset L_2(Q_i), i = 1, 2, ..., k$, where Q_i's are probability measures, and $\mathcal{F}_{i,\Theta}$ and $L_2(Q_i)$ are indexed by Θ and Q_i, respectively. Suppose that for each pairs $f_t, f_s \in \mathcal{F}_\Theta$ where $s, t \in \Theta$, there exist $f_{i,t}, f_{i,s} \in \mathcal{F}_{i,\Theta}, i = 1, 2, ..., k$ such that*

$$\sqrt{Q|f_t - g_s|^2} \leq \sum_{i=1}^{k} \sqrt{Q_i|f_{i,t} - g_{i,s}|^2}.$$

Then for all $\epsilon > 0$, it holds that

$$N(2k\epsilon, L_2(Q), \mathcal{F}_\Theta) \leq \prod_{i=1}^{k} N(\epsilon, L_2(Q_i), \mathcal{F}_{i,\Theta}).$$

Let $\mathcal{X}$ be a sample space and $\mathcal{C}$ be a collection of subsets of $\mathcal{X}$. For set $\{x_1, x_2, ..., x_n\} \subset \mathcal{X}$, define

$$\Delta_n(\mathcal{C}, \{x_1, x_2, ..., x_n\}) = |\{C \cap \{x_1, x_2, ..., x_n\} | C \in \mathcal{C}\}|,$$

where $|A|$ is the cardinality of set A. The VC index $V(\mathcal{C})$ is the smallest n such that

$$\sup_{x_1, x_2, ..., x_n \in \mathcal{X}} \Delta_n(\mathcal{C}, \{x_1, x_2, ..., x_n\}) < 2^n, \tag{0.1}$$

where the supremum is taking over all possible points in $\mathcal{X}$. We say $\mathcal{C}$ or $\{I_C | C \in \mathcal{C}\}$ is a *VC* class if $V(\mathcal{C}) < \infty$.

Lemma 0.8 (Kosorok (2008) Theorem 9.2)**.** *There exists a universal constant $K < \infty$ such that, for any VC class of sets $\mathcal{C}$, any probability measure Q and any $0 < \epsilon < 1$,*

$$N(\epsilon, \{I_C | C \in \mathcal{C}\}, L_2(Q)) \leq KV(\mathcal{C})(4e)^{V(\mathcal{C})} \epsilon^{-2(V(\mathcal{C})-1)}.$$

Lemma 0.9 (Kosorok (2008) Lemma 9.12)**.** *The class $\{x \in \mathbb{R}^d | \beta' x \leq t, \beta \in \mathbb{R}^d, t \in \mathbb{R}\}$ is a VC class with VC index d+2.*

Lemma 0.10 (Nolan and Pollard (1987) Theorem 6). *Let $\mathcal{E}$ be a class of P-degenerate kernels with envelop $F_{\mathcal{E}}$ and let $U_i, i \geq 1$ be i.i.d random vectors with distribution associated with P, then there exists a constant C which is free of n such that*

$$E(\sup_{g\in\mathcal{E}} |\sum_{i\neq j} g(U_i, U_j)| \leq CE(\theta_n + \tau_n J(\theta_n/\tau_n, L_2(T_n), \mathcal{E})), \tag{0.2}$$

where $\tau_n = 2n(T_n F_{\mathcal{E}}^2)^{1/2}$, T_n is a probability measure which may not be the same as P that is associated with U_1, ..., U_n, $T_n F_{\mathcal{E}}^2 = \int F_{\mathcal{E}}^2 dT_n$, $J(\delta, L_2(T_n), \mathcal{E}) = \int_0^\delta N(\epsilon(T_n F_{\mathcal{E}}^2)^{1/2}, L_2(T_n), \mathcal{E})d\epsilon$, for all $\delta > 0$ and $\theta_n = \frac{n}{2}\sup_{g\in\mathcal{E}}(T_n g^2)^{1/2}$.

A class of functions $\mathcal{F} \subset L_2(Q)$ is said to be Q-Donsker, if

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n f(Z_i) - Qf\right) \rightsquigarrow G(f), f \in \mathcal{F},$$

where $Z_i, i \geq 1$ are i.i.d with distribution Q and $G(f)$ is a tight Gaussian process indexed by $\mathcal{F}$ with covariance function

$$Cov(G(f), G(g)) = Q(fg) - QfQg, f, g \in \mathcal{F}.$$

Lemma 0.11 (Kosorok (2008) Theorem 2.5). *Let P be a probability measure and $\mathcal{F} \subset L_2(P)$ with envelop $F_{\mathcal{F}}$. If $PF_{\mathcal{F}}^2 \leq \infty$ and*

$$\int_0^1 \sqrt{\log \sup_Q N(\epsilon(F_{\mathcal{F}}^2)^{1/2}, L_2(Q), \mathcal{F})} d\epsilon < \infty,$$

then $\mathcal{F}$ is P-Donsker (see the next paragraph).

Assumption A1. *The cdf $F_o = F_W$ is discontinous.*

Remark 1. *Under the assumption that $Y = \beta X + W$ and A1 holds, the SMLE and the MSMLE $\hat{\beta}$ of β satisfy $P(\hat{\beta} \neq \beta$ i.o.$) = 0$ (see Yu and Wong (2002 and 2003)). Under such assumptions together with $W \perp X$, we shall establish the asymptotic distribution of $\sqrt{n}(\check{F}^* - \hat{F}_Y)$. Since n is large, WLOG, we can assume that $\hat{\beta} = \beta$ and*

$$\check{F}^*(t) = \frac{\frac{1}{n^2}\sum_{j=1}^n \sum_{i=1}^n I(Y_i + \beta X_j \leq t, X_i = 0)}{\sum_{i=1}^n I(X_i = 0)/n}$$

Let $F_{Y^*}(t, \beta) = P(W \leq t - \beta' X)$. By definition $F_{Y^*}(t, \beta) = F_Y(t)$ under the given assumptions.

Theorem 3. Suppose that $Y = \beta X + W$, $X \perp W$, and Assumption A1 holds, and thus $P(\widehat{\beta} \neq \beta_0$ i.o.$) = 0$. Moreover, suppose $P(X = 0) = p > 0$. Then

$$\sqrt{n}\left(\check{F}^*(t) - \widehat{F}_Y(t)\right) \rightsquigarrow D_1(t) - F_Y(t)D_2 - D_3(t), t \in \overline{\mathbb{R}},$$

where $D_2 = D_1(\infty)$, $D_1(t)$ and $D_3(t)$ are both Brownian bridge with zero mean and covariance

$$\begin{aligned}
Cov(D_1(t), D_1(s)) &= p^{-1}\int\Big(\int I(w+\beta_0 x \le t)dF_X(x)\int I(w+\beta_0 x\le s)dF_X(x)\Big)dF_0(w)\\
&\quad -2F_Y(t)F_Y(s) + \int F_0(t-\beta_0 x)F_0(s-\beta_0 x)dF_X(x)\\
&\quad +F_Y(t)[F_0(s)-F_Y(s)] + F_Y(s)[F_0(t)-F_Y(t)],\\
Cov(D_3(t), D_3(s)) &= F_Y(t\wedge s) - F_Y(t)F_Y(s),\\
Cov(D_1(t), D_3(s)) &= \int\int I(w+\beta_0 x\le t)I(w\le s)dF_X(x)dF_0(w) - F_Y(t)F_Y(s)\\
&\quad + \int F_0(t-\beta_0 x)F_0(s-\beta_0 x)dF_X(x) - F_Y(t)F_Y(s),
\end{aligned}$$

Remark 2. *The theorem is not valid if H_0 is not true. In particular, if $P(X=0)P(X=u_o)>0$ for a $u_o\ne 0$, $Y\ne\beta X+W$, and A1 holds, then $\exists$ a β such that $P(\hat\beta\ne\beta \text{ i.o.})=0$, where $\hat\beta$ is the SMLE. In the latter case, the asymptotic distribution of the test statistic given in the theorem is not valid and it can be shown that F_Y in $Cov(D_1(t), D_1(s)$ should be changed to F_{Y^*}, and $Cov(D_1(t),D_3(s)) = \int\int I(w+\beta x\le t)I(w\le s)dF_X(x)dF_0(w) - \underbrace{F_{Y^*}(t)F_{Y^*}(s)}_{not\ F_Y(t)F_Y(s)} +$*

$\underbrace{\int\int I(w+\beta x\le t)I(y\le s)F_{X,Y}(0,w)F_{X,Y}(x,y)}_{not\ \int F_0(t-\beta_0 x)F_0(s-\beta_0 x)dF_X(x)-F_Y(t)F_Y(s)}$, *where $dF_{X,Y}(0,w)=pdF_o(w)$ if $X\perp W$.*

In general, if H_0 is not true, then it is not clear whether the SMLE and MSMLE $\hat\beta$ always converge. Then the asymptotic distribution may not exist. Since we are testing H_0, the data may not satisfy H_0. The advantage of the bootstraping distribution of $\check F^* - \hat F_Y$ over the approach in the theorem is that it does not need that the data satisfy $Y=\beta X+W$, and the test statistic is still valid regardless whether or not H_0 holds.

Proof of Theorem 3. We only prove the case when $p=1$ and the extension to the case when $p>1$ can be done analogically. Let $u=(x,y)\in\mathbb{R}^2$ and $U_i=(X_i,Y_i)$, $i=1,2,...n$ which are i.i.d. Consider a class of functions $\mathcal{F}=\{I(y+\beta_0 x\le t)|t\in\overline{\mathbb{R}}\}$. For each g or $g_t\in\mathcal{F}$, define the kernel

$$K_g(u_1,u_2) = I(y_1\le t-\beta_0 x_2)I(|x_1|=0),$$

and corresponding symmetric kernel

$$h_g(u_1,u_2)=\frac{K_g(u_1,u_2)+K_g(u_2,u_1)}{2}.$$

Define classes of functions $\mathcal{K}_\mathcal{F}=\{K_g|g\in\mathcal{F}\}$ and $\mathcal{H}_\mathcal{F}=\{h_g|g\in\mathcal{F}\}$, and for each measurable function h on $\mathbb{R}^2\times\mathbb{R}^2$, define a stochastic process by

$$M_n(h)=\frac{2}{n(n-1)}\sum_{1\le i<j\le n}h(U_i,U_j)-E\left(h(U_1,U_2)\right), \tag{0.3}$$

and define $Ph(\cdot,u) = E(h(U_1,U_2)|U_2 = u)$, $Ph(u,\cdot) = E(h(U_1,U_2)|U_1 = u)$ and $P\otimes Ph = E(h(U_1,U_2))$. For each symmetric kernel h_g, letting

$$\widetilde{h}_g(u_1,u_2) = h_g(u_1,u_2) - Ph_g(u_1,\cdot) - Ph_g(\cdot,u_2) + P\otimes Ph_g,$$

the Hoeffding decomposition tells us that $\widetilde{h}_g$ is a degenerate kernel and it yields the decomposition

$$M_n(h_g) = M_n(\widetilde{h}_g) + \frac{2}{n}\sum_{i=1}^{n}(Ph_g(U_i,\cdot) - P\otimes Ph_g). \tag{0.4}$$

Let $\widetilde{H}_{\mathcal{F}} = \{\widetilde{h}_g | g\in\mathcal{F}\}$, then it follows that

$$\begin{aligned}
E(\sup_{g\in\mathcal{F}}|M_n(\widetilde{h}_g)|) &= E(\sup_{\widetilde{h}\in\widetilde{\mathcal{H}}_{\mathcal{F}}}|M_n(\widetilde{h})|)\\
&= \frac{2}{n(n-1)}E\left(\sup_{\widetilde{h}\in\widetilde{\mathcal{H}}_{\mathcal{F}}}|\sum_{1\le i<j\le n}\widetilde{h}(U_i,U_j) - E\left(\widetilde{h}(U_1,U_2)\right)|\right)\\
&= \frac{1}{n(n-1)}E\left(\sup_{\widetilde{h}\in\widetilde{\mathcal{H}}_{\mathcal{F}}}|\sum_{i\neq j}\widetilde{h}(U_i,U_j) - \underbrace{E\left(\widetilde{h}(U_1,U_2)\right)}_{=0 \text{ since } \widetilde{h} \text{ is degenerate}}|\right)\\
&\le \frac{C}{n(n-1)}E(\theta_n + \tau_n J(\theta_n/\tau_n, L_2(T_n), \widetilde{\mathcal{H}}_{\mathcal{F}})) \quad \text{(by (0.2) in Lemma 0.10)}\\
&\le \frac{C}{n(n-1)}E(2n + 8nJ(1/4, L_2(T_n), \widetilde{\mathcal{H}}_{\mathcal{F}}))\\
&\quad \text{(as } \sup_{h\in\widetilde{\mathcal{H}}_{\mathcal{F}}}|h| \le 4 \text{ and thus } 4\theta_n \le \tau_n \le 8n)\\
&\le \frac{2C}{n-1}E(1 + 4J(1/4, L_2(T_n), \widetilde{\mathcal{H}}_{\mathcal{F}})),
\end{aligned} \tag{0.5}$$

where $\theta_n = \frac{n}{2}\sup_{g\in\widetilde{\mathcal{H}}_{\mathcal{F}}}(T_n g^2)^{1/2}$, $\tau_n = 2n(T_n F^2_{\widetilde{\mathcal{H}}_{\mathcal{F}}})^{1/2}$, and T_n is defined in Lemma 0.10.

Next we will establish a bound of the empirical entropy $J(1/4, L_2(T_n), \widetilde{\mathcal{H}}_{\mathcal{F}})$. Let $g_i(u) = I(y + \beta_0 x \le t_i), i = 1,2$. It is not hard to verify using triangular inequality and Jensen's inequality that

$$\begin{aligned}
&\sqrt{T_n|\widetilde{h}_{g_1} - \widetilde{h}_{g_2}|^2}\\
=\ &\sqrt{T_n|(h_{g_1} - h_{g_2}) - (Ph_{g_1}(u,\cdot) - Ph_{g_2}(u,\cdot)) - (Ph_{g_1}(\cdot,u) - Ph_{g_2}(\cdot,u)) + P\otimes P(h_{g_1} - h_{g_2})|^2}\\
\le\ &\sqrt{T_n|h_{g_1} - h_{g_2}|^2} + \sqrt{T_n|Ph_{g_1}(u,\cdot) - Phg_2(u,\cdot)|^2}\\
&+\sqrt{T_n|Ph_{g_1}(\cdot,u) - Phg_2(\cdot,u)|^2} + \sqrt{P\otimes P|h_{g_1} - h_{g_2}|^2}\\
=\ &\sqrt{T_n|h_{g_1} - h_{g_2}|^2} + 2\sqrt{T_n|Ph_{g_1}(u,\cdot) - Phg_2(u,\cdot)|^2}\\
&+\sqrt{P\otimes P|h_{g_1} - h_{g_2}|^2} \quad \text{bBy the fact } h_g \text{ is symmetric)}\\
\le\ &\sqrt{T_n|h_{g_1} - h_{g_2}|^2} + 2\sqrt{T_n\otimes P|h_{g_1} - h_{g_2}|^2} + \sqrt{P\otimes P|h_{g_1} - h_{g_2}|^2}.
\end{aligned} \tag{0.6}$$

Moreover, for any probability measure $Q(u_1,u_2)$, we have

$$
\begin{aligned}
\sqrt{Q|h_{g_1}-h_{g_2}|^2} &\le \frac{1}{2}\sqrt{\int |K_{g_1}(u_1,u_2)-K_{g_2}(u_1,u_2)|^2 dQ(u_1,u_2)} \\
&\quad +\frac{1}{2}\sqrt{\int |K_{g_1}(u_2,u_1)-K_{g_2}(u_2,u_1)|^2 dQ(u_1,u_2)} \\
&\le \frac{1}{2}\sqrt{\int |I(y_1\le t_1-\beta_0 x_2)-I(y_1\le t_2-\beta_0 x_2)|^2 dQ(u_1,u_2)} \\
&\quad +\frac{1}{2}\sqrt{\int |I(y_2\le t_1-\beta_0 x_1)-I(y_2\le t_2-\beta_0 x_1)|^2 dQ(u_1,u_2)} \qquad (0.7)
\end{aligned}
$$

Therefore, (by the definition of $N(4\epsilon, L_2(T_n), \widetilde{\mathcal{H}}_{\mathcal{F}})$, (0.6)) and (0.7), we can see, for every $\epsilon > 0$, it holds that

$$
\begin{aligned}
& N(4\epsilon, L_2(T_n), \widetilde{\mathcal{H}}_{\mathcal{F}}) \\
\le\ & N(\epsilon/2, L_2(T_n), \mathcal{H}_{\mathcal{F}})N^2(\epsilon/2, L_2(T_n\otimes P), \mathcal{H}_{\mathcal{F}})N(\epsilon/2, L_2(P\otimes P), \mathcal{H}_{\mathcal{F}}) \quad \text{(by (0.6) and Lemma 0.7)} \\
\le\ & N^2(\epsilon/8, L_2(T_n), \mathcal{F})N^4(\epsilon/8, L_2(T_n\otimes P), \mathcal{F})N^2(\epsilon/8, L_2(P\otimes P), \mathcal{F}) \quad \text{(by (0.7) and Lemma 0.7)} \\
\le\ & \sup_Q N^8(\epsilon/8, L_2(Q), \mathcal{F}), \quad \text{(By taking supernum of all measures)} \qquad (0.8)
\end{aligned}
$$

where the supernum is taken over all probability measures.

Since $\mathcal{F}$ is a subset of $\{I(y+\beta x\le t)|\beta\in\mathbb{R}, t\in\mathbb{R}\}\cup\{0,1\}$ and $\{I(y+\beta x\le t)|\beta\in\mathbb{R}, t\in\mathbb{R}\}$ is a VC class with VC index 3 due to Lemma 0.9. Therefore, applying Lemma 0.8, we can see

$$
\sup_Q N(\epsilon, L_2(Q), \mathcal{F}) \lesssim \epsilon^{-4}. \qquad (0.9)
$$

Since $\sup_{h\in\widetilde{\mathcal{H}}_{\mathcal{F}}}|h|\le 4$, combining (0.8) and (0.9), we have

$$
\begin{aligned}
J(1/4, L_2(T_n), \widetilde{\mathcal{H}}_{\mathcal{F}}) &= \int_0^{1/4} \log N(4\epsilon, L_2(T_n), \widetilde{\mathcal{H}}_{\mathcal{F}}) d\epsilon \\
&\lesssim \int_0^1 -\log \epsilon d\epsilon < \infty. \qquad (0.10)
\end{aligned}
$$

Combining (0.5) and (0.10), we obtain that

$$
\sup_{g\in\mathcal{F}} \sqrt{n}|M_n(\widetilde{h}_g)| = O_p(\frac{1}{\sqrt{n}}) = o_P(1).
$$

The previous equality and (0.4) lead to

$$
\sup_{g\in\mathcal{F}} \sqrt{n}\left|M_n(h_g) - \frac{2}{n}\sum_{i=1}^n (Ph_g(U_i,\cdot) - P\otimes Ph_g)\right| = O_p(\frac{1}{\sqrt{n}}) = o_P(1). \qquad (0.11)
$$

By definition, rewrite $\check{F}^*$ as

$$\begin{aligned}\check{F}^*(t) &= \frac{1}{n}\sum_{i=1}^n \widehat{F}_0(t-\widehat{\beta}X_i)\\ &= \frac{1}{n}\sum_{i=1}^n \frac{\sum_{j=1}^n I(Y_j\le t-\widehat{\beta}X_i)I(|X_j|=0)}{\sum_{j=1}^n I(|X_j|=0)}\\ &= \frac{1}{n}\frac{\sum_{i=1}^n\sum_{j=1}^n I(Y_j\le t-\widehat{\beta}X_i)I(|X_j|=0)}{\sum_{j=1}^n I(|X_j|=0)}.\end{aligned}$$

Now let us define

$$\begin{aligned}G_n(t,\beta) &= \frac{1}{n-1}\frac{\sum_{i=1}^n\sum_{j=1}^n I(Y_j\le t-\beta X_i)I(|X_j|=0)}{\sum_{j=1}^n I(|X_j|=0)},\\ \bar{G}_n(t,\beta) &= \frac{1}{n(n-1)}\sum_{i=1}^n\sum_{j=1}^n I(Y_j\le t-\beta X_i)I(|X_j|=0).\end{aligned}$$

Moreover, direct examinations show that for $g_t(u)=I(y+\beta_0 x\le t)$

$$\begin{aligned}&|\bar{G}_n(t,\beta_0)-\frac{2}{n(n-1)}\sum_{1\le i<j\le n}h_{g_t}(U_i,U_j)|\\ =\ &|\bar{G}_n(t,\beta_0)-\frac{1}{n(n-1)}\sum_{i\neq j}K_{g_t}(U_i,U_j)|\\ =\ &\frac{1}{n(n-1)}\sum_{i=1}^n K_{g_t}(U_i,U_i)\\ \le\ &\frac{1}{n(n-1)}\sum_{i=1}^n I(|X_i|=0).\end{aligned}$$

As a consequence, we obtain

$$\sup_{t\in\overline{\mathbb{R}}}\sqrt{n}|\bar{G}_n(t,\beta_0)-\frac{2}{n(n-1)}\sum_{1\le i<j\le n}h_{g_t}(U_i,U_j)| = O_P(n^{-1/2})=o_P(1). \tag{0.12}$$

Since for $g_t(u)=I(y\le t-\beta_0 x)$, $P\otimes Pg_t=pF_Y(t)$. So in the view of (0.3), (0.4), (0.11) and (0.12), it not hard to conclude $\sqrt{n}(\bar{G}_n(t,\beta_0)-pF_Y(t))$ and $n^{-1/2}\sum_{i=1}^n(Ph_{g_t}(U_i,\cdot)-pF_Y(t))$ has the same limit distribution in terms of weak convergence as stochastic processes indexed by $\overline{\mathbb{R}}$. In the following we will establish the limit distribution of $\sqrt{n}(\bar{G}_n(t,\beta_0)-pF_Y(t))$ using $n^{-1/2}\sum_{i=1}^n(Ph_{g_t}(U_i,\cdot)-pF_Y(t))$ as a proxy process.

Next we will show the class $PH_{\mathcal{F}}:=\{Ph_g(u,\cdot)|g\in\mathcal{F}\}$ is Donsker. Similar calculation to (0.7), for any probability measure Q, it is not hard to show for $g_i(u) = I(y + \beta_0 x \le t_i)$

$$\begin{aligned}
& \sqrt{Q|Ph_{g_1}(u,\cdot) - Ph_{g_2}(u,\cdot)|^2} \\
\le & \sqrt{Q \otimes P|h_{g_1} - h_{g_2}|^2} \\
\le & \frac{1}{2}\sqrt{\int |I(y_1 \le t_1 - \beta_0 x_2) - I(y_1 \le t_2 - \beta_0 x_2)|^2 dQ \otimes P(u_1,u_2)} \\
& + \frac{1}{2}\sqrt{\int |I(y_2 \le t_1 - \beta_0 x_1) - I(y_2 \le t_2 - \beta_0 x_1)|^2 dQ \otimes P(u_1,u_2)}.
\end{aligned}$$

The previous inequality and Lemma 0.7 suggest

$$N(\epsilon, L_2(Q), P\mathcal{H}_{\mathcal{F}}) \le N^2(\epsilon/4, L_2(Q \otimes P), \mathcal{F})$$

Therefore, (0.9) and the inequality above imply

$$\begin{aligned}
\sup_Q N(\epsilon, L_2(Q), P\mathcal{H}_{\mathcal{F}}) &\le \sup_Q N^2(\epsilon/4, L_2(Q), \mathcal{F}) \\
&\lesssim \epsilon^{-8}.
\end{aligned}$$

Above covering number condition and Lemma 0.11 suggest the $P\mathcal{H}_{\mathcal{F}}$ is a Donsker class. Hence $n^{-1/2}\sum_{i=1}^n (Ph_{g_t}(U_i,\cdot) - pF_Y(t))$ weakly converges to a Gaussian process and $\sqrt{n}(\bar{G}_n(t,\beta_0) - pF_Y(t))$ converges to the same Gaussian process. Now, let us define a multi-dimensional process indexed by $\overline{\mathbb{R}}$, for $g_t(u) = I(y \le t - \beta_0 x)$, letting

$$\begin{aligned}
V_n(t) &:= (V_{1n}(t), V_{2n}, V_{3n}(t)) \\
&:= \sqrt{n}\Big(p^{-1}\bar{G}_n(t,\beta_0) - F_Y(t), \frac{1}{np}\sum_{i=1}^n I(|X_i| = 0) - 1, \frac{1}{n}\sum_{i=1}^n I(Y_i \le t) - F_Y(t)\Big), t \in \overline{\mathbb{R}},
\end{aligned}$$

and

$$\begin{aligned}
\widetilde{V}_n(t) &:= (\widetilde{V}_{1n}(t), \widetilde{V}_{2n}, \widetilde{V}_{3n}(t)) \\
&:= \sqrt{n}\Big(\frac{1}{n}\sum_{i=1}^n (p^{-1}Ph_{g_t}(U_i,\cdot) - F_Y(t)), \frac{1}{np}\sum_{i=1}^n I(|X_i| = 0) - 1, \frac{1}{n}\sum_{i=1}^n I(Y_i \le t) - F_Y(t)\Big), t \in \overline{\mathbb{R}}.
\end{aligned}$$

Previous argument suggest V_n and $\widehat{V}_n$ should have the same limit distribution if they converge weakly. But notice previous argument also shows $V_{1n}(t)$ converges weakly to a Gaussian process. The limit distribution of V_{2n} is also Gaussian distribution due to C.L.T. Similar argument can show that $V_{3n}(t)$ also converges weakly to a Gaussian process. In order to apply Lemma 0.5, we still need to check the condition: for each positive integers k, m and each $t_1, ...t_k, s_1, ..., s_m \in \overline{\mathbb{R}}$, by C.L.T, it is not difficult to show $(\widetilde{V}_{1n}(t_1), ..., \widetilde{V}_{1n}(t_k), \widetilde{V}_{2n}, \widetilde{V}_{3n}(s_1), ..., \widetilde{V}_{3n}(s_m))$ converges to a $(k+m+1)$-dimensional normal random variable $(Z^{(1)}_{t_1}, ..., Z^{(k)}_{t_k}, Z^{(2)}, Z^{(3)}_{s_1}, ..., Z^{(3)}_{s_m})$ in distribution, where

$$\begin{aligned}
Z^{(2)} &= Z^{(1)}_{\infty}, \\
Cov(Z^{(1)}_t, Z^{(1)}_s) &= p^{-1}\int\Big(\int I(w+\beta_0 x \le t)dF_X(x)\int I(w+\beta_0 x\le s)dF_X(x)\Big)dF_0(w) \\
&\quad -2F_Y(t)F_Y(s)+\int F_0(t-\beta_0 x)F_0(s-\beta_0 x)dF_X(x) \\
&\quad +F_Y(t)[F_0(s)-F_Y(s)]+F_Y(s)[F_0(t)-F_Y(t)], \\
Cov(Z^{(3)}_t, Z^{(3)}_s) &= F_Y(t\wedge s)-F_Y(t)F_Y(s), \\
Cov(Z^{(1)}_t, Z^{(3)}_s) &= \int\int I(w+\beta_0 x\le t)I(w\le s)dF_X(x)dF_0(w)-F_Y(t)F_Y(s) \\
&\quad +\int F_0(t-\beta_0 x)F_0(s-\beta_0 x)dF_X(x)-F_Y(t)F_Y(s).
\end{aligned}$$

Moreover, the equivalence of $\widetilde{V}_{1n}(t)$ and $V_{1n}(t)$ shows $(V_{1n}(t_1),...,V_{1n}(t_k),V_{2n},V_{3n}(s_1),...,V_{3n}(s_m))$ converges the same limit distribution. Now we can apply Lemma 0.5 to $V_n(t)$ and it follows that

$$V_n(t)\rightsquigarrow D(t)=\Big(D_1(t),D_2,D_3(t)\Big),\ t\in\overline{\mathbb{R}},$$

where $D_1(t),D_3(t)$ are both Gaussian processes with zero mean and covariance

$$\begin{aligned}
Cov(D_1(t),D_1(s)) &= p^{-1}\int\Big(\int I(w+\beta_0 x \le t)dF_X(x)\int I(w+\beta_0 x\le s)dF_X(x)\Big)dF_0(w) \\
&\quad -2F_Y(t)F_Y(s)+\int F_0(t-\beta_0 x)F_0(s-\beta_0 x)dF_X(x) \\
&\quad +F_Y(t)[F_0(s)-F_Y(s)]+F_Y(s)[F_0(t)-F_Y(t)], \\
Cov(D_3(t),D_3(s)) &= F_Y(t\wedge s)-F_Y(t)F_Y(s), \\
Cov(D_1(t),D_3(s)) &= \int\int I(w+\beta_0 x\le t)I(w\le s)dF_X(x)dF_0(w)-F_Y(t)F_Y(s) \\
&\quad +\int F_0(t-\beta_0 x)F_0(s-\beta_0 x)dF_X(x)-F_Y(t)F_Y(s),
\end{aligned}$$

and $D_2=D_1(\infty)$.

Now let us apply functional delta method to process $V_n(t)$. Define $\phi(v_1,v_2,v_3)=v_1/v_2-v3$. It not hard to verify the Hadamard derivative at point $(F_Y(t),1,F_Y(t))$ is $\phi_{(F_Y(t),1,F_Y(t))}(v_1,v_2,v_3)=v_1-F_Y(t)v_2-v_3$. As a consequence, by Lemma 0.6, we have

$$\begin{aligned}
&\sqrt{n}\Big(G_n(t,\beta_0)-\frac{1}{n}\sum_{i=1}^n I(Y_i\le t)\Big) \qquad (0.13)\\
. &= \sqrt{n}\Big(\frac{p^{-1}\bar{G}_n(t,\beta_0)}{\frac{1}{np}\sum_{i=1}^n I(|X_i|=0)}-\frac{1}{n}\sum_{i=1}^n I(Y_i\le t)\Big)\rightsquigarrow D_1(t)-F_Y(t)D_2-D_3(t),t\in\overline{\mathbb{R}}.
\end{aligned}$$

By Remark 2, it follows that $P(\widehat{\beta}\neq\beta_0 \text{ i.o.})=0$. Therefore (0.14) implies that

$$\sqrt{n}\Big(G_n(t,\widehat{\beta})-\widehat{F}_Y(t)\Big)\rightsquigarrow D_1(t)-F_Y(t)D_2-D_3(t),t\in\overline{\mathbb{R}}.$$

Finally, the desire result follows the above inequality and following fact

$$\sup_{t\in\overline{\mathbb{R}}} \sqrt{n}|G_n(t,\widehat{\beta}) - \check{F}^*(t)| \leq \frac{\sqrt{n}}{n-1} = o_P(1).$$

□

References

Kosorok, M. (2008). *Introduction to empirical processes and semiparametric inference*. Springer, New York.

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *The Annals of Statistics*, pages 780–799.