Technical Report for
"Testing independence and goodness -of-fit in linear regression models allowing non-existence of the mean of the response variable".

By Qiqing Yu and Ruiqi Liu

Department of Mathematical Sciences, SUNY, Binghamton, NY 13902

Corresponding author's email address: qyu@math.binghamton.edu

Current version: 11/10/2017

AMS 1991 subject classification: Primary 62J05; Secondary 62J20.

Short Title: The Marginal Distribution Tests.

Key words and phrases: linear regression model, semi-parametric model, diagnostic plot, non-parametric test.

Abstract: We propose an approach to simultaneously test the assumptions of independence and goodness-of-fit for a multiple linear regression (LR) model $Y = \beta'\mathbf{X} + W$. If $E(|Y||\mathbf{X}) = \infty$, then the claims on the distributions of all existing test statistics for the LR model are false, *e.g.*, the real sizes (*i.e.*, the probabilities of type I error) of Stute's test (1997) and Sen & Sen's test (2014) with a nominal size of 5% can be $\geq 33\%$ or even 99%. Our test is valid even if $E(|Y||\mathbf{X}) = \infty$. Thus it is more realistic than all the existing tests. Our approach is called the MD approach, as it is based on the difference between two estimators of the marginal distribution (MD) F_Y. We establish the asymptotic properties of the MD test. The simulation study suggests that the MD approach has certain advantages over existing tests even when $E(Y|\mathbf{X})$ exists, though it is not uniformly most powerful. We apply the MD approach to 3 real data sets.

1. Introduction. We propose a new approach for the diagnostic plotting method and for simultaneously testing the assumptions of independence and goodness-of-fit for a multiple linear regression model, called the marginal distribution (MD) plot and MD test, respectively. The MD approach has certain advantages over the existing methods.

Let $(\mathbf{X}_1, Y_1)$, ..., $(\mathbf{X}_n, Y_n)$ be i.i.d. observations from a joint cumulative distribution function (cdf) $F_{\mathbf{X},Y}$, with density function $f_{\mathbf{X},Y}$, where $\mathbf{X}$ is a p-dimensional random vector and Y is a response variable. Let $F_{Y|\mathbf{X}}$ and $f_{Y|\mathbf{X}}$ be the conditional cdf and density function, respectively. Data analysis depends on the regression model for $F_{\mathbf{X},Y}$.

A common regression model is the linear regression (LR) model,

$$Y = \beta'\mathbf{X} + W, \text{ where } W \text{ has the baseline cdf } F_o \ (F_W = F_o = F_{Y|\mathbf{X}}(\cdot|0)),\ \beta \in \mathcal{R}^p, \tag{1.1}$$

the p-dimensional euclidean space, and β' is the transpose of β. The coordinates of $\mathbf{X}$ can be dependent *e.g.*, $\mathbf{X} = (Z, Z^2, ..., Z^p)'$, where Z is a random variable. The LR model is often formulated by

$$Y = \alpha + \beta'\mathbf{X} + \epsilon, \text{ where } E(\epsilon|\mathbf{X}) = 0. \tag{1.2}$$

If the conditional variance $Var(W|\mathbf{X})$ does not depend on $\mathbf{X}$, it is called an ordinary linear regression (OLR) model, otherwise, it is called a weighted linear regression (WLR) model.

Remark 1. *Advantages that the LR model is specified by Eq. (1.1) rather than Eq. (1.2) are:*

(1) Eq. (1.2) but not (1.1) requires that $E(Y|\mathbf{X})$ exists;

(2) In general, β but not α is identifiable under censorship models (Yu and Wong (2002));

(3) It is often less important to estimate α than β, the effect of the covariate $\mathbf{X}$ on Y.

Under the OLR model, the semi-parametric MLE (SMLE) (if F_o is discontinuous), the least squares estimator (LSE), and the modified SMLE (MSMLE) are all consistent estimators of β (see Yu and Wong (2002, 2003 and 2004)) if $f_{\mathbf{X},Y} \in \Theta_{lse}$, where

$$\Theta_{lse} = \{F_{\mathbf{X},Y}\text{: } \Sigma_{\mathbf{X}} \text{ is non-singular and } Cov(\mathbf{X}, Y) \text{ exists}\}, \tag{1.3}$$

and $\Sigma_{\mathbf{X}}$ is the $p \times p$ covariance matrix of $\mathbf{X}$. If $E(Y|\mathbf{X})$ does not exist, but $E(\ln f_W(W))$ exists, Yu and Wong (2002) show that the MSMLE is still consistent, and if the cdf F_W is discontinuous, then the SMLE and the MSMLE $\tilde{\beta}$ satisfy $P(\tilde{\beta} \neq \beta$ infinitely often$) = 0$. However, the LSE is inconsistent if $E(|Y||\mathbf{X})$ does not exist (see Example 4.6).

The log linear regression model is a special case of $h(Y_i) = \beta' g(\mathbf{X}_i) + W_i$, where $h(\cdot)$ and $g(\cdot)$ are functions of Y_i and $\mathbf{X}_i$, respectively. By redefining Y as $h(Y)$, it can be viewed as a LR model. Other generalizations are the generalized additive model (gam) (see Hastie and Tibshirani (1990) and Wood (2006)), the partially linear model (plm) (Hardle (1994)), and the generalized partially linear single-index model (gplsim) (Carroll, *et al.* (1997)). The Cox regression model (Cox and Oakes (1984)) and the generalized linear model (Nelder and Wedderburn, (1972)) are also common regression models.

Let Θ be the collection of all possible $F_{\mathbf{X},Y}$. In order to apply a certain LR model in Eq. (1.1), say Θ_0, among so many choices for $F_{\mathbf{X},Y}$, it is important to check whether the data fit the model Θ_0. There are many model diagnostic tests for a LR model in the literature (see, for instance, Gonzalez-Manteiga and Crujeiras (2014)).

A simple test is the t-test (or F-test) (see *e.g.*, Example 2.1 or Draper and Smith (1966)). Ramsey's RASET test (Ramsey (1969)), the Harvey Collier test (1976) and the Rainbow test for linearity (Utts (1982)) are also well-known tests. Another test, called the gam test (see Wood (2006), or Section 2), is also a common test of a LR model. The aforementioned existing tests are the test of H_0^t: $\xi(\cdot) \equiv 0$, where $\xi(\mathbf{X}) = E(Y|\mathbf{X}) - \beta'\mathbf{X}$. In order to establish the distribution theories for the tests, each of these tests imposes certain regularity conditions on $F_{\mathbf{X},Y}$, which specifies a parameter space for $F_{\mathbf{X},Y}$, say Θ_p, under which the test is valid. The Θ_p depends on the specific test and is a certain common regression model that contains Θ_0. Thus $\Theta_p \neq \Theta$. If $F_{\mathbf{X},Y} \notin \Theta_p$, these tests are invalid (see Stute (1997)), in the sense that the (asymptotic) distributions specified for these tests are false.

Stute (1997) proposed the first test of the model in (1.2) that the parameter space Θ_p does not belong to any aforementioned common regression model and thus is more realistic than the existing tests before then. If W and $\mathbf{X}$ are independent ($W \perp \mathbf{X}$), the LR model is an OLR model. Many goodness-of-fit tests for an OLR model crucially use the independence of W and $\mathbf{X}$. Stute's test can not distinguish the OLR model from the WLR model. Sen and Sen (2014) proposed a test, called the SS-test hereafter, that can simultaneously check these two crucial model assumptions, *i.e.*, it tests

$$H_0\text{: } Y = \beta'\mathbf{X} + W \text{ and } W \perp \mathbf{X} \text{ against } H_1\text{: } H_0 \text{ is false.} \tag{1.4}$$

The SS-test is based on the pairwise distance between points in the sample and assuming $F_{\mathbf{X},Y} \in \Theta_{lse}$ (see (1.3)), among other regularity conditions. The latter assumptions specify the parameter space Θ_p within which their test is valid. This Θ_p is not a subset of any common regression model. Thus it is also more realistic than the existing tests of H_0 in (1.4).

Remark 2. *It is interesting to point out the following facts: (1) If H_0 is true and $F_{\mathbf{X},Y} \in \Theta_{lse}$, then most existing tests are valid. (2) If H_0 is not true and $F_{\mathbf{X},Y} \in \Theta_{lse}$, then most existing tests are invalid (see Examples 4.1 and 4.2 in Section 4), except Stute's test and the SS-test. (3) However, if H_0 is true but $E(|Y||\mathbf{X}) = \infty$, then all the existing tests including Stute's test and the SS-test are invalid (see Example 4.5 in Section 4).*

The MD approach we propose in this paper for testing H_0 in Eq. (1.4) compares two estimators of F_Y: one is the empirical distribution function (edf) and the other makes use of a consistent estimator of F_o (see (1.1) and (3.4)) and the MSMLE of β instead of the LSE unless one feels confident that $E(|Y||\mathbf{X}) < \infty$. The main reasons that the MD test can be valid for testing H_0 in Eq. (1.4) when $E(|Y||\mathbf{X}) = \infty$ are as follows.

(1) The estimator of F_o given in (3.4) does not involve estimating $E(Y|\mathbf{X} = 0)$ or $E(Y|\mathbf{X})$ as in the existing tests, including Stute's test and the SS-test.

(2) The MSMLE of β is consistent under H_o in Eq. (1.4), at least if $F_{\mathbf{X},Y} \in \Theta_m \cup \Theta_d$, where $\Theta_m = \{F_{\mathbf{X},Y} \in \Theta : E(|\ln f_{Y|\mathbf{X}}(Y|0)|) < \infty\}$ and $\Theta_d = \{F_{\mathbf{X},Y} \in \Theta :\ F_o \text{ is discontinuous}\}$ (Yu and Wong (2002)).

Thus the parameter space of the MD test is more realistic than the parameter spaces of all the existing tests, including Stute's test and the SS-test.

The paper is organized as follows. The drawbacks of most existing tests are discussed in Section 2. The MD approach and its theoretical justification, as well as its asymptotic distribution are introduced in Section 3. The simulation results on comparing various tests are presented in Section 4. Data analysis is given in Section 5. Some concluding remarks are given in Section 6. Some proofs of the lemmas and theorems are put in Appendix.

2. The drawbacks of most existing tests in the literature. In order to compare various existing tests of model (1.2), it is of interests to compare the parameter spaces Θ_p of these tests. Ideally, Θ_p is Θ, which consists of all possible $F_{\mathbf{X},Y}$.

Most existing LR model diagnostic tests actually test

$$H_0^*\colon\ \xi(\cdot) = 0 \text{ against } H_1^*\colon\ \xi(\cdot) \neq 0, \text{ where } \xi(\mathbf{X}) = E(Y|\mathbf{X}) - \beta'\mathbf{X} - \alpha \text{ and} \tag{2.1}$$

$F_{\mathbf{X},Y} \in \Theta_p$, a subset of a common regression model. In order to establish the asymptotic distribution for the test statistic, certain constraints on $\xi(\cdot)$ are needed. For example, the t-test and the F-test set $\xi(\mathbf{X}) = \theta g(\mathbf{X})$, and the parameter space is $\Theta_p = \Theta^t$, where

$$\Theta^t = \{F_{\mathbf{X},Y} : Y = \beta'\mathbf{X} + \theta' g(\mathbf{X}) + W,\ W \perp \mathbf{X},\ g(\cdot) \text{ is given } \}. \tag{2.2}$$

Thus they really test H_0^t: $\theta = 0$ against H_1^t: $\theta \neq 0$. The rejection regions of these tests are based on the (asymptotic) distributions of the test statistics, which are established based on the model Θ^t. If $F_{\mathbf{X},Y} \notin \Theta^t$, the test is invalid. It suffices to give a counterexample as follows.

Example 2.1. Consider the case that $\mathbf{X}$ is a random variable, $E(\mathbf{X}^4) < \infty$, $E(W) = \alpha$ and $\xi(x) = \theta x^2$. Denote $\Theta_0 = \{F_{\mathbf{X},Y} :\ Y = \beta\mathbf{X} + W\}$ and $\Theta_p^t = \{F_{\mathbf{X},Y} :\ Y = \beta\mathbf{X} + \theta\mathbf{X}^2 + W\}$. A t-test for the simple LR model is a test of H_0^t: $\theta = 0$ against H_1^t: $\theta \neq 0$. Given a random sample from $F_{\mathbf{X},Y}$, denote $\mathbf{Y} = (Y_1, ..., Y_n)'$ and $A = \begin{pmatrix} \mathbf{X}_1 - \overline{\mathbf{X}} & \mathbf{X}_1^2 - \overline{\mathbf{X}^2} \\ \vdots & \vdots \\ \mathbf{X}_n - \overline{\mathbf{X}} & \mathbf{X}_n^2 - \overline{\mathbf{X}^2} \end{pmatrix}_{n\times 2}$, the $n \times 2$ matrix. The LSE $\begin{pmatrix} \hat\beta \\ \hat\theta \end{pmatrix} = (A'A)^{-1}A'\mathbf{Y} = \begin{pmatrix} \overline{\mathbf{X}^2} - \overline{\mathbf{X}}(\overline{\mathbf{X}}) & \overline{\mathbf{X}^3} - \overline{\mathbf{X}^2}(\overline{\mathbf{X}}) \\ \overline{\mathbf{X}^3} - \overline{\mathbf{X}^2}(\overline{\mathbf{X}}) & \overline{\mathbf{X}^4} - \overline{\mathbf{X}^2}(\overline{\mathbf{X}^2}) \end{pmatrix}^{-1} \begin{pmatrix} \overline{\mathbf{X}Y} - \overline{\mathbf{X}}(\overline{Y}) \\ \overline{\mathbf{X}^2 Y} - \overline{\mathbf{X}^2}(\overline{Y}) \end{pmatrix}$. If $F_{\mathbf{X},Y} \in \Theta_p^t$, then conditional on $\mathbf{X}_i$'s, $COV(\begin{pmatrix} \hat\beta \\ \hat\theta \end{pmatrix})$ is $\sigma_W^2 (A'A)^{-1}$ and

$$\text{the t-test rejects } H_0 \text{ if } \frac{|\hat\theta|}{\hat S_W \sqrt{(0,1)(A'A)^{-1}(0,1)'}} > \begin{cases} 1.96 & \text{if } n \text{ is large} \\ t_{0.025,n-3} & \text{if } W \sim N(\alpha, \sigma^2), \end{cases} \tag{2.3}$$

where $S_W^2 = \frac{1}{n-3}\sum_{i=1}^n (Y_i - \overline{Y} - \hat\beta(\mathbf{X}_i - \overline{\mathbf{X}}))^2$ and the size of the test is 0.05. The validity of the t-test relies on the assumption that $F_{\mathbf{X},Y} \in \Theta_p^t$, *i.e.*, $Y = \beta\mathbf{X} + \theta\mathbf{X}^2 + W$. Otherwise, the (asymptotic) distribution of the test statistic in (2.3) is false. For instance, if $E(Y|\mathbf{X}) = X^3$ and $X \in U(-1,1)$ (the uniform distribution), then $F_{\mathbf{X},Y} \notin \Theta_p^t$ and $(\hat\beta,\hat\theta)\overset{a.s.}{\to}(\beta,\theta) = (1,0)$. Consequently, the test does not reject H_0^t, though both H_0^t and H_1^t are incorrect, as $E(Y|X) = X^3 \neq X$.

Write $\mathbf{X} = (X_1, ..., X_p)'$. The parameter space of the gam test is

$$\Theta^{plm} = \{F_{\mathbf{X},Y}\colon\ Y = \beta'\mathbf{X} + \textstyle\sum_{j=1}^q g_j(X_j) + W\}, \tag{2.4}$$

where $\beta = (0, ..., 0, \beta_{q+1}, ..., \beta_p)'$ and $q \leq p$. Thus Θ^{plm} is a special case of the plm model and $\xi(\mathbf{X}) = \sum_{j=1}^q g_j(X_j)$. The gam test fits the data to a partially linear model using the local scoring algorithm, which iteratively fits weighted additive models by backfitting. The backfitting algorithm is a Gauss-Seidel method for fitting additive models, by iteratively smoothing partial residuals. The algorithm separates the parametric from the nonparametric part of the fit, and fits the parametric part using weighted linear least squares with the backfitting algorithm. It tests H_0^{gam}: $\xi(\cdot) = 0$ v.s. H_1^{gam}: $\xi(\cdot) \neq 0$. Again the rejection region is based on the asymptotic distribution of the test statistic, which is established under the assumption that $F_{\mathbf{X},Y} \in \Theta^{plm}$. If $\mathbf{X}$ is a random variable and $\mathbf{X} \perp W$ (see Eq. (2.4)), then the gam test is a valid test of H_0 against H_1 in Eq. (1.4), provided that n is very large. Otherwise, it is not. For instance, if the data are from a Cox's model with a bivariate random vector $\mathbf{X}$ and the baseline distribution is a uniform distribution, then $E(Y|\mathbf{X})$ is not of the form of a plm model and the gam test is invalid.

It is easy to see that $\Theta \supset \Theta^{plm} \supset \Theta_p^t \supset \Theta_0$. If the data are from the Cox model with a covariate vector, then $F_{\mathbf{X},Y} \notin \Theta_p^t \cup \Theta^{plm}$, so neither of the hypotheses among H_0^t, H_1^t, H_0^{gam} and H_1^{gam} is correct. In applications, we often really do not know whether $F_{\mathbf{X},Y}$ belongs to any particular regression model Θ_p specified for those tests, just like we do not know whether $F_{\mathbf{X},Y} \in \Theta_0$. Thus most of the existing tests have such a drawback.

If the data are not from the parameter spaces specified by the existing tests then these tests are invalid, and thus these tests are just a random guessing. Actually, in such cases, it is often that these tests are worse than random guessing. In Example 4.1 of Section 4, for a sample of size 300 from the Cox model $h(t|x) = h_o(t)e^{\beta' x}$, the gam test with level 0.05 does not reject both the simple linear regression model $Y = \beta'\mathbf{X} + W$ and the weighted LR model with a probability at least 0.99. There is a good explanation for that. In view of Eq. (2.2) and (2.4), the current tests including the gam test try to find out whether $\theta = 0$ in Eq. (2.2) or $g_i(\cdot) = 0$ in Eq. (2.4). For a Cox's model, $\theta = 0$ or $g_i(\cdot) = 0$ is somewhat true. Thus the gam test is less likely to reject H_1^t: $\theta = 0$ or the LR model Θ_0.

Stute's test is based on the $\sup_x |\frac{1}{\sqrt{n}}\sum_{i=1}^n [Y_i - \hat\alpha - \hat\beta' X_i]\mathbf{1}(X_i \leq x)|$, making use of invariancy and the empirical process. Here $(\hat\alpha, \hat\beta)$ is the LSE. Notice that the model in (1.1) is different from the one in Eq. (1.2) or Eq. (1.4). Stute's test is a valid test of the model in (1.2). Sen and Sen (2014) establish the asymptotic distribution of their test under 4 scenarios based on whether $E(Y|\mathbf{X}) = \alpha + \beta'\mathbf{X}$ or $\mathbf{X} \perp \epsilon$ (see (1.2)), assuming $F_{\mathbf{X},Y} \in \Theta_{lse}$ (see (1.3)). Thus the parameter spaces Θ_p of Stute's test and the SS-test are more realistic than those of the other existing tests, as an $F_{\mathbf{X},Y}$ in these Θ_p does not have to belong to any common regression model. However, all these tests do not allow $E(|Y||\mathbf{X}) = \infty$. Our simulation study suggests that if $E(Y|\mathbf{X})$ does not exist then $P(H_1|H_0)$ (the probability of type I error) of the SS-test or Stute's test with a nominal size of 5% can be $\geq 33\%$ or even 99% (see Examples 4.4 and 4.5), as well as $\leq 5\%$. Thus a valid test allowing $E(|Y||\mathbf{X}) = \infty$ is long overdue.

3. The MD Approaches. The notations and some preliminary results are introduced in §3.1. The MD diagnostic plotting method and the MD test are introduced in §3.2, §3.3 and §3.4. Their theoretical justification is given in §3.5 and §3.6. The proofs of the lemmas and theorems are all relegated to the appendices for a better presentation.

3.1. Preliminary. Given $F_{\mathbf{X},Y} \in \Theta$, define $W = Y|(\mathbf{X} = 0)$, thus $F_W = F_o = F_{Y|\mathbf{X}}(\cdot|0)$. We first consider the test of the OLR model specified in (1.4). Thus

$$\Theta_0 = \{F_{\mathbf{X},Y} : \ Y = \beta'\mathbf{X} + W, \text{ where } W \perp \mathbf{X},\ \beta \text{ and } F_W \text{ are unknown}\}. \qquad (3.1)$$

The next lemma characterizes various common regression models and motivating the MD approach for the LR model as well as some other regression models.

Lemma 1. *If $W \perp X$, then $F_{Y|\mathbf{X}}$ is a function of (F_o, β), $F_Y(t) = E(F_{Y|\mathbf{X}}(t|\mathbf{X}))$, and*

$$F_{Y|\mathbf{X}}(t|x) = \begin{cases} 1-(1-F_o(t))^{e^{\beta' x}} & \textit{if } (\mathbf{X},Y) \textit{ follows the continuous Cox model,} \\ F_o(t-\beta' x) & \textit{if } (\mathbf{X},Y) \textit{ follows the LR model,} \\ F_o(t-g(x)) & \textit{if } (\mathbf{X},Y) \textit{ follows the gam, or plm, or gplsim model,} \end{cases}$$

where $g(\cdot)$ is unknown and is defined for various models as follows. Under the gam model, $g(x) = \sum_{j=1}^{p} f_j(x_j)$, $x = (x_1, ..., x_p)'$ and f_i is a function; under the plm model, $g(x) = \gamma' u + \sum_{j=1}^{q} f_j(z_j)$, where γ and $u \in \mathcal{R}^{p-q}$, $x' = (u', z_1, ..., z_q)$; under the gplsim model, $g(x) = f_1(\gamma' u) + \theta' z$, where $x' = (u', z')$.

For convenience, we write $F_Y(t) = F_Y(t;\beta)$, as F_Y is a function of the unknown parameter β. Given $F_{\mathbf{X},Y}$ and β, define another random variable $Y^* = \beta'\mathbf{X} + W^*$, where $F_{W^*}(\cdot) = F_{Y|\mathbf{X}}(\cdot|0)$ and $\mathbf{X} \perp W^*$. By Lemma 1, the cdf of Y^* is

$$F_{Y^*}(t)\ (= F_{Y^*}(t;\beta)) = E(F_o(t-\beta'\mathbf{X}))\ \text{(denoted also by } F^*(t) \text{ or } F^*(t;\beta)). \qquad (3.2)$$

Theorem 1. *If $F_{\mathbf{X},Y} \in \Theta_0$ (see Eq. (3.1)), then*

(a) $F_o(\cdot) = F_{Y|\mathbf{X}}(\cdot|0) = F_{Y^*|\mathbf{X}}(\cdot|0)$,

(b) $F_{Y|\mathbf{X}} = F_{Y^*|\mathbf{X}}$,

(c) $F_Y = F_{Y^*}$ $(= F^*)$.

If $F_{\mathbf{X},Y} \in \Theta \setminus \Theta_0$ then

(e) $F_o(\cdot) = F_{Y|\mathbf{X}}(\cdot|0) = F_{Y^*|\mathbf{X}}(\cdot|0)$,

(d) $F_{Y|\mathbf{X}} \neq F_{Y^*|\mathbf{X}}$,

Notice that if $F_{\mathbf{X},Y} \in \Theta_0$ as in (3.1), $E(Y|\mathbf{X})$ may not exist. By the theorem,

$$F^* = \begin{cases} F_Y & \text{if } Y = \beta'\mathbf{X} + W, \\ F_{Y^*} & \text{if } Y \neq \beta'\mathbf{X} + W. \end{cases}$$

The equation and Theorem 1 motivate the MD plot and the MD test. Given data $(\mathbf{X}_i, Y_i)$'s from $F_{\mathbf{X},Y}$, if $F_{\mathbf{X},Y} \in \Theta_0$ in (3.1), then β in $F^*(t;\beta)$ is uniquely determined by $F_{\mathbf{X},Y}$. It is often that β in $F^*(t;\beta)$ is also uniquely determined by $F_{\mathbf{X},Y}$ even if $F_{\mathbf{X},Y} \notin \Theta_0$, such as in

Examples 4.4 and 4.5 (in §4), or the case that $F_{\mathbf{X},Y} \in \Theta_{lse}$ (see (1.3)). One estimates β by the LSE if one feels confident that $\Theta_p = \Theta_{lse}$, or by the MSMLE otherwise. Consider, for example, the LSE $\hat\beta$ assuming $F_{\mathbf{X},Y} \in \Theta_{lse}$. Then β can be uniquely determined by $F_{\mathbf{X},Y}$ through $\beta = \Sigma_{\mathbf{X}}^{-1} Cov(\mathbf{X}, Y)$, as $\hat\beta = (\overline{\mathbf{X}\mathbf{X}'} - \overline{\mathbf{X}}(\overline{\mathbf{X}'}))^{-1}(\overline{\mathbf{X}'Y} - \overline{\mathbf{X}'}(\overline{Y})) \overset{a.s.}{\to} \beta$.

3.2. The MD plot. The MD plot is

to plot $y = \hat F_{Y^*}(t)$ and $y = \hat F_Y(t)$ or together with its 95% confidence band, where $\hat F_Y(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}(Y_i \le t)$, $\hat F^*(t) = \frac{1}{n}\sum_{i=1}^n \hat F_o(t - \hat\beta'\mathbf{X}_i)$, $\hat\beta$ is a consistent estimator of β *i.e.*, the LSE if one feels confident that $F_{\mathbf{X},Y} \in \Theta_{lse}$, or the SMLE if F_o is discontinuous, or the MSMLE otherwise, $\mathbf{1}(A)$ is the indicator function of an event A, the 95% confidence band is $\hat F_Y(t) \pm 1.96\sqrt{\hat F_Y(t)(1 - \hat F_Y(t))/n}$, and $\hat F_o$ is a consistent estimator of F_o to be introduced in Eq. (3.4). If the two curves are close, in particular, if the curve of $y = \hat F_{Y^*}(t)$ lies within the confidence band of $\hat F_Y$, then it suggests that the model does fit the data. If the curve of $y = \hat F_{Y^*}(t)$ lies outside the confidence band of $\hat F_Y$, then it suggests that the model does not fit the data.

The key of our MD approach is to construct an estimator of the baseline cdf F_o, say $\hat F_o$, which satisfies that for each t, $\hat F_o(t) \overset{P}{\to} F_o(t)$ $\forall$ $F_{\mathbf{X},Y} \in \Theta$. We now explain how to construct the estimators $\hat F_o$ and $\hat F^*$. Since $F_o = F_{Y|\mathbf{X}}(\cdot|0)$, it is desirable that $f_{\mathbf{X}}(0) > 0$, where $f_{\mathbf{X}}$ is the density function of $\mathbf{X}$, though this is not always true for the given $f_{\mathbf{X}}$. However, since

$$\beta'\mathbf{X} + W = \beta'(\mathbf{X} - a) + \beta' a + W, \text{ and } W \perp \mathbf{X} \text{ iff } W \perp (\mathbf{X} - a),$$

without loss of generality (WLOG), we shall assume hereafter that the zero vector satisfies

$$f_{\mathbf{X}}(0) > 0 \text{ and } Y_1, ..., Y_m \text{ are the } Y_i\text{'s where } ||\mathbf{X}_i|| \le \delta_n \text{ and } \delta_n \to 0 \ (e.g., \delta_n = cn^{\frac{-1}{3p}} \tag{3.3}$$

$c = 2r$ and r is the inter-quartile-range). Otherwise, let $\mathcal{M}$ satisfy $f_{\mathbf{X}}(\mathcal{M}) > 0$, $\check{\mathbf{X}} = \mathbf{X} - \mathcal{M}$ and $\check W = \beta'\mathcal{M} + W$, hence $Y = \beta'\mathbf{X} + W$ yields $Y = \beta'\check{\mathbf{X}} + \check W$ and $f_{\check{\mathbf{X}}}(0) = f_{\mathbf{X}}(\mathcal{M}) > 0$. Moreover, it is desirable, though not necessary, that $\mathcal{M}$ is a mode of $f_{\mathbf{X}}$, so that there are more $\mathbf{X}_i$'s in the neighborhood of $\mathcal{M}$. Recall that a (population) mode $\mathcal{M}$ of $F_{\mathbf{X}}$ satisfies that $P(||\mathbf{X} - \mathcal{M}|| < \delta) \ge P(||\mathbf{X} - a|| < \delta)$ $\forall$ small $\delta > 0$ and for all $a \in \mathcal{R}^p$, where $||\cdot||$ is a norm.

Given a random sample of size n with $\mathbf{X}_i \in \mathcal{R}^p$, in constructing $\hat{F}_o$, it is often to estimate $\mathcal{M}$ by $\hat{\mathcal{M}}$ first and then modify $\mathbf{X}_i$ by $\check{\mathbf{X}}_i = \mathbf{X}_i - \hat{\mathcal{M}}$. There are several ways to estimate $\mathcal{M}$. For instance, one can let $\hat{\mathcal{M}} = argmax_a \#\{\mathbf{X}_i : ||\mathbf{X}_i - a|| \leq \delta_n\}$, that is, the number of elements in the set $\{\mathbf{X}_i : ||\mathbf{X}_i - a|| \leq \delta_n\}$ is maximized by letting $a = \hat{\mathcal{M}}$. Another way is to first construct a $p-$ dimensional grid with $\lfloor n^{\frac{1}{3}} \rfloor$ cells as follows. Here $\lfloor x \rfloor$ is the largest integer that is not greater than x. Suppose the data are contained in a p dimensional box $B = [l_1, r_1] \times \cdots \times [l_o, r_p]$, break each interval $[l_i, r_i]$ into $\lfloor n^{\frac{1}{3p}} \rfloor$ equal intervals. Then there are roughly $\lfloor n^{\frac{1}{3}} \rfloor$ cells in the grid, say B_1, ..., $B_{\lfloor n^{\frac{1}{3}} \rfloor}$. Let B_j be the cell that the number of elements X_i's in B_j achieves the largest value among all possible cells. Then let the center of B_j be the estimator $\hat{\mathcal{M}}$. The second way is more convenient. If $n = 100$ and $p = 3$, the number of elements in the cell B_j would be at least 20. If $p \in \{1, 2\}$, then a mode can be estimated by plotting $\mathbf{X}_i$'s and finding where the data are more concentrated. Notice that both $\mathcal{M}$ and $\hat{\mathcal{M}}$ are not uniquely determined. Hereafter, WlOG, we assume $\hat{\mathcal{M}} = 0$.

Under the assumption in (3.3), the edf $\hat{F}_o$ based on Y_1, ..., Y_m is a consistent estimator of $F_o(t)$ for all $F_{\mathbf{X},Y} \in \Theta$. Let $\hat{F}^*$ be the edf based on the $n \times k$ pseudo observations $\hat{Y}_{ij} = \hat{\beta}'\mathbf{X}_i + Y_j$, $i \in \{1, 2, ..., n\}$ and $j \in \{1, ..., m\}$, where $\hat{\beta}$ is the SMLE if there exist more than 3 ties in $Y_i - \hat{\beta}'\mathbf{X}_i$'s (see Yu and Wong (2003), or Example 4.4 in Section 4), otherwise $\hat{\beta}$ is the MSMLE or the LSE. In particular,

$$\begin{aligned} \hat{F}^*(t) =& \frac{1}{n}\sum_{i=1}^{n} \hat{F}_o(t - \hat{\beta}'\mathbf{X}_i) = \frac{\frac{1}{n^2}\sum_{i,j} \mathbf{1}(Y_i + \hat{\beta}'\mathbf{X}_j \leq t, ||\mathbf{X}_i|| \leq \delta_n)}{\frac{1}{n}\sum_{k=1}^{n} \mathbf{1}(||\mathbf{X}_i|| \leq \delta)} \qquad (3.4) \\ (\hat{F}_o(t) =& \frac{\sum_i \mathbf{1}(Y_i \leq t, ||\mathbf{X}_i|| \leq \delta_n)}{\sum_i \mathbf{1}(||\mathbf{X}_i|| \leq \delta_n)} \qquad (\delta_n \text{ is as in (3.3))).} \end{aligned}$$

Remark 3. *One may wonder whether a naive estimator of F_o is the edf $\check{F}_o$ based on $\hat{W}_i$'s ($= Y_i - \hat{\beta}'\mathbf{X}_i$). This $\check{F}_o$ is a consistent estimator of F_o if H_0 in Eq. (1.4) is true. Then F^* can be estimated by $\check{F}^*(t) = \frac{1}{n}\sum_{i=1}^{n} \check{F}_o(t - \hat{\beta}'\mathbf{X}_i)$. The drawback of this naive approach is that if H_0 in Eq. (1.4) is false then $\check{F}_o$ is not consistent. In both Examples 4.1 and 4.2, $\check{F}^*$ suggests that the data fit the incorrect models Θ_0. Moreover, it requires $E(|Y||\mathbf{X}) < \infty$. Thus it does not serve our purpose of a diagnostic tool.*

If the curve of $\hat{F}_{Y^*}(t)$ lies either entirely outside or entirely inside the confidence band of $\hat{F}_Y(t)$, then the indication is quite clear. Otherwise, it is quite subjective to say whether the two curves are close. Thus it is desirable to derive certain statistical tests.

3.3. The MD tests. The MD plotting method leads to a class of tests as follows.

$$T_1 = \int |\hat{F}_Y(t) - \hat{F}^*(t)| d\hat{F}_Y(t), \ T_2 = \sup_{t \le \max_i Y_i} |\hat{F}_Y(t) - \hat{F}^*(t)|, \tag{3.5}$$

$$T_3 = \int \mathcal{W}(t)(\hat{F}_Y(t) - \hat{F}^*(t)) dG(t), \text{ or } T_4 = \int \mathcal{W}(t) |\hat{F}_Y(t) - \hat{F}^*(t)|^k dG(t),$$

where $k \ge 1$, $\mathcal{W}(\cdot)$ is a weight function, and dG is a measure, *e.g.*, dt, $d\hat{F}_o$, $d\hat{F}_Y$ and $d\hat{F}^*(t)$. The percentiles of these T_j's can be estimated by two ways:

A. Derive the asymptotic distribution of these T_i's (see Theorem 3 and §3.6);

B. Make use of the modified bootstrap method as follows.

B.1. Obtain $\hat{\beta}$, an estimator of β based on $(\mathbf{X}_i, Y_i)$'s under H_0, *e.g.*, the LSE if it is sure that $F_{\mathbf{X},Y} \in \Theta_{lse}$, otherwise, the MSMLE.

B.2. Take a random sample of size m (see Eq. (3.3)) from the $\mathbf{X}_i$'s in a neighborhood of 0 and another random sample of size $n - m$ from the $\mathbf{X}_i$'s outside the neighborhood. It yields a sample of $\mathbf{X}_i$'s, say $\mathbf{X}_1^{(1)}$, ..., $\mathbf{X}_n^{(1)}$.

B.3. If there is no tie in $\hat{W}_i$'s ($= Y_i - \mathbf{X}_i'\hat{\beta}$'s), construct a continuous distribution function $\tilde{F}_o$ satisfying $\tilde{F}_o(t) = \hat{F}_o(t)$ at the discrete points of $\hat{F}_o$. Otherwise, set $\tilde{F}_o = \hat{F}_o$.

B.4. Generate a random sample of size n from $\tilde{F}_o$, say, $W_1^{(1)}$, ..., $W_n^{(1)}$.

B.5. Let $Y_i^{(1)} = \hat{\beta}'\mathbf{X}_i^{(1)} + W_i^{(1)}$, $i = 1$, ..., n.

B.6. Now, obtain an MD test statistic value, say $T^{(1)}$, based on $(X_i^{(1)}, Y_i^{(1)})$'s and Eq.(3.5).

B.7. Repeat the previous 6 steps a large number of times, say 100 times, obtain $T^{(j)}$ for $j = 1$, ..., 100. Thus the desired percentile can be estimated by these $T^{(j)}$'s.

The MD tests are valid tests of H_0^{MD}: $F_Y = F_{Y^*}$ against H_1^{MD}: $F_Y \neq F_{Y^*}$, where Y^* is defined in §3.1. Thus the probability of type II error for testing H_0 in (1.4) may be large if H_0 is not true but $F_{Y^*} = F_Y$. Their parameter space is at least $\Theta^{MD} = \Theta_{lse} \cup \Theta_m \cup \Theta_d$ to be proved in §3.5. In view of Lemma 1, the baseline cdf $F_o(t) = F_{Y|\mathbf{X}}(t|0)$ is well defined for each $F_{\mathbf{X},Y} \in \Theta$.

3.4. About the WLR model. We can also consider another test of a WLR model

$$H_0\colon\ Y = \beta'\mathbf{X} + W, \text{ where } W = \alpha + g(\mathbf{X})\epsilon, \text{ and } \epsilon \perp \mathbf{X}, \tag{3.6}$$

ϵ is a random variable with variance $\sigma_\epsilon^2 = 1$, and $g(\cdot)$ is a function. Define $Y_i^w = Y_i/g(\mathbf{X}_i)$, then $Y_i^w = \beta'\mathbf{X}_i/g(\mathbf{X}_i) + \alpha/g(\mathbf{X}_i) + \epsilon_i^w$, where $\epsilon_i^w = (W_i - \alpha)/g(\mathbf{X}_i)$. Under the WLR model, there exist estimators of (α, β) and $g(\mathbf{X}_i)$ in the literature (see Draper and Smith (1966)), say $(\hat\alpha, \hat\beta)$ and $\hat g(\mathbf{X}_i)$. Define Y_{ij}^* by

$$Y_{ij}^* = \hat\beta'\mathbf{X}_i + Y_j/\hat g(\mathbf{X}_i),\ i = 1, ..., n \text{ and } j = 1, ..., m. \tag{3.7}$$

Then the MD plot and MD test for the WLR model can be derived in a similar way as those for the OLR model in (1.4). We skip the details.

3.5. Theoretical justification of the MD approach. It can be shown that under certain condition, $\hat F^*$ is consistent, thus the MD plot makes sense, and the MD test is a consistent test of H_0^{MD}, that is, the probability of type II error $P(H_0|H_1) \to 0$ if $F_Y \neq F_{Y^*}$. Moreover, under certain regularity conditions, we establish the asymptotic distribution of the process $\sqrt{n}(\hat F^* - \hat F_Y)$ so that the asymptotic distribution is of the MD tests follow.

Lemma 2. *Suppose* $\sqrt{n}P(||\mathbf{X}|| \leq \delta_n) \to \infty$, *then* $\sup_{t\in\mathcal{R}} |\hat F_o(t) - F_o(t)| = o_p(1)$.

Remark 4. *If* 0 *is a mode of* $\mathbf{X}$ *and* $\delta_n = 2rn^{\frac{-1}{3p}}$ *(see (3.3)), then* $\sqrt{n}P(||\mathbf{X}|| \leq \delta_n) \to \infty$. *To establish the convergence of* $\hat F^*(t)$, *the following regularity assumption may be needed.*

$$\sup_{t\in\mathcal{R}} |\frac{1}{n}\sum_{j=1}^{n}[F_o(t - \hat\beta'\mathbf{X}_j) - F_o(t - \beta'\mathbf{X}_j)]| = o_p(1), \tag{3.8}$$

and $\hat\beta \overset{a.s.}{\to} \beta$ *or* $P(\hat\beta \neq \beta$ *infinitely often*$) = 0$ *(satisfied by the SMLE and the MSMLE if* F_o *is discontinuous (see Yu and Wong (2002, 2003, 2004))).*

Lemma 3. *Eq. (3.8) is satisfied if either (i)* F_o *is continuous and* $\hat\beta \overset{a.s.}{\to} \beta$, *or (ii)* $F_{\mathbf{X},Y} \in \Theta_0 \cup \Theta_d$ *and* $\hat\beta$ *is the SMLE or MSMLE, or (iii)* $F_{\mathbf{X},Y} \in \Theta_{lse}$ *and* $\hat\beta$ *is the LSE.*

Notice that $F_{\mathbf{X},Y}$ does not need to belong to Θ_0 in cases (i) and (iii) of the lemma.

Theorem 2. *If $(\mathbf{X}_i, Y_i)$, $i = 1, \ldots, n$, are a random sample from $F_{\mathbf{X},Y} \in \Theta$ and if zero vector is a mode of $\mathbf{X}$, then for all $\beta \in \mathcal{R}^p$,*

$$\sup_{t\in\mathcal{R}} |\frac{1}{n}\sum_{i=1}^{n} \hat{F}_o(t-\beta'\mathbf{X}_i) - E(F_o(t-\beta'\mathbf{X}))| = o_p(1). \tag{3.9}$$

Moreover, assume that either of the three sufficient conditions stated in Lemma 3 holds, then

$$\sup_{t\in\mathcal{R}} |\hat{F}^*(t;\hat{\beta}) - F_{Y^*}(t;\beta)| = \sup_{t\in\mathcal{R}} |\frac{1}{n}\sum_{i=1}^{n} \hat{F}_o(t-\hat{\beta}'\mathbf{X}_i) - E(F_o(t-\beta'\mathbf{X}))| = o_p(1). \tag{3.10}$$

Theorem 3. *Suppose that $Y = \beta X + W$, F_o is discontinuous, $\hat{\beta}$ is either the SMLE or the MSMLE, $\mathbf{X} \perp W$ and $P(X = 0) = q > 0$. Then*

$$\sqrt{n}\Big(\hat{F}^*(t) - \hat{F}_Y(t)\Big) \text{ converges in distribution to } D_1(t) - F_Y(t)D_2 - D_3(t), t \in [-\infty, \infty]$$

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

$$\begin{aligned} Cov(D_1(t), D_1(s)) =& q^{-1}\int\Big(\int \mathbf{1}(w+\beta x \le t)dF_X(x)\int \mathbf{1}(w+\beta x \le s)dF_X(x)\Big)dF_o(w) \\ & - 2F_Y(t)F_Y(s) + \int F_o(t-\beta x)F_o(s-\beta x)dF_X(x) \\ & + F_Y(t)[F_o(s) - F_Y(s)] + F_Y(s)[F_o(t) - F_Y(t)], \end{aligned} \tag{3.11}$$

$$\begin{aligned} Cov(D_3(t), D_3(s)) =& F_Y(t\wedge s) - F_Y(t)F_Y(s), \\ Cov(D_1(t), D_3(s)) =& \int\int \mathbf{1}(w+\beta x \le t)\mathbf{1}(w \le s)dF_X(x)dF_o(w) - F_Y(t)F_Y(s) \\ & + \int F_o(t-\beta x)F_o(s-\beta x)dF_X(x) - F_Y(t)F_Y(s). \end{aligned}$$

The proofs of Theorems 2 and 3 put in Appendix I and Appendix II, respectively. In Theorems 2 and 3, we make use of some simple assumptions. But we believe that these assumptions can be weakened. For instance, the condition of $\sqrt{n}P(||\mathbf{X}|| \le \delta_n) \to \infty$ in Theorem 2 can be replaced by $\sqrt{n}\delta_n \to \infty$, provided that $f_{\mathbf{X}}$ is both continuous and positive at the origin. The condition $P(\mathbf{X} = 0) > 0$ in Theorem 3 can be replaced by $P(\mathbf{X} = a) > 0$

for some a due to (3.4). The other assumptions in Theorem 3 can also be weakened, but the proofs would be much longer and more difficult.

Remark 5. *The main idea of the proof of Theorem 3 can be explained as follows. Under the assumptions in Theorem 3, F_o is discontinuous and $Y = \beta X + W$. Then the SMLE and the MSMLE $\hat\beta$ of β satisfy $P(\hat\beta \neq \beta$ i.o.$) = 0$ (Yu and Wong (2002 and 2003)). As a consequence, $P(\hat\beta = \beta$ if n is large enough$)$=1. If n is large, then WLOG, we can assume $\hat\beta = \beta$ and thus*

$$\begin{aligned}\hat F^*(t) - \hat F_Y(t) =& \frac{\frac{1}{n^2}\sum_{j=1}^n \sum_{i=1}^n \mathbf{1}(Y_i + \beta X_j \le t, X_i = 0)}{\frac{1}{n}\sum_{k=1}^n \mathbf{1}(X_k = 0)} - \frac{1}{n}\sum_{i=1}^n \mathbf{1}(Y_i \le t)\\ &\overset{def}{=} G(U,V,Z), \text{ where } G(U,V,Z) = \tfrac{U}{V} - Z,\end{aligned} \tag{3.12}$$

and U, V and Z are defined in an obvious way. U is a typical U-statistic. V and Z are two sample means of i.i.d. random variables with finite expectations. One can show that $G(U,V,Z) \overset{\cdot a.s.\cdot}{\to} G(E((U,V,Z))) = \frac{F_Y(t)q}{q} - F_Y(t)$. The asymptotic variance of the process $\sqrt{n}(\hat F^*(t) - \hat F_Y(t))$ can be derived by the delta method, which can be simplified as follows.

$$\begin{aligned}\sigma^2 =& \frac{E(F_o(t - [(\beta'\mathbf{X}_2) \vee (\beta'\mathbf{X}_3)])) - F_Y^2(t)}{q} - E(F_o^2(t - \beta'\mathbf{X})) - 2E(F_o(t - (\beta'\mathbf{X}) \vee 0))\\ &+ 2F_o(t)F_Y(t) + F_Y(t), \text{ where } \mathbf{X}_2 \text{ and } \mathbf{X}_3 \text{ are i.i.d. from } \mathbf{X}.\end{aligned}$$

The covariance can also be simplified but is skipped for simplicity.

Remark 6. *Eq. (3.9) and Eq. (3.10) are the theoretical justification of the MD approach. First, Eq. (3.10) implies that under certain regularity conditions, e.g. the 3 sufficient conditions in Lemma 3, $\hat F^*(\cdot, \hat\beta)$ is a consistent estimator of F_{Y^*} That is if n is large and H_0 is true, the curves of $\hat F_Y$ and $\hat F_{Y^*}$, which are equal, should be very close. On the other hand, if $F_{\mathbf{X},Y} \notin \Theta_0$ it is likely (though not always) that $F_Y \neq F_{Y^*}$ and thus $\hat F_Y$ and $\hat F_{Y^*}$ are not close. Thus the MD plot is a good diagnostic method for the model Θ_0 in (1.4) or (3.6).*

3.6. Justification of the bootstrap approach. There are two approaches to determine the percentiles of the MD test statistics T_i's: One is to establish the asymptotic distribution of T_i's under H_0, making use of Theorem 3, that is, making use of formula (3.11) or (3.12). The alternative is to make use of the bootstrap method as in §3.3. The first approach is actually the common approach in constructing a test. However, there is a drawback in this approach, namely, the distribution of the test may be false if H_0 is false, *e.g.*, the gam test is invalid if $F_{\mathbf{X},Y}$ belongs to a Cox's model and thus it makes type II error with probability $\geq 90\%$ (see Example 4.1). In our case, formula (3.11) is false if $Y \neq \beta'\mathbf{X} + W$ or $W \not\perp \mathbf{X}$.

In contrast, the advantage of bootstrapping distribution of $\hat{F}^* - \hat{F}_Y$ is that its distribution is valid even if H_0 is false. Thus, in our simulation, we did not make use of Theorem 3. The main idea of the bootstrapping MD test is that one can generate a pseudo random sample of the regression data based on the original sample of $\mathbf{X}_i$'s and based on the sample $Y_1, ..., Y_m$ in the neighborhood of a support point of $F_{\mathbf{X}}$, say 0. In view of Theorem 2, this pseudo random sample satisfies the following properties:

1) If the model fits the data, the pseudo random sample is from a model which is approximately the same as the true model Θ_0 specified by H_0 in Eq. (1.4) (due to Eq. (3.9)).
2) Otherwise, $\forall\ F_{\mathbf{X},Y} \in \Theta$, the pseudo random sample is from a model which is *exactly* the same as the model $Y_o = \hat{\beta}'\mathbf{X} + W_o$ with the baseline cdf $F_{W_o} = \hat{F}_o \approx F_o$ and $W_o \perp \mathbf{X}$.
3) Moreover, if $F_{\mathbf{X},Y}$ satisfies the assumptions in Theorem 2, then the pseudo random sample is from a model which is approximately the same as the model $Y^* = \beta'\mathbf{X} + W^*$ with the baseline cdf $F_{W^*} = F_o$, $W^* \perp \mathbf{X}$, and $\hat{F}^*$ approximate the same as $F^*(\cdot;\beta)$ with $\beta = \lim \hat{\beta}$ a.s. (due to Eq. (3.10)).

Property 3 is interesting, but not important. Properties 1 and 2 are important. In particular, if the data do not fit the model in H_o, then it is often that $\hat{F}_Y$ and $\hat{F}^*$ are quite different. Thus if $F_Y \neq F_{Y^*}$, or $P(F_Y \neq \hat{F}^*) = 1$ as in property 2, $P(H_0|H_1) \to 0$, as $n \to \infty$. That is, the MD test is a consistent model test of H_0^{MD}: $F_Y = F_{Y^*}$ for all $F_{\mathbf{X},Y}$, or at least for all $F_{\mathbf{X},Y}$ under certain regularity conditions.

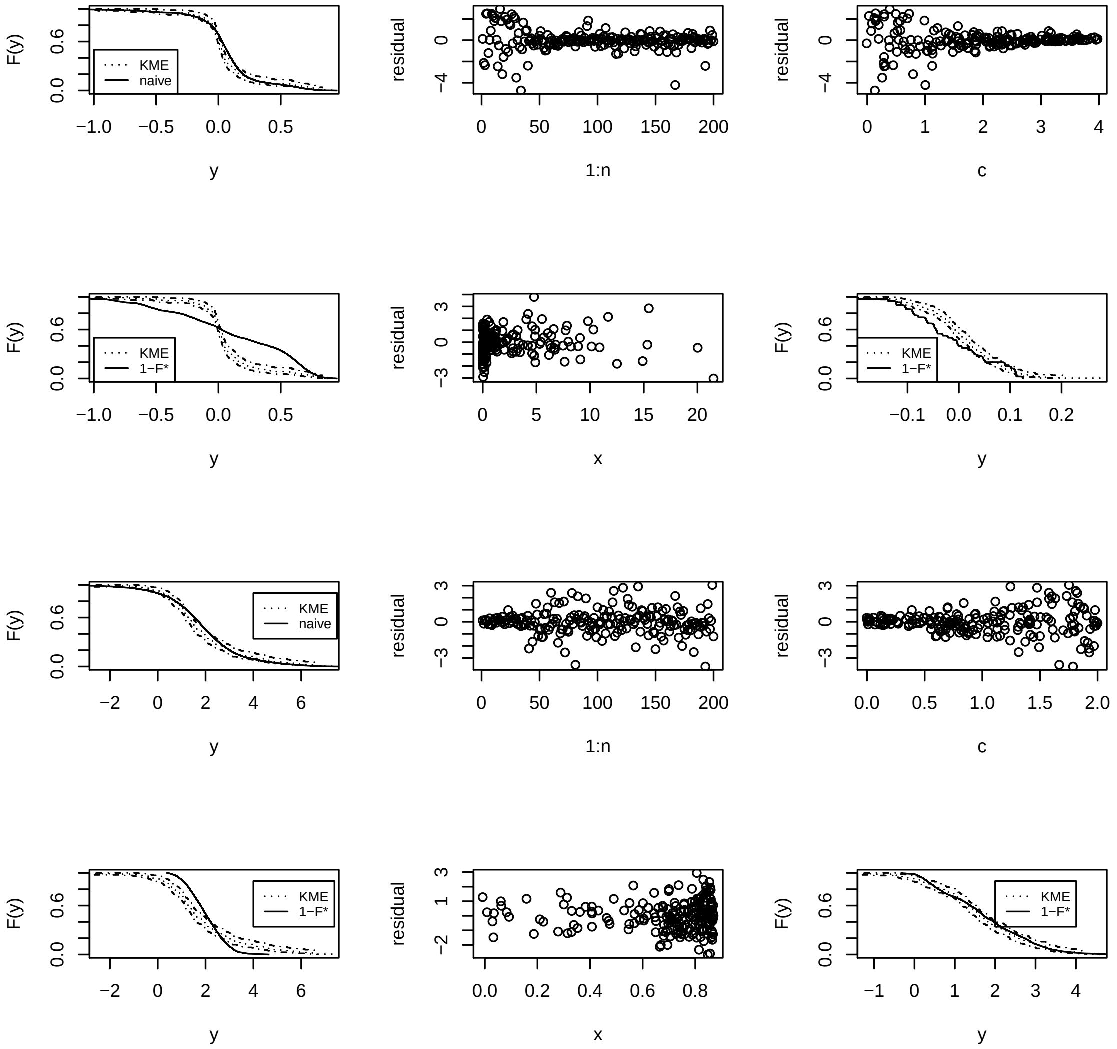
4. Simulation Results. We compare the MD tests T_1 and T_2 (see Eq. (3.5)) to the SS-test, or the gam test (using R package“gam”), or the t-test (see Example 2.1), which are all relevant for testing H_0 in Eq. (1.4) or (3.6), under different parameter spaces Θ_p. Since Stute's test is not relevant to H_0, we did not compare it, though our simulation results indicate that its size with a nominal size of 0.05 may be $\geq 0.38\%$ if $E(Y|\mathbf{X})$ does not exist. In the simulation study in Sen and Sen (2014), the data are all from the LR models that may or may not satisfy H_0 in Eq. (1.4). In addition to the models similar to those in Sen and Sen (2014), we also consider models that are from the Cox model, or $E(Y|\mathbf{X})$ may not exist, or $\mathbf{X} \not\perp W$. In all cases, the replication is 5000. We only report the results on T_1, as T_1 is better than T_2 most of the time in our simulation. $\hat{P}(H_1|H_0)$ and $\hat{P}(H_0|H_1)$ are the estimated $P(H_1|H_0)$ and $P(H_0|H_1)$, respectively. We estimate β by the LSE in the first 3 examples, and by the SMLE or the MSMLE in Examples 4.4 and 4.5, as $E(Y|\mathbf{X})$ does not exist only in Examples 4.4 and 4.5, and it is faster to compute the LSE. We ignore MD plots in Examples 4.4, 4.5 and 4.6 to cut the length of the paper.

Example 4.1. We generated data (X_i, Y_i), $i = 1, ..., n$ from the Cox model $h(t|X) = h_o(t)\exp(X)$, where $h_o = 1(t \geq 0)$, $X \sim U(-4/k, 4)$, $k \approx n^{0.7}$, and n is between 60 and 300. We fitted the data to the OLR or WLR model $Y = \beta X + W$.

The Cox model does not belong to any LR model. The Cox model can be viewed as a partially linear regression model or a gam model, but the gam test is invalid, as $X \not\perp Y - \beta X - E(Y|X)$. The t-test is also invalid, as discussed in Example 2.1.

For such data with a sample size $n = 200$, the residual plots (see panels (1,2) and (1,3) in Figure 1) and the MD plot (see panel (2,1)) suggest that the OLR model may not fit the data, but a WLR model with a weight function $\sqrt{|(X-4)^3\mathbf{1}(X < 3.7) + (X-4.5)^3\mathbf{1}(X \geq 3.7)|}$ might work (see the residual plot in panel (2,2)). However, the MD plot (see panel (2,3)) suggests that the WLR model does not fit the data neither. Thus the MD plots are better.

We applied the MD test, the SS-test and the gam test to these models. Since neither the OLR model nor the WLR model fit the data, $\hat{P}(H_0|H_1)$ and $\hat{P}(H_0^2|H_1^1)$ are presented in the top half of Table 1. The simulation results suggest that the MD test T_1 and the SS-test perform very well for testing the incorrect OLR model, even when $n = 60$. The MD test can detect that the data do not fit the WLR model for large sample sizes such as $n \geq 200$. The gam test almost never rejects the OLR model and the WLR model for the data from the Cox model with the sample sizes $n \leq 300$. The SS-test cannot detect the incorrect WLR model for $n \leq 300$. Here T_1 for the WLR model is the modified one specified in §3.4 and Eq. (3.7).



Cox data naive $1-\check{F}^*$ plot for OLR (*i*, residual) (X_i,residual) for OLR
MD plot for OLR (X_i,residual) for WLR MD plot for WLR
WLR data naive $1-\check{F}^*$ plot for OLR (*i*, residual) (X_i,residual) for OLR
MD plot for OLR (X_i,residual) for WLR MD plot for WLR

Figure 1. Residuals and MD plots under the Cox Model or the WLR model

For comparison sake, we also generated random samples from another WLR model:

$$Y = X + W, \text{ where } W \perp X,\ W \sim N(1, X + 0.3) \text{ and } X \sim U(0, 2).$$

Under this model, we carried out two sets of simulation studies. We first fitted the data to the OLR model $Y = \beta X + W$, where $W \perp X$. The residual plots (see panels (3,2) and (3,3)) and the MD plot (see panel (4,1)) of Figure 1 suggest that the OLR model does not fit the data, but a WLR model might work (see panels (4,2) and (4,3)). The naive estimator $\check{F}^*$ (see Remark 3) suggests that the data from the Cox model and from the WLR model all fit the OLR model (see panels (1,1) and (3,1)). Thus it is useless. We also applied the same three tests to the WLR model. Since the data were from the WLR model, thus we computed $\hat{P}(H_0|H_1)$ for fitting the OLR model and $\hat{P}(H_1^2|H_0^2)$ for fitting the WLR model, where H_0^2: the model is the WLR model v.s. H_1^2: H_0^2 is not true. The simulation results are presented in the bottom half of Table 1. The nominal sizes of these tests are 0.05.

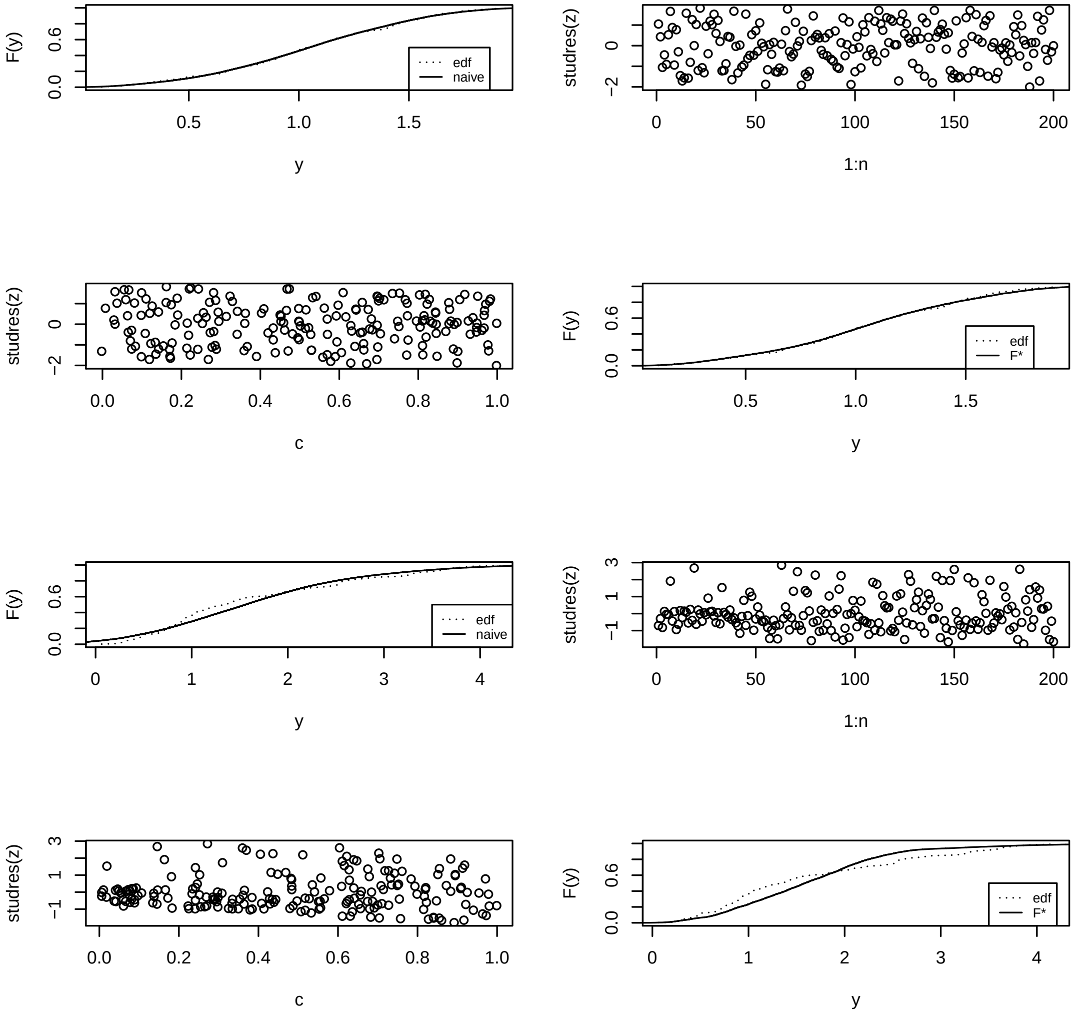
Test model:		OLR			WLR		
Data	Test:	T_1	SS	gam	T_1	SS	gam
	n	$\hat{P}(H_0 \mid H_1)$			$\hat{P}(H_0^2 \mid H_1^2)$		
Cox	60	0.01	0.06	1.00	0.78	0.96	1.00
	100	0.00	0.00	1.00	0.55	0.96	1.00
	200	0.00	0.00	1.00	0.19	0.95	1.00
	300	0.00	0.00	1.00	0.02	0.94	1.00
		$\hat{P}(H_0 \mid H_1)$			$\hat{P}(H_1^2 \mid H_0^2)$		
WLR	60	0.03	0.00	0.59	0.04	0.05	0.08
	120	0.00	0.00	0.60	0.04	0.05	0.07

Table 1. Simulation Results in Example 4.1

The simulation results suggest that the SS-test is a little bit more powerful than the MD test if the data are from the WLR model, but the MD test is more powerful than the SS-test if the data are from the Cox model.

Example 4.2. Consider two OLR models: (1) $Y = C\beta + W$, and (2) $Y = C\beta + \theta Z_3 + W$, where $\beta = 1$, $\theta = 3$, $W \sim U(0,1)$, $C \sim U(-1/k, 1)$ and $Z_1 \sim bin(1, (C + 1/k) * 0.9)$, $k \approx n^{0.7}$, $Z_2 \sim U(0,1)$, $Z_3 = Z_2 Z_1$ and $W \perp (C, Z_1, Z_2)$. We generated random samples (C_i, Y_i)'s from the first model and another random samples (C_i, Y_i)'s (without Z_{3i}'s) from the second model with $n \in \{60, 200\}$. The first sample satisfies the OLR model (1.4), but the second sample does not satisfy the OLR model with $X = C$, as C and Z_3 are correlated and thus $C \not\perp (3Z_3 + W)$. We fit these two data sets to the first model.

In Figure 2, we present the various plots with $n = 200$. Notice from Figure 2 that the residual plots in panels (1,2), (2,1), (3,2) and (4,1) suggest both data sets fit Model 1. The MD plots (see panels (2,2) and (4,2)) suggest the first data set but not the second data set fits Model 1. The naive plot $\check{F}^*$ in panel (3,1) (see Remark 3) is not as good as the MD plot in panel (4,2).



Model 1 data: naive $\check{F}^*$ plot $(i$, residual)
$(C_i$, residual) MD plot
Model 2 data: naive $\check{F}^*$ plot $(i$, residual)
$(C_i$, residual) MD plot

Figure 2. Residuals and MD plots for fitting Model 1

We also compared various tests: the MD test T_1, the gam test, the SS-test and the t-test. The first data set is from Model 1, thus we computed $\hat{P}(H_1|H_0)$. The second data set is not from Model 1, thus we computed $\hat{P}(H_0|H_1)$. The t-test and the gam test are invalid, even though Model 2 is an OLR model. The simulation results are given in Table 2.

	model (1)				model (2)			
	$\hat{P}(H_1|H_0)$				$\hat{P}(H_0|H_1)$			
n	T_1	SS	gam	t-test	T_1	SS	gam	t-test
60	0.04	0.05	0.00	0.05	0.24	0.35	0.96	> 0.96
200	0.05	0.05	0.00	0.05	0.01	0.00	0.96	> 0.96

Table 2. Simulation Results in Example 4.2

The four tests all have sizes smaller than the nominal size 0.05. The t-test is not applicable if the data are from the second model (see Example 2.1). Nevertheless, we present the rate of the t-test not rejecting $\theta = 0$ (though $P(H_0|H_1) = 1$, as the correct model is Model 2, not Model 1). The gam test performs quite poorly. It simply cannot detect the wrong model in this case. The MD test and the SS-test perform well for $n = 60$ and extremely well for $n = 200$. It seems that the MD test is more powerful than the SS-test in this case if n is moderate, but not for a large n.

Example 4.3. The MD approach is based on the semi-parametric method and it is expected to work well when the sample sizes are large. We now consider a case that the sample size is not very large. We generated data with the sample sizes $n = 20$ and 50 from the OLR model: $Y = X + W$, where $X \sim U(0, 1)$ and $F_W(t) = 1 - e^{-5t}$, $t > 0$.

For this data set, we may test three hypotheses:

Case 1: H_0: $Y = \beta X + W$ v.s. H_1: $Y \neq \beta X + W$.

Case 2: H_0^1: $\alpha = 0$ v.s. H_1^1: $\alpha \neq 0$.

Case 3: H_0^2: $\beta = 0$ v.s. H_1^2: $\beta \neq 0$.

That is, in either of the three cases, the correct assumption is H_0: the model is $Y = X + W$ with $E(W) = \alpha$, and both H_i^1 and H_1^2 are special cases of H_0. Notice that in both cases, the null hypotheses are false, as $\alpha = 0.2$ and $\beta = 1$.

The MD plot (see panel (2,3)) and the residual plots in Figure 3 suggest that the model $Y = \beta X + W$ is appropriate. In particular, the curve $1 - \hat{F}^*$ lies within the confidence band of $1-\hat{F}_Y$. The scatter plot of (X_i, Y_i)'s (see panel (1,1)) suggests that $\beta \neq 0$ but $\alpha = 0$. The MD plot in panel (2.2) both suggest that $\beta \neq 0$, as the curve $1-\hat{F}^*$ lies outside the confidence band of $1-\hat{F}_Y$. But the MD plot in panel (2.1) raises some doubt on whether $\alpha = 0$. The t-test, the MD test and the SS-test are all valid. Notice that the data fit the model in H_0. We present the estimates of the probabilities of errors in Table 3. The simulation suggests that the MD approach works even for small sample sizes, but the SS-test is more powerful.

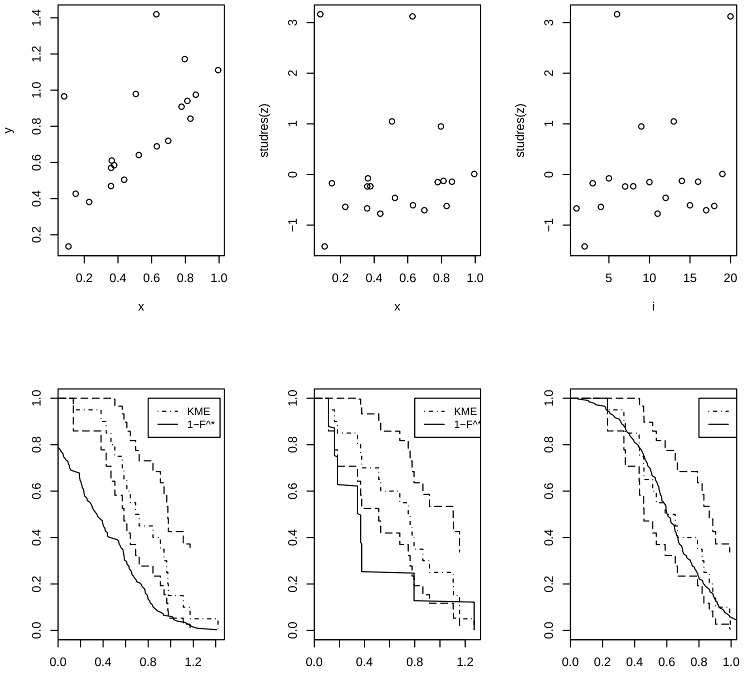


Figure 3. MD plots for a sample size $n = 20$

n	$\hat{P}(H_1\|H_0)$			$\hat{P}(H_0^1\|H_1^1)$			$\hat{P}(H_0^2\|H_1^2)$		
	T_1	SS	t	T_1	SS	t	T_1	SS	t
20	0.004	0.002	0.05	0.043	0.028	0.041	0.286	0.033	0.050
50	0.003	0.000	0.05	0.001	0.000	0.071	0.145	0.000	0.064

Table 3. Simulation Results in Example 4.3

Example 4.4. Let Z be a Cauchy random variable with the density function $f_Z(t) = \frac{1}{\pi(1+t^2)}$,

$$F_X(t) = (0.5 + 0.5F_{|Z|}(t))\mathbf{1}(t \geq 0), \text{ and } F_W(t) = 0.5\mathbf{1}(t \geq 0) + 0.5F_Z(t). \qquad (4.1)$$

Consider Model 1: $Y = X + W$, where $X \perp W$. Then $E(W)$ and $\sigma_{\mathbf{X}}$ do not exist, but $E(\ln f_W(W))$ exists. Let (X_1, Y_1), ..., (X_n, Y_n) be i.i.d. from the model $Y = X + W$, where $n \in \{25, 50, 100, 200\}$. It is proved in Appendix that the SMLE and the MSMLE are $\beta = 1$ with probability 1 (w.p.1) if n is large enough. The simulation study suggests that the LSE of β is not consistent as expected. The upper half of Table 4 presents the simulation results on the estimates $\hat{P}(H_1^1|H_0^1)$, of the MD test T_1, the gam test and the SS-test. Our simulation results suggest that with the nominal size 0.05, T_1 has $P(H_1|H_0) \leq 0.05$, but the SS-test and the gam test have $P(H_1|H_0)$ converging to 0.33 and 0.9+, respectively. Thus the results suggest that the MD test is a valid test for H_0 in Eq. (1.4), but not the SS-test and the other existing tests. Since the SS-test has a $\hat{P}(H_1|H_0)$ converging to 0.33, the SS-test cannot be a test of both H_0 in (1.4) and the assumption $E(|Y||\mathbf{X}) < \infty$. It is interesting to see that the MD approach works well even with a sample size as small as 25.

		$\hat{P}(H_1^1 \mid H_0^1)$			$\hat{P}(H_0^2 \mid H_1^2)$		
		T_1	gam	SS	T_1	T_1	T_1
	n				$b = 1$	$b = 10$	$b = 300$
Ex. 4.4	25	0.03	0.66	0.31	0.95	0.94	0.86
	50	0.03	0.78	0.33	0.93	0.58	0.17
	100	0.03	0.86	0.33	0.81	0.27	0.01
	200	0.04	0.91	0.33	0.65	0.15	0.00
Ex. 4.5	25	0.04	1.00	0.98	0.96	0.94	0.90
	50	0.04	1.00	0.99	0.93	0.90	0.67
	100	0.03	1.00	1.00	0.91	0.77	0.46
	200	0.04	1.00	1.00	0.82	0.74	0.27

Table 4. Simulation Results in Examples 4.4 and 4.5

We also generated random sample (X_i, Y_i)'s from Model 2: $Y = b\exp(X) + W$, and test the model $Y = \beta X + W$, where $F_X(t) = 0.4\mathbf{1}(t \geq 1) + 0.12\mathbf{1}(t \geq 0) + 0.48F_{|Z|}(t)$, F_W is as in Eq. (4.1) and $b \in \{1, 10, 300\}$. The SMLE or MSMLE $\hat{\beta} = b(e-1) \neq b$ if n is large (see Appendix). Our simulation study on T_1 yields $\hat{P}(H_0^2|H_1^2)$ given at the top right of Table 4. Though $\hat{P}(H_0^2|H_1^2)$ has not reached 0 yet when $n = 200$ and $b = 1$ or 100, the tendency indicates $\hat{P}(H_0^2|H_1^2) \to 0$, and the results suggest that the MD test is consistent. However, the convergence speed depends on the model, the larger the b, the faster. This is also why we replace f_X in Model 2. In view of $\hat{P}(H_1^1|H_0^1)$ for the gam test and the SS-test, there is no point to compute $\hat{P}(H_0^2|H_1^2)$ for the other existing tests. **Example 4.5.** The models in Example 4.4 satisfy $E(|\ln f_W(W)|) < \infty$. Under this assumption, it is proved that the MSMLE of β is consistent. Now let the two models in Example 4.4 remain the same except that F_W is replaced by $F_W(t) = 0.5[\mathbf{1}(t \geq 0) + \int_e^t s^{-1}(\ln s)^{-2} ds \mathbf{1}(t > e)]$. Then both $E(W)$ and $E(\ln f_W(W))$ do not exist. We generate (X_i, Y_i) from these two models, and test H_0 as in Eq. (1.4). The simulation results are presented in the bottom half of Table 4. The results suggest that the MD tests are valid even if $E(|\ln f_W(W)|) = \infty$, though $\hat{P}(H_0^2|H_1^2)$ converges slower than those in Example 4.4. Notice that $\hat{P}(H_1^1|H_0^1)$ for the gam test and the SS-test with nominal size 0.05 converges to 1.00, even worse than the case of Example 4.4.

Example 4.6. We generated data from the model $Y = X + W$, where F_W is as in Eq. (4.1) and $X \sim bin(1, 0.5)$. For $n \in \{100, 1000\}$, the sample means (SE's) of the LSE are 5.85 (61412) and 6.10 (144392), respectively. It suggests that the LSE does not converges.

5. Data Analysis. We present data analysis of three sets of real data. The MD plots and residual plots of the three data sets are presented in Figure 4. We use the MSMLE in the MD approach.

Example 5.1. Divusa Data set (Faraway (2016), p.30). The data set records divorce rates Y in the USA from 1920 to 1996 (X), with $n = 77$. The fitted values and response variables roughly lie along the 45 degree line (see panel (1,1)), while the residual plot (1,2) suggests the violation of linearity assumption or time correlation of residuals. The MSMLE $\hat{\beta} = 0.1$. The MD plot does not support the linear model (see panel (1,3)). The MD tests, the SS-test and the gam test all reject H_0 in (1.4) with p-values ≈ 0.00.

Example 5.2. Sat Data set (Faraway (2016), p.86). The data were collected to study the relationship test score Y and the salary X with $n = 50$. Both fitted values plot and residual plot (see panels (2,1) and (2,2)) suggest that the linearity assumption may not be true, agreeing with the MD plot (see panel (2,3)). The MSMLE $\hat{\beta} = -6.431$. The p-values are 0.05, 0.01, 0.13 and 0.34 for T_1, T_2, the gam test and the SS-test, respectively. Thus the MD tests reject H_0 as in (1.4), but not the gam test and the SS-test. The MD tests are better than the gam test and the SS-test here.

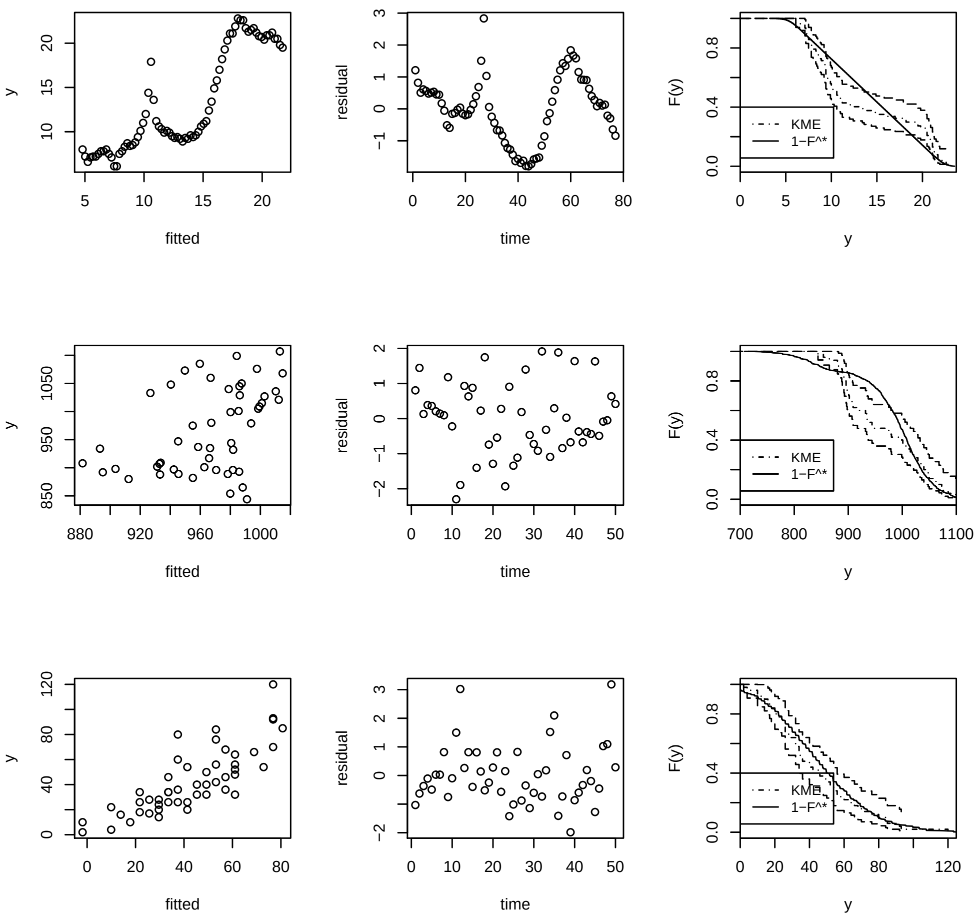


Figure 4. Diagnostic plots for 3 real data sets

Example 5.3. Cars Data set (Ezekiel (1930)). The data give the speed X of cars and the distances Y taken to stop and were recorded in the 1920s with $n = 50$. The MD plot (see panel (3,3)) and the plot in panels (3,1) and (3,2) suggest that the linearity assumption is correct. The p-values are 0.66, 0.68, 0.11 and 0.86 for T_1, T_2, the gam test, and the SS-test, respectively. They all support the LR model. It is worth mentioning that there are 26 distinct values of $Y_i - \tilde{\beta}X_i$'s and 34 ties. For instance, there are 8 ties at 8. The SMLE and MSMLE $\tilde{\beta} = 2$ and the LSE $\hat{\beta} = 3.96$.

It is worth mentioning that the MD plots (see panels (1,3), (2,3) and (3,3)) clearly indicate that $\hat{F}^*$ is either within or mostly outside the confidence band of the edf. The gam test and the SS-test reject H_0: $Y = \beta\mathbf{X} + W$ only in Example 5.1, whereas the MD tests reject H_0 in Examples 5.1 and 5.2. Moreover, the MD tests agree with the MD plots in these three data sets. But the gam test and the SS-test disagree with the residual plots and MD approaches in Example 5.2. Thus in these examples, the MD approaches perform better than the residual plots, the SS-test and the gam test.

6. Concluding Remark. It is well known that if $F_{\mathbf{X},Y} \in \Theta_{lse}$, then most existing tests are valid if $F_{\mathbf{X},Y} \in \Theta_0$, except some parametric tests. It is interesting to notice from Examples 4.4 and 4.5 that the gam test and the SS-test (as well as Stute's test) can have a size much greater than the nominal size 0.05 if H_0 is true but $E(|Y||\mathbf{X}) = \infty$. Thus the distributions claimed about these tests are false if H_0 is true and $E(Y|\mathbf{X})$ does not exist. But the MD test is a valid test of H_0 against H_1 in Eq. (1.4), even if either $E(Y|\mathbf{X})$ does not exist or H_0 is false. Notice that our simulation results indicates that none of the MD test and the SS-test is uniformly more powerful than the other in the cases that the SS-test is valid. Example 4.1 suggests that the MD tests are more powerful if the data are from the Cox model. On the other hand, if $F_Y = F_{Y^*}$ but H_0 in (1.4) fails, then the SS-test is more powerful than the MD test. The MD test is a valid test of H_0^{MD}: $F_Y = F_{Y^*}$, though it is not a consistent test of H_0 in (1.4) if $F_{Y|\mathbf{X}} \neq F_{Y^*|\mathbf{X}}$ but $F_Y = F_{Y^*}$.

7. Appendix I. We present the proofs of the lemmas and theorems in Section 3 and the proofs of some examples in Section 4 here.

Proof of Lemma 1. $F_Y(t) = E(F_{Y|\mathbf{X}}(t|\mathbf{X}))$ is trivially true. Notice that Cox's model is specified by $h_{Y|\mathbf{X}}(t|x) = h_o(t)e^{\beta' x}$, and

$$S_{Y|\mathbf{X}}(t|x) = \exp(-\int_{u\leq t} h_{Y|\mathbf{X}}(u|x)du) = \exp(-\int_{u\leq t} h_o(u)du e^{\beta' x}) = (S_o(t))^{e^{\beta' x}}.$$

Under the LR model $Y = \beta'\mathbf{X} + W$, $F_W(t) = F_o(t) = F_{Y|\mathbf{X}}(t|0)$. Since $W \perp \mathbf{X}$ by assumption and $W = Y - \beta'\mathbf{X}$, it is easy to yield $F_{Y|\mathbf{X}}(t|x) = F_o(t - \beta' x)$.

The proofs for the gam model and the plm model are similar and are skipped. □

Proof of Theorem 1. Notice that $F_{Y|\mathbf{X}}$ and $F_{Y^*|\mathbf{X}}$ are two conditional cdf and both conditional on the same random variable $\mathbf{X}$. $F_{Y^*|\mathbf{X}}$ is the cdf satisfying $E(Y|\mathbf{X}) = \beta'\mathbf{X}$ with parameter β. If $Y = \beta'\mathbf{X} + W$, then statements (a), (b) and (c) are trivially true by Lemma 1 and Eq. (3.2). Otherwise, statement (d) is also trivially true and statement (f) follows from Lemma 1, as $Y^* = \beta'\mathbf{X} + W$. If statement (e) is false, then $F_{Y|\mathbf{X}}(t|\mathbf{X}) = F_{Y^*|\mathbf{X}}(t|\mathbf{X})\ \forall\ t$. It follows that $f_{Y|\mathbf{X}}(t|\mathbf{X}) = f_{Y^*|\mathbf{X}}(t|\mathbf{X})\ \forall\ t$ and

$$f_{\mathbf{X},Y}(\mathbf{X}, t) = f_{Y|\mathbf{X}}(t|\mathbf{X}) f_{\mathbf{X}}(\mathbf{X}) = f_{Y^*|\mathbf{X}}(t|\mathbf{X}) f_{\mathbf{X}}(\mathbf{X}) = f_{\mathbf{X},Y^*}(\mathbf{X}, t),\ \forall\ t.$$

Notice that $f_{\mathbf{X},Y}$ and $f_{\mathbf{X},Y^*}$ are two density functions. The last equation contradicts the assumption that $(\mathbf{X}, Y)$ is not from the model $Y = \beta'\mathbf{X} + W$. Thus statement (e) holds. □

The next lemma is needed in the proofs of Lemmas 2 and 3. Its proof will be given later.

Lemma 4. Suppose that G is a cdf, $G_n(t) \to G(t)$ and $G_n(t-) \to G(t-)$ pointwisely, $G_n \geq 0$, and G_n is non-decreasing for $n \geq 1$. Then $\sup_{t\in\mathcal{R}} |G_n(t) - G(t)| = o(1)$.

Proof of Lemma 3. Denote $F_1(t) = E(F_o(t - \beta'\mathbf{X}))$. By the strong law of large numbers (SLLN), $\frac{1}{n}\sum_{i=1}^n F_o(t - \beta'\mathbf{X}_i) \overset{a.s.}{\to} F_1(t)$. Denote $\check{F}_n(t) = \frac{1}{n}\sum_{i=1}^n F_o(t - \hat{\beta}'\mathbf{X}_i)$. $\check{F}_n$ and F_1 are both cdf's. We shall prove the next three statements one by one:

$$\text{(a) } \check{F}_n(t) \overset{a.s.}{\to} F_1(t) \text{ for each } t;\ \text{(b) } F_1(t) \text{ is continuous; (c) } \sup_t |\check{F}_n(t) - F_1(t)| \overset{a.s.}{\to} 0. \quad (8.1)$$

$\forall\ \epsilon > 0$, $\exists\ N > 0$ such that $F_o(-N) + 1 - F_o(N) < \epsilon$ and $P(||\mathbf{X}|| > N) < \epsilon$.

Notice that F_o is uniformly continuous on the closed interval $[-2N, 2N]$. Thus

$\exists\ \delta \in (0, N)$ such that $|F_o(x) - F_o(y)| < \epsilon$ whenever $|x - y| < \delta$ and $x, y \in [-2N, 2N]$.

Let $\Omega_o = \{\hat{\beta} \to \beta, \frac{1}{n}\sum_{i=1}^n \mathbf{1}(||\mathbf{X}_i|| > N) \to P(||\mathbf{X}|| > N)\}$. By assumption and the SLLN, $P(\Omega_o) = 1$. Given $\omega \in \Omega_o$, abusing notation, write $\check{F}_n(t;\omega) = \check{F}_n(t)$ and $\mathbf{X}_i(\omega) = \mathbf{X}_i$, then $\exists$ n_ω such that $|(\hat{\beta} - \beta)'\mathbf{X}_i| < \delta$ whenever $n \geq n_\omega$ and $||\mathbf{X}_i|| \leq N$.

Let $n \geq n_\omega$, $u_i = t - \hat{\beta}'\mathbf{X}_i$, $v_i = t - \beta'\mathbf{X}_i$, $\mathcal{A}_{1i} = \{||\mathbf{X}_i|| \leq N$, and u_i or $v_i \in [-N, N]\}$, $\mathcal{A}_{2i} = \{||\mathbf{X}_i|| \leq N$, and u_i and $v_i \notin [-N, N]\}$, and $\mathcal{A}_{oi} = \{||\mathbf{X}_i|| > N\}$. Then

$$|F_o(t - \hat{\beta}'\mathbf{X}_i) - F_o(t - \beta'\mathbf{X}_i)|$$

$$\begin{cases} < \epsilon & \text{if } \mathbf{1}(\mathcal{A}_{1i}) = 1 \text{ (as } |u_i - v_i| \le \delta \text{ and } |u_i| \vee |v_i| \le 2N) \\ = |F_o(u_i) - F_o(v_i)| < \epsilon & \text{if } \mathbf{1}(\mathcal{A}_{2i}) = 1 \text{ and } u_i,\, v_i < -N \text{ (as } F_o(u_i),\, F_o(v_i) < \epsilon) \\ = |1 - F_o(u_i) - (1 - F_o(v_i))| < \epsilon & \text{if } \mathbf{1}(\mathcal{A}_{2i}) = 1 \text{ and } u_i,\, v_i > N \\ \le 1 & \text{otherwise.} \end{cases}$$

$$\begin{aligned} => \quad |\frac{1}{n}\sum_{i=1}^{n}(F_o(t - \hat{\beta}'\mathbf{X}_i) - F_o(t - \beta'\mathbf{X}_i))| \le & \frac{1}{n}\sum_{i=1}^{n}[\epsilon\mathbf{1}(\mathcal{A}_{1i} \cup \mathcal{A}_{1i}) + \mathbf{1}(A_{oi})] \\ < & \epsilon + \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(||X_i|| > N) \qquad (n \ge n_\omega) \\ \le & 2\epsilon \text{ if } n_\omega \to \infty. \end{aligned} \tag{8.2}$$

Since w is arbitrary in Ω_o, $P(\Omega_o) = 1$, and ϵ is arbitrary, statement (a) in (8.1) holds. That is, $P(\Omega_1) = 1$, where $\Omega_1 = \{\omega : \check{F}_n(t;\omega) \to F_1(t) \ \forall \ t\}$. If F_1 is continuous and $\omega \in \Omega_1$, then $\sup_t |\check{F}_n(t;\omega) - F_1(t)| \to 0$ by Lemma 4. Thus statement (c) in (8.1) holds, and it further yields Eq. (3.8). It remains to prove (b) in (8.1). A heuristic arguement is as follows.

$$|F_1(t) - F_1(t-)| = \int F_o(t - \beta'\mathbf{x}) - F_o((t - \beta'\mathbf{x})-)dF_{\mathbf{X}}(\mathbf{x}) = \int 0 dF_{\mathbf{X}}(\mathbf{x}) = 0 \ \ \forall \ t,$$

where $F_1(t-) = \lim_{x \uparrow t} F_1(x)$. A rigorous proof is similar to the proof in (8.2). For a given t, if $|t - y| \le \delta$, then $|u - v| \le \delta$, where $u = t - \beta'\mathbf{X}$ and $v = y - \beta'\mathbf{X}$. Let $A_1 = \{||\mathbf{X}|| \le N, |u| \wedge |u| \le N\}$, $A_2 = \{||\mathbf{X}|| \le N, |u| \wedge |v| > N\}$, and $A_o = \{||\mathbf{X}|| > N\}$.

$$\begin{aligned} |F_1(t) - F_1(y)| = & \int F_o(t - \beta'\mathbf{x}) - F_o(y - \beta'\mathbf{x})dF_{\mathbf{X}}(\mathbf{x}) \\ \le & \int_{A_1 \cup A_2 \cup A_o} |F_o(t - \beta'\mathbf{x}) - F_o(y - \beta'\mathbf{x})| dF_{\mathbf{X}}(\mathbf{x}) \\ \le & \int_{A_1} \epsilon dF_{\mathbf{X}}(\mathbf{x}) + \int_{A_2} \epsilon dF_{\mathbf{X}}(\mathbf{x}) + \int_{A_o} dF_{\mathbf{X}}(\mathbf{x}) \\ \le & 2\epsilon. \ \ \square \end{aligned}$$

Proof of Lemma 2. Define $a_n(t,s) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(Y_i \le t, ||\mathbf{X}_i|| \le s)$,
$b_n(s) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(||\mathbf{X}_i|| \le s)$, $a(t,s) = P(Y \le t, ||\mathbf{X}|| \le s)$, $b(s) = P(||\mathbf{X}|| \le s)$.
So $\hat{F}_o(t) = \frac{a_n(t,\delta_n)}{b_n(\delta_n)}$ and $F_o(t) = \lim_{n\to\infty}\frac{a(t,\delta_n)}{b(\delta_n)}$.

By Lemma 4, we have

$$\lim_{n\to\infty}\sup_{t\in\mathcal{R}}|\frac{a(t,\delta_n)}{b(\delta_n)} - F_o(t)| = 0. \tag{8.3}$$

By the empirical process results (see Kosorok (2009)),

$$\sup_{t\in\mathcal{R},s\in\mathcal{R}} |a_n(t,s) - a(t,s)| = O_p(1/\sqrt{n})$$

Notice by the fact $a_n(t,\delta_n) \leq b_n(\delta_n)$, we have

$$|\frac{a_n(t,\delta_n)}{b_n(\delta_n)} - \frac{a(t,\delta_n)}{b(\delta_n)}| \leq |\frac{b_n(\delta_n) - b(\delta_n)}{b(\delta_n)}| + |\frac{a_n(t,\delta_n) - a(t,\delta_n)}{b(\delta_n)}|.$$

We can bound the first term by Chebyshev inequality, $P(|\frac{b_n(\delta_n)-b(\delta_n)}{b(\delta_n)}| > \eta) \leq \frac{1}{\eta^2 nb(\delta_n)}$. The supremum of the second term is of order, $\sup_{t\in cR} |\frac{a_n(t,\delta_n)-a(t,\delta_n)}{b(\delta_n)}| = O_p(\frac{1}{\sqrt{n}b(\delta_n)})$. So we have

$$\sup_{t\in\mathcal{R}} |\hat{F}_o(t) - \frac{a(t,\delta_n)}{b(\delta_n)}| = o_p(1), \tag{8.4}$$

provided $\sqrt{n}b(\delta_n) \to \infty$. Eq. (8.3) and Eq. (8.4) yield $\sup_{t\in\mathcal{R}} |\hat{F}_o(t) - F_o(t)| = o_p(1)$. □

Proof of Theorem 2. By Lemma 2, Eq. (3.9) follows from

$$\sup_{t\in\mathcal{R}} |\frac{1}{n}\sum_{i=1}^{n}[\hat{F}_o(t-\beta\mathbf{X}_i) - F_o(t-\beta\mathbf{X}_i)]| \leq \sup_{t\in\mathcal{R}} |\hat{F}_o(t) - F_o(t)| = o_p(1).$$

Moreover, by Lemma 2, we have

$$\sup_{t\in\mathcal{R}} |\frac{1}{n}\sum_{i=1}^{n}[\hat{F}_o(t-\hat{\beta}\mathbf{X}_i) - F_o(t-\hat{\beta}\mathbf{X}_i)]| \leq \sup_{t\in\mathcal{R}} |\hat{F}_o(t) - F_o(t)| = o_p(1).$$

Also by Eq. (3.8),

$$\sup_{t\in\mathcal{R}} |\frac{1}{n}\sum_{i=1}^{n}[F_o(t-\hat{\beta}\mathbf{X}_i) - F_o(t-\beta\mathbf{X}_i)]| = o_p(1).$$

The Central Limit Theorem and the previous two inequalities yield

$$\sup_{t\in\mathcal{R}} |\frac{1}{n}\sum_{i=1}^{n}[\hat{F}_o(t-\hat{\beta}\mathbf{X}_i) - E(F_o(t-\beta'\mathbf{X}))| = o_p(1).$$

Notice that a sufficient condition of Eq. (3.8) is $F_{\mathbf{X},Y} \in \Theta_{lse}$.

If Y is discontinuous, and $F_{\mathbf{X},Y} \in \Theta_0$, then the SMLE $\hat{\beta}$ satisfies that $P(\hat{\beta} = \beta) \to 1$. Hence, Eq. (3.10) follows directly from Eq. (3.8). □

Proof of Lemma 4. For each $\eta > 0$, define $D = \{t \in \mathcal{R} : G_n(t) - G(t-) > \eta\}$. Since G is non-decreasing and bounded, there are finite elements in D. Consider the following separation on the whole real line $t_0 < t_1 < t_2 < ... < t_M$, s.t

(a) $G(t_0) < \eta$.

(b) $1 - G(t_M) < \eta$.

(c) Every element in D should be one of t_k's.

(d) $G(t_k-) - G(t_{k-1}) < \eta$ for all $k = 0, 1, ..., M$.

So $\sup_{t\in\mathcal{R}} |G_n(t) - G(t)| \leq \max\{T_1, T_2, T_3, T_4\}$, where

$T_1 = \max_{0\leq k\leq M-1} \sup_{t\in(t_k,t_{k+1})} |G_n(t) - G(t)|$,

$T_2 = \max_{0\leq k\leq M} |G_n(t_k) - G_(t_k)|$,

$T_3 = \sup_{t\leq t_0} |G_n(t) - G_(t)|$,

$T_4 = \sup_{t>t_M} |G_n(t) - G_(t)|$.

By monotonicity,

$\sup_{t\in(t_k,t_{k+1})} |G_n(t) - G(t)|$

$\leq \max\{|G_n(t_{k+1}-) - G(t_k)|, |G(t_{k+1}-) - G_n(t_k)|$

$\leq \max\{|G_n(t_{k+1}-) - G(t_{k+1}-)| + |G(t_{k+1}-) - G(t_k)|, |G(t_{k+1}-) - G(t_k)| + |G(t_k) - G_n(t_k)|\}$

$\leq 2\eta$, provided n is large enough.

Hence $T_1 \leq 2\eta$, when n is large enough. By the convergence conditions, $T_2 \leq \eta$, when n is large enough. Notice T_3 can be bounded by

$\max\{G_n(t_0), G(t_0)\} \leq \max\{|G_n(t_0) - G(t_0)| + G(t_0)\} \leq 2\eta$, provided n is large enough.

Similar argument yields $T_4 \leq 2\eta$ when n is large. Since η is arbitrary, the lemma holds. □

Proofs in Examples 4.4 and 4.5. The SMLE and the MSMLE are proposed by Yu and Wang (2002). The SMLE and the MSMLE of β under the model $Y = \beta'\mathbf{X} + W$ maximize $\mathcal{L} = \prod_{i=1}^n f(Y_i - a'\mathbf{X}_i)$ and $\mathcal{L}_m = \prod_{i=1}^n \frac{F(Y_i - a'\mathbf{X}_i + h) - F(Y_i - a'\mathbf{X}_i - h)}{2h}$, respectively, over all possible cdfs F and all possible values of $a \in \mathcal{R}^p$, where $f(t) = F(t) - F(t-)$, $h = o(1/n)$ is predetermined (*e.g.*, $h = n^{-1/5}$), as W_i's are i.i.d., and $W_i = Y_i - \beta'\mathbf{X}_i$ by the assumption. If the solution of the SMLE is not uniquely determined, Yu and Wong suggested to choose the one which is closest to the LSE. For illustration, we shall show that the SMLE of β is 1 and $b(e-1)$ for the data from Models 1 and 2 in these two examples, respectively.

For each a, $\mathcal{L}$ is maximized by $f(t) = \hat{f}_a(t)$, the density function of the edf based on $Y_i - a'\mathbf{X}_i$, $i \in \{1,...,n\}$. If the data are from Model 1, then $Y_i = X_i + W_i$, $P(W_i = 0) = 0.5$ and $P(W_i \neq W_j | W_i W_j \neq 0) = 1$. Then $\mathcal{L} = \prod_{i=1}^{n_0} f(X_i(1-a)) \prod_{i>n_0} f(W_i + (1-a)X_i)$, where $n_0 = \sum_{i=1}^{n} \mathbf{1}(W_i = 0)$. If n is large, then $n_0 \approx n/2$ and $\mathcal{L}$ achieves its maximum $(\frac{n_0}{n})^{n_0}(\frac{1}{n})^{n-n_0}$ at $a=1$ w.p.1. That is, $\mathcal{L}$ is maximized by the SMLE $\hat{\beta}=1$.

If the data are from Model 2, *i.e.*, $Y = be^X + W$, then (X_i, Y_i) takes 6 types of values: $(0,b)$, $(1,be)$, $(0,b+w)$, $(1,be+w)$, (x,be^x), and (x,be^x+w), where $x,w \notin \{0,1\}$. $Y_i - aX_i \in \{b, be-a, b+W_i, be+W_j-a, be^{X_k} - aX_k, be^{X_h} + W_k - aX_h\}$, where X_i, X_j, X_k and X_h are distinct observations ($\notin \{0,1\}$). If $a = b(e-1)$, then the first 2 elements equal b, and $P(Y_i - b(e-1)X_i = b) \approx 0.4 + 0.12$, whereas the last four elements, $b + W_i$, $W_j + b$, $b\exp(X_k) - b(e-1)X_k$, $b\exp(X_h) + W_h - b(e-1)X_h$ are distinct for a reason similar to Model 1. Thus one can show that the SMLE of β is $b(e-1)$ if n is large.

Notice that $\mathcal{L}_m = \prod_{i=1}^{n}(F(Y_i - aX_i + h) - F(Y_i - aX_i - h)) \times (2h)^{-n}$ and if h is very small, then $F(Y_i - aX_i + h) - F(Y_i - aX_i - h)) = f_a(Y_i - aX_i)$ Thus one can show also that the MSMLE of β is the same as the SMLE if n is large in these two examples. For a more rigorous proof, see Yu and Wong (2002).

8. Appendix II. Proof of Theorem 3. The proof of Theorem 3 can be found in Liu and Yu (2017) (see http://www.math.binghamton.edu/ftp/disc.pdf).

9. References.

Carroll, R.J. Fan, J. Gijbels, I. and Wand, M.P. (1997). Generalized Partially Linear Single-Index Models. *Journal of American Statistical Society.* 92 477-489.

Cox, D.R. and Oaks, D. (1984). Analysis of Survival Data. *Chapman and Hall.* N.Y.

Draper, N. and Smith, H. (1966). Applied Regression Analysis. (2nd ed.) *John Wiley & Sons.* N.Y.

Ezekiel, M. (1930). Methods of Correlation Analysis. *Wiley.* N.Y.

Faraway, J. (2016). Functions and Data sets for Books by Julian Faraway. *Package "faraway".* 30-30. <jjf23@bath.ac.uk>.

Gonzalez-Manteiga, W. and Crujeiras, R. M. (2013). An updated review of Goodness-of-Fit tests for regression models *Test* 22 (3) 361-411

Hardle, W. (1994). Applied nonparametric regression. Cambridge Unv. Press.

Harvey, A.C. (1976). Estimating regression models with multiplicative heteroscedasticity. *Econonometrica.* 44, 461 - 465.

Hastie, T. J. and Tibshirani, R. J. (1990). Generalized Additive Models. *Chapman & Hall/CRC.* N.Y.

Kosorok, M.R. (2009). Introduction to Empirical Processes and Semiparametric Inference. Spring, N.Y.

Nelder, J. and Wedderburn, R. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A.* 135, 370 - 384.

Ramsey, J. B. (1969). Tests for Specification Errors in Classical Linear Least Squares Regression Analysis. *Journal of the Royal Statistical Society* Series B. 31 (2), 350 - 37.

Sen, A.; and Sen, B. (2014). Testing independence and goodness-of-fit in linear models *Biometrika* 101(4) 927-942

Stute, W. (1997). Nonparametric model checks for regression. *Annals of Statistics.* 25 (2) 613-641.

Utts, J.M. (1982). The Rainbow Test for Lack of Fit in Regression. *Communications in Statistics - Theory and Methods.* 11, 1801 - 1815.

Wood, S.N. (2006). Generalized additive models, an introduction with R. *Chapman &*

Hall/CRC Press. N.Y.

Yu, Q.Q. and Wong, G.Y.C. (2002). Asymptotic properties of a modified semi-parametric MLE in linear regression analysis with right-censored data. *Acta Mathematica Sinica*, 18 405-416.

Yu, Q.Q. and Wong, G.Y.C. (2003). Asymptotic properties of the generalized semi-parametric MLE in linear regression. *Statistica Sinica*, 13, 311-326.

Yu, Q.Q. and Wong, G.Y.C. (2004). Modified semi-parametric MLE in linear regression analysis with complete data or right-censored data. *Technometrics*, 47 34-42.

Yu, Q.Q. and Liu, R.Q. (2017). Appendix II of the paper with the title "Testing independence and goodness -of-fit in linear regression models allowing non-existence of the mean of the response variable". http://people.math.binghamton.edu/qyu/ftp/disc.pdf