Summary on the simulation studies.
There are many regression models:
the linear regression models,
the logistic regression models,
the generalized linear (regression) models,
the generalized additive models, etc..
Given a data set, one needs to check which model fits the data. This is to test
H_0: the data fits a given model, *e.g.*, $E(Y|X) = \beta' X$, v.s. H_1: H_0 is false.
If H_0 and H_1 are not properly designed, then
the previous 3 model checking tests can be misleading as in simulations 3 and 4 of Ex. 6.
The existing model checking tests are the tests of
H_0^t: $\xi(\cdot) = 0$, v.s. H_1^t: $\xi(\cdot) \neq 0$, where $\xi(\mathbf{X}) = E(Y|\mathbf{X}) - \beta'\mathbf{X}$ has a certain form.
For instance, $\xi = \theta g(\mathbf{X})$ in the 3 aforementioned model checking tests. 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. For instance, in Example 6,

$$\Theta_p = \{F_{\mathbf{X},Y} : \ Y = \alpha + \beta X + \theta \sin X + \epsilon, \ \epsilon \sim N(0, \sigma^2), \ X \perp \epsilon\}.$$

Thus $\Theta_p \neq \Theta$, the family of all cdfs $F_{\mathbf{X},Y}$. If $F_{\mathbf{X},Y} \notin \Theta_p$, these tests are <u>*invalid*</u> in the sense that the (asymptotic) distributions specified for these tests are false.
In simulation 1, H_0 is true and the model assumptions holds,
the test can either reject H_0 or do not reject. But $P(H_1|H_0) = 0.05$.
In simulation 2, H_1 is true, and the assumptions for the t-test hold.
The test rejects H_0 with probability $\to 1$ as $n \to \infty$ ($P(H_0|H_1) \to 0$, a consistent test).
In simulations 3 and 4, both H_0 and H_1 are false,
the test can reject H_0 with probability 0 or 0.96.
In Example 7, both H_0 and H_1 are false, as $E(W|X)$ does not exist.
An estimate of $P(H_0|H_1)$ is ≈ 0.8.
Remark. Type I error, denoted by $H_1|H_0$, implies that H_0 is true. In Simulation 1 of Ex.6, it is true that $P(H_1|H_0) = 0.05$.
Type II error, denoted by $H_0|H_1$, implies that H_1 is true. In Simulation 2 of Ex.6, it is true that $P(H_1|H_0) \approx 0.33$.
In Simulations 3 and 4 of Ex.6 and in Example 7, neither H_0 nor H_1 is true. Thus neither $P(H_1|H_0)$ nor $P(H_0|H_1)$ is a proper terms.

Appendix. The marginal distribution (MD) approach. We shall introduce a new approach for model checking, which can do better than the previous tests most of the time.
A.1. Preliminary. We assume that
$(\mathbf{X}_1, Y_1)$, ..., $(\mathbf{X}_n, Y_n)$ are i.i.d. observations from $F_{\mathbf{X},Y}$, with density function $f_{\mathbf{X},Y}$, where $\mathbf{X}$ is a p-dimensional random vector and Y is a response variable.
Let $F_{Y|\mathbf{X}}$ be the conditional cdf with density function $f_{Y|\mathbf{X}}$. Denote $F_o = F_{Y|\mathbf{X}}(\cdot|\mathbf{0})$, which is called the baseline cdf of $F_{Y|\mathbf{X}}$.
A common regression model is the linear regression (LR) model,

$$Y = \beta'\mathbf{X} + W, \text{ where } (F_W = F_o), \ \beta \in \mathcal{R}^p \tag{1.1}$$

(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 (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, there are several consistent estimators of β if $F_{\mathbf{X},Y} \in \Theta_{lse}$, where

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

and $\Sigma_{\mathbf{X}}$ is the $p \times p$ covariance matrix of $\mathbf{X}$. They include
the semi-parametric MLE (SMLE) (if F_o is discontinuous),
the modified SMLE (MSMLE) (see Yu and Wong (2002, 2003 and 2004)),
the least squares estimator (LSE) and
the quantile or median regression estimator.
Yu and Wong (2002) show that
the MSMLE is still consistent if $E(\ln f_W(W))$ exists, and
the SMLE and the MSMLE $\tilde{\beta}$ satisfy $P(\tilde{\beta} \neq \beta \text{ infinitely often}) = 0$ if the cdf F_W is discontinuous.

However, the LSE is inconsistent if $E(|Y||\mathbf{X}) = \infty$.

Given $F_{\mathbf{X},Y} \in \Theta$, the family of all joint cdf of $(\mathbf{X}, Y)$, $F_o = F_{Y|\mathbf{X}}(\cdot|\mathbf{0})$ is well defined, even if $(\mathbf{X}, Y)$ does not satisfy the linear regression model in H_0: $Y = \beta'\mathbf{X} + W$, where W is a random variable that its mean may not exist. We first consider the test of H_0. Let

$$\Theta_0 = \{F_{\mathbf{X},Y} : \ Y = \beta'\mathbf{X} + W, \text{ where } W \perp \mathbf{X}, \ \beta \text{ and } F_W \text{ are unknown}\} \tag{2.1}$$

($F_W = F_o$). Then H_0: $F_{\mathbf{X},Y} \in \Theta_0$. The next lemma characterizes various LR model and motivating the MD approach for the LR model.

Lemma 1. *$F_{Y|\mathbf{X}}$ is a function of (F_o, β), $F_Y(t) = E(F_{Y|\mathbf{X}}(t|\mathbf{X}))$. If $F_{\mathbf{X},Y} \in \Theta_0$, then $F_{Y|\mathbf{X}}(t|x) = F_o(t - \beta'x)$.*