Specification testing for binary choice model via maximum score

Yuta Ota and Taisuke Otsu

January 9, 2026

KEO Discussion Paper No. 194

KEO Discussion Paper No. 194

SPECIFICATION TESTING FOR BINARY CHOICE MODEL VIA MAXIMUM SCORE

YUTA OTA AND TAISUKE OTSU

Abstract. This paper proposes a Hausman-type statistic to the test specification of a parametric binary choice model by comparing the maximum likelihood estimator and the maximum score estimator. Although the convergence rates are different, it is still meaningful to compare these estimators to detect misspecification of parametric models. A simulation study illustrates that the proposed test offers better size properties than the conventional information matrix test, and exhibits reasonable power against common forms of misspecification, such as heavy-tailed distributions and heteroskedasticity.

1. Hausman type test

This paper is concerned with specification testing of parametric binary choice models, such as probit and logit, which are commonly applied in empirical research. Suppose we observe a random sample $\{Y_i, X_i\}_{i=1}^n$ of (Y, X), where $Y \in \{0, 1\}$ is a binary dependent variable and $X \in \mathbb{R}^k$ is a vector of covariates. The binary choice model of interest is

$$Y = \mathbb{I}\{1 + X'\beta + U \geq 0\}, \tag{1}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function, β is a k-dimensional vector of parameters, and U is an unobservable error term. We normalize the intercept to be 1, but other normalizations can be also applied. To estimate the parameters β, researchers commonly employ a parametric model on the error term, such as the homoskedastic probit ($U|X \sim N(0, \eta)$) and homoskedastic logit ($U|X \sim \mathit{Logistic}(0, \eta)$), and implement the method of maximum likelihood. It is known that although the maximum likelihood estimator is consistent and asymptotically optimal when the parametric distributional form of $U|X$ is correctly specified, it is generally inconsistent for β when the parametric model is misspecified (White, 1982). For example, the probit maximum likelihood estimator using $U|X \sim N(0, \eta)$ is generally inconsistent under not only non-normal errors but also heteroskedastic errors. Since the probit and logit methods are widely applied in empirical research, it is of substantial interest to check the validity of specified parametric models.

In this paper, we maintain the conditionally zero median assumption on the error term:

$$\mathrm{Med}(U|X) = 0, \tag{2}$$

and consider specification testing of a parametric model

$$\mathbb{H}_0 : U|X \sim F_\eta(\cdot|X),$$

where $F_\eta(\cdot|X)$ is the conditional distribution satisfying (2) with parameters η. Obviously the probit or logit satisfies (2). The alternative hypothesis is $\mathbb{H}_1 : \mathbb{H}_0$ is not true (but (2) holds true). Under certain regularity conditions, the maximum likelihood estimator $\hat{\beta}_{\mathrm{ML}}$ satisfies

$$\begin{aligned} n^{1/2}(\hat{\beta}_{\mathrm{ML}} - \beta) &\overset{d}{\to} N(0, V_{\mathrm{ML}}) \quad \text{under } \mathbb{H}_0, \\ n^{1/2}(\hat{\beta}_{\mathrm{ML}} - \beta^*) &\overset{d}{\to} N(0, V_{\mathrm{ML}}^*) \quad \text{under } \mathbb{H}_1, \end{aligned} \tag{3}$$

where β^* is the probability limit of $\hat{\beta}_{\mathrm{ML}}$ under $\mathbb{H}_1$ (called the pseudo-true value), and V_{ML} and V_{ML}^* are asymptotic variances. Under $\mathbb{H}_1$, β^* is generally different from β in (1) (see, White, 1982).

The main idea of this paper is to compare the maximum likelihood estimator $\hat{\beta}_{\mathrm{ML}}$ against a semiparametric estimator that is consistent for β under the median restriction in (2) without any parametric restriction on the distribution form of the error term. To this end, we employ the maximum score estimator (Manski, 1975), that is

$$\hat{\beta}_{\mathrm{MS}} = \arg\max_{\beta} \sum_{i=1}^{n} \left[Y_i \mathbb{I}\{1 + X_i'\beta \geq 0\} + (1 - Y_i)\mathbb{I}\{1 + X_i'\beta < 0\} \right].$$

Under (2) and mild regularity conditions, $\hat{\beta}_{\mathrm{MS}}$ is consistent for β and exhibits the so-called cube root asymptotics (Kim and Pollard, 1990):

$$n^{1/3}(\hat{\beta}_{\mathrm{MS}} - \beta) \overset{d}{\to} \mathcal{Z}, \tag{4}$$

where $\mathcal{Z}$ is the minimizer of a Gaussian process; see Kim and Pollard (1990, Theorem 1.1) for the definition. It should be note that $\hat{\beta}_{\mathrm{MS}}$ is consistent for β regardless of $\mathbb{H}_0$ or $\mathbb{H}_1$, but converges slower than $\hat{\beta}_{\mathrm{ML}}$. We argue that although the convergence rates are different, it is still meaningful to compare these estimators to detect misspecification of parametric models. In particular, we propose the following Hausman type statistic:

$$T = ||n^{1/3}(\hat{\beta}_{\mathrm{MS}} - \hat{\beta}_{\mathrm{ML}})||.$$

One attractive feature of this statistic is that it does not require any tuning constants, such as bandwidth or series length. Note that under $\mathbb{H}_0$, (3) and (4) imply

$$T = ||n^{1/3}(\hat{\beta}_{\mathrm{MS}} - \beta) - n^{-1/6}\{n^{1/2}(\hat{\beta}_{\mathrm{ML}} - \beta)\}|| \overset{d}{\to} ||\mathcal{Z}||.$$

On the other hand, under $\mathbb{H}_1$, it can be written as

$$T = ||n^{1/3}(\hat{\beta}_{\mathrm{MS}} - \beta) + n^{1/3}(\beta - \beta^*) - n^{-1/6}\{n^{1/2}(\hat{\beta}_{\mathrm{ML}} - \beta)\}||,$$

and thus T diverges as far as $\beta^* \neq \beta$. Therefore, the main result of this paper is summarized as follows.

Proposition. *Consider the setup of this section. Suppose (3) and (4) hold true. Let* $c_{1-\alpha}$ *be the* $(1-\alpha)$*-th quantile of* $||\mathcal{Z}||$. *Then* $\mathbb{P}\{T \leq c_{1-\alpha}\} \to 1-\alpha$ *under* $\mathbb{H}_0$, *and* T *diverges to infinity under* $\mathbb{H}_1$ *as far as* $\beta^* \neq \beta$.

Since $\mathcal{Z}$ is the limiting distribution of the maximum score estimator, several methods are available to compute the critical value $c_{1-\alpha}$, such as the subsampling (by Delgado, Rodríguez-Poo and Wolf, 2001) and bootstrap (Cattaneo, Jansson and Nagasawa, 2020). In our simulation study, we illustrate by the conventional subsampling method to estimate $c_{1-\alpha}$ and briefly mention the results for the bootstrap method.

Remark. [Comparison with information matrix test] A standard specification test in the maximum likelihood framework is the information matrix test (IMT) proposed by White (1982), which effectively tests the null hypothesis of the information matrix equality $\mathbb{H}_0^{\mathrm{IM}} : \mathbb{E}\left[\frac{\partial \ell(Y|X;\theta)}{\partial\theta}\frac{\partial \ell(Y|X;\theta)}{\partial\theta'}\right] = -\mathbb{E}\left[\frac{\partial^2 \ell(Y|X;\theta)}{\partial\theta\partial\theta'}\right]$ for the conditional log-likelihood function $\ell(Y|X;\theta)$ of the parameters $\theta = (\beta', \eta')'$. In terms of local power, the IMT dominates our test, since the IMT can have non-trivial power against the Pitman-type local alternatives at the rate $n^{-1/2}$ (instead of $n^{-1/3}$); see Newey (1985). However, despite its theoretical appeal, the IMT has seen limited application in practice (Golden et al. 2012). White's original (full) IMT is analytically and computationally burdensome due to the requirement for third derivatives. Simulation evidence further suggests that both the full IMT and the outer product of gradients (OPG)-based implementation that avoids explicit third derivatives, following Chesher (1983) and Lancaster (1984), can exhibit erratic finite-sample behavior, and size distortions tend to worsen as the degrees of freedom increase (Taylor 1987; Orme 1990; Davidson and MacKinnon 1992; Stomberg and White 2000). These concerns have motivated continued methodological developments aimed at improving the practical performance of IMT-based tests (Golden et al. 2016).

In our simulation study focusing on the probit model presented below, we also find severe size distortions even in the case of $\dim(\theta) = 3$ and low size-adjusted power (due to the adjustment required for the substantial over-rejection), whereas our proposed method effectively avoids such over-size. Therefore, at least, our test can serve as a useful complement to the IMT.

2. Simulation

We conduct a simulation study to evaluate the finite sample performance of our proposed test. We focus on the probit model as a representative case and assess its empirical size and power against several common forms of misspecification. The data generating process for our simulation is based on the latent variable model $Y_i = \mathbb{I}\{1 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + U_i \geq 0\}$ with the true parameters $(\beta_1, \beta_2) = (1, 1)$ and the sample sizes $n \in \{300, 500, 1000, 2000, 5000\}$. For $k = 1, 2$, the covariate $X_{k,i}$ is independently generated from one of the following distributions: the symmetric $N(0,1)$ or $N(0.5,1)$, or the asymmetric $\mathit{Gumbel}(0,1)$ or $\mathit{Gumbel}(0.5,1)$. The cases with mean 0.5 create settings with greater data imbalance where the proportion of $Y_i = 1$ is higher. We consider four specifications for the error term U_i: (i) the null distribution with $U_i \sim N(0,1)$ (correct specification), (ii) an alternative with $U_i \sim Logistic(0,1)$ (distributional misspecification), (iii) an alternative with $U_i \sim Cauchy(0,1)$ (heavy-tailed misspecification), and (iv) an alternative with $U_i = \epsilon_i\sqrt{1+(X_{1,i}+X_{2,i})^2}$ and $\epsilon_i \sim N(0,1)$ (heteroskedasticity). The nominal significance level is $\alpha = 0.05$. For each case, the results are based on $1,000$ Monte Carlo replications.

First, we evaluate the size under the correct specification ($U_i \sim N(0,1)$). To implement the maximum score estimator, we employ a grid search. The critical values are estimated via subsampling with a block size set to 30% of the original sample size, following Delgado, Rodríguez-Poo and Wolf (2001). We also confirmed that the results are robust to alternative block sizes of 20% and 40%. We compare our test with the IMT using a commonly used OPG-based implementation (Chesher 1983; Lancaster 1984).

Table 1 reports the size properties of both tests The results show that the IMT is noticeably oversized. Although this distortion is mitigated as the sample size increases, the size remains above the 5% nominal level even at $n = 5000$. Furthermore, this tendency toward over-rejection becomes more pronounced with greater data imbalance. [1] In contrast, our proposed method consistently exhibits an empirical size below 0.05 across all settings, suggesting that our method is more suitable for practical applications.

n	IMT (OPG-based)				Maximum Score			
	$N(0,1)$	$N(0.5,1)$	$G(0,1)$	$G(0.5,1)$	$N(0,1)$	$N(0.5,1)$	$G(0,1)$	$G(0.5,1)$
300	0.531	0.669	0.644	0.864	0.016	0.011	0.009	0.007
500	0.396	0.528	0.556	0.778	0.003	0.005	0.005	0.007
1000	0.262	0.364	0.410	0.564	0.008	0.002	0.003	0.000
2000	0.179	0.241	0.280	0.422	0.005	0.005	0.003	0.003
5000	0.089	0.129	0.170	0.227	0.008	0.007	0.006	0.008

Note: G denotes the Gumbel distribution.

Table 1. Empirical size under correct specification ($U_i \sim N(0,1)$).

To evaluate the power properties meaningfully despite the size distortions of the IMT, we present the size-adjusted power in Figure 1.

For the Logistic alternative, our method exhibits a power of approximately 0.2 across all settings, whereas the IMT shows negligible power, close to zero. While a power of around 0.2 is not particularly high, it represents a reasonable performance given the close distributional similarity between the Logistic and Normal distributions.

Regarding the Cauchy alternative, our method functions effectively even with moderate sample sizes; for instance, it achieves a power of 0.50 for $X \sim Gumbel(0,1)$ at $n = 1000$. Except for the case where $X \sim N(0,1)$, the power increases with sample size, exceeding 0.7 at $n = 5000$. Interestingly, we observe a tendency where the power increases as data imbalance becomes more severe. This

[1] The degree of data imbalance, measured by the proportion of $Y_i = 1$ (average of 1,000 replications), is essentially unchanged across sample sizes. The proportions are 0.72 for $X \sim N(0,1)$, 0.88 for $X \sim N(0.5,1)$, 0.86 for $X \sim G(0,1)$, and 0.95 for $X \sim G(0.5,1)$.

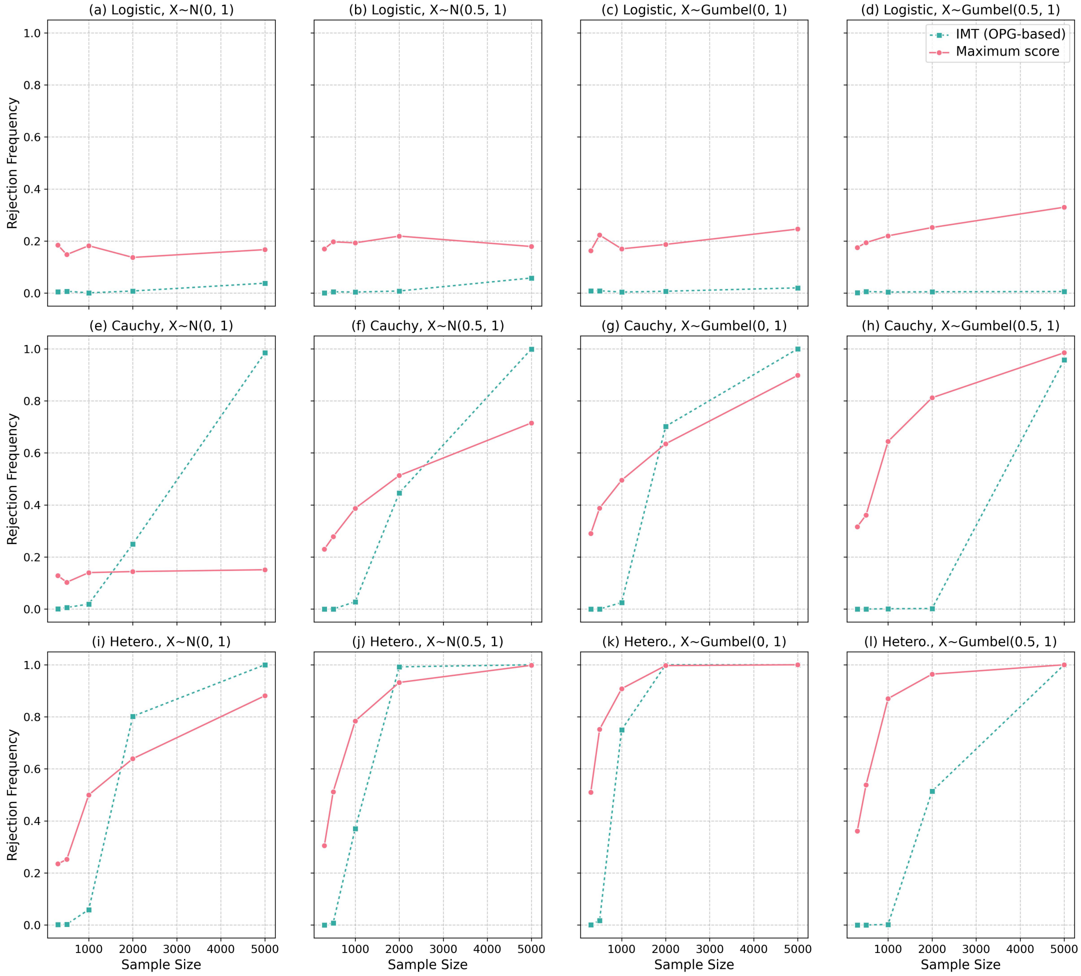

FIGURE 1. Empirical size-adjusted power under alternatives. The figure shows rejection frequencies against sample size (on a logarithmic scale) for each combination of the alternative and the distribution of X.

could be attributed to the fact that the maximum likelihood estimates are more likely to deviate from the true values under high imbalance, thereby facilitating detection. On the other hand, the IMT has little detection power at $n \leq 1000$. It does not perform well until the sample size becomes sufficiently large ($n \geq 2000$), occasionally surpassing our method. Unlike our approach, however, the IMT struggles with high data imbalance.[2] In the case of $X \sim \mathit{Gumbel}(0.5, 1)$, the IMT fails to perform adequately unless $n = 5000$, and its power remains consistently lower than that of our method across all sample sizes.

[2] The degree of data imbalance under the alternative specifications is measured by the average proportion of $Y_i = 1$ (averaged over 1,000 replications). The values for $X \sim N(0,1)$, $X \sim N(0.5,1)$, $X \sim \mathit{Gumbel}(0,1)$, and $X \sim \mathit{Gumbel}(0.5,1)$ are, respectively: 0.67–0.68, 0.82, 0.81, and 0.90 for Logistic; 0.67, 0.79, 0.78, and 0.87 for Cauchy; and 0.72, 0.83, 0.82, and 0.88 for the heteroskedasticity case. In the Logistic case, the 0.67–0.68 range reflects minor sampling fluctuations across n.

Similar trends are observed for the heteroskedasticity alternative. Our method maintains relatively high power even at small sample sizes ($n \leq 1000$); for example, it yields a power of 0.75 for $X \sim N(0.5, 1)$ at $n = 500$. In contrast, the IMT often shows power below 0.05 in these settings. As the sample size increases ($n \geq 2000$), the IMT's power improves, in some instances outperforming our method.

In addition to the results above, we also computed the power based on subsampling using the same 30% block size. Although the power is lower compared to the size-adjusted power, and the detection of the subtle Logistic alternative becomes negligible, the test still exhibits respectable power against other distinct alternatives. For example, for sample sizes $n = (300, 500, 1000, 2000, 5000)$, the power values are $(0.23, 0.23, 0.25, 0.41, 0.78)$ in the Cauchy case with $X \sim Gumbel(0.5, 1)$, and $(0.11, 0.18, 0.41, 0.71, 0.90)$ in the heteroskedasticity case with $X \sim N(0.5, 1)$. Given that these power levels are achieved while maintaining the size control shown in Table 1, the proposed method is considered sufficiently useful for practical applications.

Overall, the simulation results demonstrate that our proposed test is effective, even though it tends to be conservative with an empirical size often falling below 0.05. It shows particularly high power against alternatives often discussed in economics, such as heavy-tailed distributions and heteroskedasticity. When compared to the IMT (OPG-based) in terms of size-adjusted power, our method proves to be advantageous especially when the sample size is not large or the data imbalance is high. These characteristics complement the weaknesses of the IMT, which performs poorly under these conditions.[3]

References

[1] Chesher, A. (1983) The information matrix test: Simplified calculation via a score test interpretation, *Economics Letters*, 13, 45-48.

[2] Cattaneo, M. D., Jansson, M. and K. Nagasawa (2020) Bootstrap-based inference for cube root asymptotics, *Econometrica*, 88, 2203-2219.

[3] Davidson, R. and J. G. MacKinnon (1992) A new form of the information matrix test, *Econometrica*, 60, 145-157.

[4] Delgado, M. A., Rodríguez-Poo, J. M. and M. Wolf (2001) Subsampling inference in cube root asymptotics with an application to Manski's maximum score estimator, *Economics Letters*, 73, 241-250.

[5] Golden, R. M., Henley, S. S., White, H. and T. M. Kashner (2012) New directions in information matrix testing: eigenspectrum tests, *In Recent advances and future directions in causality, prediction, and specification analysis: Essays in honor of Halbert L. White Jr*, 145-177, New York, NY: Springer New York.

[6] Golden, R. M., Henley, S. S., White, H. and T. M. Kashner (2016) Generalized information matrix tests for detecting model misspecification, *Econometrics*, 4, 46.

[7] Kim, J. and D. Pollard (1990) Cube root asymptotics, *Annals of Statistics*, 18, 191-219.

[8] Lancaster, T. (1984) The covariance matrix of the information matrix test, *Econometrica*, 52, 1051-1054.

[3] Although the subsampling method proves to be an effective approach for practical applications, more accurate approximation methods such as the bootstrap-based method by Cattaneo, Jansson and Nagasawa (2020) can also be employed. In our preliminary simulation study, we find that their bootstrap approximation exhibits better size properties than subsampling, but becomes less powerful for some settings, such as the Cauchy alternative.

[9] Manski, C. F. (1975) Maximum score estimation of the stochastic utility model of choice, *Journal of Econometrics*, 3, 205-228.

[10] Newey, W. K. (1985). Maximum likelihood specification testing and conditional moment tests, *Econometrica*, 1047-1070.

[11] Orme, C. (1990) The small-sample performance of the information-matrix test, *Journal of Econometrics*, 46, 309-331.

[12] Stomberg, C. and H. White (2000) Bootstrapping the information matrix test, *University of California, San Diego Department of Economics Discussion Paper*.

[13] Taylor, L. W. (1987) The size bias of White's information matrix test, *Economics Letters*, 24, 63–67.

[14] White, H. (1982) Maximum likelihood estimation of misspecified models, *Econometrica*, 50, 1-25.

Faculty of Economics, Keio University, 2-15-45 Mita, Minato-ku, Tokyo 108-8345, Japan.
Email address: yuta-ota@keio.jp

Department of Economics, London School of Economics, Houghton Street, London, WC2A 2AE, UK.
Email address: t.otsu@lse.ac.uk