SPECIFICATION TESTING FOR BINARY CHOICE MODEL VIA MAXIMUM SCORE

YUTA OTA AND TAISUKE OTSU

Abstract. This paper proposes a Hausman type statistic to test specification of a parametric binary choice model by comparing the maximum likelihood estimator and 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 illustrates that the proposed test is powerful against common forms of misspecification, such as heteroskedasticity and heavy-tailed distributions, while also maintaining reasonable power in distinguishing between closely related distributions, such as the normal and logistic.

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 ($U|X \sim Logistic(0, \eta)$), 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. 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

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

where β^* is a vector of the pseudo-true values, and V_{ML} and V^*_{ML} are asymptotic variances. Under $\mathbb{H}_1$, $\hat{\beta}_{\mathrm{ML}}$ is generally inconsistent and $\beta^* \neq \beta$.

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):

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

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

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

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 $\mathbb{H}_1$.*

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.

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 $Y_i = \mathbb{I}\{1 + \beta X_i + U_i \geq 0\}$ with the true parameter $\beta = 1$ and the sample sizes $n \in \{500, 1000, 5000, 10000, 50000, 100000\}$. The covariate X_i is drawn from both symmetric ($N(0,1)$ and $N(0.5,1)$) and asymmetric ($Gumbel(0,1)$ and $Gumbel(0.5,1)$) distributions. 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) null with $U_i \sim N(0,1)$ (correct specification), (ii) alternative with $U_i \sim Logistic(0,1)$ (distributional misspecification), (iii) alternative with $U_i \sim Cauchy(0,1)$ (heavy-tailed misspecification), and (iv) alternative with $U_i = \epsilon_i\sqrt{1+X_i^2}$ 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.

To implement the maximum score estimator, a grid search is employed. To evaluate the potential performance of our test, we first use an "oracle" critical value. This value, serving as a close approximation of the true asymptotic critical value, is the 95th percentile of the statistic's empirical distribution under the null hypothesis ($U_i \sim N(0,1)$), obtained from $1,000$ replications with a large sample size of $n = 1,000,000$. This allows us to assess the statistic's behavior independently of the critical value estimation method. Next, to demonstrate the test's performance in a practical setting, we implement it using the subsampling method with $1,000$ subsamples. We set the block size to 20% of the original sample size, following Delgado, Rodríguez-Poo and Wolf (2001), and confirmed that the results are robust to other choices of 10% and 30%. The empirical size is the rejection frequency under $\mathbb{H}_0$, while the power is the rejection frequency under the alternatives. We present the results based on the oracle critical values in Figure 1 and those based on the subsampling method in Figure 2.

We first discuss the performance of our test using the oracle critical values, as shown in Figure 1. The test exhibits well-controlled size properties, with the empirical size remaining close to the 5% nominal level across almost all designs. Turning to the power, as expected, the rejection frequency for all alternatives consistently increases with the sample size. A clear pattern also emerges in the relative performance. The test is most powerful against the heteroskedasticity alternative; at $n = 10,000$, its power is already high, ranging from approximately 79% to 100% across the four designs, and it reaches 100% in all cases at $n = 100,000$. The performance against the Cauchy

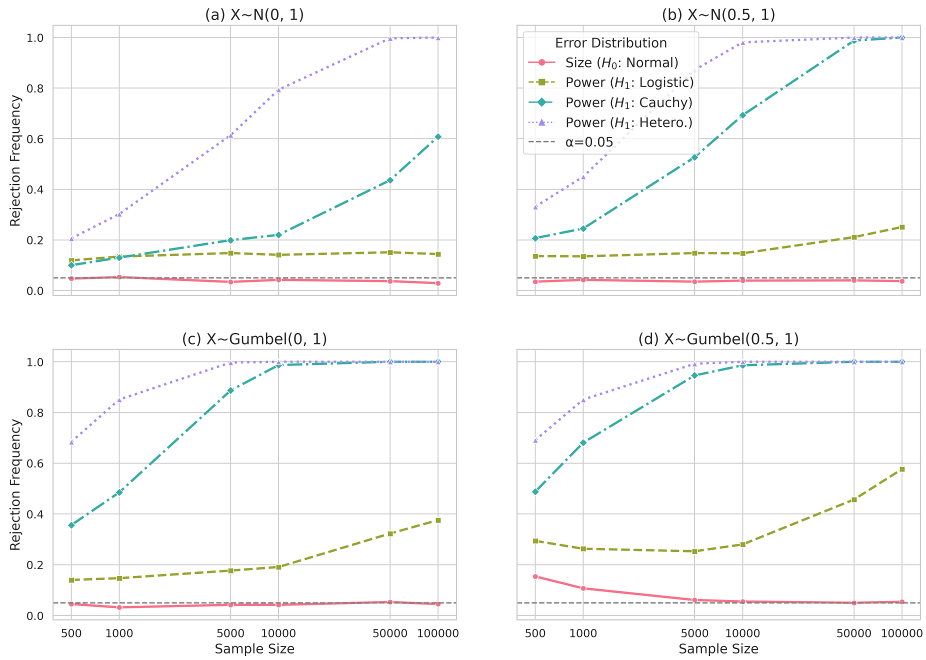

FIGURE 1. Empirical size and power with oracle critical values. The figure shows rejection frequencies against sample size on a logarithmic scale. Each panel corresponds to a different distribution for the covariate X. Lines represent the empirical size and power against three alternatives. The dashed line indicates the 5% nominal significance level.

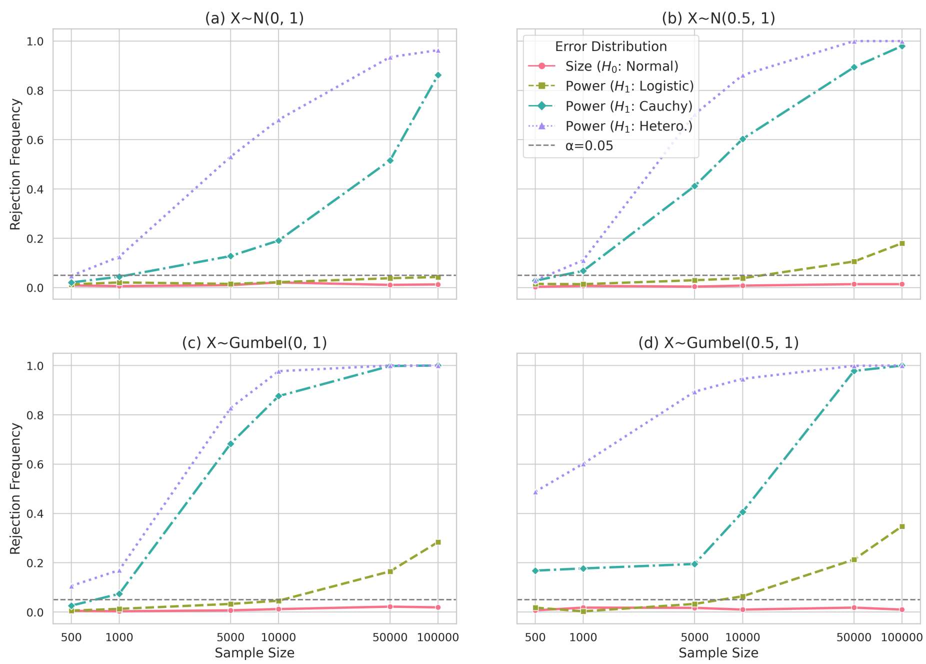

FIGURE 2. Empirical size and power with subsampling critical values. The experimental setup is identical to that of Figure 1, except that the critical values are obtained using the subsampling method.

alternative is also strong, with the power ranging from approximately 22% to 99% at $n = 10,000$ and improving to approximately from 61 to 100% at $n = 100,000$. Finally, against the logistic alternative, a challenging case due to its distributional similarity to the null, the test maintains reasonable power. At $n = 10,000$, its power ranges from approximately 14% to 28%, which increases to from approximately 14 to 58% at $n = 100,000$. Overall, these results demonstrate that the test is effective against a diverse range of misspecifications relevant to economics, though its power varies depending on their specific type.

The power depends not only on the misspecification of the error term, but also on the distribution of the regressor and the degree of data imbalance. Its power is notably higher when X is asymmetrically distributed (Gumbel) compared to the symmetric (Normal) case. For instance, at $n = 500$, the power against the heteroskedasticity and Cauchy alternatives reaches approximately 68% and 36%, respectively, when $X \sim \mathit{Gumbel}(0, 1)$, substantially exceeding approximately 21% and 10% observed when $X \sim N(0, 1)$. Moreover, the test's power tends to increase with greater data imbalance. For example, against the Cauchy alternative at $n = 10,000$, the power exceeds approximately 69% when $X \sim N(0.5, 1)$ but only approximately 22% when $X \sim N(0, 1)$. Similarly, for the logistic alternative at $n = 100,000$, the power increases from approximately 38% when $X \sim \mathit{Gumbel}(0, 1)$ to approximately 58% when $X \sim \mathit{Gumbel}(0.5, 1)$. A plausible explanation is that the probit model is less stable under data imbalance, making deviations from the null hypothesis easier to detect.

Next, we turn to the results using the practical subsampling method, presented in Figure 2. The overall trends are qualitatively similar to those observed with the oracle results. Reading the size, it becomes conservative when using subsampling, with empirical sizes well below the 5% nominal level in all cases. This indicates that the test is not prone to over-rejection in practice. Turning to the power, it is generally lower than that of the oracle counterpart. However, the test remains effective in most cases. The power fails to increase only for the logistic alternative when $X \sim N(0, 1)$; in all other scenarios, it improves as the sample size grows, becoming comparable to the oracle levels. This demonstrates that our test, when implemented with the feasible subsampling procedure, is a practical and valuable tool.[1]

Overall, the simulation results demonstrate that our proposed test is effective. It shows particularly high power against alternatives often discussed in economics, such as heteroskedasticity and heavy-tailed distributions, while maintaining reasonable power against closer alternatives.

References

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

[1]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.

[2] 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.

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

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

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