OXFORD BULLETIN OF ECONOMICS AND STATISTICS, 74, 1 (2012) 0305-9049
doi: 10.1111/j.1468-0084.2011.00654.x

Alternative Versions of the RESET Test for Binary Response Index Models: A Comparative Study*

ESMERALDA A. RAMALHO and JOAQUIM J. S. RAMALHO

Departament of Economics and CEFAGE-UE, Universidade de Évora, Portugal
(e-mails: ela@uevora.pt; jsr@uevora.pt)

Abstract

Binary response index models may be affected by several forms of misspecification, which range from pure functional form problems (e.g. incorrect specification of the link function, neglected heterogeneity, heteroskedasticity) to various types of sampling issues (e.g. covariate measurement error, response misclassification, endogenous stratification, missing data). In this article we examine the ability of several versions of the RESET test to detect such misspecifications in an extensive Monte Carlo simulation study. We find that: (i) the best variants of the RESET test are clearly those based on one or two fitted powers of the response index; and (ii) the loss of power resulting from using the RESET instead of a test directed against a specific type of misspecification is very small in many cases.

I. Introduction

In the econometric analysis of binary responses, parametric single index models are typically employed. These models rely on the assumption of a Bernoulli distribution with mean μ for the response y conditional on the covariates x, where $\mu = G[h(x\theta)]$, $G(\cdot)$ is a cumulative density function and $h(x\theta)$ is an index function in x and the vector of parameters of interest θ. Consistent estimation of θ requires μ to be correctly specified. However, misspecification of μ may arise for a variety of reasons. On the one hand, the assumed cumulative density function $G(\cdot)$ or the index function $h(\cdot)$ may not describe properly the target population. On the other hand, even in cases where the specification chosen for $G[h(x\theta)]$ is in fact appropriate for describing the population of interest, often θ cannot be consistently estimated from the available data set due to sampling issues of which the practitioner is unaware (e.g. measurement error in one or more covariates, misclassification of the outcome variable, non-ignorable missing data, endogenous stratification; see *inter alia* Chesher, 1991; Hausman, Abrevaya and Scott-Morton, 1998; Ramalho and Smith, 2011;

*The authors thank the editor and the referees for helpful comments. Aspects of this research were presented at the 25th Annual Congress of the European Economic Association, Glasgow, the 16th International Conference on Computing in Economics and Finance, London, and the 3rd International Conference of the ERCIM Working Group on Computing & Statistics, London. Financial support from Fundação para a Ciência e a Tecnologia is gratefully acknowledged (grant PTDC/ECO/64693/2006).
JEL Classification numbers: C12, C15, C25.

and Imbens, 1992, respectively). Therefore, when employing parametric models for binary data, it is essential to test the correct specification of μ.

There are two distinct sets of tests that may be applied to assess the specification of μ: (i) general tests for model misspecification, where no specific alternative hypothesis is specified; and (ii) specific tests, which are usually based on the formulation of an alternative parametric model. The former tests are sensitive to a wider variety of departures from the postulated model, while the latter are potentially more powerful when the alternative model is correctly specified but otherwise tend to have low power. Since empirical researchers often do not have any idea about the kind of misspecification that may affect their model and given the great variety of potential misspecification sources, general specification tests are much more commonly applied to test the specification of μ in binary regression models. In fact, apart from the heteroskedasticity test proposed by Davidson and MacKinnon (1984), specific tests for binary models are very rarely applied in empirical work.

In the context of linear regression models, the most widely used general specification test is Ramsey's (1969) Regression Specification Error Test (RESET), which consists of a mere joint significance test for some fitted powers of $x\theta$. As noted by Pagan and Vella (1989) and Peters (2000), RESET-type tests may also be used in binary and other nonlinear single index models. Therefore, due to its simplicity and ease of implementation, in the last decade the RESET test has also become the most popular general specification test for binary and other parametric models.[1] However, while in the linear setting the size and power of the RESET test have been extensively investigated by Monte Carlo methods, in the binary response framework very little is known about its finite sample properties.[2] In fact, to the best of our knowledge, only Thomas (1993) has analysed the performance of the RESET test in the binary setting and only through a very small-scale Monte Carlo study, which was limited to the logit model and a very specific pattern of misspecification.

The main aim of this article is precisely to carry out an in-depth investigation of the finite sample behaviour of the RESET test in the binary response framework. To this end, as tractable analytical power comparisons are not available, we perform an extensive Monte Carlo simulation study that examines, under many different scenarios, the finite sample performance of several versions of the RESET test that differ on the number of powers included as test variables. We consider some of the most popular parametric models for binary responses (logit, probit, cauchit, loglog) and a wide variety of data generating processes in order to investigate the ability of the test variants to detect not only pure functional form problems (misspecification of $G(\cdot)$ or $h(\cdot)$) but also the existence of sampling problems. In each case, the finite sample power of the RESET test is compared with that of a test specifically designed to detect the kind of misspecification simulated.

The remainder of the article is organized as follows. Section II describes the notational framework of the article and discusses the main consequences of various forms of misspecification that may affect binary regression models. In section III, some variants of the RESET test are discussed as well as the specific tests that will be included in the Monte Carlo simulation study described in section IV. Finally, section V concludes.

[1]For some time, other popular general specification test for binary models was the information matrix test introduced by White (1982). However, due to its poor finite sample properties, this test is now rarely applied.

[2]For Monte Carlo studies on the behaviour of the RESET test in the linear framework see, for example, Ramsey and Gilbert (1972), Godfrey and Orme (1994), Leung and Yu (2000) and Hatzinikolaou and Stavrakoudis (2006).

II. Some specification issues in binary models

Consider a sample of $i = 1, \ldots, N$ individuals and let $y = \{0, 1\}$ be the response variable of interest and x a vector of p exogenous variables. The conditional expected value of y given x is defined as

$$\mu \equiv E\left(y \,|\, x, \theta\right) = G\left[h(x\theta)\right] \tag{1}$$

Consistent maximum likelihood (ML) estimation of θ requires in general that the assumed structural model $G[h(x\theta)]$ is in fact a suitable description of the behaviour of the population of interest and that a data set that effectively reflects the characteristics of the target population is available.[3]

Next, we give some examples of misspecification problems that commonly affect binary models. The impact of each of these forms of misspecification in the conditional mean of y given x is illustrated in Figure 1 for simulated samples of 10,001 observations where a probit model with a linear index, a single covariate x_1 and $\theta = 1$, that is $h(x\theta) = x_1$, is taken as a reference. x_1 is a sequence on the interval $[-3, 3]$, except in the case of omission of variables and covariate measurement error where a normal distribution with zero mean and variance one was used for generating it. Despite the simplicity of these examples, the diversity of possible consequences produced by the various forms of misspecification are clearly illustrated in Figure 1.

Misspecification of the structural model

Misspecification of the structural model may be due to an incorrect choice of the 'link' function $G(\cdot)$ or to an incorrect choice of how and which explanatory variables should appear in the index function $h(\cdot)$.

Example 1. Incorrect link function

Despite the popularity of the logit and the probit specifications for $G(\cdot)$, which are given by, respectively, $e^{h(x\theta)}/[1 + e^{h(x\theta)}]$ and $\Phi[h(x\theta)]$, in some cases there may be other models that provide a better description of the data. For example, the cauchit (also known as arc tangent), defined as $0.5 + \pi^{-1} \arctan[h(x\theta)]$, is appropriate for cases where the shape of μ presents fatter tails, and the loglog and complementary loglog, defined as $e^{e^{-h(x\theta)}}$ and $1 - e^{-e^{h(x\theta)}}$, are suitable for cases where asymmetric functional forms are required. The first graph of Figure 1 shows the differences between these five link functions. Note that while the symmetric cauchit, logit and probit models approach 0 and 1 at the same rate, the asymmetric cloglog (loglog) model increases slowly (sharply) at small values of $G(\cdot)$ and sharply (slowly) when $G(\cdot)$ is near 1.

Example 2. Omission of relevant covariates

The omission of a relevant explanatory variable in models for binary data leads, in general, to inconsistent estimation of θ. In effect, when some relevant variables w are omitted from $G[h(\cdot)]$, the conditional mean of the response given the included covariates x is given by

[3]Naturally, consistent estimation of θ is also possible if the structural model is appropriately adapted to reflect the fact that the sampled and target populations may be different.

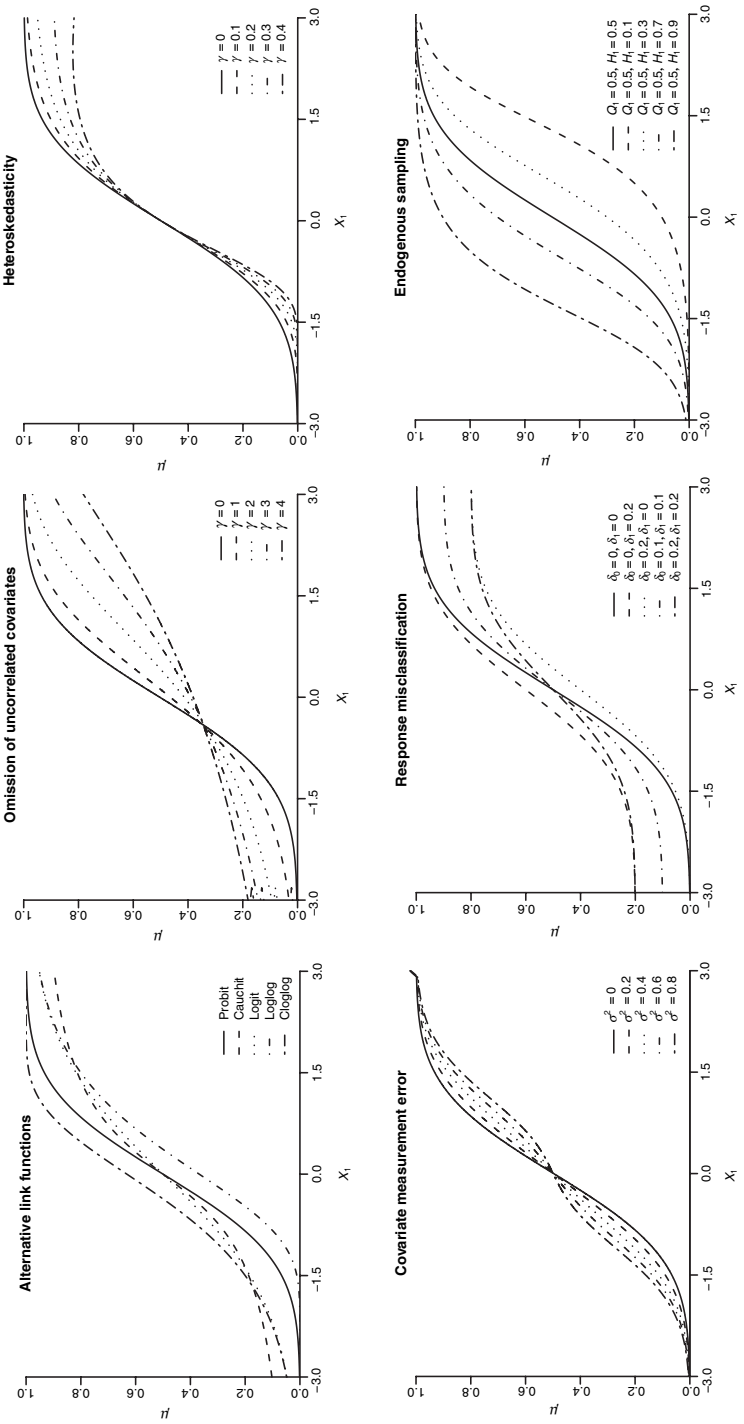

Figure 1. Examples of misspecification of binary regression models (base model: probit)

$$\mu = \int_W G[h(x, w, \theta, \gamma)] f(w \mid x)\, \mathrm{d}w, \tag{2}$$

where γ is the vector of parameters associated to w and $f(w \mid x)$ is the conditional density function of w given x. In contrast to linear models, where the omission of w is innocuous in cases where x and w are uncorrelated, even in such a case equation (2) differs in general from the naive version of the conditional mean $G[h(x\theta)]$.[4] This case is represented in the second graph of Figure 1, which considers an example where a relevant variable w, distributed as a displaced exponential with variance one and generated independently from x_1, is omitted. It is clear that μ is no longer symmetric around zero and presents fatter tails than the probit benchmark, the amount of dispersion depending on the weight of the omitted variable on the index, which is determined by γ.

Example 3. Nonlinear index misspecified as linear due to heteroskedasticity

An obvious source of misspecification is the omission of nonlinear terms in the index. This omission may be the result of the presence of heteroskedasticity, a problem which again is innocuous for consistent estimation of θ in linear models but not in this setting. Consider a linear latent model $y^* = x\theta + u$, where u is a variate with zero mean and variance $s(x, \gamma)$ defined in such a way that when $\gamma = 0$, $s(x, \gamma) = 1$. Define the observed binary outcome as $y = 1$ ($y = 0$) if $y^* > 0$ ($y^* \leq 0$). Clearly, the functional form implied by this formulation is

$$\mu = G\left[\frac{x\theta}{\sqrt{s(x, \gamma)}}\right] \tag{3}$$

and using the linear index $x\theta$, overlooking the nonlinearities induced by heteroskedasticity, leads to inconsistent estimation of θ; see Davidson and MacKinnon (1984) and Yatchew and Griliches (1985).

Figure 1 contains an illustration of equation (3) for the case where the skedastic function is $s(x_1, \gamma) = e^{2\gamma x_1}$. Again, the symmetric characteristic of the probit is distorted and the variability of y given x is increased.

Misspecification due to observation problems

In some cases, the population of interest is properly described by the functional form chosen for $G[h(x\theta)]$ but the available data set, due to some sampling issues, provides a distorted representation of $G[h(x\theta)]$. In this subsection, we briefly analyse three potentially variance increasing and/or shape distorting sources of misspecification that are related to the observation process: covariate measurement error, response misclassification and endogenous sampling. We focus on cases where the index is linear, $h(x\theta) = x\theta$, to simplify the notation. In all the examples that follow, the functional form appropriate for the data is written as a function of $G(x\theta)$, so that the distortions created by the three sampling issues become apparent and the mechanism that governs them may be analysed in a simple way.

[4]The consistency of the ML estimator is not affected only when $\theta = 0$. See *inter alia* Ramalho and Ramalho (2010).

Example 4. Covariate measurement error

The effects of the presence of measurement error in continuous covariates may be examined by using Chesher's (1991) small parameter asymptotic approximations. Assume that we observe an error-prone version x^* of the covariates x according to $x^* = x + u$, where u is a p-dimensional vector of unobservable measurement errors, which have an unknown continuous joint distribution $f(u)$. Assume also that x and u are independently distributed, $E(u) = 0$, and $E(uu') = \Sigma = [\sigma_{jk}]$, where Σ is a positive semi-definite $p \times p$ matrix. The approximation for a small error variance for μ is

$$\mu = G(x^*\theta)\left[1 + \sigma_{jk} m^{jk}(y, x^*)\right] + o(\Sigma), \tag{4}$$

where $m^{jk}(y,x^*) = 0.5[l^{jk}_{y|x}(x^*\theta) + l^{j}_{y|x}(x^*\theta) l^{k}_{y|x}(x^*\theta) + 2 l^{j}_{y|x}(x^*\theta) l^{k}_{x}(x^*)]$, superscripts denote derivatives with respect to the latent covariates which are mismeasured, subscripts indicate elements of vectors, $l_{y|x}(x^*\theta) = \ln G(x^*\theta)$, $l_x(x^*) = \ln f(x^*)$, $o(\Sigma)$ is such that $\lim_{\max(\sigma_{jj})\to 0} \frac{o(\Sigma)}{\max(\sigma_{jj})} = 0$, and the Einstein summation convention from 1 to p is to be performed over indices that appear both as superscripts and subscripts; see Chesher (1991) for details. For the particular case where only one covariate, say x_t, is error-prone, $m^{jk}(y,x^*)$ of equation (4) simplifies to $\frac{0.5\theta_t^2}{G(x^*\theta)}[\nabla^2_{x\theta} G(x^*\theta) + \frac{2}{\theta_t} \nabla_{x\theta} G(x^*\theta) l_X(x^*)]$, where $\nabla_{x\theta}$ denotes derivative with respect to $x\theta$ and θ_t is the coefficient associated to x_t.

It is clear from equation (4) that the term $\sigma_{jk} m^{jk}(y,x^*)$ reflects the distortions caused by the presence of measurement error. Only in absence of measurement error, as $x = x^*$ and $\sigma_{jk} = 0$, the functional form is reduced to the model $G(x\theta)$ maintained in the population of interest. Figure 1 contains an illustration of equation (4) for a probit model for five different magnitudes of the variance of the measurement error. Although the symmetric property of the original probit curve is preserved, it is clear that covariate measurement error induced dispersion, which becomes more substantial as the variance of the measurement error grows.

Example 5. Response misclassification

The consequences of response misclassification may be simply formalized following Cox and Snell (1989), pp. 122–123. Define two parameters, δ_1 and δ_0, as the probability of observing 1 (0) when the actual response is 0 (1). The probability of observing $y = 1$ given x may be written as $\Pr(y = 1 \mid x) = (1 - \delta_0) G(x\theta) + \delta_1 [1 - G(x\theta)]$, which gives rise to

$$\mu = \delta_1 + \left(1 - \delta_0 - \delta_1\right) G(x\theta), \tag{5}$$

where $0 \leq \delta_0, \delta_1 \leq 1$ and, for identification matters, $\delta_0 + \delta_1 \leq 1$; see also Hausman *et al.* (1998).

The functional form equation (5) reduces to $G(x\theta)$ only in absence of misclassification, such that $\delta_0 = \delta_1 = 0$. Figure 1 shows that, similarly to covariate measurement error, this kind of measurement error induces dispersion. However, now the symmetry of the probit curve is preserved only in the case designated in the literature as random misclassification, which is characterized by $\delta_0 = \delta_1$. For $\delta_0 \neq \delta_1$, various forms of asymmetry are created according to the magnitude of both δ_0 and δ_1, which govern, respectively, the right and the left tail of the curve.

Example 6. Endogenous sampling

Endogenous or response-based sampling is common when the variable of interest is binary, either as a consequence of an endogenous stratified (or a choice-based) sampling design, where the proportion of each response in the sample is fixed by design, or due to the presence of missing data on both y and x (case usually designated as unit non-response) governed by a non-ignorable response mechanism that depends on y. Define H and Q as the proportion of individuals for which $y = 1$ in the sample and in the population, respectively. The sampling conditional probability of observing 1 given x is

$$\mu = \frac{H}{Q}\left[\frac{1-H}{1-Q} + \left(\frac{H}{Q} - \frac{1-H}{1-Q}\right) G(x\theta)\right]^{-1} G(x\theta); \tag{6}$$

see *inter alia* Manski and McFadden (1981).

The functional form that describes the observed data, equation (6), only reduces to $G(x\theta)$ in two cases: (i) the data is self-weighting or missing completely at random, that is $H = Q$; and (ii) $G(\cdot)$ is a logit (although in this case equation (6) is a logit with an intercept displaced in $\ln\left(\frac{Q}{H}\frac{1-H}{1-Q}\right)$). The distortions imposed by this sampling problem are illustrated in Figure 1. Clearly, μ becomes asymmetric in all cases. When $H > Q$ ($H < Q$), the proportion of 1's is inflated (depressed) in the sample, relative to the population. Therefore, the curve is shifted to the left (right), which implies that μ goes more rapidly (slowly) to one than $G(x\theta)$.

III. Specification tests for binary regression models

This section briefly discusses some alternative specification tests suitable to test the null hypotheses that $G[h(x\theta)]$ is an appropriate specification for $E(y \mid x)$. For simplicity, assume that $h(x\theta) = x\theta$ under the null hypothesis, i.e. $H_0 : E(y \mid x) = G(x\theta)$. All tests described next are implemented as Lagrange Multiplier (LM) statistics for the omission of a set of artificial regressors z from $G(\cdot)$. We compute these statistics from auxiliary regressions of the type proposed by Davidson and MacKinnon (1984), who showed that, in the binary response framework, an LM statistic for the omission of z with good small sample properties is given by $LM = ESS$, where ESS is the explained sum of squares of the auxiliary regression

$$\tilde{u} = \tilde{g}x^{*}\delta + error, \tag{7}$$

where $g = \nabla_{x\theta} G(x\theta)$, $\hat{u} = Y - \hat{G}$, $\tilde{u} = \hat{u}\hat{\omega}$, $\tilde{g} = \hat{g}\hat{\omega}$, $\hat{\omega} = [\hat{G}(1-\hat{G})]^{-0.5}$, $x^{*} = (x', z')$ and $\hat{}$ indicates evaluation under H_0 at $\hat{\delta} = (\hat{\theta}, 0)$.

Following Wooldridge (2002), we suggest an integrated approach to construct the artificial regressor z, which may be applied both in tests against general and specific alternatives. Let $\mu = F[G(x\theta), \alpha]$ be the model maintained under H_1, which reduces to $G(x\theta)$ for some particular value of the vector α. As shown in Wooldridge (2002), p. 464, the artificial regressors can be straightforwardly calculated as $z = \nabla_{\alpha}\hat{\mu}\hat{g}^{-1}$.

In the next two sections, for each of the test in analysis, the three features required for its implementation are described: the null hypothesis in test, the alternative model, and the composition of the vector z. First, we describe the alternative versions of the RESET. Then, we examine tests designed to be sensitive to each of the misspecification problems considered in section II.

TABLE 1

Specific tests

Specific test for:	$H_1: E(Y \mid X) =$	H_0:	z
Link function	$T(x\eta)$	$G(x\theta)$	$(\hat{T} - \hat{G})\hat{g}^{-1}$
Omitted variables	$G(x\theta + \omega\gamma)$	$\gamma = 0$	ω
Heteroskedasticity	$\mu = G\left[\frac{x\theta}{e^{x_1\gamma}}\right]$	$\gamma = 0$	$-x\hat{\theta}x_1$
Covariate measurement error	$G(x^*\theta)[1 + \sigma^2 m(y, x^*)] + o(\sigma^2)$	$\sigma^2 = 0$	$0.5\hat{\theta}_1^2[\nabla_{x\theta}\hat{g} + \frac{2}{\theta_1} l_{X_1}(x_1^*)\hat{g}]\hat{g}^{-1}$
Response misclassification	$\delta_1 + (1 - \delta_0 - \delta_1)G(x\theta)$	$\delta_0 = \delta_1 = 0$	$z_1 = -\hat{G}\hat{g}^{-1},\ z_2 = (1 - \hat{G})\hat{g}^{-1}$
Endogenous sampling	$\frac{H}{Q}\left[\frac{1-H}{1-Q} + \left(\frac{H}{Q} - \frac{1-H}{1-Q}\right)G(x\theta)\right]^{-1} G(x\theta)$	$H = Q$	$\hat{G}(1 - \hat{G})\hat{g}^{-1}$

The RESET

The RESET, instead of being derived to test against a particular alternative model, is based on the idea that any index model of the form $E(y \mid x) = F(x\theta)$ can be arbitrarily approximated by $G[x\theta + \sum_{j=1}^{J} \gamma_j (x\theta)^{j+1}]$ for J large enough. Therefore, testing the hypothesis $H_0: E(y \mid x) = G(x\theta)$ is equivalent to test for $H_0: \gamma = 0$ in the augmented model $E(y \mid x, z) = G(x\theta + z\gamma)$, where $z = [(x\hat{\theta})^2, \ldots, (x\hat{\theta})^{J+1}]$. As the first few terms in the expansion are the most important, in practice, the more popular versions of the test use $J \leq 3$. According to the number of test variables included, different is the variant of the RESET.[5] In this article we consider five variants of the test, designated as RESETJ, for $J = \{1, 2, 3, 4, 5\}$.

Some specific tests

In contrast to the RESET, the tests based on specific alternative models are designed to be sensitive to particular forms of misspecification. Therefore, they are expected to be more powerful than those derived against general alternatives and, thus, suitable to be used as benchmarks for the finite sample power behaviour of the general RESET test. The information required to implement these tests is summarized in Table 1.

To test two alternative specifications for the link function, say $G(\cdot)$ and $T(\cdot)$, one against the other, we consider the P test developed by Davidson and MacKinnon's (1981) for testing non-nested hypothesis. For all the other examples of misspecifications, we consider specific tests based on the general models of section II, except for the case of omitted variables, where certainly a good benchmark for the RESET test is provided by a direct LM test for the relevancy of the omitted variable w. The tests considered for heteroskedasticity, covariate measurement error, and response misclassification were originally proposed by, respectively, Davidson and MacKinnon (1984), Chesher (1991) and Copas (1988), while the test for endogenous sampling is new.[6]

[5]A well known alternative version of the RESET for linear models where the test variables are different from the ones considered here is that of Thursby and Schmidt (1977), where $z = [x_1^2, \ldots, x_k^2, \ldots, x_1^{J+1}, \ldots, x_k^{J+1}]$ for cases where the regressors do not include dummy variables.

[6]Note that Ramalho and Smith (2011) had already proposed a test to detect non-ignorable discrete choice non-response. However, while their test was derived in the generalized method of moments framework, the test proposed in this article is a simple LM test constructed in the ML setup based on model (6). The major difference is that the

IV. A Monte Carlo simulation study

This section presents an extensive Monte Carlo simulation study on the finite sample performance of five versions of the RESET test that differ in the number of test variables, which ranges from one (RESET1) to five (RESET5). In the power analysis, in each example, we consider also a specific LM test, derived from the parametric model that governs the simulated data, as a benchmark for the performance of the RESET test.[7]

The finite sample properties of the tests are expected to differ according to the structural model from which the data are generated. Additionally, the power of the test certainly will depend also on the mechanism responsible by the deviations from the postulated model. Therefore, in all the examples simulated, we consider four alternative links for binary data (cauchit, logit, probit and loglog) and assume a linear index with two covariates, $h(x\theta) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$, in most cases. As in Santos Silva (2001), x_1 is generated as a standard normal variate (with one exception, described later on), and x_2 is generated as a Bernoulli variate with mean 2/3. We set $\theta_2 = 1$ and consider several values for both θ_0 or θ_1 in order to control the percentage of zeros and ones of y and the contribution of x_1 for the variance of the index, respectively. We consider also several values for the parameters that define the misspecification mechanisms, as explained below. Given the substantial amount of results produced, we summarize them in Figures 2–11. All experiments are based on 10,000 replications. In most cases, we consider sample sizes of $N = 500$ and $N = 5{,}000$.

Size properties of alternative RESET tests

In this section we examine the size performance of the different RESET variants in analysis. Figures 2 and 3 display the percentage of rejections of H_0 for a nominal level of 5% when this hypothesis is indeed true (the horizontal lines represent the limits of a 95% confidence interval for the nominal size). In Figure 2, we consider $N = 500$, $\theta_0 = \{-2, -1.5, \ldots, 2\}$ and $\theta_1 = \{-2.5, -2, \ldots, 2.5\}$, while in Figure 3, for four different θ vectors, we represent the empirical size of the tests for $N = \{500, 1{,}000, \ldots, 4{,}500, 5{,}000\}$.

Figure 2 suggests that, in general, the empirical sizes of RESET1 and RESET2 are not significantly different from the nominal level of 5% (most cases) or are slightly undersized (e.g. loglog model for $\theta_1 = 1$). In contrast, the remaining RESET variants appear to be unreliable in many cases, especially in cauchit and logit models or when the model is poorly identified (θ_1 is close to zero): in the former case they tend to be oversized, while in the latter they are clearly undersized. These findings are corroborated by Figure 3: while the RESET versions based on 3 or more powers are still oversized in many cases even when $N = 5{,}000$, both RESET1 and RESET2 display an appropriate behaviour for almost all of the sample sizes simulated.[8] Thus, in which regards the size properties of RESET

former test is derived from the sampling joint density function of the response and the covariates and the latter is based on the sampling density function of the response conditional on the covariates.

[7]Note that these specific tests are expected to have low or no power in cases where the alternative specification is incorrect but the investigation of their robustness in these cases is out of the scope of this article.

[8]Actually, according to some additional experiments not reported in the article, in some cases only for sample sizes as large as 50,000 do the RESET4 and, mainly, RESET5 variants exhibit empirical sizes that are not significantly different from the nominal ones at a 5% level. Probably, this has to do with the fact that the fifth and sixth powers of the fitted values may be very high and distort somehow the behaviour of the test in finite samples.

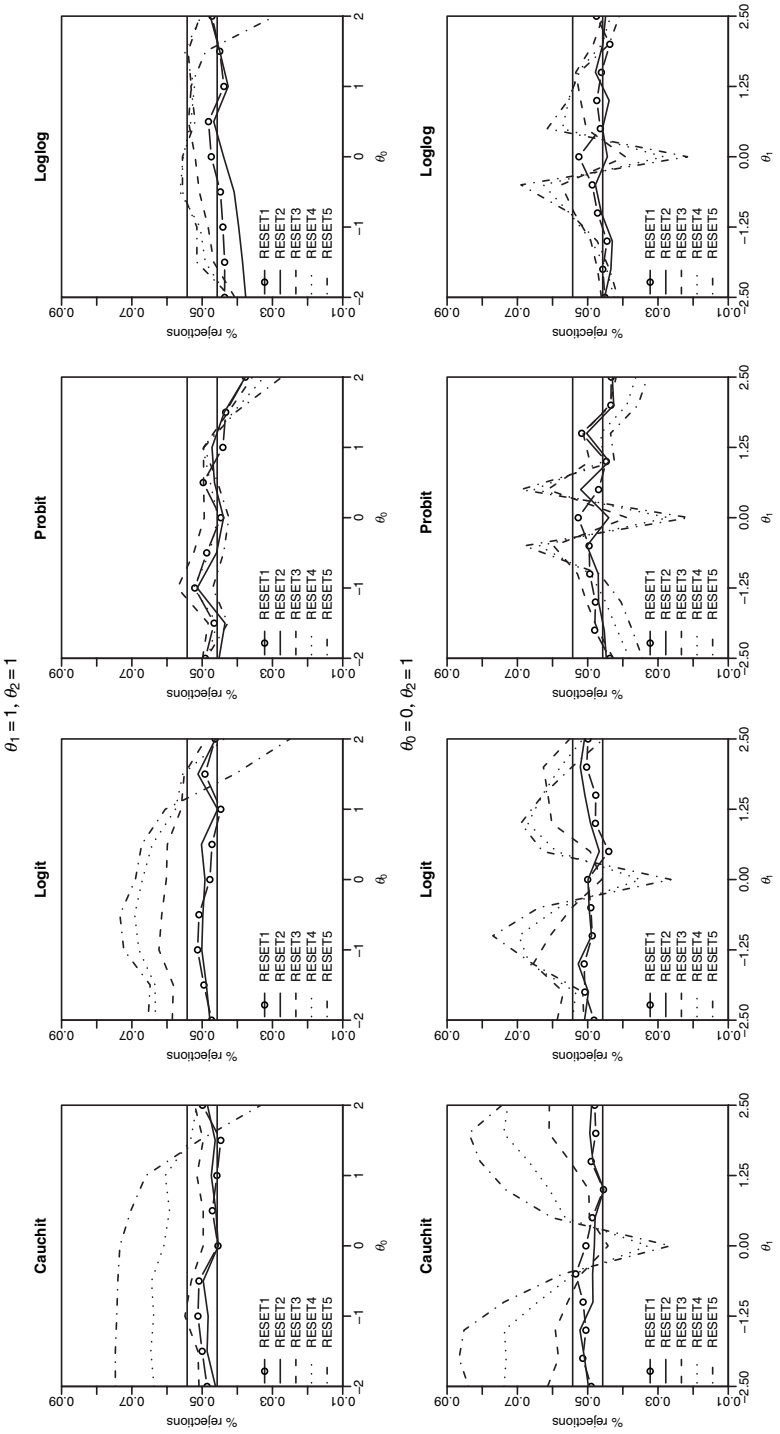

Figure 2. Empirical size ($N = 500$)

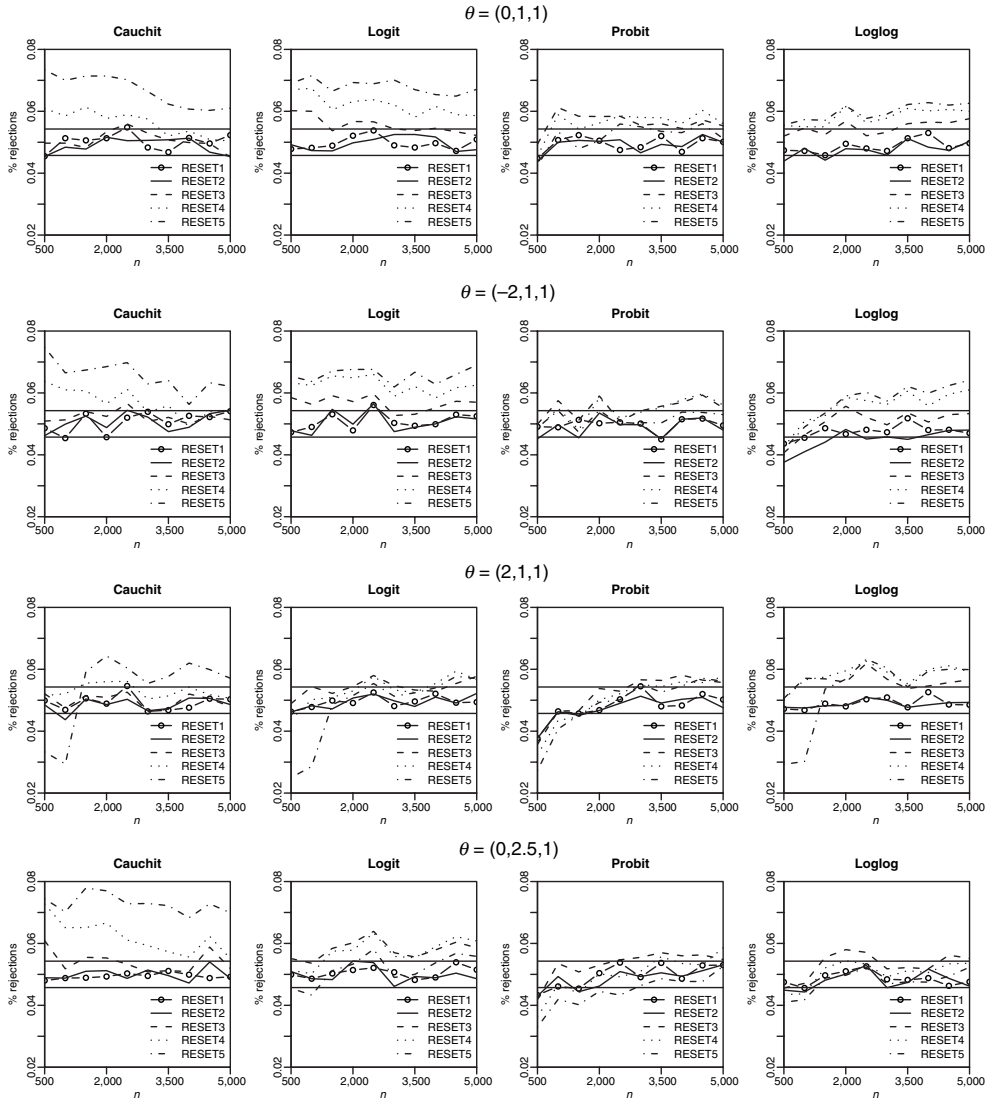

Figure 3. Empirical size for different sample sizes

tests, it is clearly preferable to compute versions that use only one or two fitted powers of the response index.

Power properties of alternative RESET tests

In this section we investigate the power properties of the five RESET variants using simulated data for each one of the six types of misspecification sources described in section II.

Misspecification of the link function

Figures 4 and 5 show the ability of both RESET and *P* non-nested tests to detect departures from the true link function. In each case, the null hypothesis corresponds to an incorrect link

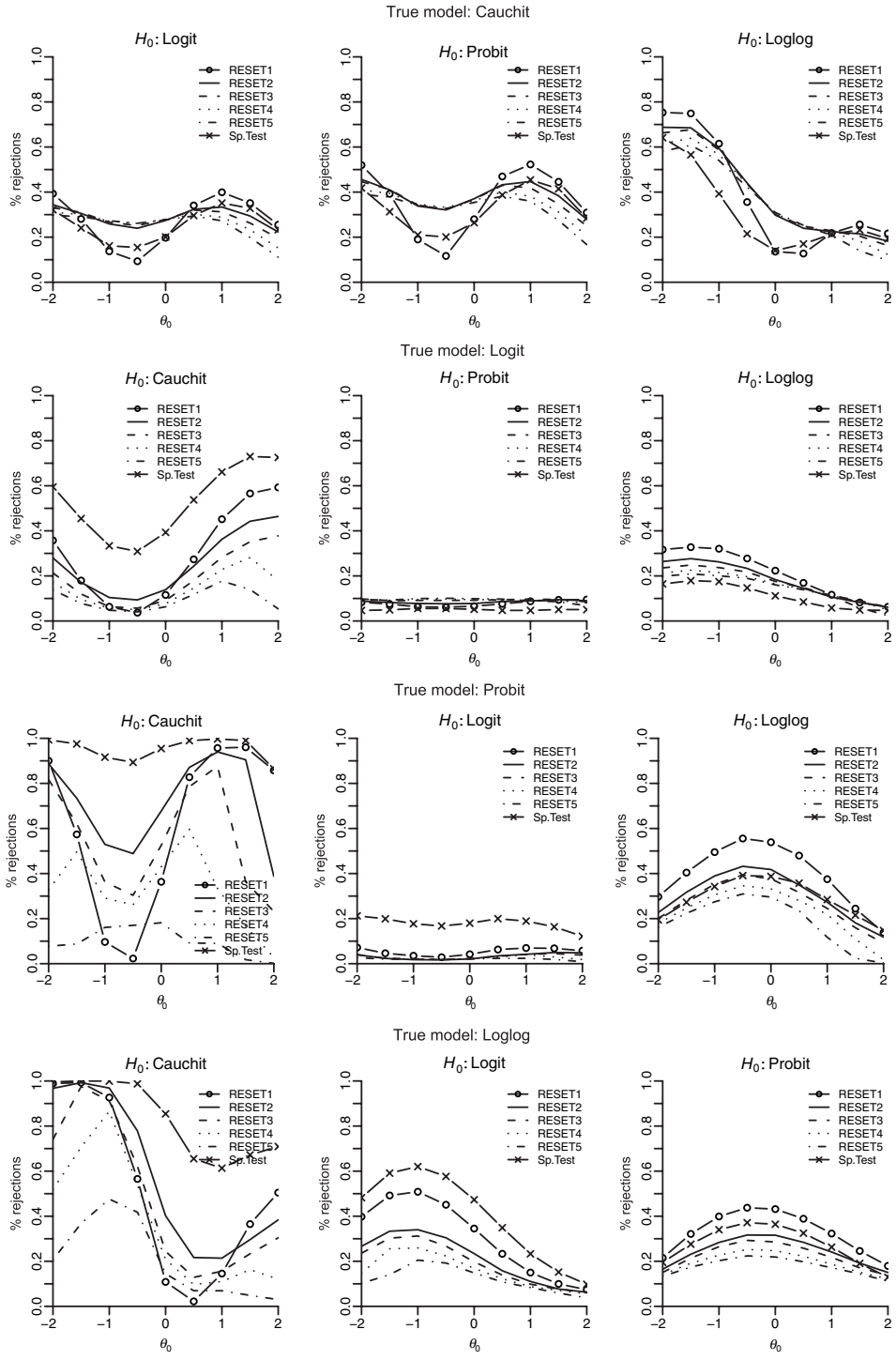

Figure 4. Empirical power – misspecification of the link function ($\theta_1 = 1, \theta_2 = 1; N = 500$)

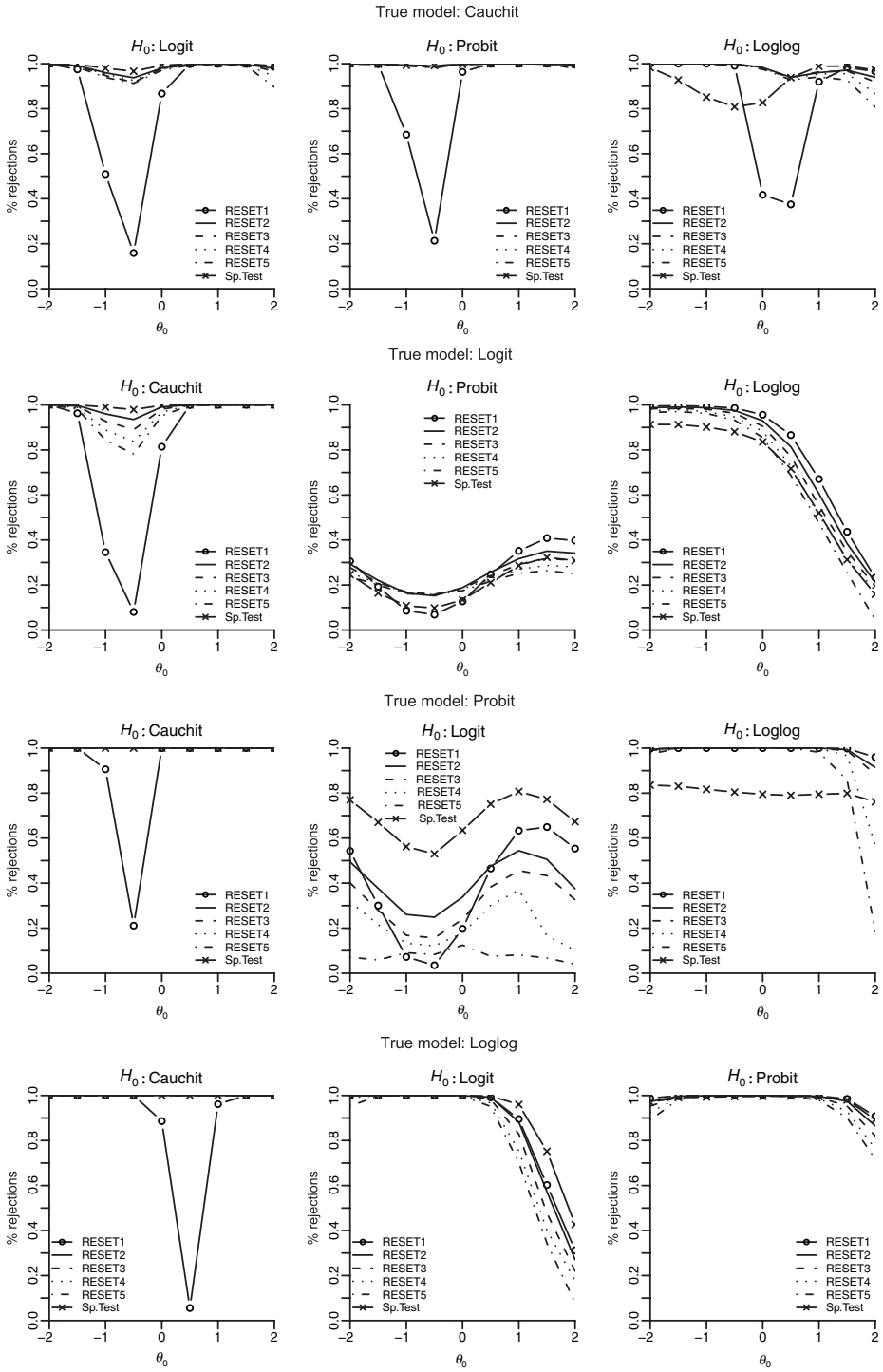

Figure 5. Empirical power – misspecification of the link function ($\theta_1 = 1, \theta_2 = 1; N = 5,000$)

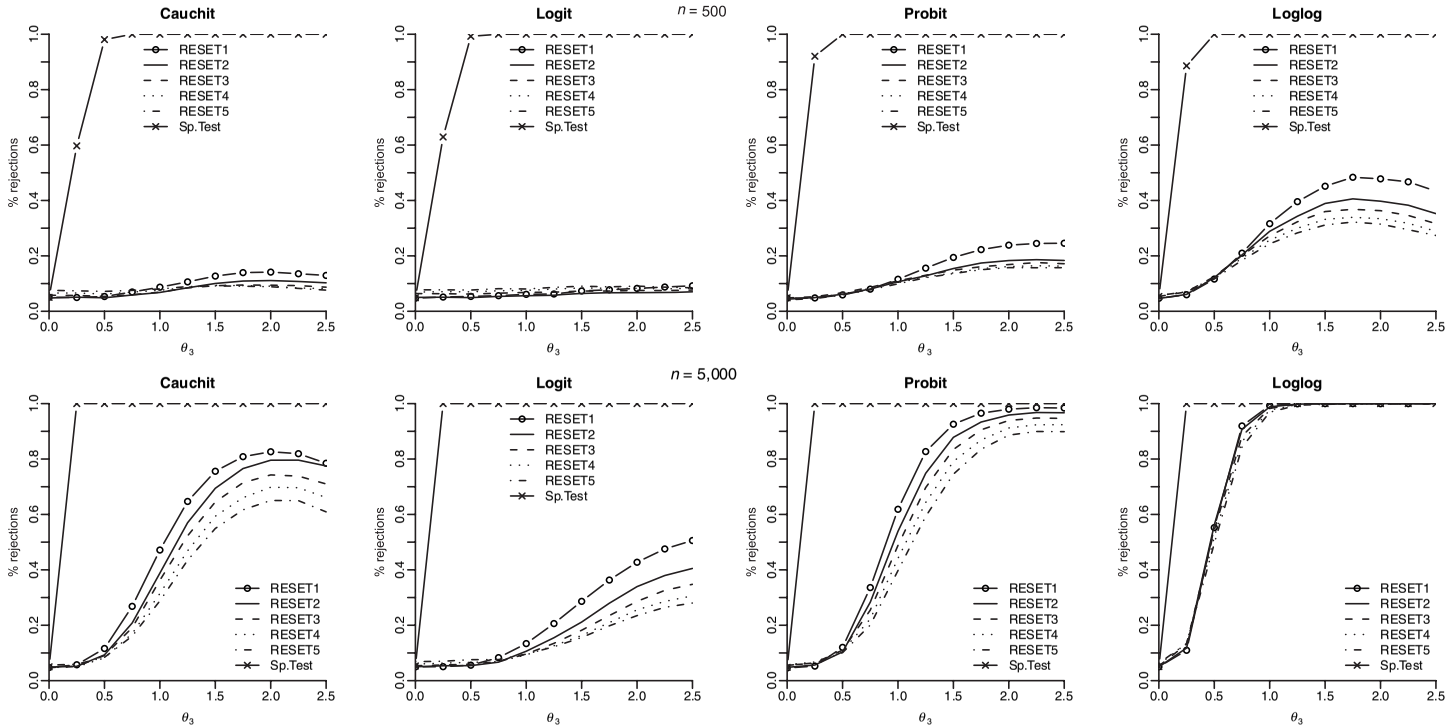

Figure 6. Empirical power – omission of an uncorrelated covariate

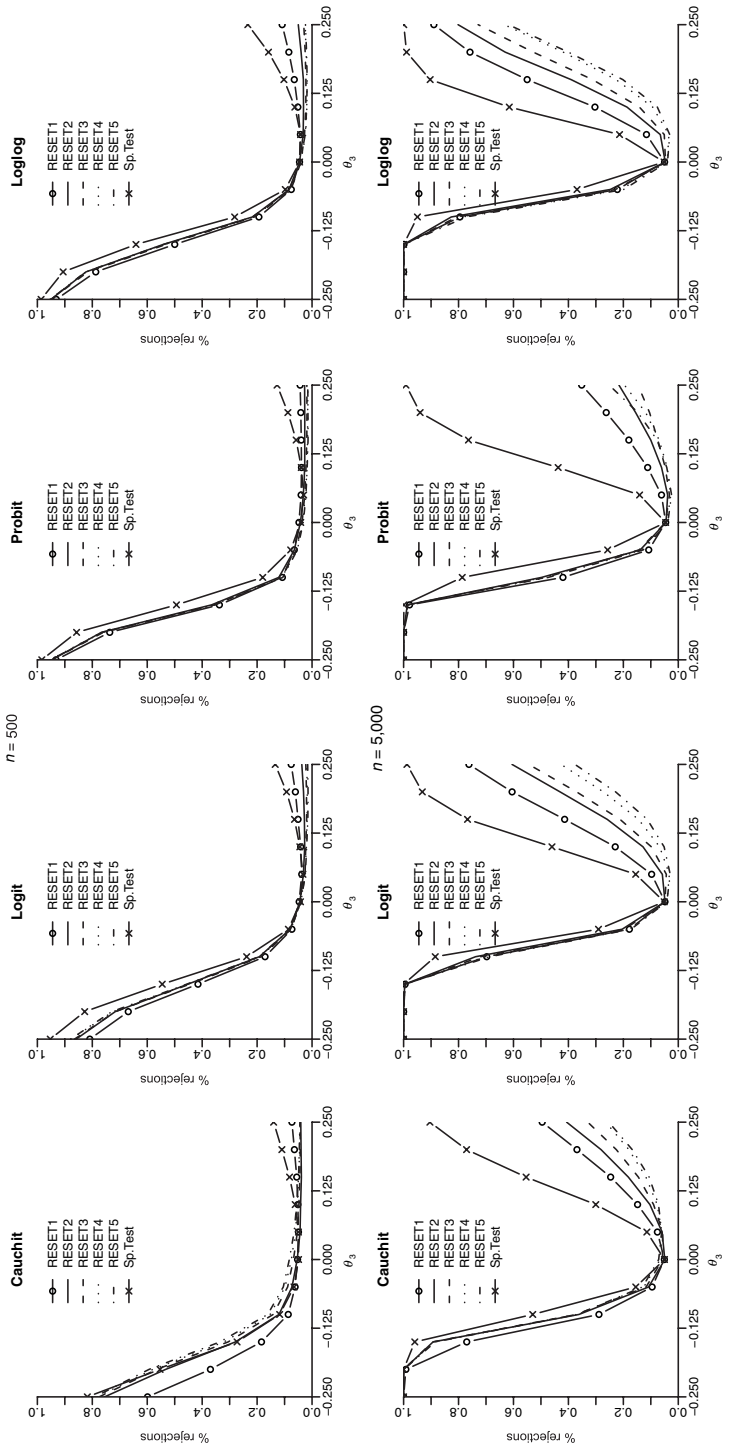

Figure 7. Empirical power – omission of a quadratic term of an included covariate

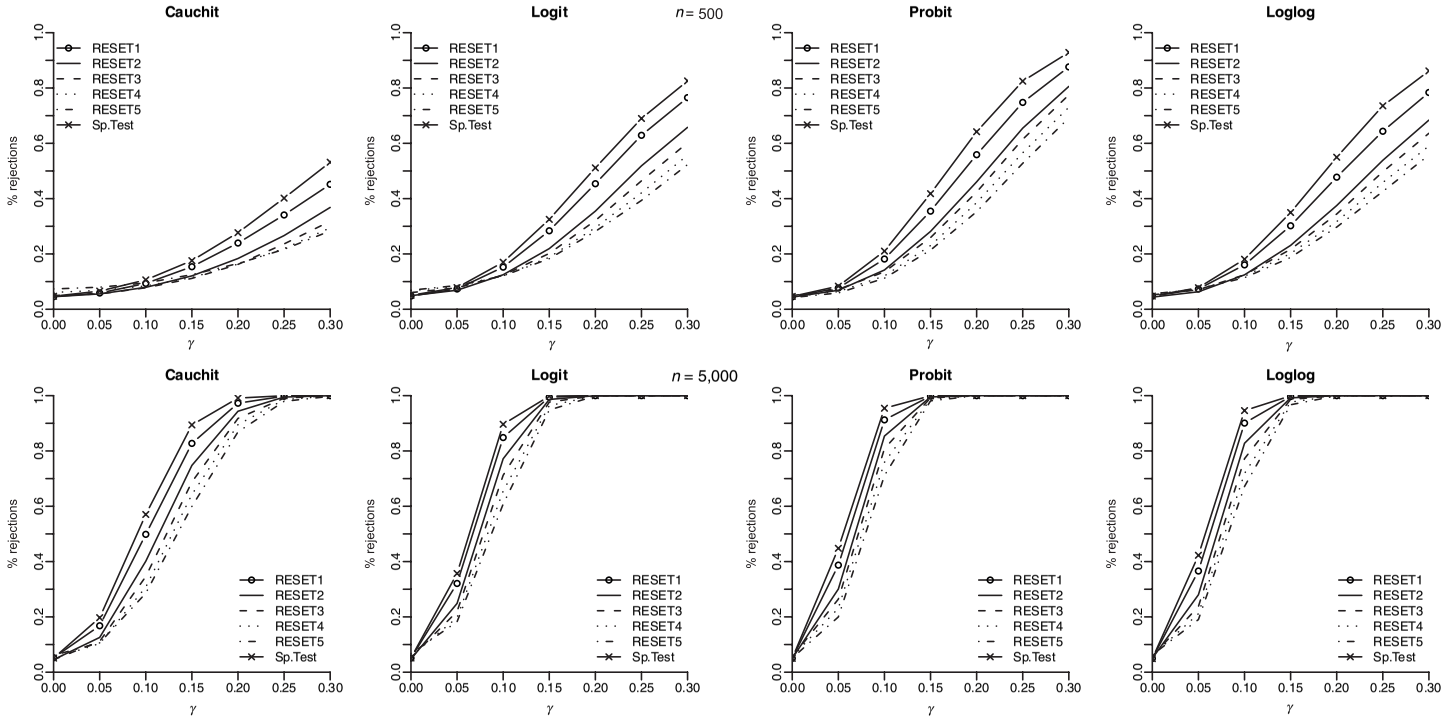

Figure 8. Empirical power – heteroskedasticity

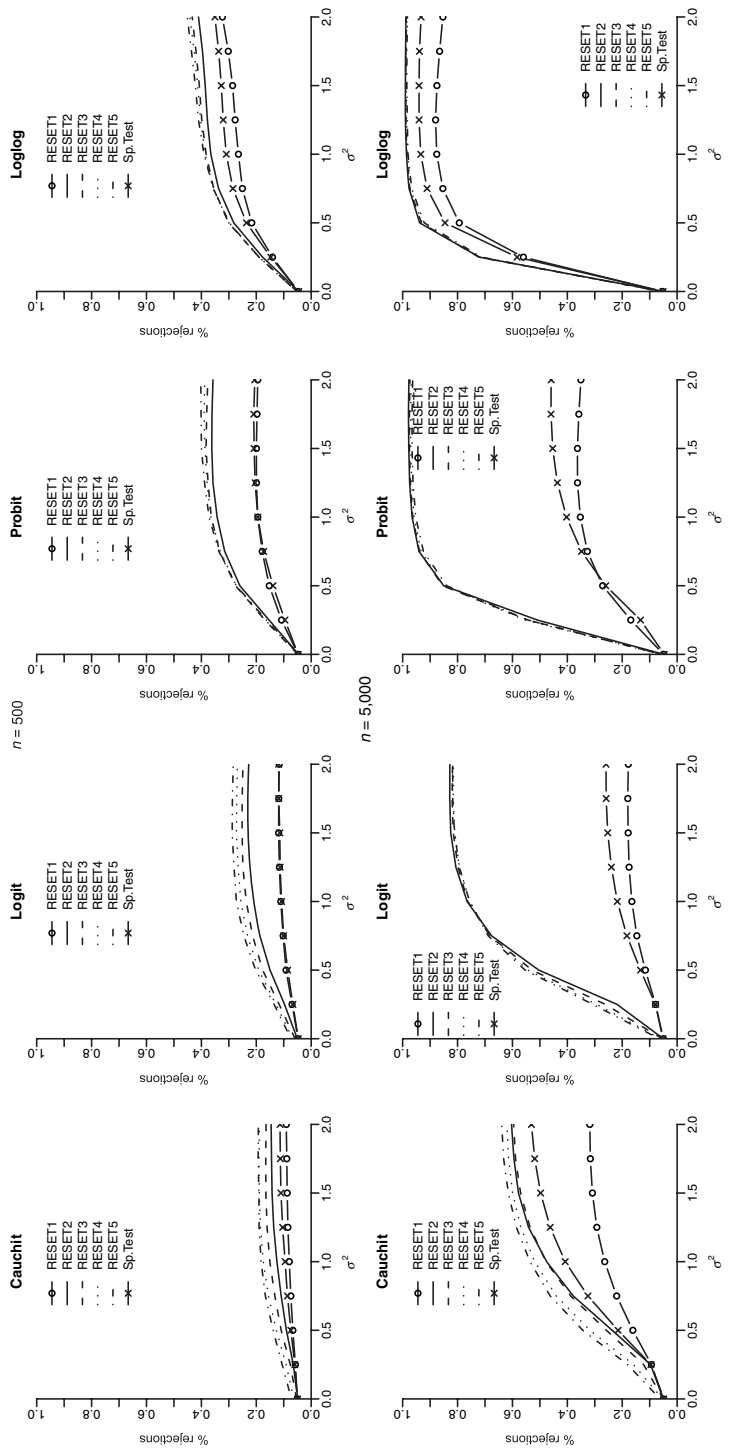

Figure 9. Empirical power – covariate measurement error

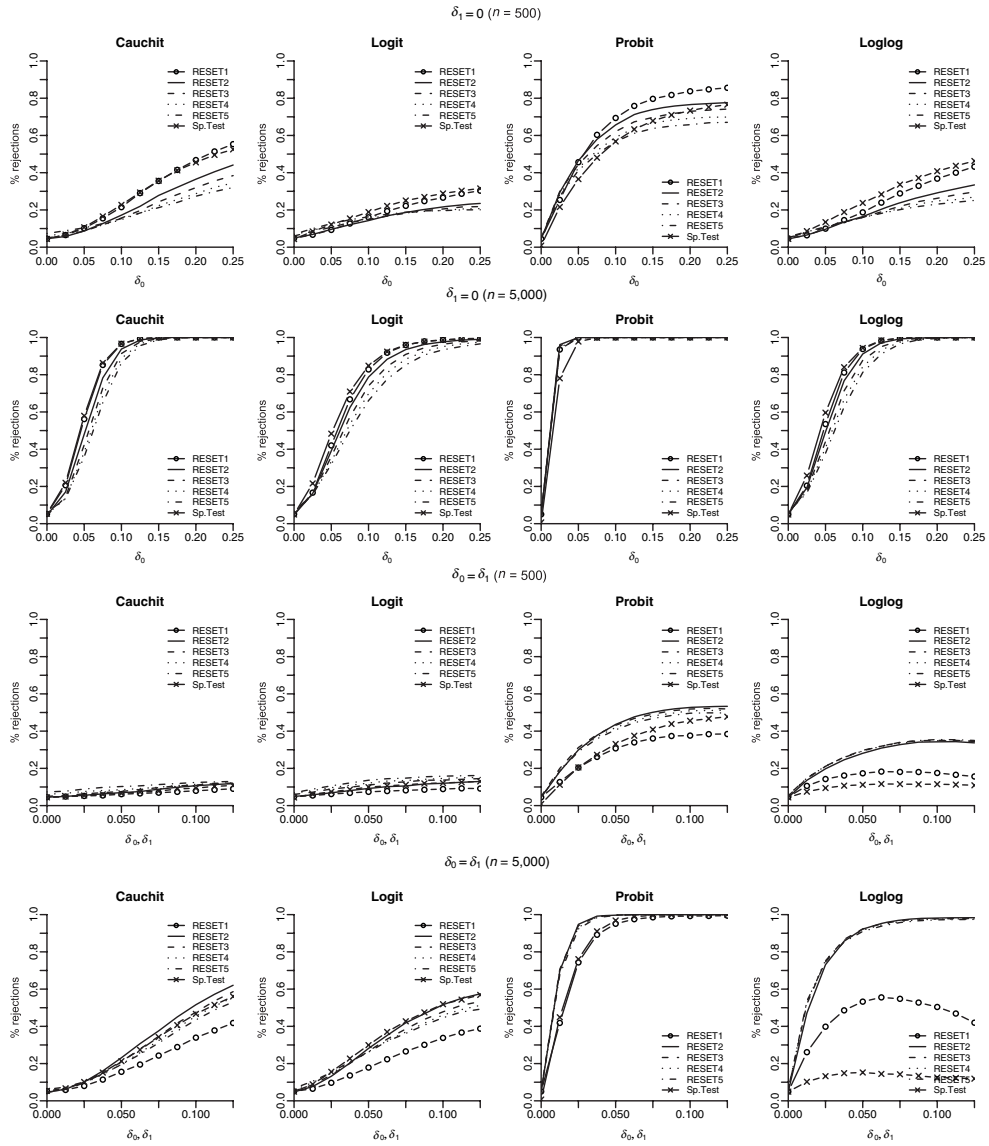

Figure 10. Empirical power – response misclassification

function, which in the case of the P test is assessed against the true specification. Clearly the estimated power of the tests reflects the degree of similarity between the shapes of the assumed and the true link functions, see the first graph of Figure 1. For example, when the choice is between the three symmetric models, the tests, in general: (i) have more power to distinguish between the heavy-tailed cauchit and the other models than for distinguishing between logit and probit models; and (ii) have lower power when θ_0 approaches $-2/3$ (the mean of $h(x\theta)$ approaches zero), since around this value of θ_0 the three functions are very similar.

In all cases, a more powerful RESET test is obtained if we use in their computation two instead of a higher number of powers. The RESET1 version, on the other hand, does

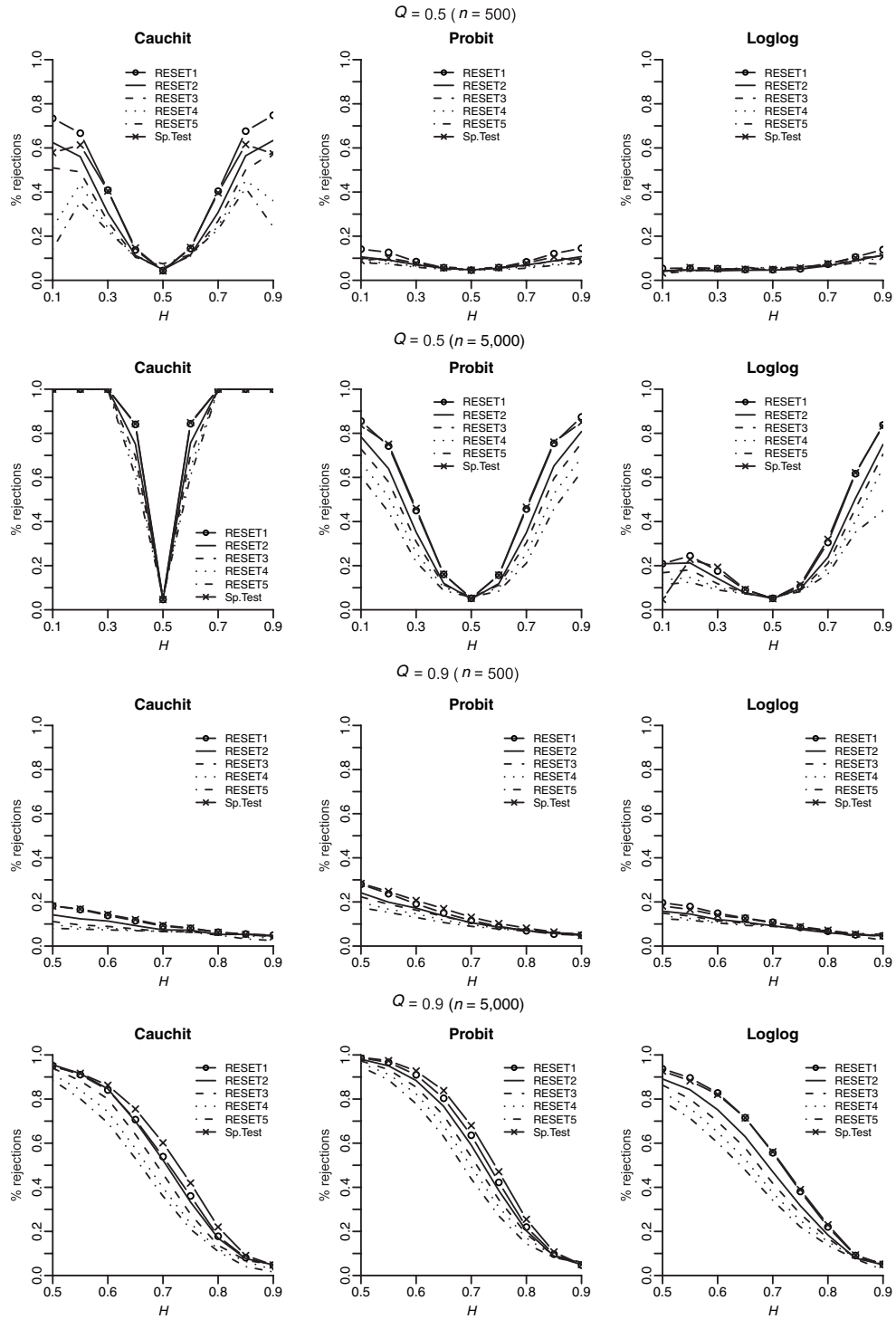

Figure 11. Empirical power – endogenous stratified sampling

not display an uniform behaviour. Indeed, while in some cases its power is larger than that of RESET2, in other cases its power is the lowest of all versions (e.g. when the cauchit is one of the alternative links and the mean of $h(x\theta)$ is not far away from zero). Comparing the RESET and the P tests, we find that in some cases the latter is much more powerful (e.g. H_0: cauchit) but in others it occurs the opposite (e.g. H_0: loglog).

Misspecification of the index function

In Figures 6 and 7 we analyse the power of the tests when some relevant covariates are omitted from the index model. In Figure 6, the omitted variable, x_3, is uncorrelated with the included regressors, being generated as a displaced exponential variate with variance one. In Figure 7, the nonlinear variable x_1^2 is omitted and x_1 is generated as a displaced exponential variate with variance one.[9] In both cases, we set $\theta = (0, 1, 1)$ and compute the percentage of rejections of the null hypothesis for different values of the parameter θ_3 associated to either x_3 or x_1^2. In the former case we consider $\theta_3 = \{0, 0.25, \ldots, 2.5\}$, while in the latter $\theta_3 = \{-0.25, -0.2, \ldots, 0.25\}$.

Again, in general, increasing the number of test variables in the computation of the RESET test diminishes its power. This conclusion is now valid even when RESET1 is included in the comparison. In fact, in these examples, this is the most powerful RESET version in most cases (the only exceptions occur when misspecification is due to the omission of a quadratic term and θ_3 is negative). Unlike the previous experiments, the loss of power resulting from using the RESET test instead of a specific test may be enormous, especially in the case of uncorrelated covariates. Nevertheless, note that even in this case the RESET test is consistent, unlike what happens in linear regression models where it has no power against this type of misspecification. On the other hand, the lower power displayed by the RESET test in the logit case is certainly related to the robustness of this model to the omission of uncorrelated covariates; see Ramalho and Ramalho (2010).[10]

In Figure 8 we consider another type of misspecification of the index model, which is now due to heteroskedasticity of the form $s(x_1, \gamma) = e^{2\gamma x_1}$, with $\gamma = \{0, 0.05, \ldots, 0.3\}$. The conclusions are very similar to those obtained in the previous experiments since an identical ranking of the RESET versions was achieved. The main difference is that now the loss of power relative to the specific test is less important.

Misspecification due to observation problems

Finally, we analyse the power of the RESET alternatives when the misspecification results from some sampling problems. First, in Figure 9, we consider the case of covariate measurement error. We consider a data generating process where only the observation of x_1 is affected by the measurement error u, so the data is generated using $h(x\theta) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$ but estimation is based on $h(x^*\theta) = \theta_0 + \theta_1 x_1^* + \theta_2 x_2$, where $x_1^* = x_1 + u$ and $\theta = (0, 1, 1)$. We

[9] In this case, we cannot generate x_1 as a standard normal variate as we do in all the other experiments carried out in this article. In fact, as noted by a referee, in that case x_1 and x_1^2 would be uncorrelated and, hence, the omitted variable would be uncorrelated with the included predictors in both setups of this section.

[10] Note also that, as some graphs of Figure 6 suggest, the power of the tests decreases for high values of θ_3. See Savin and Wurtz (1999) for an explanation of this peculiar feature of binary response models that arises when the probability of all outcomes zero or all outcomes one approaches the unity.

generate u from a Student-t distribution with five degrees of freedom and consider several values for the variance of the measurement error, $\sigma^2 = \{0, 0.25, \ldots, 2\}$.

In this case, the results obtained are very different from those of previous experiments. Now, the ranking of the RESET tests is completely reversed: inclusion of more test variables gives rise to a more powerful statistic. In particular, the RESET1 version exhibits much lower power than the other variants. Moreover, the power of RESET tests (apart from RESET1) is clearly superior to that of the specific test, which suggests that the test that we are using as benchmark for RESET tests is of poor quality, at least when applied to binary regression models.[11] Note that, unlike all the other cases analysed in this article, this is the only experimental design where the alternative hypothesis underlying the specific test does not correspond exactly to the true data generating process, but merely to the small error variance approximation given by equation (4).

In Figure 10 we analyse two patterns of response misclassification. Again, we set $\theta = (0, 1, 1)$. In the first set of experiments only ones are misclassified as zeros ($\delta_1 = 0$ and $\delta_0 \neq 0$) and in the second the probability of misclassifying a one or a zero is identical ($\delta_0 = \delta_1$). As in the previous case, there is no clear superiority of the specific test relative to the best RESET variants, particularly when the probabilities of misclassification are identical. On the other hand, the characteristics of RESET1 and RESET2 identified in most of the previous experiments are again apparent. Indeed, while RESET2 exhibits in most cases a superior performance relative to alternatives based on a higher number of test variables, RESET1 is sometimes the most powerful test ($\delta_1 = 0$) and other times the least powerful of all RESET versions ($\delta_0 = \delta_1$).

The problem of endogenous sampling is examined in Figure 11. For two different proportions of ones in the population, $Q = 0.5$ and $Q = 0.9$, we simulate cases where the corresponding proportion in the sample, H, takes several values: $H = \{0.1, 0.2, \ldots, 0.9\}$ and $H = \{0.5, 0.55, \ldots, 0.9\}$, respectively. We set $\theta_1 = \theta_2 = 1$ and choose θ_0 in order to produce the values fixed for Q. Naturally, in these final experiments we do not consider the logit case, given its robustness to the problem in analysis. Now, using a higher number of test variables in the computation of the RESET test leads to a decrease of its power. The specific test is clearly the most powerful test but the difference to the best RESET versions is unimportant.

V. Concluding remarks

In this article we examined the ability of several versions of the RESET test to detect various types of misspecification in binary regression models. In terms of size performance, we found that both RESET1 and RESET2 have in general suitable size properties, while the other RESET variants display actual sizes which are too often significantly different from the nominal ones. In terms of power, RESET2 exhibits in all cases but one (covariate

[11] To the best of our knowledge, the test proposed by Chesher (1991) is the only inference procedure sensitive to measurement error that: (i) it is sufficiently general to be applied to any nonlinear model and, consequently, to all binary models considered in this article; and (ii) it does not require additional information on, for example, the variance or the distribution of the measurement error and/or the existence of a validation sample. Our Monte Carlo results suggest that this greater flexibility may compromise the power of the test in such a serious way that it is preferable to apply omnibus tests like the RESET. Clearly, the derivation of more powerful tests for detecting covariate measurement error in binary regression models is an important issue for future research.

measurement error) a superior power performance than other alternatives based on a higher number of test variables. Moreover, even in the case of covariate measurement error, the loss of power of RESET2 relative to the other versions is minimal in most cases. On the other hand, the power behaviour of RESET1 is not uniform at all. Indeed, while in most cases its power is the largest of all RESET versions (e.g. misspecification of the index function, endogenous stratified sampling and some cases of misspecification of the link function and misclassification), in others its power is much lower than the other RESET variants (e.g. other cases of misspecification of the link function and misclassification, covariate measurement error). Overall, our results show that there is no reason for empirical researchers to employ other RESET statistics besides RESET1 or RESET2.

In comparison with tests specifically constructed to assess a particular type of misspecification, the loss of power suffered by RESET1 and RESET2 is very small in many cases (e.g. heteroskedasticity, all sampling problems). The only cases where the loss of power may be substantial occur when the misspecification is due to the omission of covariates, especially when they are uncorrelated with the included regressors, and in some cases of misspecification of the link function. Thus, in the absence of reliable information about a plausible alternative model, RESET1 and RESET2 are clearly good alternatives for testing the specification of binary regression models.

Given that the power performance of the RESET1 and RESET2 statistics is often very distinct, it would be very useful to have a single RESET statistic combining the sometimes very powerful performance of RESET1 with the more uniform behaviour of RESET2. There is an area of econometrics where the issue of combining different versions of one test into a single statistic is frequently addressed. Indeed, when a nuisance parameter is present only under the alternative hypothesis, as each value of the nuisance parameter gives rise to a different test statistic, it is usual to use a single test statistic that summarizes the information provided by all possible test versions according to a suitable criterion (e.g. the supremum of the test variants); see *inter alia* Andrews and Ploberger (1994) and Hansen (1996). As the choice of the number of powers to include in the RESET procedure may be seen as an analogous problem to that of the choice of an arbitrary value for the nuisance parameter, we are currently examining the use of supremum-type RESET tests. Some preliminary Monte Carlo analysis revealed a very promising finite sample performance for a bootstrap-based supremum-RESET test.

Another approach for combining variants of general specification tests into a single statistic is that proposed by Aerts, Claeskens and Hart (1999). These authors developed a test statistic that, similarly to the RESET case, uses sequences of nested orthogonal series estimators to detect departures from the null model but, in contrast to RESET tests, does not require the number of terms used in the approximation to be set *a priori*, being defined by some model selection criteria (e.g. the Akaike Information Criterion). Aerts *et al.* (1999) were able to derive the asymptotic distribution of their test statistic, which may be an important advantage relative to the application of supremum-type RESET statistics. Indeed, our preliminary research suggests that bootstrap methods will be typically required to approximate the distribution of the supremum statistics. However, a problem with the tests proposed by Aerts *et al.* (1999), which explains why, to the best of our knowledge, they have never been applied in the econometrics literature, is that there is no natural way to choose the sequence of nested models required to implement the test when

the base model has more than one covariate (the expansion is based on x and not on $x\hat{\theta}$ as in the RESET case). Hence, an effective comparison of the performance of their test with some supremum variant of the RESET test will also imply the investigation of what kind of sequences deliver best power properties for the Aerts *et al.* (1999) test.

Another avenue for future research is the possibility of using different expansions in the construction of the RESET test. In fact, the test by Aerts *et al.* (1999) is based on Fourier instead of polynomial expansions. However, in the RESET case, to the best of our knowledge, only DeBenedictis and Giles (1998, 1999) have considered such hypothesis, proposing a Fourier-based RESET test. In a small Monte Carlo simulation study, they found promising results for their RESET version in the linear regression framework. Given the limited evidence provided so far, none of which is for binary parametric models, the investigation of the finite sample performance of such RESET variant is clearly another interesting research topic.

Final Manuscript Received: December 2010

References

Aerts, M., Claeskens, G. and Hart, J. D. (1999). 'Testing the fit of a parametric function', *Journal of the American Statistical Association*, Vol. 94, pp. 869–879.

Andrews, D. W. K. and Ploberger, W. (1994). 'Optimal tests when a nuisance parameter is present only under the alternative', *Econometrica*, Vol. 62, pp. 1383–1414.

Chesher, A. (1991). 'The effect of measurement error', *Biometrika*, Vol. 78, pp. 451–462.

Copas, J. B. (1988). 'Binary regression models for contaminated data', *Journal of the Royal Statistical Society B*, Vol. 50, pp. 225–265.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*, 2nd edn. Chapman and Hall, London.

Davidson, R. and MacKinnon, J. G. (1981). 'Several tests for model specification in the presence of alternative hypotheses', *Econometrica*, Vol. 49, pp. 781–793.

Davidson, R. and MacKinnon, J. G. (1984). 'Convenient specification tests for logit and probit models', *Journal of Econometrics*, Vol. 25, pp. 241–262.

DeBenedictis, L. F. and Giles, D. E. A. (1998). 'Diagnostic testing in econometrics: variable addition, RESET, and Fourier approximations', in Ullah A. and Giles D. E. A. (eds), *Handbook of Applied Economic Statistics*, Marcel Dekker, New York, pp. 383–417.

DeBenedictis, L. F. and Giles, D. E. A. (1999). 'Robust specification testing in regression: the FRESET test and autocorrelated disturbances', *Journal of Quantitative Economics*, Vol. 15, pp. 67–75.

Godfrey, L. G. and Orme, C. D. (1994). 'The sensitivity of some general checks to omitted variables in the linear model', *International Economic Review*, Vol. 35, pp. 489–506.

Hansen, B. E. (1996). 'Inference when a nuisance parameter is not identified under the null hypothesis', *Econometrica*, Vol. 64, pp. 413–430.

Hatzinikolaou, D. and Stavrakoudis, A. (2006). 'Empirical size and power of some diagnostic tests applied to a distributed lag model', *Empirical Economics*, Vol. 31, pp. 631–643.

Hausman, J. A., Abrevaya, F. and Scott-Morton, F. M. (1998). 'Misclassification of the dependent variable in a discrete-response setting', *Journal of Econometrics*, Vol. 87, pp. 239–269.

Imbens, G. (1992). 'An efficient method of moments estimator for discrete choice models with choice-based sampling', *Econometrica*, Vol. 60, pp. 1187–1214.

Leung, S. F. and Yu, S. (2000). 'How effective are the RESET tests for omitted variables', *Communications in Statistics – Theory and Methods*, Vol. 29, pp. 879–902.

Manski, C. F. and McFadden, D. (1981). 'Alternative estimators and sample designs for discrete choice analysis', in Manski C. F. and McFadden D. (eds), *Structural Analysis of Discrete Data with Econometric Applications*, MIT Press, Cambridge, MA, pp. 2–50.

Pagan, A. and Vella, F. (1989). 'Diagnostic tests for models based on individual data: a survey', *Journal of Applied Econometrics*, Vol. 4, S29–S59.
Peters, S. (2000). 'On the use of the RESET test in microeconometric models', *Applied Economics Letters*, Vol. 7, pp. 361–365.
Ramalho, E. A. and Ramalho, J. J. S. (2010). 'Is neglected heterogeneity really an issue in binary and fractional models? A simulation exercise for logit, probit and loglog models', *Computational Statistics and Data Analysis*, Vol. 54, pp. 987–1001.
Ramalho, E. A. and Smith, R. J. (2011). 'Discrete Choice Nonresponse', *Review of Economic Studies* (forthcoming).
Ramsey, J. B. (1969). 'Tests for specification errors in classical linear least-squares regression analysis', *Journal of the Royal Statistical Society B*, Vol. 31, pp. 350–371.
Ramsey, J. B. and Gilbert, R. F. (1972). 'A Monte Carlo study of some small sample properties of tests for specification error', *Journal of the American Statistical Association*, Vol. 67, pp. 180–186.
Santos Silva, J. M. C. (2001). 'A score test for non-nested hypothesis with applications to discrete data models', *Journal of Applied Econometrics*, Vol. 16, pp. 577–597.
Savin, N. E. and Wurtz, A. H. (1999). 'Power of tests in binary response models', *Econometrica*, Vol. 67, pp. 413–421.
Thomas, J. M. (1993). 'On testing the logistic assumption in binary dependent variable models', *Empirical Economics*, Vol. 18, pp. 381–392.
Thursby, J. G. and Schmidt, P. (1977). 'Some properties of tests for specification error in a linear regression model', *Journal of the American Statistical Association*, Vol. 72, pp. 635–641.
White, H. (1982). 'Maximum likelihood estimation of misspecified models', *Econometrica*, Vol. 50, pp. 1–25.
Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.
Yatchew, A. and Griliches, Z. (1985). 'Specification error in probit models', *Review of Economics and Statistics*, Vol. 67, pp. 134–139.