Computational Statistics and Data Analysis 54 (2010) 987–1001

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Is neglected heterogeneity really an issue in binary and fractional regression models? A simulation exercise for logit, probit and loglog models

Esmeralda A. Ramalho, Joaquim J.S. Ramalho *

Department of Economics, Universidade de Evora, Largo dos Colegiais 2, 7000-803 EVORA, Portugal
CEFAGE-UE, Universidade de Evora, Largo dos Colegiais 2, 7000-803 EVORA, Portugal

ARTICLE INFO

Article history:
Received 6 November 2008
Received in revised form 3 October 2009
Accepted 12 October 2009
Available online 21 October 2009

ABSTRACT

Theoretical and simulation analysis is performed to examine whether unobserved heterogeneity independent of the included regressors is really an issue in logit, probit and loglog models with both binary and fractional data. It is found that unobserved heterogeneity has the following effects. First, it produces an attenuation bias in the estimation of regression coefficients. Second, although it is innocuous for logit estimation of average sample partial effects, it may generate biased estimation of those effects in the probit and loglog models. Third, it has much more deleterious effects on the estimation of population partial effects. Fourth, it is only for logit models that it does not substantially affect the prediction of outcomes. Fifth, it is innocuous for the size of Wald tests for the significance of observed regressors but, in small samples, it substantially reduces their power.

1. Introduction

In economics, researchers are often interested in explaining a limited dependent variable, Y, as a function of a set of explanatory variables, X. Because of the bounded nature of the variable of interest, linear specifications often provide an inadequate description of the conditional mean of Y, $E(Y|X)$, since no restriction is imposed on the range of values taken by the predicted outcome. Moreover, when interest lies in the conditional probability of Y, $\Pr(Y|X)$, nonlinear models are typically used. While the omission of relevant explanatory variables that are independent of the included regressors is relatively innocuous in linear models, it generally causes inconsistency in the estimation of the parameters of interest in nonlinear models (see *inter alia* Gourieroux, 2000, pp. 32–33). In this paper we examine the consequences of the presence of that type of unobserved heterogeneity in logit, probit and loglog models for binary and fractional or proportionate data.

To the best of our knowledge, there are very few studies on the consequences of unobserved heterogeneity in binary and fractional regression models. Moreover, the few studies undertaken have assumed restrictive conditions or consider only the effects of neglected heterogeneity on particular aspects of those models. For example, Lee (1982) derived conditions under which the omission of an orthogonal explanatory variable would not cause bias in the estimation of the remaining parameters of a binary logit model. However, those conditions are too stringent to be of practical use. Yatchew and Griliches (1985) showed that for a binary probit model with a normally distributed omitted variable, the estimators for the parameters of the included variables suffer from attenuation bias. Wooldridge (2002, 2005), under similar assumptions, demonstrated that this bias does not affect the consistent estimation of the partial effect of the observed regressors on the outcome. Cramer (2003, 2007) considered the binary logit model and proved formally that the same bias attenuation would occur

* Corresponding author at: Department of Economics, Universidade de Evora, Largo dos Colegiais 2, 7000-803 EVORA, Portugal. Tel.: +351 266 740 894; fax: +351 266 740 807.
E-mail addresses: jsr@uevora.pt, jjs.ramalho@gmail.com (J.J.S. Ramalho).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2009.10.012

in this context if the distribution of the omitted variables is such that their relegation to the disturbance term of the latent regression underlying the logit model does not change its logistic distribution, which is also a strong assumption. However, Cramer (2007) also presents a small simulation study that reveals that a particular partial effect, the average sample effect, is quite insensitive to the inconsistency of the parameters of interest, even when the logit shape of the conditional distribution is severely affected. In the context of generalized linear models (which include the models analysed in this paper as particular cases), Neuhaus and Jewell (1993) restricted their attention to the case of a single observed covariate.

Given that the estimation of partial effects is often the main aim of empirical work and given that in nonlinear models the analysis of the magnitude of regression coefficients is not relevant *per se*, both Wooldridge (2002) and Cramer (2007) suggest that, similarly to what happens in linear models, unobserved heterogeneity is not an important issue in, respectively, binary probit and logit models. However, it is not clear whether the robustness of the binary logit model revealed by the simulation study of Cramer (2007) extends to the binary probit model (or, in fact, to any other binary or fractional model) because no similar analysis has been carried out for the latter model. Moreover, there are other quantities of interest in empirical work that have not been considered by those authors. One example is outcome prediction, which is relevant not only for the analysis of binary and fractional data but also in the estimation of multi-part models that require binary outcome prediction in the first stage. Testing the significance of the observed covariates is clearly another relevant issue for practitioners.

In order to address these questions, we consider the theoretical framework of Wooldridge (2002) and Cramer (2007) and extend their results to other quantities of interest and models. However, given that a more general theoretical approach does not seem to be feasible, we also conduct an extensive Monte Carlo study that extends the findings of the cited papers in several directions. On the one hand, in addition to the binary logit and probit models, we also consider an alternative asymmetric specification, the loglog model, and, in each case, both binary outcomes, where interest lies in modelling $\Pr(Y|X)$, and fractional responses, where the main purpose is modelling $E(Y|X)$; see Papke and Wooldridge (1996) for a seminal paper on the so-called fractional regression model, see Simas et al. (forthcoming) for alternative models for fractional responses and see Ramalho et al. (forthcoming) for a comprehensive survey of this subject. On the other hand, we examine the consequences of neglected heterogeneity over the performance of standard estimators for those models at various levels: (i) the magnitude and direction of the parameters of interest; (ii) the two common forms of calculating partial effects considered separately by Wooldridge (2002) and Cramer (2007); (iii) the prediction of outcomes; and (iv) the size and power of Wald tests for the significance of the included regressors. In all cases, we consider several patterns of neglected heterogeneity by assuming various alternative distributions for the omitted variables and assigning different weights to their relative importance.

This paper is organized as follows. In Section 2 we establish the framework of the paper, discussing analytically the consequences of neglected heterogeneity in binary regression models. The Monte Carlo study used to assess the performance of naive estimators in both binary and fractional regression models is carried out in Section 3. Section 4 concludes the paper.

2. Framework

Consider a random sample of $i = 1, \ldots, N$ individuals, let Y be the binary or fractional variable of interest, defined, respectively, as $Y = \{0, 1\}$ and $Y \in [0, 1]$, and let X_1 and X_2 be, respectively, k_1- and k_2-vectors of explanatory variables. We denote by θ_1 and θ_2 the k_1- and k_2-vectors of parameters associated with X_1 and X_2, respectively, and assume that there are no relevant explanatory variables other than those included in X_1 and X_2, that X_1 contains an intercept term, that X_2 is not observed and that X_1 and X_2 are independent. We also assume that

$$E\left(Y|X_1 = x_1, X_2 = x_2\right) = G\left(x_1\theta_1 + x_2\theta_2\right), \tag{1}$$

where $G(x\theta)$ is defined as $e^{x\theta}/(1+e^{x\theta})$, $\Phi(x\theta)$, and $e^{e^{-x\theta}}$ for, respectively, the logit, probit, and loglog models. Note that in the binary case $G(\cdot)$ also equals $\Pr(Y = 1|X_1 = x_1, X_2 = x_2)$.

2.1. The effects of neglected heterogeneity on parameter estimation

By a simple application of the law of iterated expectations, it follows that

$$E\left(Y|X_1\right) = E_{X_2}\left[G\left(x_1\theta_1 + x_2\theta_2\right)\right] = \int_{\mathcal{X}_2} G\left(x_1\theta_1 + x_2\theta_2\right) f_{X_2}\left(x_2\right) \mathrm{d}x_2, \tag{2}$$

where $\mathcal{X}_2$ and $f_{X_2}(x_2)$ denote, respectively, the sample space and the marginal distribution of X_2. Because, in general, $E(Y|X_1) \neq G(x_1\theta_1)$, naive estimation based on $G(x_1\theta_1)$ will not produce consistent estimators of θ_1. In fact, omission of X_2 tends to bias θ_1 towards zero, as shown by Yatchew and Griliches (1985) and Wooldridge (2002) for a particular binary probit model, as shown by Cramer (2007) for a specific binary logit model, and as shown by Neuhaus and Jewell (1993) for any generalized linear model based on a log concave density function (which binary and fractional logit, probit and loglog models exhibit) with a single observed covariate. However, as we show next by retracing the arguments of Yatchew and Griliches (1985), Wooldridge (2002) and Cramer (2007), it is not possible to prove formally that this attenuation effect is the consequence of neglected heterogeneity under any circumstances.

For simplicity, consider the following latent regression equation:

$$y^* = x_1\beta_1 + x_2\beta_2 + u, \tag{3}$$

where y^* is not observed, x_1 includes a unit variable, x_2 contains a single explanatory variable that is uncorrelated with x_1 and u is a random disturbance that is uncorrelated with the regressors. Instead of observing y^*, we observe the binary variable y, which takes the value 1 if $y^* > 0$ and takes the value 0 otherwise. Assume that u has a mean of zero and a variance of σ_u^2 and let its standardized distribution be H. When X_2 is observed, we have:

$$\begin{aligned} E\left(Y|X_1, X_2\right) &= \Pr\left(Y = 1|X_1, X_2\right) \\ &= \Pr\left(u > -x_1\beta_1 - x_2\beta_2|X_1, X_2\right) \\ &= 1 - \Pr\left(u \le -x_1\beta_1 - x_2\beta_2|X_1, X_2\right) \\ &= 1 - H\left(-x_1\frac{\beta_1}{\sigma_u} - x_2\frac{\beta_2}{\sigma_u}\right) \\ &= G\left(x_1\frac{\beta_1}{\sigma_u} + x_2\frac{\beta_2}{\sigma_u}\right), \end{aligned} \tag{4}$$

where $G(\cdot)$ is the complementary function of $H(\cdot)$. When u has a symmetric distribution, $G(\cdot) \equiv H(\cdot)$. It is well known that the parameters β_1 and β_2 are not separately identified from σ_u. Let $\theta_1 = \beta_1/\sigma_u$.

Assume now that X_2 is not observed and has a mean of zero and a variance of τ^2. Then, the composite error $u^* = x_2\beta_2 + u$ is independent of x_1 and has a variance of $\sigma_{u^*}^2 = \beta_2^2\tau^2 + \sigma_u^2$. Denote the standardized distribution of u^* by H^*. In this setting, it follows that:

$$\begin{aligned} E\left(Y|X_1\right) &= \Pr\left(Y = 1|X_1\right) \\ &= \Pr\left(u^* > -x_1\beta_1|X_1\right) \\ &= 1 - \Pr\left(u^* \le -x_1\beta_1|X_1\right) \\ &= 1 - H^*\left(-x_1\frac{\beta_1}{\sigma_{u^*}}\right) \\ &= G^*\left(x_1\frac{\beta_1}{\sigma_{u^*}}\right). \end{aligned} \tag{5}$$

Let $\theta_1^* = \beta_1/\sigma_{u^*}$. Clearly, one cannot evaluate the effects on parameter estimation of omitting X_2 unless one assumes that $H = H^*$, i.e. the distribution of X_2 must be such that its inclusion in the error term does not change the distribution of the disturbance. If we make this assumption, then $G = G^*$ and, by comparing (4) and (5), we find that:

$$\theta_1^* = \frac{\sigma_u}{\sigma_{u^*}}\theta_1. \tag{6}$$

Given that $\sigma_{u^*} > \sigma_u$ (unless $\beta_2 = 0$ or $\tau^2 = 0$), in general $|\theta_1^*| < |\theta_1|$, which implies that, under the assumptions made, the omission of an explanatory variable produces an attenuation bias in the estimation of the coefficients of the observed covariates.

In this proof, the crucial assumption is that $H = H^*$, which researchers in this field typically make, with the exception of Neuhaus and Jewell (1993), who, however, use a geometric approach that applies only to models with a single observed covariate. Indeed, both Yatchew and Griliches (1985) and Wooldridge (2002, 2005) assumed that both u and X_2 are normally distributed, which implies that u^* also has a normal distribution. On the other hand, in his proof of the existence of an attenuation bias in the logit model, rather than specifying the distribution of X_2, Cramer (2007) assumed that both u and u^* had a logistic distribution. However, in practice, it is extremely unlikely that $H = H^*$. Moreover, for fractional regression models, which cannot be written in latent form, finding a corresponding similar proof does not seem feasible. Therefore, in the Monte Carlo simulation study carried out in the next section, we investigate whether Eq. (6), which applies only to specific binary regression models, holds approximately for cases in which $H \neq H^*$ and for fractional regression models.

2.2. The effects of neglected heterogeneity on partial effects

For empirical analysis based on nonlinear models, the main focus is not on the analysis of the magnitude of the regression coefficients, but on consistent estimation of the partial effects. In applied work, the two standard measures of partial effects in nonlinear models are the average sample effect (*ASE*), which is the mean of the partial effects calculated independently for each individual in the sample, and the population partial effect (*PPE*), which is calculated for specific values of the covariates. As discussed in detail by Wooldridge (2002), in the presence of neglected heterogeneity, of interest are the partial effects averaged across the population distribution of the omitted variables.

Consider again the model described by (1) and assume that X_2 is not observed. In this setting, for the covariate x_{1j}, those partial effects are defined by

$$ASE = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial E\left(Y_i|X_{1i}\right)}{\partial x_{1j}} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial E_{X_2}\left[G\left(x_{1i}\theta_1 + x_{2i}\theta_2\right)\right]}{\partial x_{1j}} \tag{7}$$

and, with evaluation at a given point $X_1 = \bar{x}_1$ (e.g. the mean of the observed regressors), by:

$$PPE = \frac{\partial E\left(Y|X_1 = \bar{x}_1\right)}{\partial x_{1j}} = \frac{\partial E_{X_2}\left[G\left(\bar{x}_1\theta_1 + x_2\theta_2\right)\right]}{\partial x_{1j}}. \tag{8}$$

Because both effects depend on X_2, the following naive estimators should be inconsistent:

$$\widehat{ASE^n} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial G\left(x_{1i}\hat{\theta}_1^n\right)}{\partial x_{1j}} \tag{9}$$

and

$$\widehat{PPE^n} = \frac{\partial G\left(\bar{x}_1\hat{\theta}_1^n\right)}{\partial x_{1j}}, \tag{10}$$

where $\hat{\theta}_1^n$ denotes the naive estimator of θ_1. This is because $\hat{\theta}_1^n$ is inconsistent and because $G(\cdot)$ is in general misspecified. However, when $H = H^*$ both $\widehat{ASE^n}$ and $\widehat{PPE^n}$ provide consistent estimates of *ASE* and *PPE*, respectively. To see this, consider again the example discussed in the previous section. Given (2) and (5), it follows that for binary regression models:

$$E\left(Y|X_1\right) = E_{X_2}\left[G\left(x_1\theta_1 + x_2\theta_2\right)\right] = G^*\left(x_1\frac{\beta_1}{\sigma_{u^*}}\right). \tag{11}$$

Hence:

$$PPE = \frac{\partial E_{X_2}\left[G\left(\bar{x}_1\theta_1 + x_2\theta_2\right)\right]}{\partial x_{1j}} = \frac{\partial G^*\left(\bar{x}_1\frac{\beta_1}{\sigma_{u^*}}\right)}{\partial x_{1j}}. \tag{12}$$

Therefore, as when $H = H^*$, $G = G^*$ and $\hat{\theta}_1^n$ converges to $\theta_1^* = \beta_1/\sigma_{u^*}$, it follows that, under this assumption, $\widehat{PPE^n}$ is a consistent estimator for *PPE*. A similar proof can be applied to *ASE*.

Using similar arguments, Wooldridge (2002) was the first to demonstrate that, in the binary probit model with a normally distributed omitted variable, the bias in the estimation of θ_1 does not carry over to estimation of the *PPE*. Cramer (2007) showed that the same conclusion holds for logit models when the logit shape of $E(Y|X_1, X_2)$ of (1) is preserved in $E(Y|X_1)$ of (2). Both these findings are supported by Stoker (1986), who showed that misspecification of the functional form in single index models does not affect the estimation of average behavioural derivatives. By simulation, Cramer (2007) also showed that, for logit models, even when $E(Y|X_1)$ deviates significantly from the logit functional form assumed for $E(Y|X_1, X_2)$, the *ASE* is relatively robust to neglected heterogeneity. In Section 3 we investigate whether this robustness of naive partial effects extends to other models and applies in more general settings.

2.3. The effects of neglected heterogeneity on predicted outcomes

We also examine whether naive predictions of $E(Y|X_1)$ or $\Pr(Y|X_1)$, based on the misspecified functional form $G(x_1\theta_1)$ evaluated at the inconsistent estimator $\hat{\theta}_1^n$, are reliable. Existing studies have not addressed this issue. However, outcome prediction, besides being a relevant matter *per se*, is also the basis for the estimation of partial effects in multi-part models, the first stage of which typically requires the estimation of a binary model. Because X_2 is not observed, as in the case of the partial effects discussed above, the main interest is outcome prediction averaged across the population distribution of the omitted variables.

From (11), the assumptions required for consistent estimation of partial effects are clearly still needed: only if $H = H^*$ does $G(\bar{x}_1\hat{\theta}_1^n)$ consistently predict $E(Y|X_1)$. Therefore, in a probit model with normally distributed heterogeneity or in the special logit model considered by Cramer (2007), neglected heterogeneity is not a problem for outcome prediction. In our Monte Carlo study, we focus on cases in which $H \neq H^*$.

2.4. The effects of neglected heterogeneity on Wald tests

Because testing the significance of the impact of a particular covariate, say X_{1j}, on the outcome variable is one of the main aims of any empirical study, we next evaluate the effects of neglected heterogeneity on significance tests. In particular, we examine the application of the widely used Wald test to assess the individual significance of the parameters on the observed regressors in the presence of unobserved heterogeneity. The extension to Wald tests of the joint significance of those parameters is straightforward.

When there are no omitted variables, the Wald statistic for assessing $H_0 : \theta_{1j} = 0$ is given by $W = \hat{\theta}_{1j}/\sqrt{\hat{V}(\hat{\theta}_{1j})}$, where $\hat{V}(\hat{\theta}_{1j})$ denotes an estimate of the variance of $\hat{\theta}_{1j}$, and converges to a standard normal distribution. Let $g(z) = \partial G(z)/\partial z$

and

$$a_i = \frac{g\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)^2}{G\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)\left[1 - G\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)\right]}. \tag{13}$$

For binary data, given model (1), and because we are assuming that X_1 and X_2 are independent, a consistent estimator of the covariance matrix of $\hat{\theta}$ is given by:

$$\hat{V}\left(\hat{\theta}\right) = \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} a_i x'_{1i} x_{1i} & 0 \\ 0 & \frac{1}{N}\sum_{i=1}^{N} a_i x'_{2i} x_{2i} \end{bmatrix}^{-1}. \tag{14}$$

Therefore:

$$\hat{V}\left(\hat{\theta}_1\right) = \left(\frac{1}{N}\sum_{i=1}^{N} a_i x'_{1i} x_{1i}\right)^{-1} \tag{15}$$

and

$$\hat{V}\left(\hat{\theta}_{1j}\right) = \left[\frac{1}{N}\sum_{i=1}^{N} a_i x_{1ji}^2 - \frac{1}{N}\sum_{i=1}^{N} a_i x_{1ji} x_{1'i}\left(\frac{1}{N}\sum_{i=1}^{N} a_i x'_{1'i} x_{1'i}\right)^{-1}\frac{1}{N}\sum_{i=1}^{N} a_i x_{1ji} x'_{1'i}\right]^{-1}, \tag{16}$$

where $x_{1'i}$ excludes x_{1ji} from x_{1i}. Similarly, when X_2 is omitted, the naive significance test that X_{1j} has no effect is given by $W^n = \hat{\theta}_{1j}^n / \sqrt{\hat{V}\left(\hat{\theta}_{1j}^n\right)}$, where:

$$\hat{V}\left(\hat{\theta}_{1j}^n\right) = \left[\frac{1}{N}\sum_{i=1}^{N} a_i^n x_{1ji}^2 - \frac{1}{N}\sum_{i=1}^{N} a_i^n x_{1ji} x_{1'i}\left(\frac{1}{N}\sum_{i=1}^{N} a_i^n x'_{1'i} x_{1'i}\right)^{-1}\frac{1}{N}\sum_{i=1}^{N} a_i^n x_{1ji} x'_{1'i}\right]^{-1} \tag{17}$$

and $a_i^n = g(x_{1i}\hat{\theta}_1^n)^2 / G(x_{1i}\hat{\theta}_1^n)[1 - G(x_{1i}\hat{\theta}_1^n)]$.

Under the assumptions made previously, i.e. that the distribution of the neglected heterogeneity is such that $H = H^*$, there is a case, $\theta_{1j} = 0$, in which neglected heterogeneity does not generate any bias. Indeed, in such a case the existence of an attenuation bias implies that both $\hat{\theta}_{1j}$ and $\hat{\theta}_{1j}^n$ are consistent estimators of θ_{1j} and, therefore, the size of any significance test should remain unaffected by unobserved heterogeneity; see also Lagakos and Schoenfeld (1984), who discuss this issue in the context of score tests in proportional-hazards regression models in which the coefficient being tested for significance is attached to a binary variable. Subsequently, we examine by simulation the consequences of neglected heterogeneity for the size of Wald tests when $H \neq H^*$.

Lagakos and Schoenfeld (1984) showed that the power of a score significance test for a binary included variable may be substantially lower in the presence of omitted covariates. In our framework, we suspect that neglected heterogeneity may cause some power loss in the application of the Wald test. In fact, although no general power comparison between W and W^n seems to be feasible, in the special case of the logit model such a comparison is straightforward, provided that we assume again that $H = H^*$. Indeed, for this model it is well known that $g(z) = G(z)[1 - G(z)]$, which implies that the quantities a_i and a_i^n simplify to $g(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2)$ and $g(x_{1i}\hat{\theta}_1^n)$, respectively. Because $\partial G(x\theta)/\partial x_{1j} = \theta_{1j} g(x\theta)$, in the logit model we have

$$\begin{aligned} W &= \hat{\theta}_{1j}\sqrt{\hat{V}\left(\hat{\theta}_{1j}\right)^{-1}} \\ &= \hat{\theta}_{1j}\left[\frac{1}{N}\sum_{i=1}^{N} a_i x_{1ji}^2 - \frac{1}{N}\sum_{i=1}^{N} a_i x_{1ji} x_{1'i}\left(\frac{1}{N}\sum_{i=1}^{N} a_i x'_{1'i} x_{1'i}\right)^{-1}\frac{1}{N}\sum_{i=1}^{N} a_i x_{1ji} x'_{1'i}\right]^{0.5} \\ &= \sqrt{\hat{\theta}_{1j}}\left\{\frac{1}{N}\sum_{i=1}^{N}\frac{\partial G\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)}{\partial x_{1j}} x_{1ji}^2 - \frac{1}{N}\sum_{i=1}^{N}\frac{\partial G\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)}{\partial x_{1j}} x_{1ji} x_{1'i}\right. \\ &\quad \left. \times\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial G\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)}{\partial x_{1j}} x'_{1'i} x_{1'i}\right]^{-1}\frac{1}{N}\sum_{i=1}^{N}\frac{\partial G\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)}{\partial x_{1j}} x_{1ji} x'_{1'i}\right\}^{0.5}. \end{aligned} \tag{18}$$

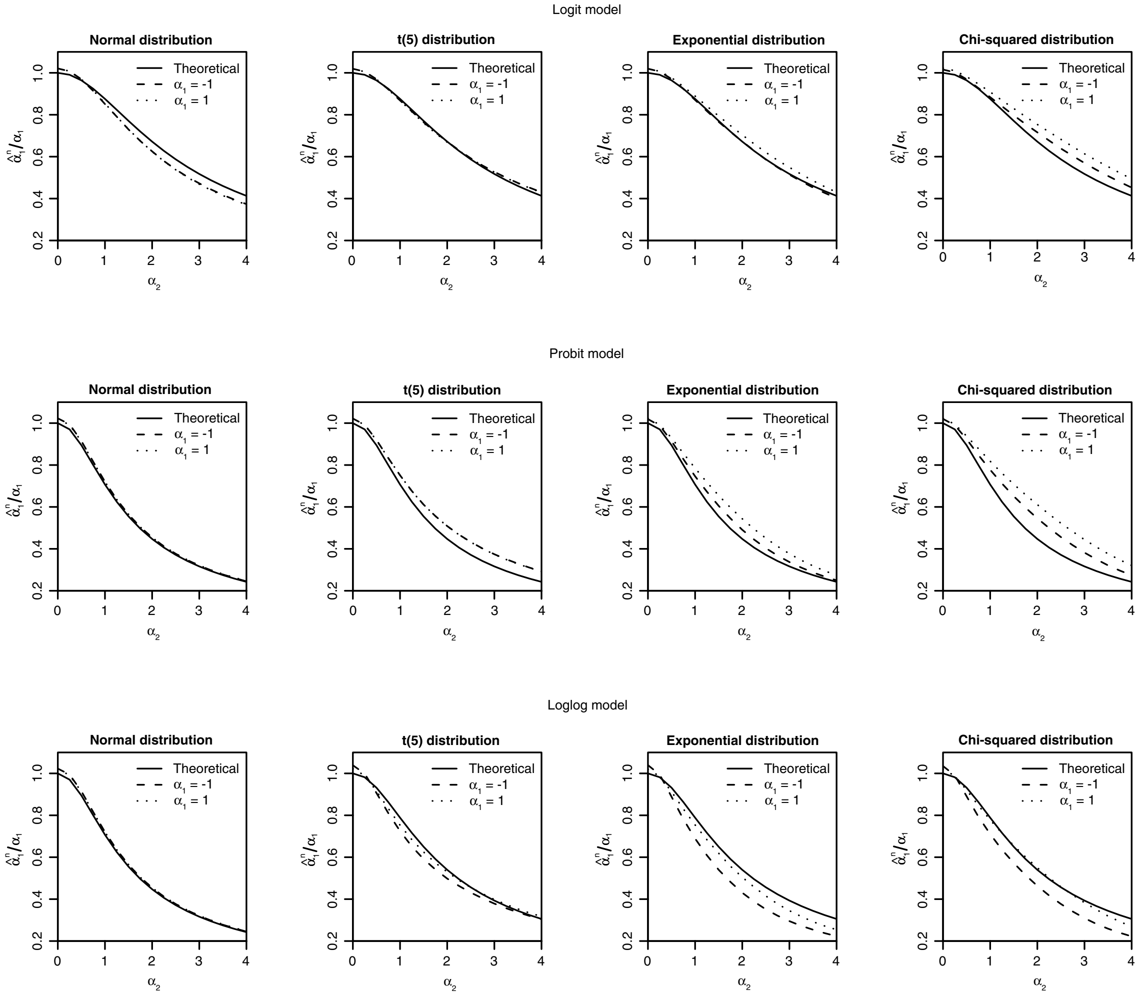

Fig. 1. Attenuation bias of parameter estimates in binary regression models.

Similarly, we have:

$$W^n = \sqrt{\hat{\theta}_{1j}^n} \left\{ \frac{1}{N} \sum_{i=1}^{N} \frac{\partial G\left(x_{1i}\hat{\theta}_{1j}^n\right)}{\partial x_{1j}} x_{1ji}^2 - \frac{1}{N} \sum_{i=1}^{N} \frac{\partial G\left(x_{1i}\hat{\theta}_{1j}^n\right)}{\partial x_{1j}} x_{1ji} x_{1'i} \right.$$
$$\left. \times \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\partial G\left(x_{1i}\hat{\theta}_{1j}^n\right)}{\partial x_{1j}} x_{1'i}' x_{1'i} \right]^{-1} \frac{1}{N} \sum_{i=1}^{N} \frac{\partial G\left(x_{1i}\hat{\theta}_{1j}^n\right)}{\partial x_{1j}} x_{1ji} x_{1'i}' \right\}^{0.5}. \tag{19}$$

Thus, because both $\partial G(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2)/\partial x_{1j}$ and $\partial G(x_{1i}\hat{\theta}_{1j}^n)/\partial x_{1j}$ converge to the same quantity, $\partial E_{X_2}[G(x_1\theta_1 + x_2\theta_2)]/\partial x_{1j}$, see (12), and $\hat{\theta}_{1j}$ and $\hat{\theta}_{1j}^n$ converge to θ_{1j} and θ_{1j}^*, respectively, it follows from (6), (18) and (19) that:

$$\frac{W^n}{W} = \sqrt{\frac{\hat{\theta}_{1j}^n}{\hat{\theta}_{1j}}} \rightarrow \sqrt{\frac{\sigma_u}{\sigma_{u^*}}}. \tag{20}$$

Hence, assuming $H = H^*$, in a logit model the naive Wald test W^n is smaller than W by a factor given by the square root of the attenuation factor that relates $\hat{\theta}_1^n$ to $\hat{\theta}_1$. This implies that, in small samples, unobserved heterogeneity may reduce the power of Wald tests.

In the Monte Carlo study that follows, we investigate the size and power properties of naive Wald statistics under general patterns of heterogeneity.

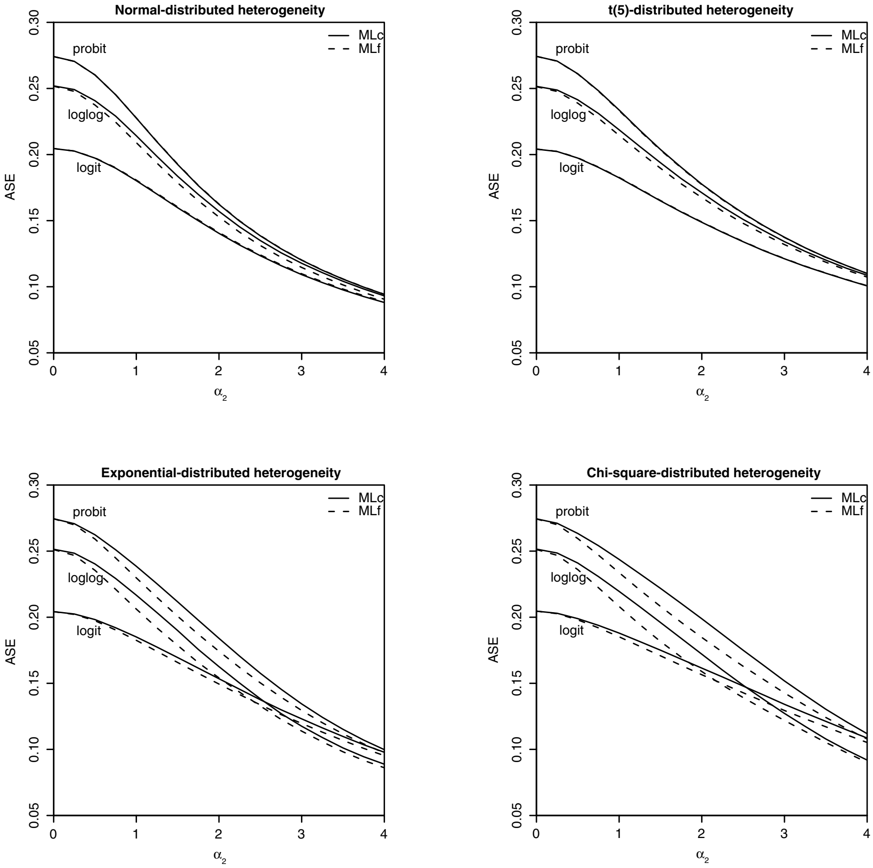

Fig. 2. Average sample effects for binary regression models ($\alpha_1 = 1$).

3. A Monte Carlo simulation study

In this section we present an extensive Monte Carlo simulation study for binary and fractional logit, probit and loglog models. All experiments are based on a simple two-variable equation given by:

$$E\left(Y|X_1, X_2\right) = G\left(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2\right), \tag{21}$$

where $\alpha_0 = 0$, α_2 ranges from 0 to 4 in steps of 0.25 and α_1 takes different values across the different experiments. Our aim is to analyse the effects of omitting X_2 on the estimation of α_1 and related statistics. Note that $\alpha_2 = 0$ corresponds to the case in which there is no neglected heterogeneity and note also that larger values of α_2 imply a larger amount of heterogeneity.

In all experiments, X_1 is generated from a mixture of normal distributions, where the variate is $N(-1, 1)$ with probability 0.7 and $N(2.333, 1)$ with probability 0.3, and X_2 is generated from the $\mathcal{N}(0, 1)$, t_5, *Exponential*(1) and $\chi^2_{(1)}$ distributions. Both variables are scaled to have means of zero and unit variances. The choice of an asymmetric distribution for X_1 was made to avoid the reflection property about the origin that would affect the sampling distributions of the estimators of α_1; see Chesher and Peters (1994) and Chesher (1995) for a discussion of the design of Monte Carlo simulation studies.

We generate Y as a Bernoulli (binary case) or a beta (fractional case) variate with a mean given by the logit, probit or loglog functional form, with the shape parameter of the beta distribution fixed at 1; see *inter alia* Ramalho et al. (forthcoming) for the mean-dispersion parametrization of the beta distribution used in the generation of data. In the former case, the parameters of interest are estimated by maximum likelihood (ML), while in the latter we use the quasi-maximum likelihood (QML) method, both of which are the standard ways of dealing with the respective types of data. In both cases, we estimate *full* and *curtailed* versions of the models, i.e. models *with* and *without* X_2. Because the full version of the model yields consistent estimators for all the quantities of interest, this model is used as a reference to evaluate the consequences of neglected heterogeneity.

All experiments were repeated 5000 times using the statistical package *R* and, given the substantial volume of results produced in each experiment, we summarize them in figures. In all cases except for the experiments regarding the Wald tests, given the similarity of the results obtained, we only report those relative to binary models; full results are available from the authors on request. Apart from the last experiment, in which several samples sizes were used, in all the remaining cases, the sample size is $N = 200$.

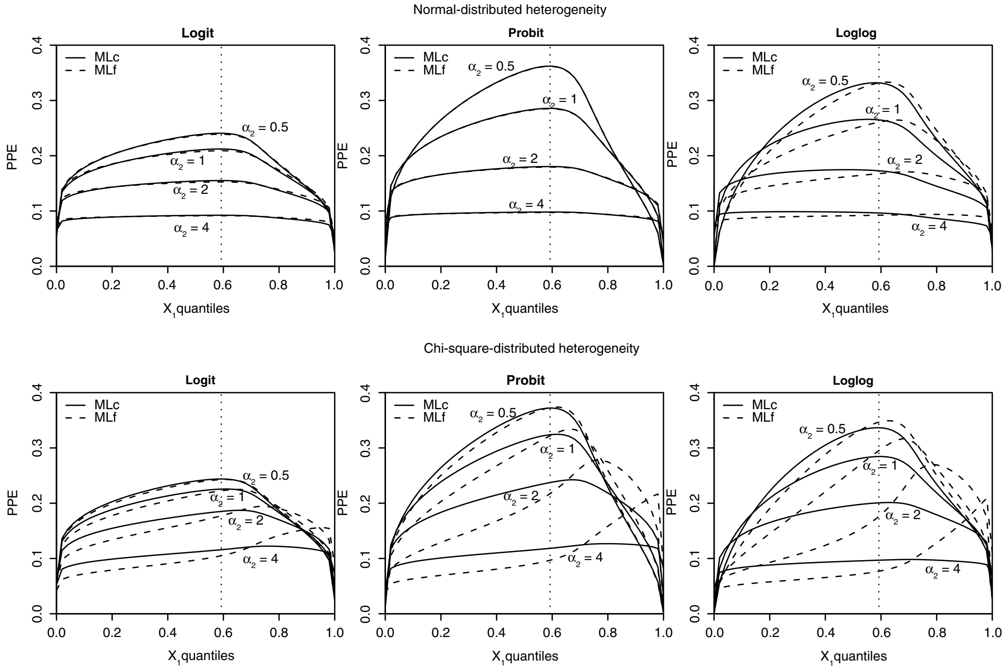

Fig. 3. Population partial effects for binary regression models ($\alpha_1 = 1$).

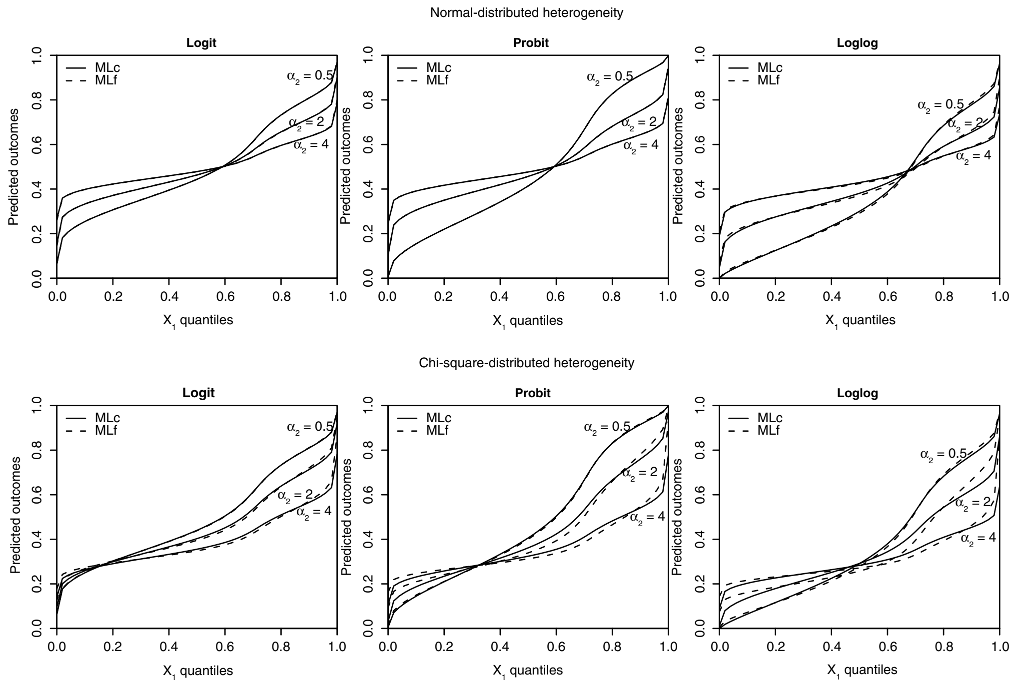

Fig. 4. Predicted outcomes for binary regression models ($\alpha_1 = 1$).

3.1. Attenuation bias in the parameter estimates

Under certain conditions, we proved above that neglected heterogeneity generates an attenuation bias in the naive estimation of the parameters of the observed regressors. Given that our Monte Carlo study incorporates only one observed covariate, it follows from the findings of Neuhaus and Jewell (1993) that an attenuation bias will be present in all the

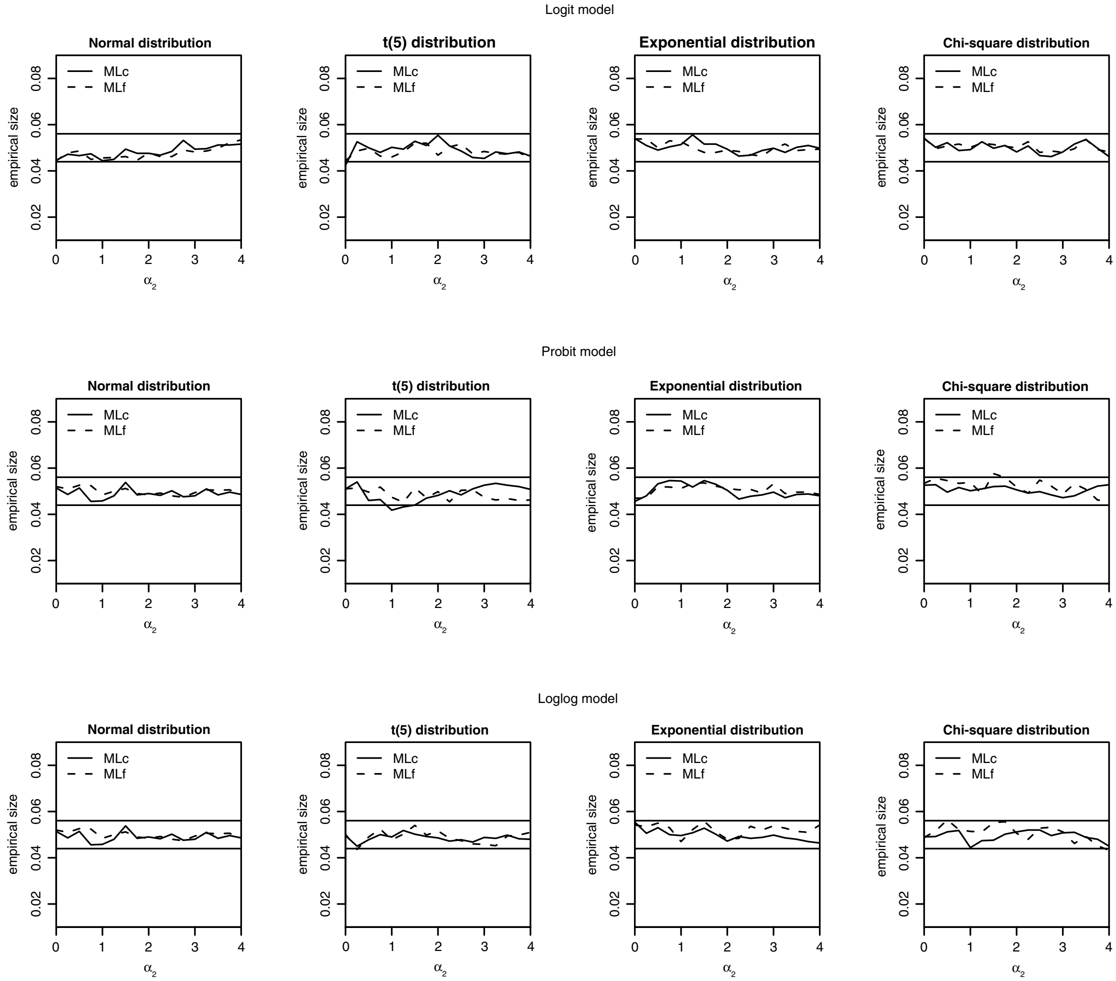

Fig. 5. Empirical size for binary regression models ($N = 200$).

models simulated. However, this bias may differ substantially from that predicted by (6). This is because the assumptions on which its derivation is based are not satisfied in 11 of the 12 models simulated. Therefore, the main aim of our first set of experiments is to examine whether Eq. (6) measures appropriately the extent of the bias caused by neglected heterogeneity when $H \neq H^*$. Fig. 1 displays the values of the ratio $\hat{\alpha}_1^n/\alpha_1$ for two different values of α_1 (-1 and 1) for each one of the 17 values of α_2 simulated. In this figure we also display (using the solid line) the value of the ratio α_1^*/α_1, obtained from (6).

Clearly, in all cases, $\hat{\alpha}_1^n$ tends towards zero, with its absolute bias increasing as α_2 (which measures the extent of the heterogeneity) increases. In the loglog models and, not surprisingly, in the probit models with normally distributed heterogeneity, and in the logit model with t_5 distributed heterogeneity, Eq. (6) approximates the attenuation bias well. However, in some cases, there are important deviations. For example, when X_2 has an exponential or chi-squared distribution, $\hat{\alpha}_1^n$ is not, in general, as biased as predicted by (6) in the logit and probit models, whereas for the loglog model the attenuation effect is amplified relative to (6). Note also that in some cases the actual bias depends on the value of α_1, whereas (6) is not a function of that parameter. Therefore, because there are many cases in which the extent of that bias is not perfectly approximated by (6), we next investigate the consequences of this for the estimation of marginal effects and the prediction of outcomes when $H \neq H^*$.

3.2. Partial effects

Using the set up from the previous section, in Fig. 2, we display the mean across the replications of the *ASE* estimated for the case in which $\alpha_1 = 1$. For the curtailed model we estimate the *ASE* as in (9). For the full model, we use (7), in which the expectation $E(Y|X_1)$ is calculated by integration as in (2) with $f_{X_2}(x_2)$ replaced by the density used to generate X_2. This figure clearly shows that in the logit case, the ML estimation results are similar whether based on the full (MLf) or curtailed (MLc) equations (the largest bias is 3.6% for $\alpha_2 = 2.75$ in the chi-squared case). Thus, as already noted by Cramer (2007), logit analysis of the *ASE* is robust to neglected heterogeneity.

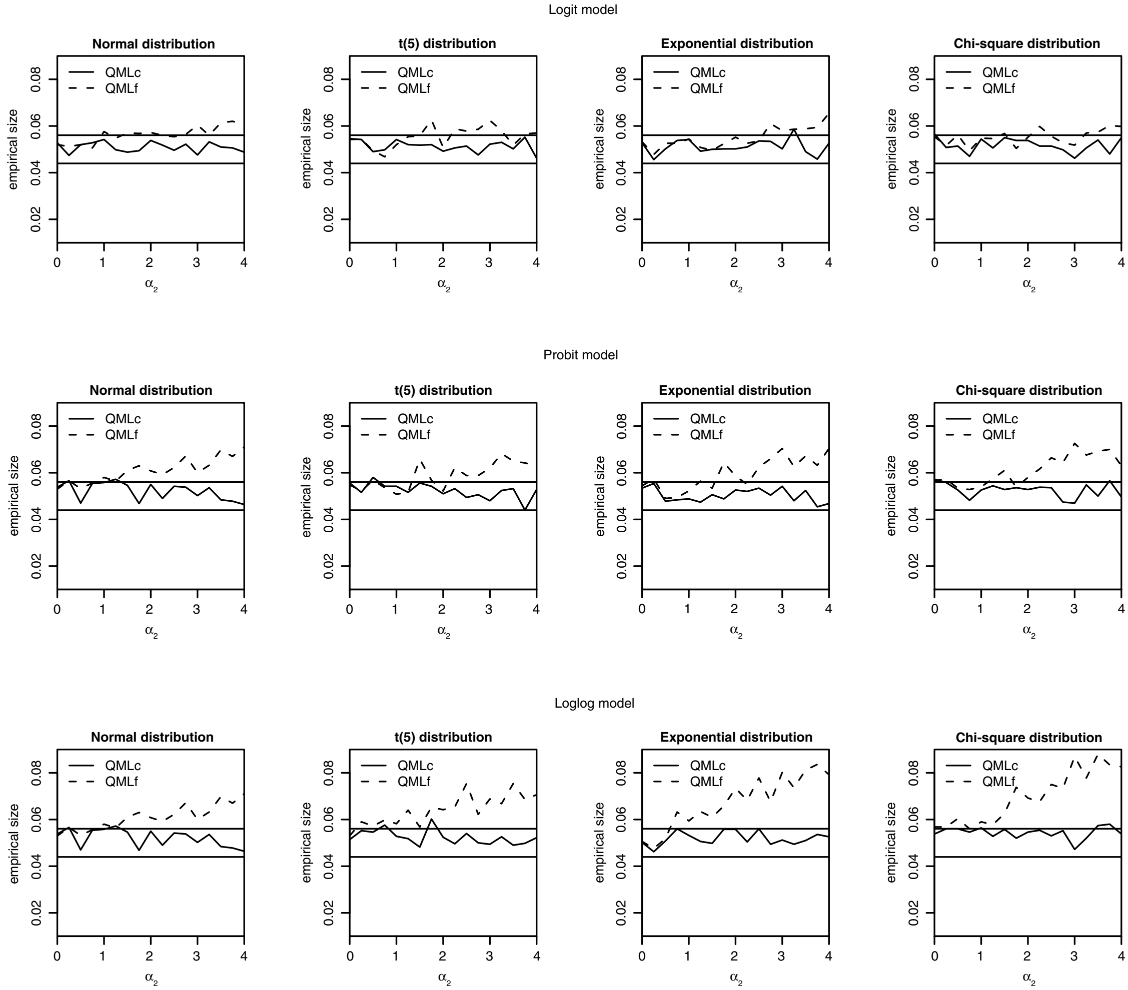

Fig. 6. Empirical size for fractional regression models ($N = 200$).

In the probit model, with a symmetrically distributed omitted variable, the estimated *ASE*s for each equation are almost identical, whereas with an asymmetric X_2, the estimated *ASE*s differ significantly from each other, by up to 7.7% (for $\alpha_2 = 2.25$ in the chi-squared case). For the loglog model, the consequences of neglected heterogeneity are somewhat similar to those found for the probit model: for a symmetric X_2, the bias is minimal (no more than 3%), but for asymmetric unobserved heterogeneity the *ASE* tends to be overestimated (by up to 8.3%, for $\alpha_2 = 1.75$ in the chi-squared case).

It is worth noting that the bias increases with the level of unobserved heterogeneity but only up to a certain point. This may be because of the negligible contribution of X_1 to the variation in $E(Y|X_1)$ when α_2 is very large (the marginal effect of X_1 tends towards zero as α_2 increases). For example, when $\alpha_2 = 4$, the weight attached to the variance of the term $\alpha_2 x_2$ in the total variance of the index $(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2)$ is 94%.

We computed the *PPE* from (10) for the curtailed equation and from (8) for the full model. In both cases, the *PPE*s were evaluated at the means and the $\{0, 0.02, 0.04, \dots, 0.98, 1\}$ quantiles of X_1. Fig. 3 shows the results obtained for $\alpha_1 = 1$ and $\alpha_2 = 0.5$, 1, 2 and 4 when X_2 is generated according to a normal and a chi-squared distribution. The dotted line indicates the mean of X_1. For cases in which X_2 is normally distributed, both the logit and the probit estimators are clearly unaffected by neglected heterogeneity. However, in the chi-squared case, while for small amounts of heterogeneity ($\alpha_2 = 0.5$) the bias in the estimation of the *PPE*s is insubstantial (up to 2.0% for the logit and up to 6.5% for the probit), for large amounts of heterogeneity ($\alpha_2 = 4$) the bias reaches 28.9% (for the logit model) or 50.0% (for the probit model), even when the analysis is restricted to the 0.05–0.95 quantile range. For the loglog model, the bias is in general substantial, reaching 17.4% for normally distributed heterogeneity and reaching 82.6% in the chi-squared case, in both cases for $\alpha_2 = 2$ and again restricting the analysis to the 0.05–0.95 quantile range.

The bias of the various estimators is much smaller when the *PPE*s are evaluated at the mean of X_1. For example, for the symmetric X_2 case, the maximum bias in the loglog model is 4.2%. Nevertheless, there are substantial biases for the chi-squared case: up to 9.8%, 21.4% and 25.1% for the logit, probit and loglog models, respectively.

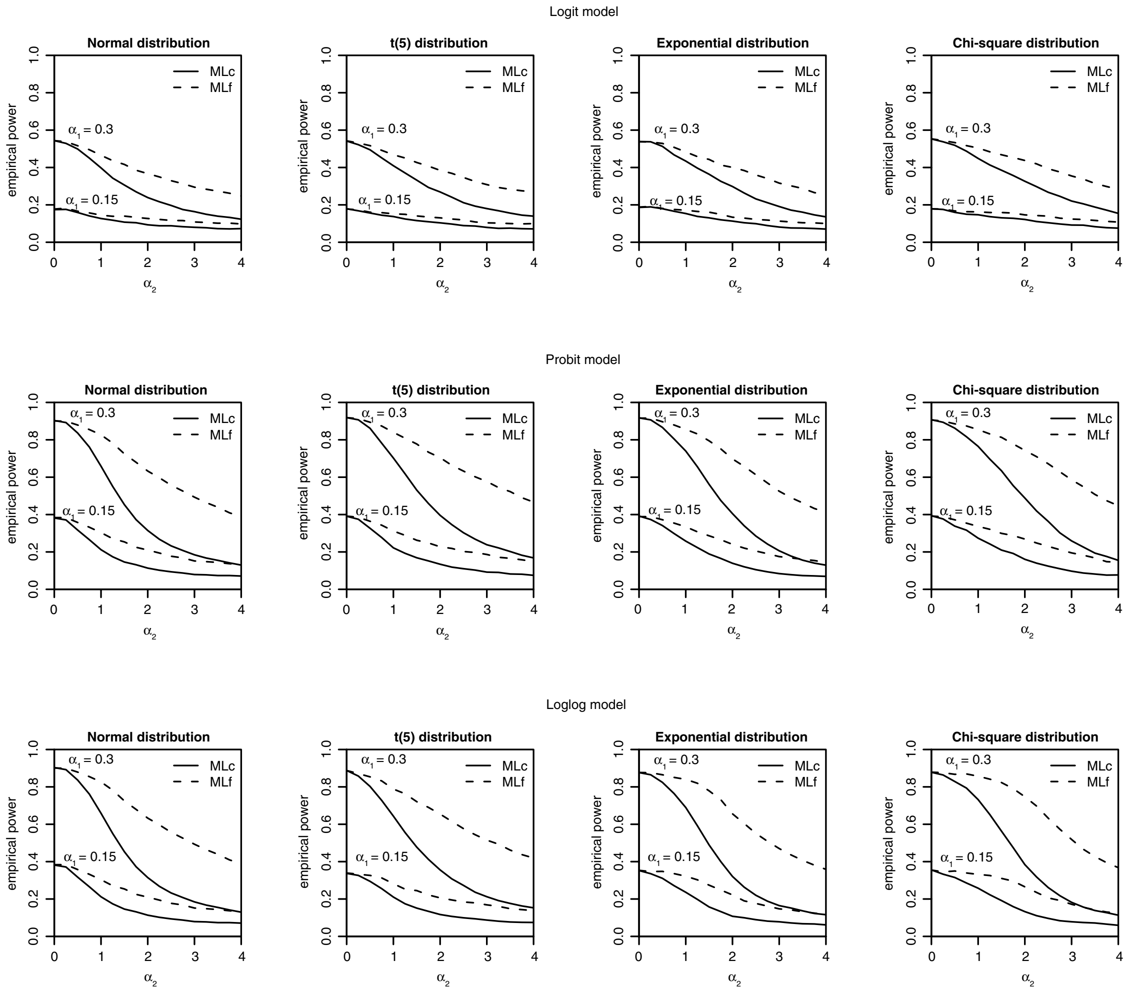

Fig. 7. Empirical power for binary regression models ($N = 200$).

Overall, we can draw three main conclusions from the results obtained in this section. First, the logit model produces more robust estimates of partial effects than do the probit and loglog models. Second, when one's aim is to estimate average partial effects, which is usually the case in empirical work (in most cases, practitioners report only average partial effects), computing *ASE*s is better than computing *PPE*s evaluated at the means of the regressors, because the former appear much more robust to neglected heterogeneity. A similar finding was reported by Ramalho et al. (forthcoming), who found that computation of *ASE*s is relatively robust to functional form misspecification in the framework of fractional regression models, whereas estimation of *PPE*s evaluated at the means of the covariates may be severely biased. Under neglected heterogeneity, computing *PPE*s for an individual with specific characteristics may be very unreliable.

3.3. Predicted outcomes

Fig. 4 illustrates the effects of the omission of X_2 on the prediction of $E(Y|X_1)$ through a simulation design similar to that used for the *PPE*s. For the full model, the prediction is based on (21), and for the curtailed equation, we used the naive estimator $G(\hat{\alpha}_0^n + \hat{\alpha}_1^n x_1)$.

Clearly, unobserved heterogeneity is relatively harmless in logit models: the maximum bias in the 0.05–0.95 quantile range is 5.0% (for $\alpha_2 = 2$). The probit model is also robust to the omission of variables when the distribution of X_2 is symmetric, but displays more important distortions when X_2 is asymmetric (with a bias of up to 15.8% for $\alpha_2 = 2$). The loglog model is relatively robust to unobserved heterogeneity when X_2 has a normal distribution but otherwise produces a bias which reaches a maximum of 23.7% (for $\alpha_2 = 2$). Hence, for outcome prediction, unobserved heterogeneity resulting from the omission of independent explanatory variables is not an issue only for the logit model. Nevertheless, our results suggest that when $H \neq H^*$, the consequences of using a misspecified model G^* are much more serious for the calculation of *PPE*s (which require computation of the derivatives of G^*) than for outcome prediction.

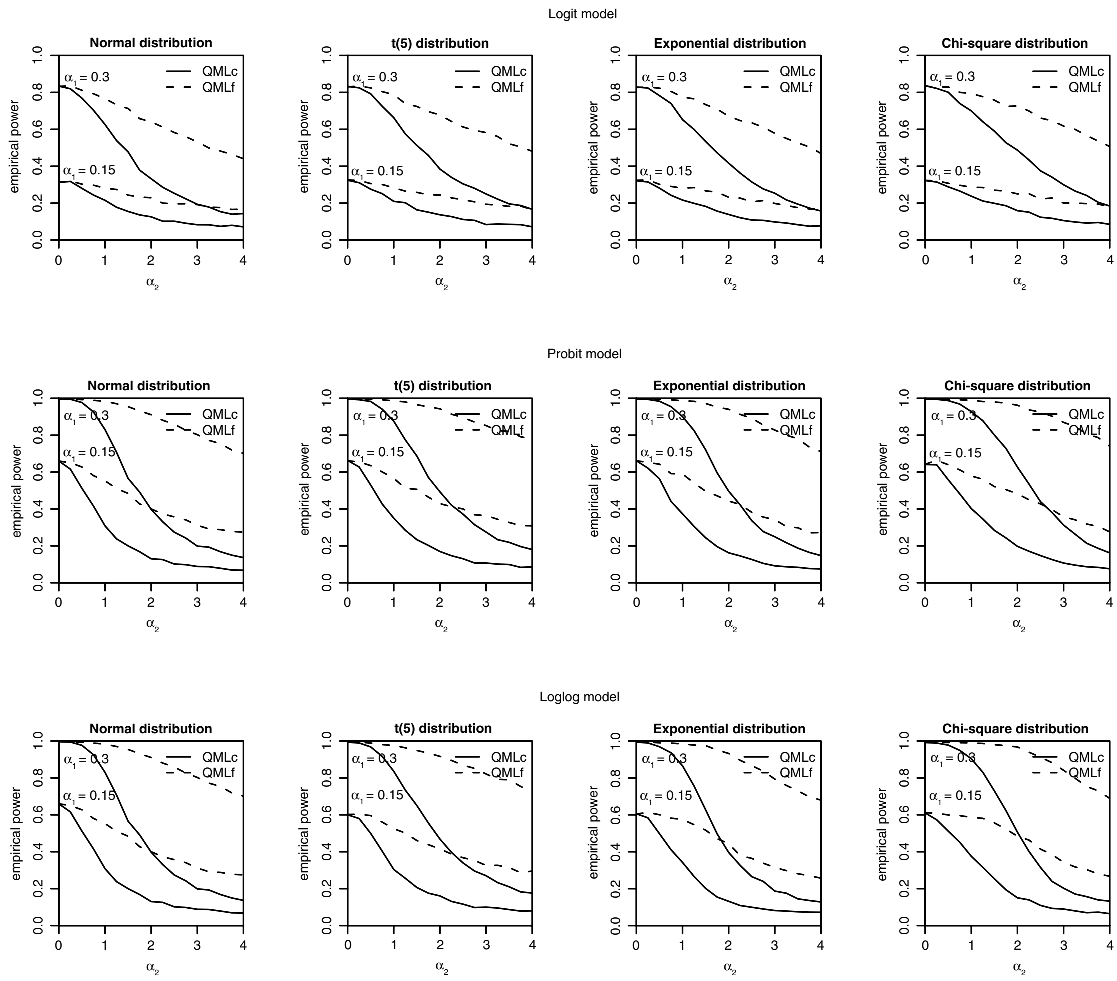

Fig. 8. Empirical power for fractional regression models ($N = 200$).

3.4. The size and power of Wald tests for the significance of observed regressors

In our final set of experiments, we investigate the size and power of naive (Q)ML-based Wald tests for assessing the statistical significance of observed regressors: i.e. we examine their capability of testing the null hypothesis $H_0 : \alpha_1 = 0$ both when it is true and when it is false. Figs. 5 and 6 display the percentage of rejections of H_0 for a nominal level of 5% when this hypothesis is true (the horizontal lines represent the limits of a 95% confidence interval for the nominal size). This percentage is similar for the curtailed and full models in the binary case, being always close to the nominal level of 5%. For fractional data, for which we use robust estimation of the standard errors because we are performing QML estimation, the empirical size of the Wald test based on the naive estimator is even closer to the nominal size than that based on the full equation. Therefore, these results show clearly that the size properties of the Wald test for $\alpha_1 = 0$ are robust to the presence of neglected heterogeneity.

With regard to the power properties of the Wald test, Figs. 7 and 8 illustrate a very different scenario. In this case, there is a substantial decline in the percentage of rejections of the false H_0 as the level of heterogeneity increases. This decay seems to be more substantial, in relative terms, in the probit and loglog models, the higher is α_1, and with fractional data.

To check whether Eq. (20), which was derived for binary logit models under the assumption $H = H^*$, approximates other models well, in Figs. 9 and 10 we represent three W^n/W ratios: that given by (20) (the solid line) and two others that are given by the mean across replications of that ratio for the two values of α_1 simulated.

For binary models, according to Fig. 9, Eq. (20) seems to be a reasonable approximation not only for the logit model on which that equation is based, but also for all the other cases. In fact, comparing Figs. 1 and 9 reveals that both cases produce similar patterns. By contrast, for fractional regression models the attenuation bias in the estimation of the Wald statistic is much larger, which explains why the loss of power detected in Fig. 8 is more substantial for these models. Clearly, Eq. (20) is not a good approximation when robust sandwich-type variance estimators are used, even in the logit case.

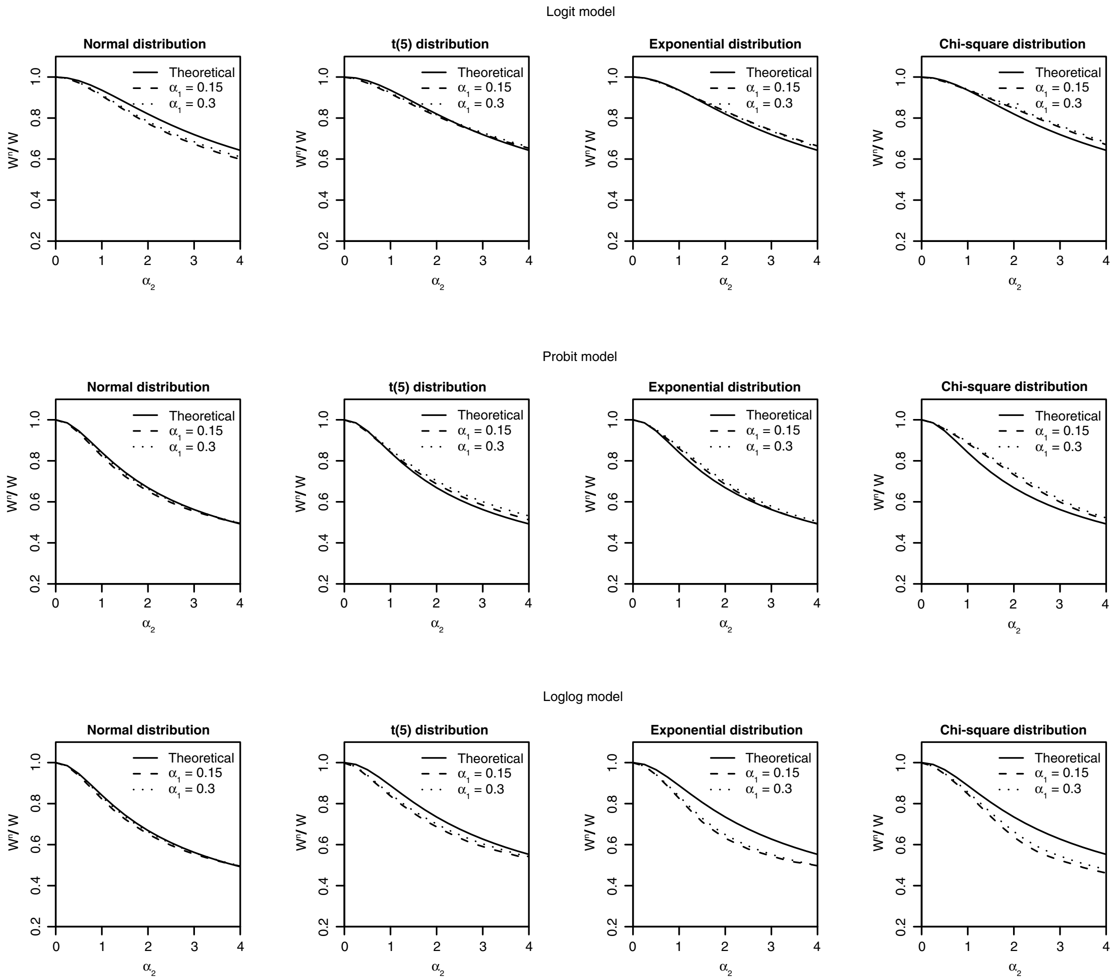

Fig. 9. Attenuation bias of Wald statistics in binary regression models.

A further investigation of the power of naive Wald tests was conducted. For the chi-squared distribution only, and with $\alpha_1 = 0.15$, which produced the poorest power performance of all the cases illustrated so far, we ran experiments for $N = \{200, 500, 1000, 2500, 5000\}$. Fig. 11 shows that in each case, the power of the test increases substantially as N increases. Given these results, it seems that we can trust the outcome of a naive Wald test that reveals that a given explanatory variable is significant. Conversely, the opposite conclusion may be simply the consequence of the omission of a relevant variable, unless the sample size is large and/or the amount of heterogeneity is small.

4. Conclusion

It is well known that the omission of orthogonal relevant variables in nonlinear models generates inconsistent estimates of the parameters on the included regressors. However, recent work on the probit and logit models by Wooldridge (2002, 2005) and Cramer (2003, 2007), respectively, shows that, in some cases, the bias does not carry over to the marginal effects of those regressors on the outcome and that, hence, neglected heterogeneity may not be really an issue in, at least, binary logit and probit models. In this paper, we demonstrated analytically that, under similar assumptions to those made by those authors, their results can be extended to all binary data models. Moreover, based on the same set of assumptions, we showed analytically that outcome prediction in any binary regression model is not affected by neglected heterogeneity and that, in the specific case of the binary logit model, Wald tests are biased towards zero.

Given that the theoretical analysis undertaken in this paper requires strong assumptions, we also performed an extensive Monte Carlo simulation study to consider more general forms of heterogeneity. We found that, in general, unobserved heterogeneity that is independent of the included covariates has the following effects. First, it produces an attenuation bias in the estimation of regression coefficients. Second, it is relatively innocuous for logit estimation of the average sample effect (*ASE*), but may bias estimation in the probit and loglog models. Third, it has much more deleterious effects on the estimation of population partial effects (*PPE*s) than on the estimation of *ASE*s. Fourth, only in the logit model does unobserved

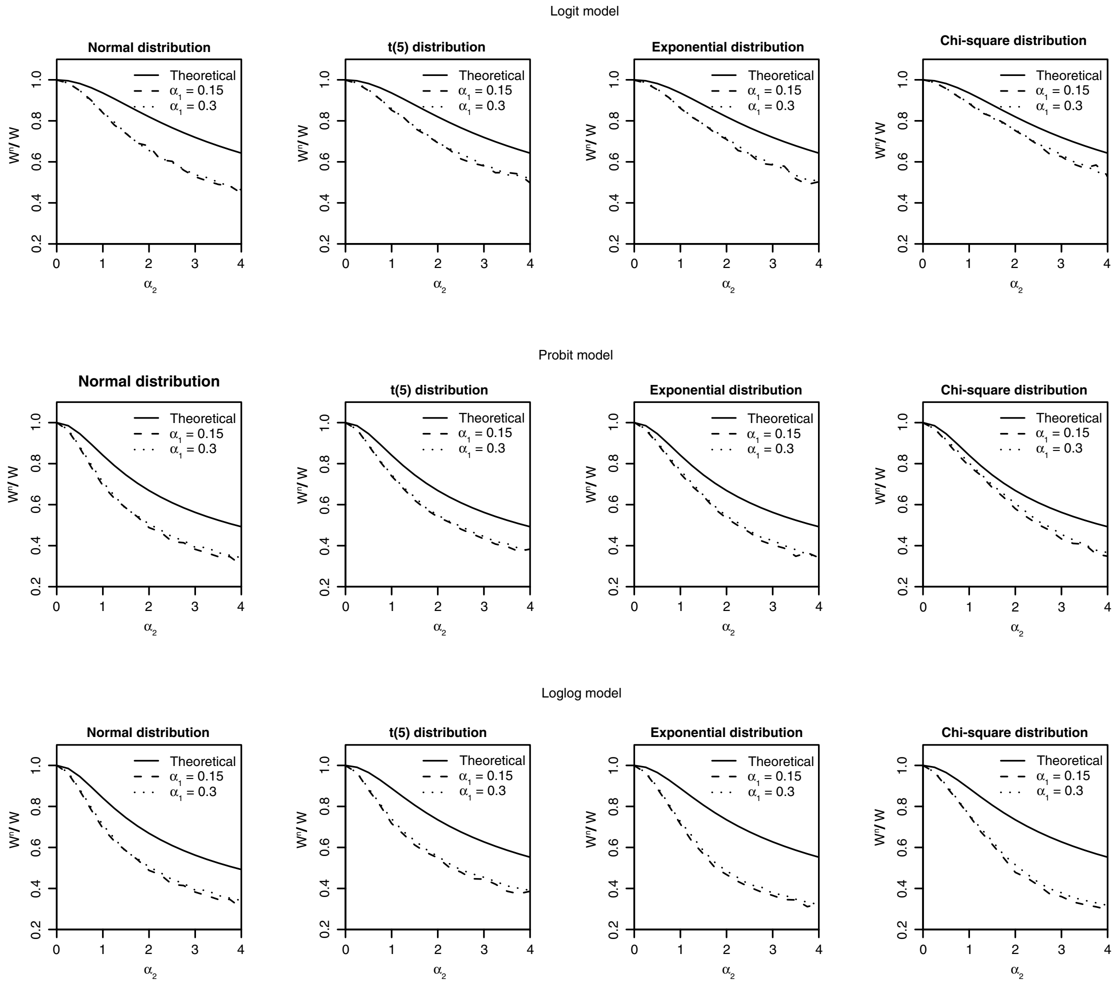

Fig. 10. Attenuation bias of Wald statistics in fractional regression models.

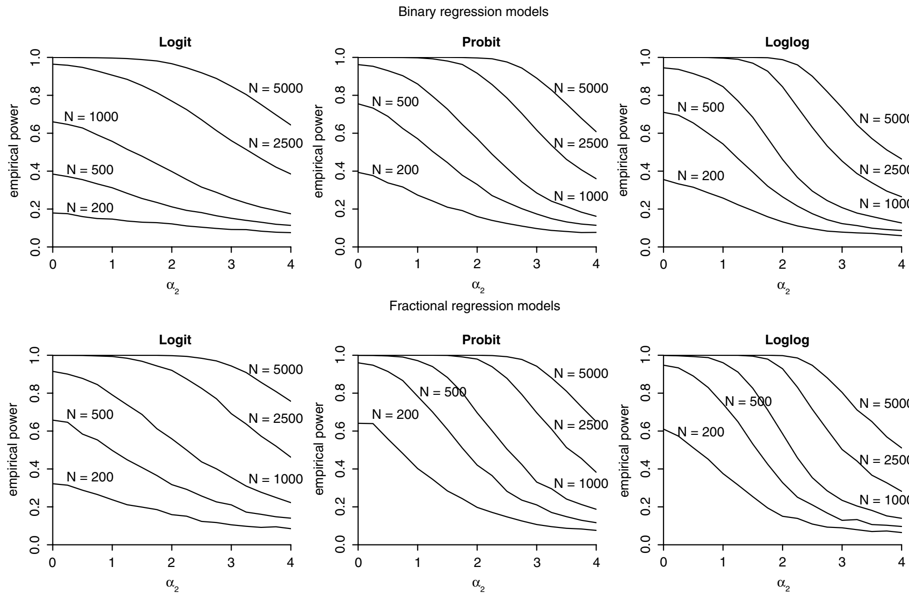

Fig. 11. Empirical power—different sample sizes (chi-square distributed heterogeneity; $\alpha_1 = 0.15$).

heterogeneity not substantially affect the prediction of outcomes. Fifth, its effect is innocuous for the size of Wald tests of the significance of the observed regressors but, in small samples, it reduces their power substantially.

Overall, our results imply that unobserved heterogeneity is not a relevant problem in any of the nonlinear models considered in this paper if the aim of the analysis is simply to determine the direction of the partial effects of the covariates. In addition, in the logit case, neglected heterogeneity is also relatively innocuous for outcome prediction and the calculation of *ASE*s. These are reassuring and useful results for practitioners because standard approaches to dealing with unobserved heterogeneity are not entirely satisfactory. They often require strong distributional assumptions for the unobservable variables that generate poorly fitting models, or ones that are too complex to be widely used by applied economists. Many of these models require the application of nonparametric techniques that often need substantial programming experience.

Another important implication of our results is that it is extremely important to test the general specification of the functional form adopted for the model. If the functional form of a binary regression model is correctly specified (which means that $H = H^*$), then neither estimation of the partial effects nor outcome prediction is affected by the presence of neglected heterogeneity. In such a case, the only relevant problem that remains is the poor power of the Wald test in small samples. However, if all variables are statistically significant or if the sample is large, then even that is not really a problem. For a comparison of various functional form tests for binary and fractional regression models, see Ramalho and Ramalho (2009) and Ramalho et al. (forthcoming), respectively.

Acknowledgements

The authors thank an associate editor and a referee for their valuable suggestions and remarks, which have substantially improved the paper. We are also grateful to João M.C. Santos Silva for helpful comments. Financial support from Fundação para a Ciência e a Tecnologia is gratefully acknowledged (grant PTDC/ECO/64693/2006).

References

Chesher, A., 1995. A mirror image invariance for M-estimators. Econometrica 63 (1), 207–211.
Chesher, A., Peters, S., 1994. Symmetry, regression design, and sampling distributions. Econometric Theory 10, 116–129.
Cramer, J.S., 2003. Logit Models from Economics and other Fields. Cambridge University Press, Cambridge.
Cramer, J.S., 2007. Robustness of logit analysis: Unobserved heterogeneity and mis-specified disturbances. Oxford Bulletin of Economics and Statistics 69 (4), 545–555.
Gourieroux, C., 2000. Econometrics of Qualitative Dependent Variables. Cambridge University Press, Cambridge.
Lagakos, S.W., Schoenfeld, D.A., 1984. Properties of proportional-hazards score tests under misspecified regression models. Biometrics 40, 1037–1048.
Lee, L.F., 1982. Specification error in multinomial logit models. Journal of Econometrics 20, 197–209.
Neuhaus, J.M., Jewell, N.P., 1993. A geometric approach to assess bias due to omitted covariates in generalized linear models. Biometrika 80, 807–815.
Papke, L.E., Wooldridge, J.M., 1996. Econometric methods for fractional response variables with an application to 401(k) plan participation rates. Journal of Applied Econometrics 11 (6), 619–632.
Ramalho, E.A., Ramalho, J.J.S., 2009. Alternative versions of the RESET test for binary response index models: A comparative study, mimeo.
Ramalho, E.A., Ramalho, J.J.S., Murteira, J., 2009. Alternative estimating and testing empirical strategies for fractional regression models. Journal of Economic Surveys, forthcoming (doi:10.1111/j.1467-6419.2009.00602.x).
Simas, A.B., Barreto-Souza, W., Rocha, A.V., 2009. Improved estimators for a general class of beta regression models. Computational Statistics and Data Analysis, forthcoming (doi:10.1016/j.csda.2009.08.017).
Stoker, T.M., 1986. Consistent estimation of scaled coefficients. Econometrica 54 (6), 1461–1481.
Wooldridge, J.M., 2002. Econometric Analysis of Cross Section and Panel Data. MIT Press, Cambridge, MA.
Wooldridge, J.M., 2005. Unobserved heterogeneity and estimation of average partial effects. In: Andrews, D.W.K., Stock, J.H. (Eds.), Identification and Inference for Econometric Models. Cambridge University Press, Cambridge, pp. 27–55.
Yatchew, A., Griliches, Z., 1985. Specification error in probit models. Review of Economics and Statistics 67 (1), 134–139.