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

Esmeralda A. Ramalho [1] *and Joaquim J. S. Ramalho*[2]

[1] Departamento de Economia, Universidade de Évora and CEFAGE – UE
[2] Departamento de Economia, Universidade de Évora and CEFAGE – UE

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

Esmeralda A. Ramalho and Joaquim J.S. Ramalho

Department of Economics, Universidade de Évora and CEFAGE-UE

This draft: May 2009

Abstract

In this paper we examine theoretically and by simulation whether or not unobserved heterogeneity independent of the included regressors is really an issue in logit, probit and loglog models with both binary and fractional data. We found that unobserved heterogeneity: (i) produces an attenuation bias in the estimation of regression coefficients; (ii) is innocuous for logit estimation of average sample partial effects, while in the probit and loglog cases there may be important biases in the estimation of those quantities; (iii) has much more destructive effects over the estimation of population partial effects; (iv) only for logit models does not affect substantially the prediction of outcomes; and (v) is innocuous for the size and consistency of Wald tests for the significance of observed regressors but, in small samples, reduces their power substantially.

Keywords: binary models, fractional models, neglected heterogeneity, partial effects, prediction, Wald tests.

JEL Classification: C12, C13, C15, C25.

1

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. Due to 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 examining the consequences of unobserved heterogeneity in binary and fractional regression models. Moreover, the few studies undertaken have assumed very restrictive conditions or were only concerned with the effects of neglected heterogeneity on particular aspects of those models. For example, Lee (1982) derived conditions under which 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 that bias does not affect the consistent estimation of the partial effect of the observed regressors on the outcome. Finally, Cramer (2003, 2007) considered the binary logit model and proved formally that the same bias attenuation would occur in this context if the distribution of the omitted variables is such that their relegation to the disturbance term of the latent regression equation that originates the logit model does not change its logistic distribution, which is also a very strong assumption. However, this last author presents also a small simulation study which reveals that a particular partial effect, the average sample effect, is quite insensitive to the inconsistency of the parameters of interest, even in cases where the logit shape of the

2

conditional distribution is severely affected.[1]

Given that calculation of partial effects is often the main aim of empirical work and 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) since 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 which 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 examine these questions, we consider the theoretical framework of Wooldridge (2002) and Cramer (2007) and extend their results for other quantities of interest and models. However, given that a more general theoretical approach does not seem to be feasible, in this paper we conduct also 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 consider also 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 $E(Y|X)$.[2] 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

---

[1] See also the work by Neuhaus and Jewell (1993) in the area of generalized linear models, which include the models analyzed in this paper as particular cases. However, their analysis was restricted to the case of a single observed covariate.

[2] See Papke and Wooldridge (1996) for a seminal paper on the so-called fractional regression model and Ramalho, Ramalho and Murteira (2009) for a comprehensive survey on this subject.

several patterns of neglected heterogeneity by assuming various alternative distributions for the omitted variables and assigning different weights to their relative importance.

The 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 simulation study to assess the performance of naive estimators in both binary and fractional regression models is carried out in Section 3. Section 4 concludes.

2 Framework

Consider a random sample of $i = 1, ..., N$ individuals and let Y be the binary or fractional variable of interest, defined, respectively, as $Y = \{0, 1\}$ and $Y \in [0, 1]$, and X_1 and X_2 be, respectively, k_1- and k_2-vectors of explanatory variables. 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. Assume also that X_1 contains an intercept term, X_2 is not observed and X_1 and X_2 are independent. Finally, 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 $G\left(x\theta\right)$ is defined as $e^{x\theta}/\left(1 + e^{x\theta}\right)$, $\Phi\left(x\theta\right)$, and $e^{e^{-x\theta}}$ for, respectively, logit, probit, and loglog models. Note that in the binary case $G\left(\cdot\right)$ also equals $\Pr\left(Y = 1|X_1 = x_1, X_2 = x_2\right)$.

2.1 Effects of neglected heterogeneity on parameter estimation

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

$$E\left(Y|X_1\right) = E_{X_2}\left[G\left(x_1\theta_1 + x_2\theta_2\right)\right] = \int_{\mathcal{X}_2} G\left(x_1\theta_1 + x_2\theta_2\right) f_{X_2}\left(x_2\right) dx_2, \tag{2}$$

where $\mathcal{X}_2$ and $f_{X_2}\left(x_2\right)$ denote, respectively, the sample space and the marginal distribution of X_2. As, in general, $E\left(Y|X_1\right) \neq G\left(x_1\theta_1\right)$, naive estimation based on $G\left(x_1\theta_1\right)$ will not produce consistent estimators for θ_1. In fact, it seems that omission of X_2 will bias θ_1 towards zero, as shown by Yatchew and Griliches (1985) and Wooldridge (2002) for a particular binary probit model, by Cramer (2007) for a peculiar binary logit model, and by Neuhaus and Jewell (1993) for any generalized linear model based on a log concave density function (which is

4

the case of the binary and fractional logit, probit and loglog models) with a single observed covariate. Howewer, as we show next, retracing the arguments of Yatchew and Griliches (1985), Wooldridge (2002) and Cramer (2007), it is not possible to prove formally that this attenuation effect will be 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 y^*, we observe the binary variable y, which takes the value 1 if $y^* > 0$ and the value 0 otherwise. Assume that u has mean zero and variance σ_u^2 and denote its standardized distribution by H. When X_2 is observed, it follows that:

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

where $G\left(\cdot\right)$ is the complementary function of $H\left(\cdot\right)$. When u has a symmetric distribution, $G\left(\cdot\right) \equiv H\left(\cdot\right)$. As it is well known, 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 mean zero and variance τ^2. Then, the composite error $u^* = x_2\beta_2 + u$ is independent of x_1 and has variance $\sigma_{u*}^2 = \beta_2^2\tau^2 + \sigma_u^2$. Denote the standardized distribution of u^* by H^*. In this setting, it follows that:

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

5

Let $\theta_1^* = \beta_1/\sigma_{u^*}$. Clearly, we cannot evaluate the effects of omitting X_2 over parameter estimation unless we assume 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, comparing (4) and (5), we find that

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

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

In this proof, the crucial assumption is that $H = H^*$. Actually, most of the papers cited above made this assumption.[3] Indeed, both Yatchew and Griliches (1985) and Wooldridge (2002, 2005) assumed that both u and X_2 are normally distributed, which implies that u^* has also a normal distribution. On the other hand, in his proof of the existence of an attenuation bias in the logit model, Cramer (2007) did not specify the distribution of X_2 but 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, no similar proof seems to be feasible. Therefore, in the Monte Carlo simulation study carried out in the next section, we investigate whether equation (6), which applies only to very specific binary regression models, also holds approximately for cases where $H \neq H^*$ and for fractional regression models.

2.2 Effects of neglected heterogeneity on partial effects

For empirical analysis based on nonlinear models, the focus is not so much the analysis of the magnitude of the regression coefficients, but consistent estimation of partial effects. The two most usual forms of measuring partial effects in nonlinear models in applied work 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

---

[3]The exception is Neuhaus and Jewell (1993). However, their geometric approach applies only to models with a single observed covariate.

(2002), in presence of neglected heterogeneity we are usually interested in calculating 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, considering evaluation at a given point $X_1 = \bar{x}_1$ (e.g. the mean of the observed regressors), by:

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

As both effects depend on X_2, the naive estimators

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

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, should be inconsistent, since $\hat{\theta}_1^n$ is inconsistent and $G\left(\cdot\right)$ is in general misspecified. However, when $H = H^*$ both $\widehat{ASE^n}$ and $\widehat{PPE^n}$ provide consistent estimates for ASE and PPE, respectively. Indeed, consider again the example discussed in the previous section. Using (2) and (5), we know 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 may be performed for ASE.

Wooldridge (2002), using similar arguments, 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 the estimation of the PPE. Cramer (2007) showed that the same

conclusion holds for logit models in the particular case where the logit shape of $E(Y|X_1, X_2)$ of (1) is preserved in $E(Y|X_1)$ of (2).[4] This last author also shows by simulation that, for logit models, even in cases where $E(Y|X_1)$ 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 may be extended to other models and more general settings.

2.3   Effects of neglected heterogeneity on predicted outcomes

In this paper we examine also 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. So far, the literature has been silent about 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 where the first stage usually requires the estimation of a binary model. Because X_2 is not observed, the main interest is outcome prediction averaged across the population distribution of the omitted variables, just like discussed above for partial effects.

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

2.4   Effects of neglected heterogeneity on Wald tests

Finally, as testing the significance of the impact of a particular covariate 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 associated to the observed regressors in presence of unobserved heterogeneity.

When there are no omitted variables, the Wald statistic for assessing $H_0 : \theta_{1j} = 0$ is

---

[4]These findings are supported by a former work by Stoker (1986), who showed that misspecification of the functional form in single index models does not affect the estimation of average behavioral derivatives.

given by $W = \hat{\theta}_{1j} \left/ \sqrt{\hat{V}\left(\hat{\theta}_{1j}\right)} \right.$, where $\hat{V}\left(\hat{\theta}_{1j}\right)$ denotes an estimate of the variance of $\hat{\theta}_{1j}$, and converges to a standard normal distribution. For binary data, considering again model (1), it follows that

$$\hat{V}\left(\hat{\theta}_{1j}\right) = \left[\frac{1}{N} \sum_{i=1}^{N} \frac{g\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)^2 e_{ij}}{G\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)\left[1 - G\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)\right]} \right]^{-1} \quad (13)$$

where $g\left(z\right) = \partial G\left(z\right)/\partial z$ and e_{ij} is the relevant element of $x_i'x_i$. Hence,

$$W = \sqrt{ \frac{1}{N} \sum_{i=1}^{N} \frac{\hat{\theta}_{1j}^2 g\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)^2 e_{ij}}{G\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)\left[1 - G\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)\right]} } . \quad (14)$$

When X_2 is omitted, the naive significance test of no effect of X_{1j} is given by

$$W^n = \sqrt{ \frac{1}{N} \sum_{i=1}^{N} \frac{\left(\hat{\theta}_{1j}^n\right)^2 g\left(x_{1i}\hat{\theta}_{1j}^n\right)^2 e_{ij}}{G\left(x_{1i}\hat{\theta}_{1j}^n\right)\left[1 - G\left(x_{1i}\hat{\theta}_{1j}^n\right)\right]} } , \quad (15)$$

since we are assuming that X_1 and X_2 are independent.

Under the assumptions made previously, i.e. the distribution of the neglected heterogeneity is such that $H = H^*$, there is a case, $\theta_{1j} = 0$, where neglected heterogeneity does not originate 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 where the included variable for which the significance is tested is binary. Later on, we will examine by simulation the consequences of neglected heterogeneity over the size of Wald tests when $H \neq H^*$.

Lagakos and Schoenfeld (1984) showed also that the power of a score significance test for a binary included variable may be substantially reduced in the presence of omitted covariates. In our framework, we may also 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, there is a special case, the logit model, where such 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 statistics (14) and (15) may be simplified to

$$W = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_{1j}^2 g\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)e_{ij}} = \sqrt{\hat{\theta}_{1j}}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{\partial G\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)}{\partial x_{1j}}e_{ij}} \qquad (16)$$

and

$$W^n = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_{1j}^n\right)^2 g\left(x_{1i}\hat{\theta}_{1j}^n\right)e_{ij}} = \sqrt{\hat{\theta}_{1j}^n}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{\partial G\left(x_{1i}\hat{\theta}_1^{jn}\right)}{\partial x_{1j}}e_{ij}}, \qquad (17)$$

respectively, since $\partial G(x\theta)/\partial x_{1j} = \theta_{1j}g(x\theta)$. Thus, as both $\partial G\left(x_{1i}\hat{\theta}_1 + x_{2i}\hat{\theta}_2\right)\big/\partial x_{1j}$ and $\partial G\left(x_{1i}\hat{\theta}_{1j}^n\right)\big/\partial x_{1j}$ 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), (16) and (17) that

$$\frac{W^n}{W} = \sqrt{\frac{\hat{\theta}_{1j}^n}{\hat{\theta}_{1j}}} \rightarrow \sqrt{\frac{\sigma_u}{\sigma_{u^*}}}. \qquad (18)$$

Hence, assuming $H = H^*$, in a logit model the naive Wald test W^n is depressed relative to W by the square root of the attenuation factor that relates $\hat{\theta}_1^n$ to $\hat{\theta}_1$.[5] This implies that, in fact, in small samples unobserved heterogeneity may reduce the power of Wald tests. However, from (18) it is also evident that under neglected heterogeneity the Wald test retains its consistency.

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

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 bear on a simple two-variable equation,

$$E(Y|X_1, X_2) = G(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2), \qquad (19)$$

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 analyze the effects of omitting X_2 on the estimation

---

[5]Note that this implies that the same relationship holds for the ratio of the standard errors of $\hat{\theta}_1^n$ and $\hat{\theta}_1$.

of α_1 and related statistics. Note that $\alpha_2 = 0$ corresponds to the case where there is no neglected heterogeneity and that larger values of α_2 imply a larger amount of heterogeneity.

In all experiments, X_1 is generated from a mixtures 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 mean zero and variance one. The choice of an asymmetric distribution for X_1 was made to avoid the reflection property about the origin that would affect the sampling distribution of the estimators of α_1; see Chesher and Peters (1994) and Chesher (1995) for a discussion on the design of Monte Carlo simulation studies.

We generate Y as a Bernoulli (binary case) or a beta (fractional case) variate with mean given by the logit, probit, or loglog functional form and the shape parameter of the beta distribution fixed at 1.[6] 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, which are the standard ways of dealing with each type of data. In both cases, we estimate *full* and *curtailed* versions of the models, i.e. models *with* and *without* X_2. The full version of the model yields consistent estimators for all the quantities of interest and, hence, it will be 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 amount of results produced in each experiment, we summarized them in figures. In most cases (the only exceptions are the experiments regarding the Wald tests), given the similarity of the results obtained, only those relative to binary models are reported.[7] Apart from the last experiment, where several samples sizes were considered, in all the remaining cases the sample size is $N = 200$.

3.1   Attenuation bias in the parameter estimates

Under some special conditions, we proved above that an attenuation bias is imposed by neglected heterogeneity over naive estimation of the parameters of the observed regressors. As in this Monte Carlo study we consider only one observed covariate, according to the

---

[6]See *inter alia* Ramalho, Ramalho and Murteira (2009) for the mean-dispersion parametrization of the beta distribution used in the generation of data.

[7]Full results are available from the authors upon request.

findings of Neuhaus and Jewell (1993), we know for sure that an attenuation bias will be present in all the models simulated. However, this bias may differ substantially from that predicted by (6), since the assumptions made in its derivation are not met in 11 out of the 12 models simulated. Therefore, the main aim of our first set of experiments is to examine whether equation (6) measures appropriately the extent of the bias caused by neglected heterogeneity when $H \neq H^*$. Figure 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 display also (solid line) the value of the ratio α_1^* / α_1, obtained from (6).

Figure 1 about here

Clearly, in all cases, $\hat{\alpha}_1^n$ is depressed towards zero, its absolute bias increasing as α_2 (i.e. the extent of heterogeneity) increases. Equation (6) gives often a very good approximation to the attenuation bias (e.g. loglog and, obviously, probit models with normal-distributed heterogeneity and logit model with t_5-distributed heterogeneity) but in some cases there are some important deviations. For example, when X_2 has an exponential or chi-square distribution $\hat{\alpha}_1^n$ is not, in general, as biased as predicted by (6) in the logit and probit models, while 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, while (6) is not a function of that parameter. Therefore, as the extent of that bias is not perfectly approximated by (6) in many cases, next we investigate the consequences of this fact over the calculation of marginal effects and prediction of outcomes when $H \neq H^*$.

3.2 Partial effects

Using the same setup of the previous section, in Figure 2 we display the mean across the replications of the ASE estimated for the case $\alpha_1 = 1$. For the curtailed model we estimate the ASE as in (9), while for the full model we use (7), where 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 shows clearly that in the logit case ML estimation based on the full (MLf) or the curtailed (MLc) equations leads to very similar results (the largest bias is 3.6% for $\alpha_2 = 2.75$ in the chi-square case). Thus, as already noted by Cramer (2007), logit analysis of the ASE is very robust to neglected heterogeneity.

12

Figure 2 about here

In the probit model, considering a symmetrical-distributed omitted variable, the ASEs estimated for each equation are also almost identical, while for asymmetric X_2 the deviations between them are no longer insignificant, achieving a maximum of 7.7% ($\alpha_2 = 2.25$, chi-square case). With regard to the loglog model, the consequences of neglected heterogeneity are somewhat similar to those found for the probit model: while for symmetric X_2 the bias is minimal (always inferior to 3%), for asymmetric unobserved heterogeneity the ASE is often somewhat overestimated (maximum bias: 8.3% for $\alpha_2 = 1.75$ in the chi-square case).

Finally, note that the bias increases with the level of unobserved heterogeneity but only until a certain point, which may be explained by the little importance of X_1 in the variation of $E(Y|X_1)$ when α_2 is very large (the marginal effect of X_1 tends to zero as α_2 increases). For example, when $\alpha_2 = 4$ the weight of 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%.

In which concerns the PPEs, we calculated them as in (10), for the curtailed equation, and (8), for the full model. In both cases, the PPE was evaluated at the mean and the $\{0, 0.02, 0.04, ..., 0.98, 1\}$ quantiles of X_1. Figure 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-square distribution. The dotted line indicates the mean of X_1. For cases where X_2 is normal-distributed, both the logit and the probit estimators are clearly unaffected by neglected heterogeneity. However, in the chi-square case, while for small amounts of heterogeneity ($\alpha_2 = 0.5$) the bias in the estimation of PPEs is not that relevant (maximum bias of 2.0% for the logit and 6.5% for the probit), for large amounts of heterogeneity ($\alpha_2 = 4$) the bias may achieve a maximum value of 28.9% (logit model) or 50.0% (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, achieving a maximum of 17.4% for normal-distributed heterogeneity and 82.6% for the chi-square case, in both cases for $\alpha_2 = 2$ and again restricting the analysis to the 0.05-0.95 quantile range.

Figure 3 about here

When we consider the evaluation of the PPE at the mean of X_1, the bias of the various estimators is much smaller. For example, for the symmetric X_2 case, the maximum bias in

the loglog model is now 4.2%. Nevertheless, for the chi-square case the bias may still be substantial: the maximum bias for the logit, probit and loglog models is, respectively, 9.8%, 21.4% and 25.1%.

Overall, the results obtained in this section allows us to achieve three main conclusions. First, the logit model produces more robust estimates of partial effects than probit or loglog models. Second, when our interest is the calculation of average partial effects, which is usually the case in empirical work (in most cases, practitioners report only average partial effects), it is preferable to compute ASEs instead of PPEs evaluated at the mean of the regressors, since the former appears to be clearly much more robust to neglected heterogeneity.[8] Finally, under neglected heterogeneity, computation of PPEs for an individual with specific characteristics may be very unreliable.

3.3 Predicted outcomes

Figure 4 illustrates the effects of the omission of X_2 in the prediction of $E(Y|X_1)$ through a simulation design similar to that used for the PPEs. For the full model the prediction is based on (19), while 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% ($\alpha_2 = 2$). The probit model is also robust to the omission of variables when the distribution of X_2 is symmetric, but displays more important distortions in cases where X_2 is asymmetric (maximum bias: 15.8% for $\alpha_2 = 2$). Finally, the loglog model is relatively robust to unobserved heterogeneity when X_2 has a normal distribution but displays some bias in the other case, achieving a maximum bias of 23.7% ($\alpha_2 = 2$). Hence, for outcome prediction, unobserved heterogeneity resulting from the omission of independent explanatory variables does not seem to be a relevant issue only in logit models. Nevertheless, note that our results suggest that when $H \neq H^*$, the consequences of using a misspecified model G^* are much more serious for calculation of PPEs (which require the computation of derivatives of G^*) than for outcome prediction.

---

[8]A similar finding was reported by Ramalho, Ramalho and Murteira (2009), who found that computation of ASEs is relatively robust to functional form misspecification in the framework of fractional regression models, while estimation of PPEs evaluated at the mean of the covariates may be severely biased.

3.4   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 ability for testing the null hypothesis $H_0 : \alpha_1 = 0$ both when it is true and false. Figures 5-6 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). This percentage is very similar for the curtailed and full models in the binary case, being always very near to the nominal level of 5%. For fractional data, where we use robust estimation of standard errors since 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 very robust to the presence of neglected heterogeneity.

With regard to the power properties of the Wald test, Figures 7-8 illustrate a very different scenario. In this case, we observe an important decay on 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, in cases where α_1 is larger, and with fractional data.

In order to check whether equation (18), which was derived for binary logit models under the assumption $H = H^*$, provides also a good approximation for other models, in Figures 9-10 we represent three W^n/W ratios: that given by (18) (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, see Figure 9, equation (18) seems to be a reasonable approximation. In fact, comparing Figures 1 and 9, a very similar pattern was obtained. In 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 Figure 8 is more substantial for these models. Clearly, equation (18) is not a good approximation when robust sandwich-type variance estimators are used.

A further investigation on the power of naive Wald tests was conducted. Only for the chi-square distribution and for the value of $\alpha_1 = 0.15$, which led to the poorest power performance of all the cases illustrated in the previous figures, we run experiments for $N = \{200, 500, 1000, 2500, 5000\}$. Figure 11 shows that in all cases the power of the test increases substantially as N increases, which confirms that the Wald test is still consistent in presence of omitted variables, as discussed in Section 2. 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. The opposite conclusion may be simply the consequence of the omission of relevant variables, 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 causes inconsistency in the estimation of the parameters of interest associated with the included regressors. However, some 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 effect 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 imposed by those authors, their results can be extended to any other model for binary data. Moreover, we showed that, while other features like outcome prediction are also robust to neglected

heterogeneity, Wald tests for the individual significance of an included covariate are biased towards the non-rejection of the null hypothesis of non-significance.

Given that the theoretical analysis undertaken in this paper requires strong assumptions, we performed also an extensive Monte Carlo simulation study considering more general forms of heterogeneity. We found that, in general, unobserved heterogeneity independent of the included covariates: (i) produces an attenuation bias in the estimation of regression coefficients; (ii) is relatively innocuous for logit estimation of the ASE, while in the probit and loglog cases there may be important biases in its estimation; (iii) has much more destructive effects over the estimation of PPEs than ASEs; (iv) only for logit models does not affect substantially the prediction of outcomes; and (v) is innocuous for the size and the consistency of Wald tests for the significance of the observed regressors but, in small samples, 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 obtaining 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 ASEs.[9] These are, we think, very comforting and useful results for practitioners since the usual ways of dealing with unobserved heterogeneity are not entirely satisfactory, requiring strong distributional assumptions for the unobservables which often give rise to a model that does not describe properly the data, or are too complex to be widely used by applied economists, often requiring the utilization of nonparametric techniques which frequently cannot be computed without 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.[10] Indeed, if the test indicates that the functional form of our binary regression model is correctly specified (which means that $H = H^*$), then we know that calculation of partial effects and outcome prediction

---

[9]Note that this unique property of robustness of the logit model is not totally unexpected. In fact, this model is also robust to other problems, like endogenous stratification and nonignorable missing data, that, in general, cause the inconsistency of the estimators based on other models; see, respectively, Hsieh, Manski and McFadden (1985) and Ramalho and Smith (2003).

[10]For a comparison of various functional form tests for binary and fractional regression models see Ramalho, Ramalho and Murteira (2009) and Ramalho and Ramalho (2009).

is not 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 the sample is very large, then even that is not really a problem.

References

Chesher, A. (1995), "A mirror image invariance for m-estimators", *Econometrica*, 63(1), 207-211.

Chesher, A. and 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, Cambridge University Press.

Cramer, J.S. (2007), "Robustness of logit analysis: unobserved heterogeneity and misspecified disturbances", *Oxford Bulletin of Economics and Statistics*, 69(4), 545-555.

Gourieroux, C. (2000), *Econometrics of Qualitative Dependent Variables*, Cambridge, Cambridge University Press.

Hsieh, D.A., Manski, C.F. and McFadden, D. (1985), "Estimation of response probabilities from augmented retrospective observations", *Journal of the American Statistical Association*, 80, 651-662.

Lagakos, S.W. and 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. and 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. and 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., and 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. and Murteira, J. (2009),"Alternative estimating and testing empirical strategies for fractional regression models", *Journal of Economic Surveys*, forthcoming.

Ramalho, E.A., and Smith, R.J. (2003), "Discrete Choice Nonresponse", Centre for Micro-data Methods and Practice, I.F.S. and U.C.L.. http://cemmap.ifs.org.uk/wps/cwp0307.pdf

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*, Cambridge, MIT Press.

Wooldridge, J.M. (2005), "Unobserved heterogeneity and estimation of average partial effects", in D.W.K. Andrews and J.H. Stock (eds.) *Identification and Inference for Econometric Models*, Cambridge, Cambridge University Press, 27-55.

Yatchew, A. and Griliches, Z. (1985), "Specification error in probit models", *Review of Economics and Statistics*, 67(1), 134-139.

Figure 1: Attenuation bias of parameter estimates in binary regression models

Figure 2: Average sample effects for binary regression models ($\alpha_1 = 1$)

Normal–distributed heterogeneity

probit

loglog

logit

MLc
MLf

ASE

α_2

t(5)–distributed heterogeneity

probit

loglog

logit

MLc
MLf

ASE

α_2

Exponential–distributed heterogeneity

probit

loglog

logit

MLc
MLf

ASE

α_2

Chi–square–distributed heterogeneity

probit

loglog

logit

MLc
MLf

ASE

α_2
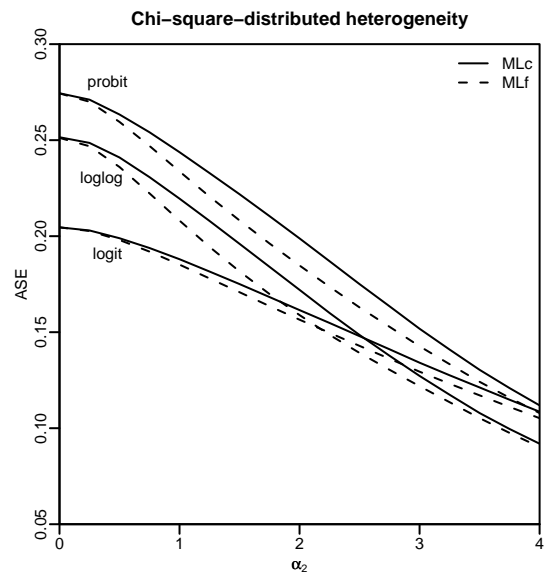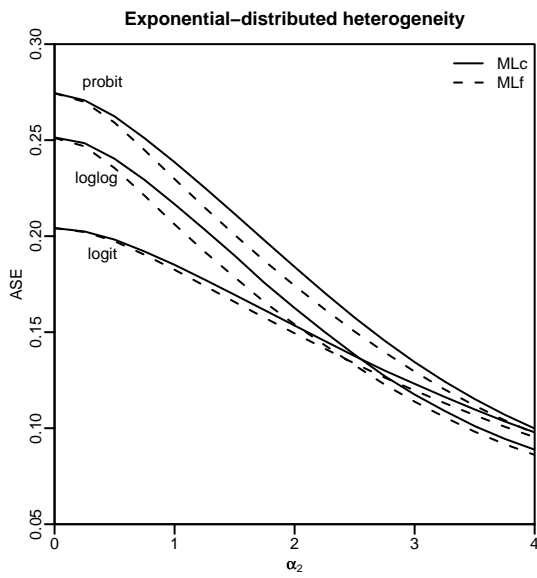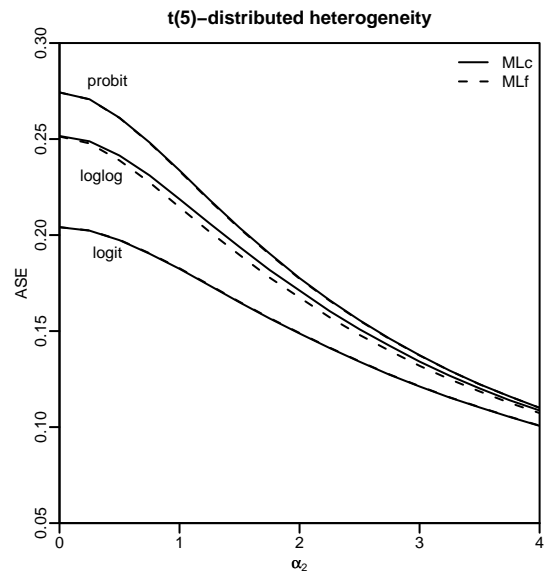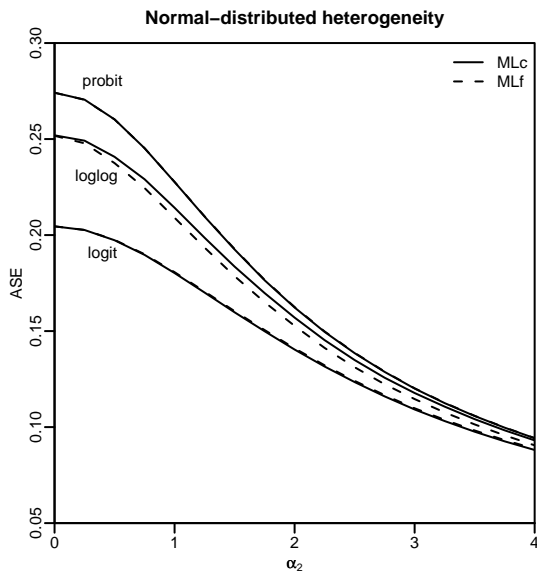
Figure 3: Population partial effects for binary regression models ($\alpha_1 = 1$)

Normal−distributed heterogeneity

Chi−square−distributed heterogeneity
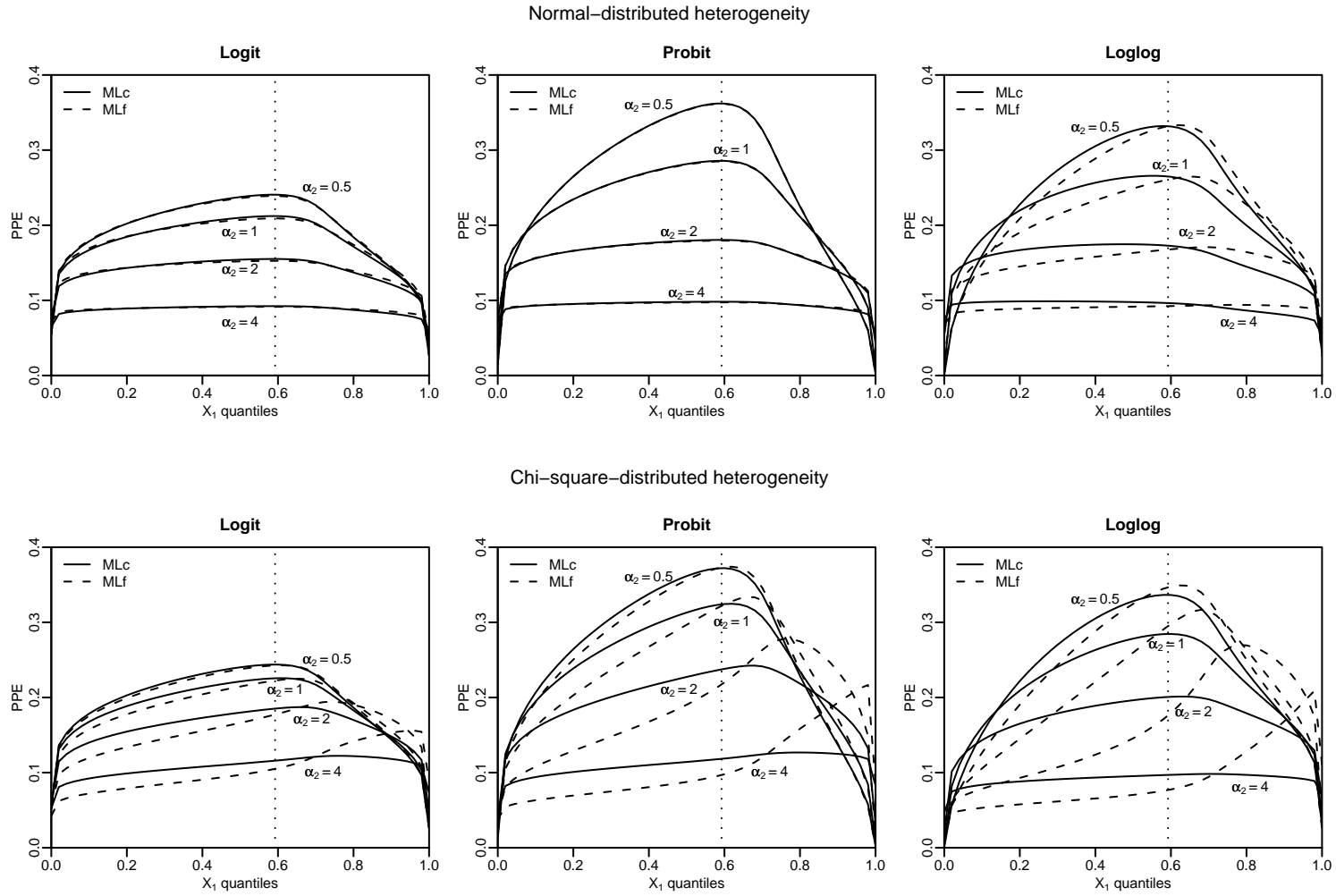
Figure 4: Predicted outcomes for binary regression models ($\alpha_1 = 1$)

Normal−distributed heterogeneity

Chi−square−distributed heterogeneity
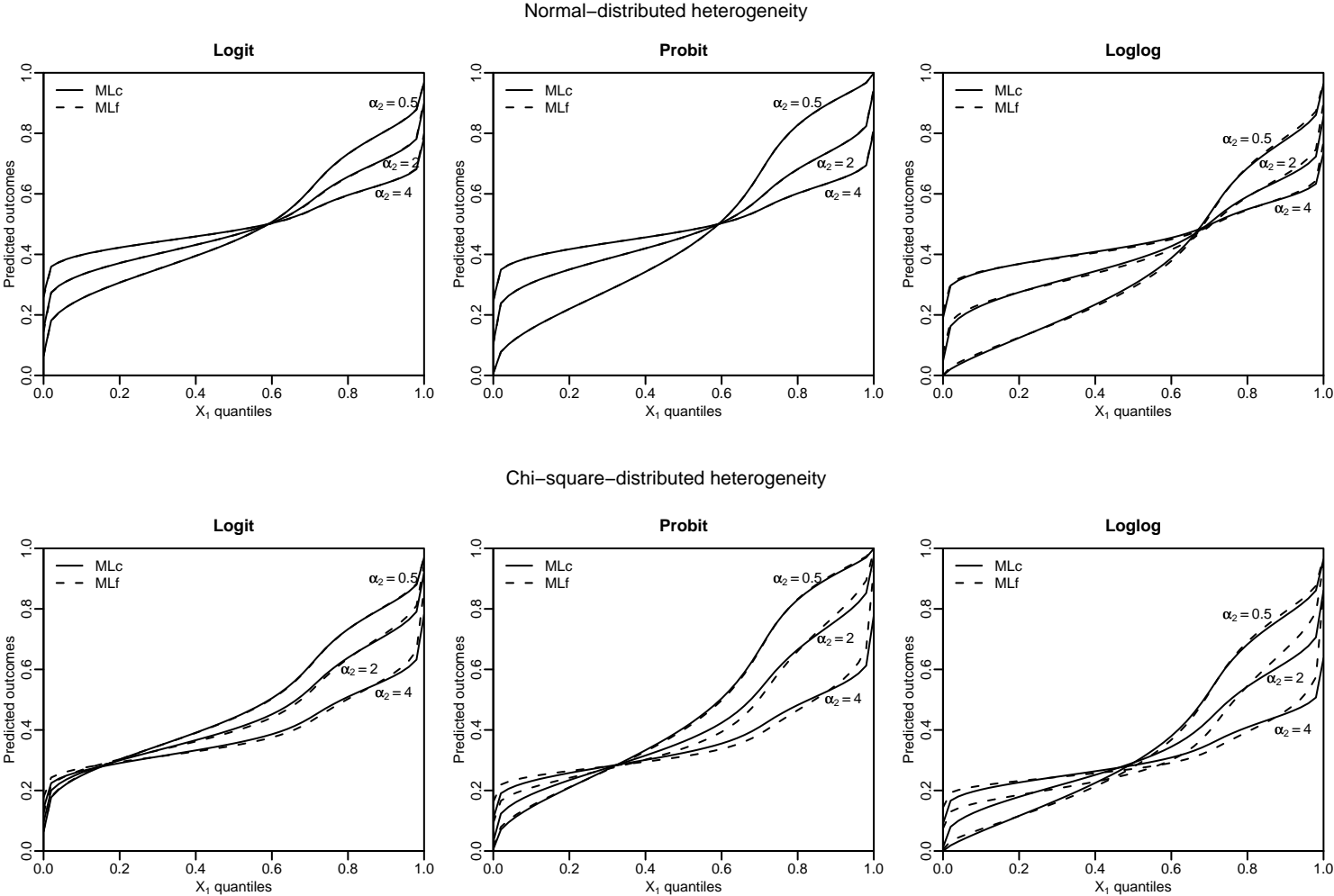
Figure 5: Empirical size for binary regression models (N = 200)
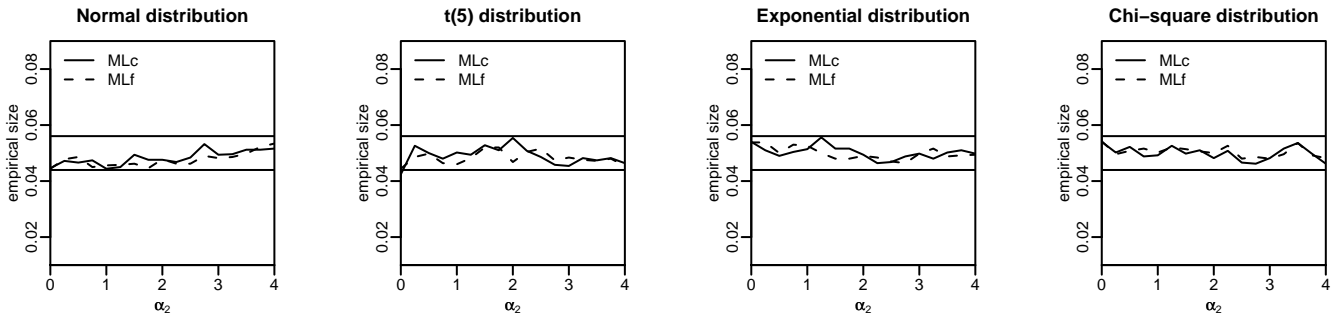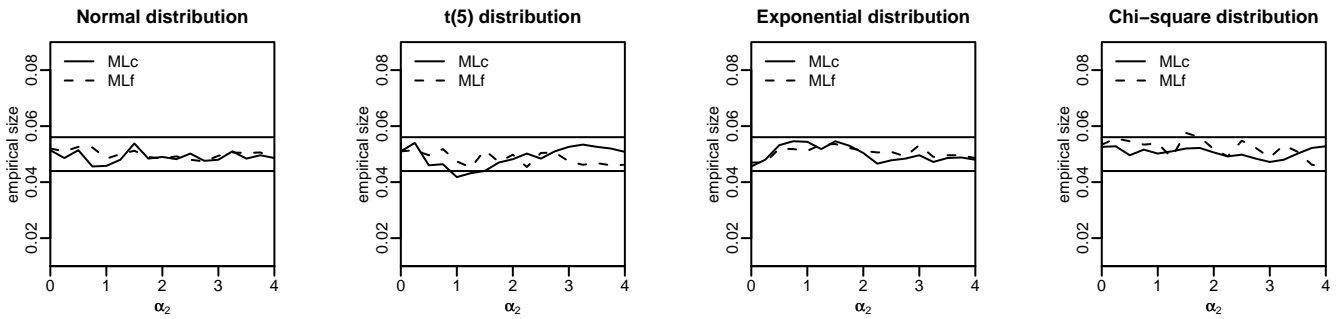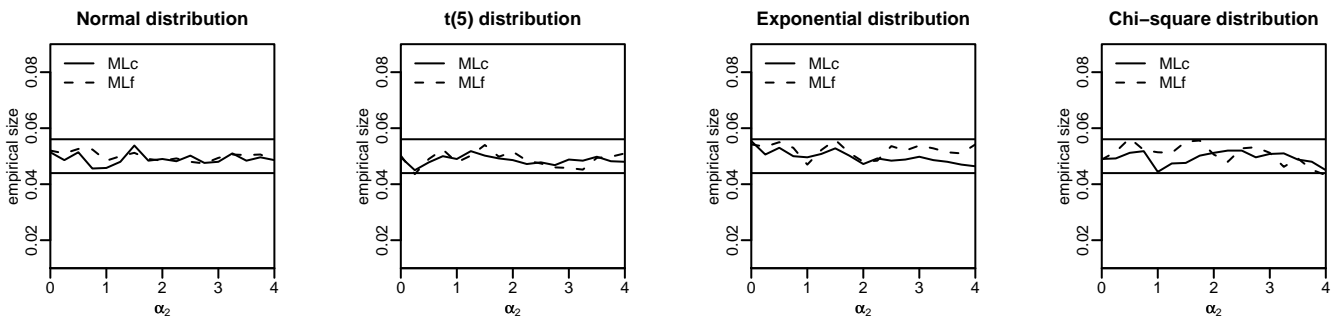
Figure 6: Empirical size for fractional regression models (N = 200)
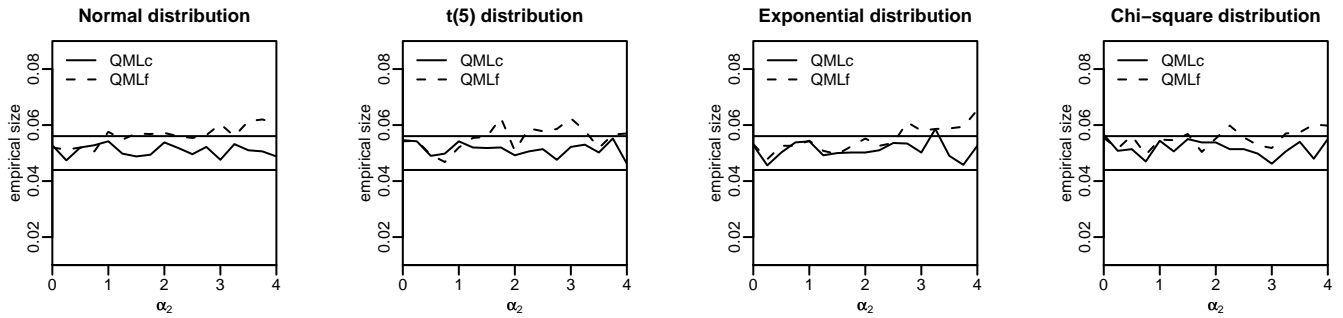
Logit model
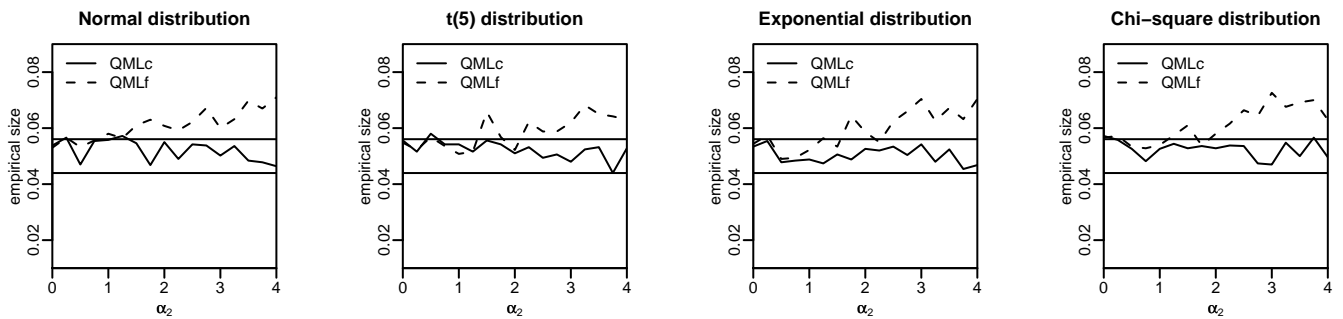
Probit model
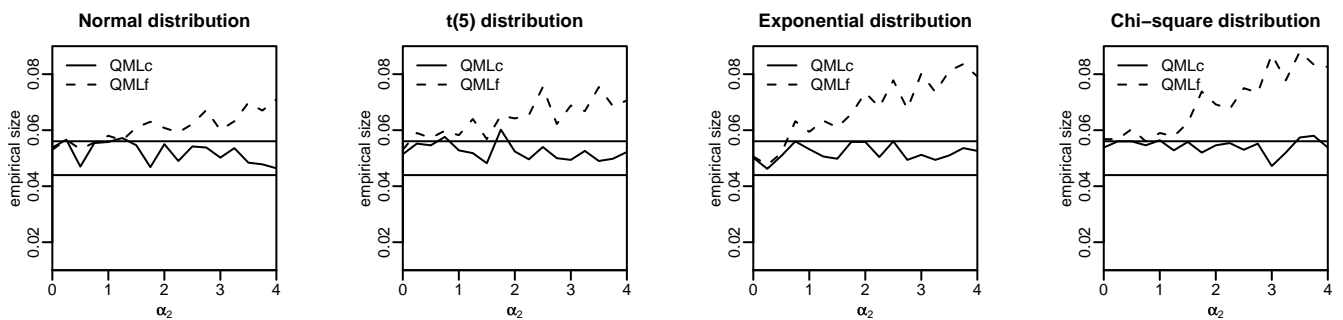
Loglog model

Figure 7: Empirical power for binary regression models (N = 200)
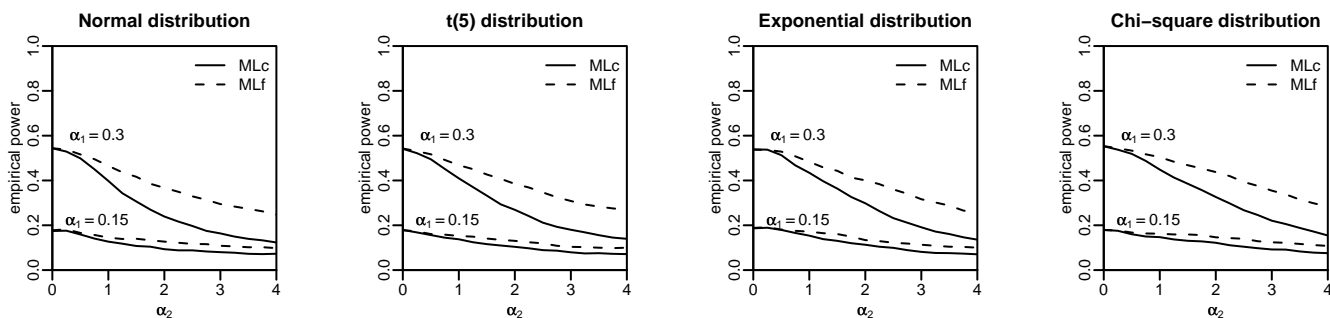
Logit model
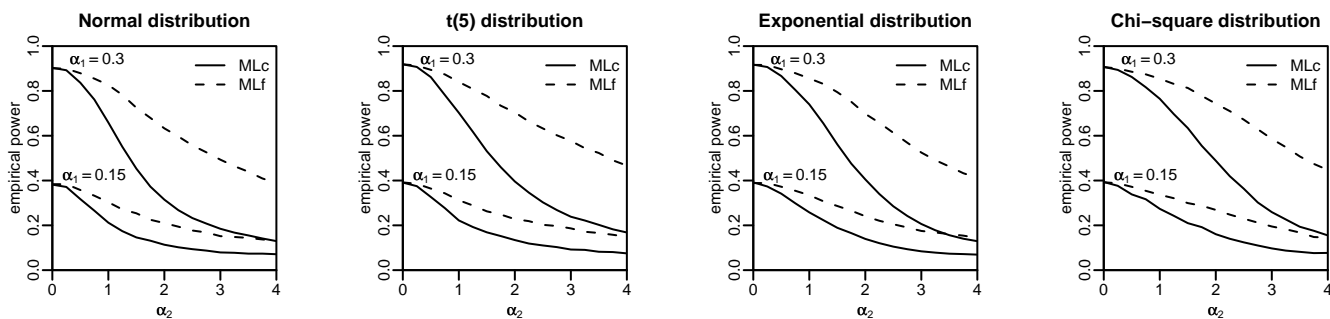
Probit model
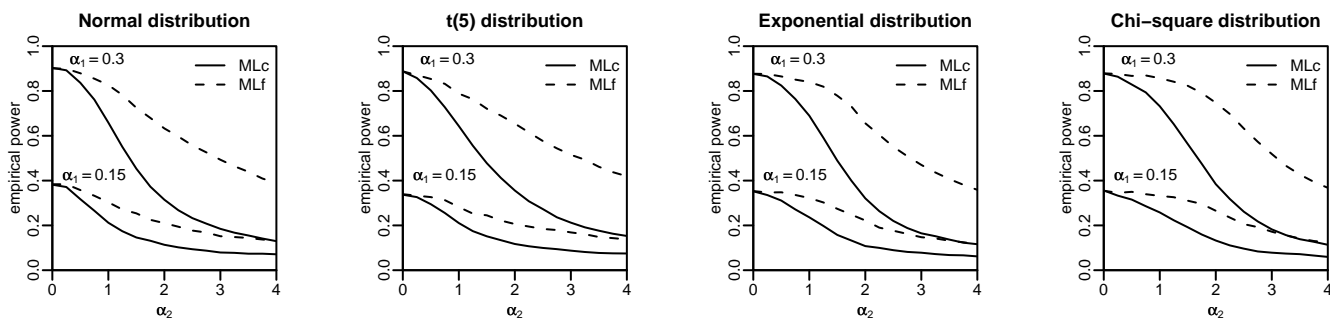
Loglog model
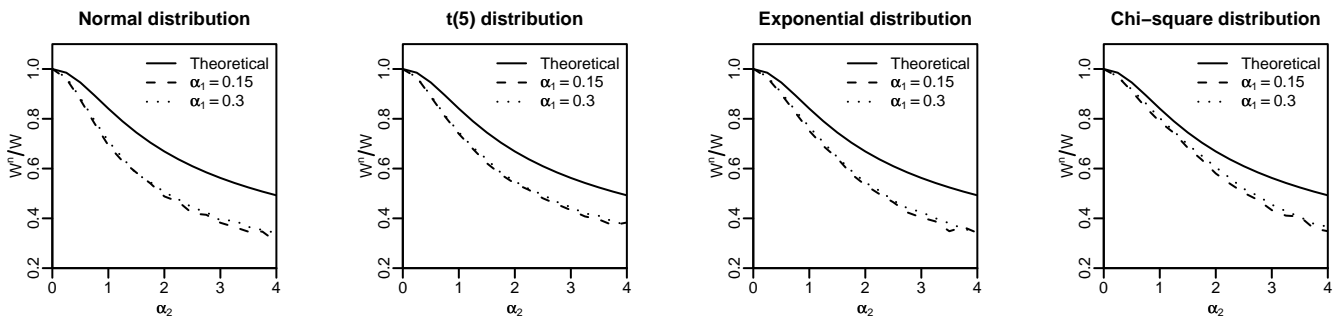
Figure 8: Empirical power for fractional regression models (N = 200)
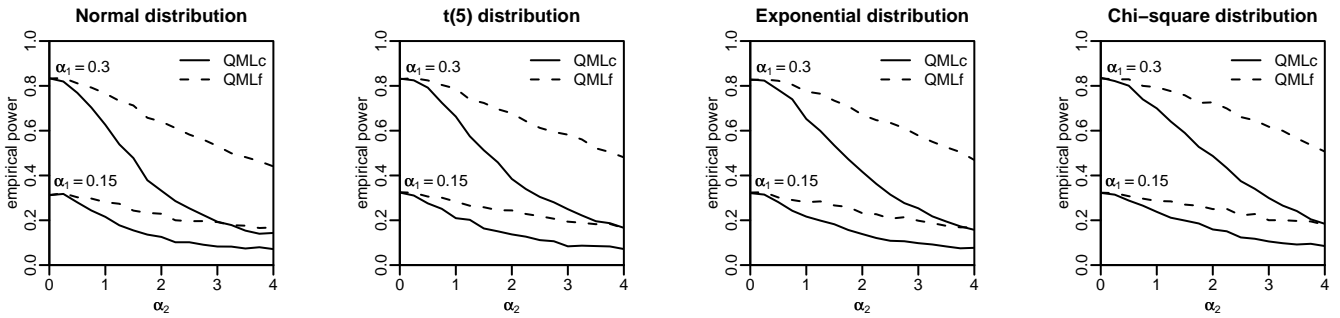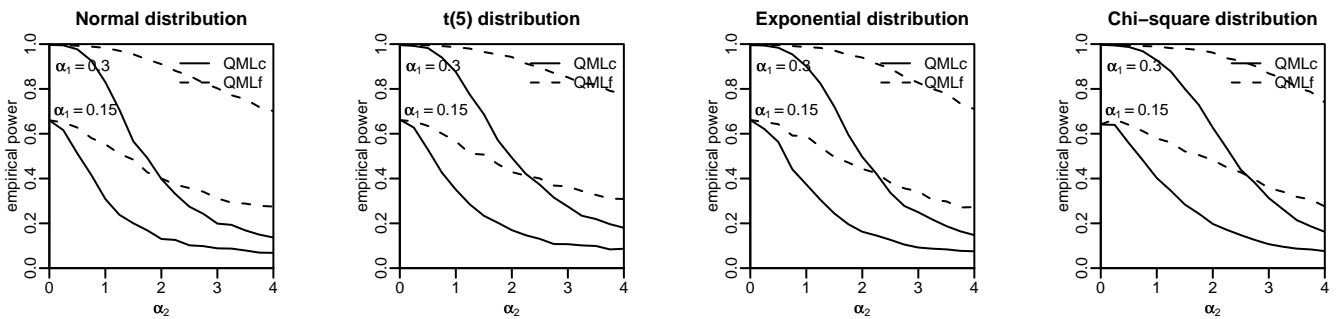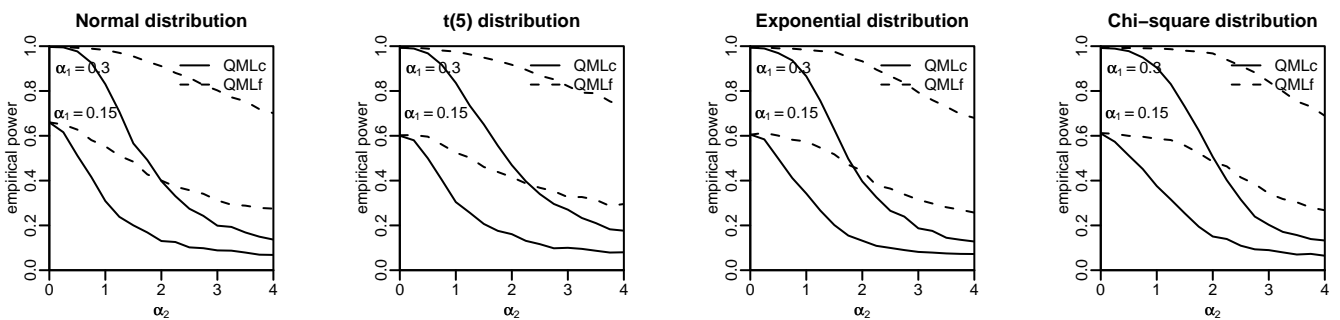
Figure 9: Attenuation bias of Wald statistics in binary regression models

Logit model
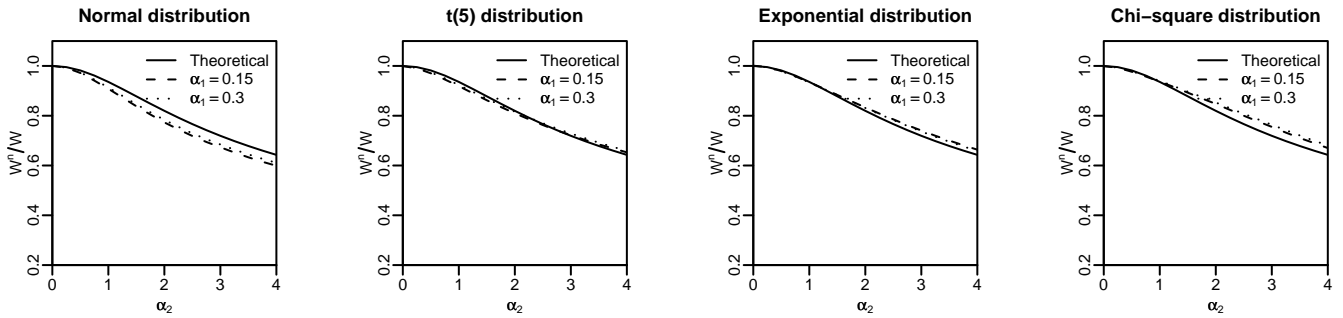
Probit model
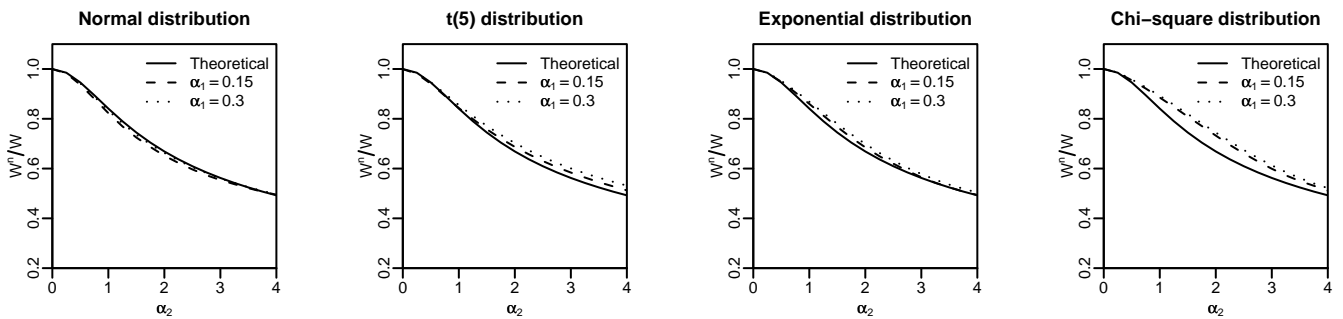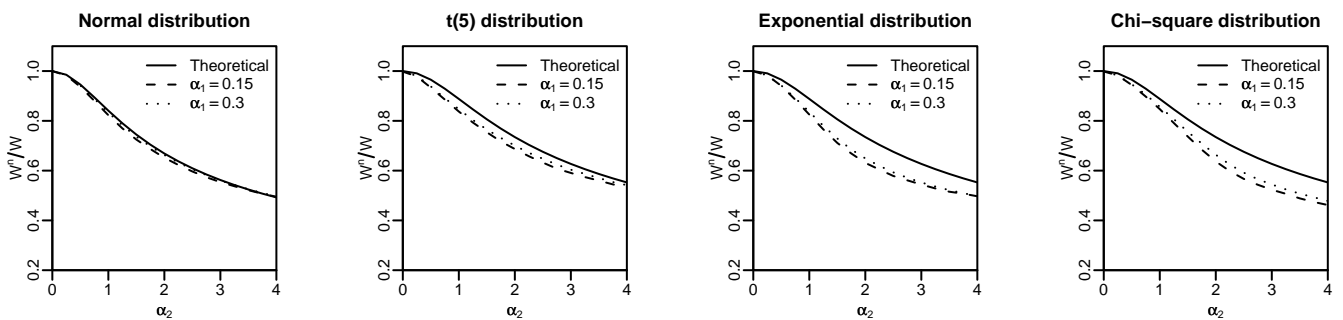
Loglog model

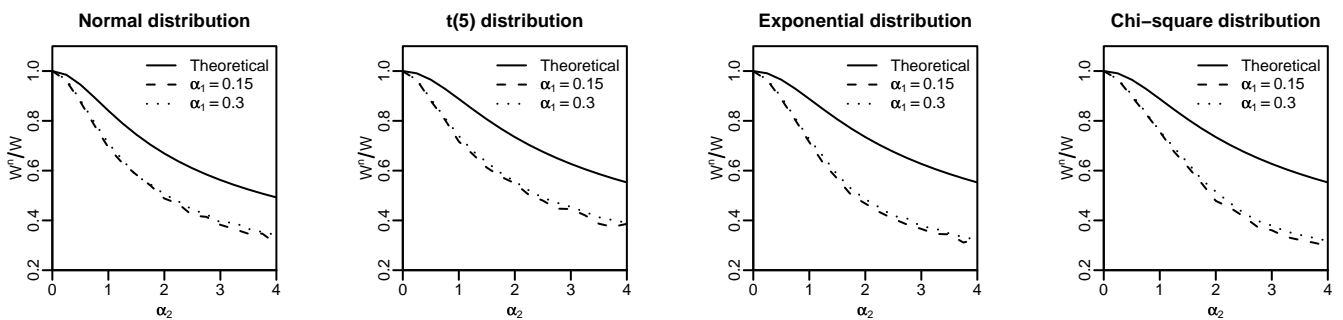Figure 10: Attenuation bias of Wald statistics in fractional regression models
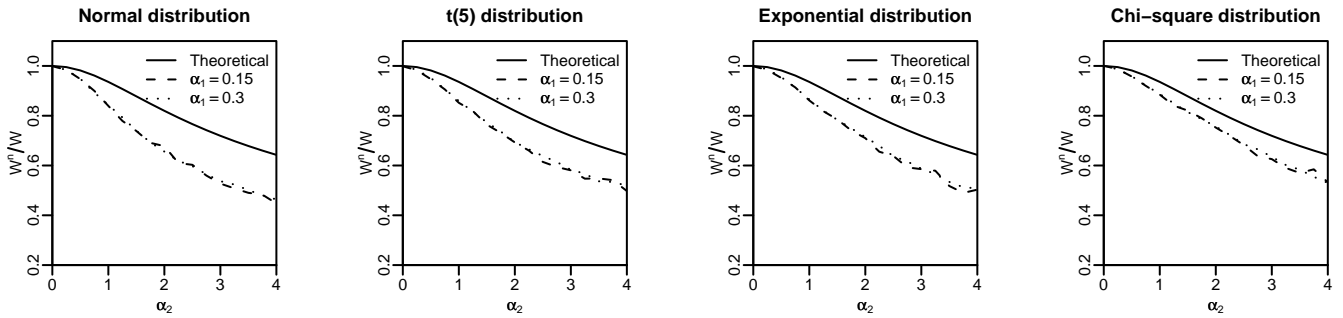
Figure 11: Empirical power – different sample sizes (chi–square–distributed heterogeneity; $\alpha_1 = 0.15$)