SUPPLEMENTARY MATERIAL

Flexible Tweedie regression models for continuous data

NAME 1[a] and NAME 2[b]

[a]ADDRESS 1; [b]ADDRESS 2.

ARTICLE HISTORY
Compiled January 16, 2017

ABSTRACT
Supplementary material for the article "Flexible Tweedie regression models for continuous data" by NAME 1 and NAME 2 submitted to Journal of Statistical Computation and Simulation. In this supplementary material we present an extra simulation study to explore the flexibility of Tweedie regression models to deal with heavy-tailed data as generated by the t-Student and slash distributions. Furthermore, we present two data analyses illustrating the application of Tweedie regression models for highly right-skewed and symmetric positive data. Finally, two extra figures to illustrate the results from the simulation study presented in the paper are presented.

1. Simulation study: robustness of Tweedie regression models

In this Section we present a simulation study that was conducted to evaluate the robustness of the Tweedie regression models in the case of model misspecification by heavy tailed distributions. We generated 1000 data sets considering four sample sizes $100, 250, 500$ and 1000 following two heavy tailed distributions, namely, t-Student and slash. The parametrization adopted was the one implemented in the `R` package `heavy` [4]. For both distributions, we designed three simulation scenarios according to the amount of variation introduced in the data. We defined, small, medium and large amount of variation data sets generated using dispersion parameter equals to 100, 500 and 1000, respectively. In order to simulate challenge data sets, we used 2 degrees of freedom. The mean structure was specified as in the Section 4 (see, main article). In the case of heavy tailed distributions, we expect negative values for the power parameter. Thus, we fitted the Tweedie regression models by using the quasi- and pseudo-likelihood approaches.

In order to compute the empirical efficiency of the quasi- and pseudo-likelihood estimators, we fitted t-Student regression models along with the logarithm link function, as implemented in the package `gamlss`(family `TF`) [5]. Although, of the extensive literature on robust estimation methods, in this paper we adopted the t-Student regression models, since it is a frequent choice for the analysis of heavy tailed data [2] and can be fitted using the orthodox maximum likelihood method. Furthermore, since there is no software available for fitting slash regression models using logarithm link function, the t-Student regression models were used as the base of comparison for both t-Student and slash data sets. Fig. S1 shows the bias plus and minus the standard error for the regression parameters by estimation methods, sample size and simulation scenarios.

CONTACT NAME 1. Email: xx@xx.xx

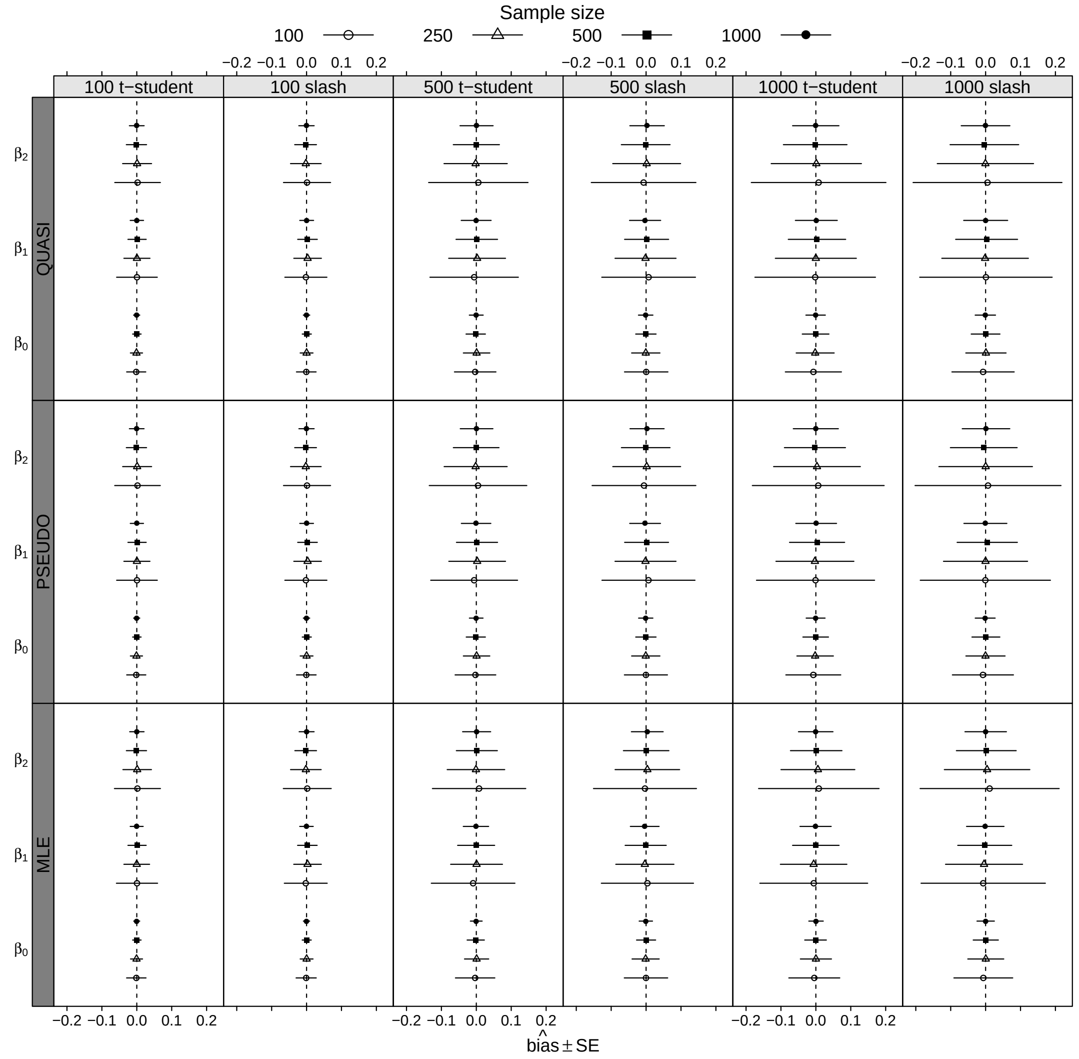

Figure S1.: Bias and confidence interval by estimation methods (quasi-likelihood (QMLE), pseudo-likelihood (PMLE) and maximum likelihood (MLE)), sample size and simulation scenarios.

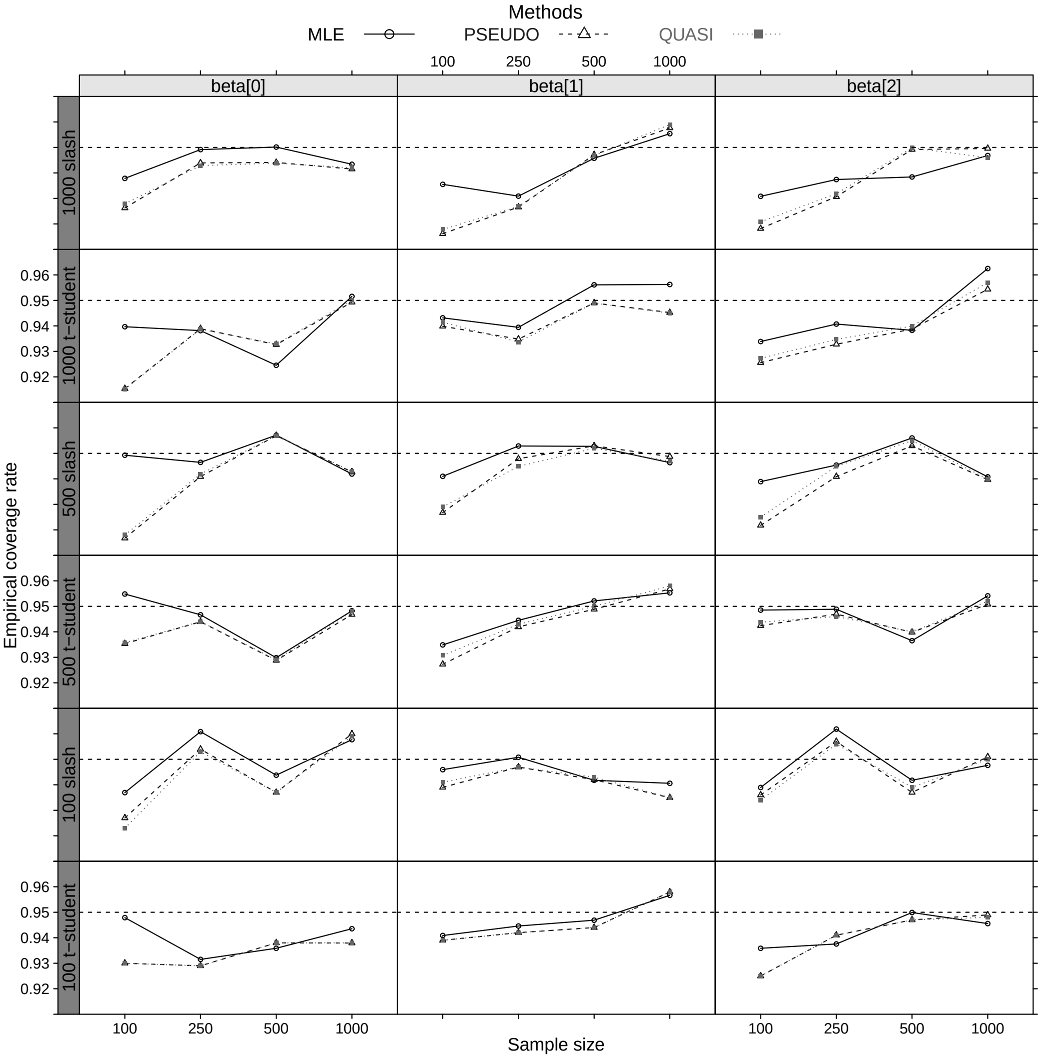

Figure S2.: Coverage rate for regression parameters by estimation methods (quasi-likelihood (QMLE), pseudo-likelihood (PMLE) and maximum likelihood (MLE)), sample size and simulation scenarios.

The results presented in Fig. S1 show that the three estimation methods provide unbiased and consistent estimates of the regression parameters in all simulation scenarios. As expected, the standard errors associated with the regression parameters increase while the amount of variation introduced in the data increases. Fig. S2 presents the coverage rate by estimation methods, sample size and simulation scenarios.

The empirical coverage rate presented values close to the nominal specified level of 95% for all estimation methods and simulation scenarios. The MLE method presented coverage rate closer to the nominal level than the QMLE and PMLE methods, however, the difference is no larger than 3%. The coverage rate of the QMLE and PMLE were virtually the same for all regression parameters, sample size and simulation scenarios. Finally, Fig. S3 presents the empirical efficiency of the QMLE and PMLE estimators for the regression parameters. The empirical efficiency was computed as the ratio between the variance of the MLE obtained by fitting the t-Student regression models and the variance of the QMLE and PMLE estimators obtained by fitting the Tweedie

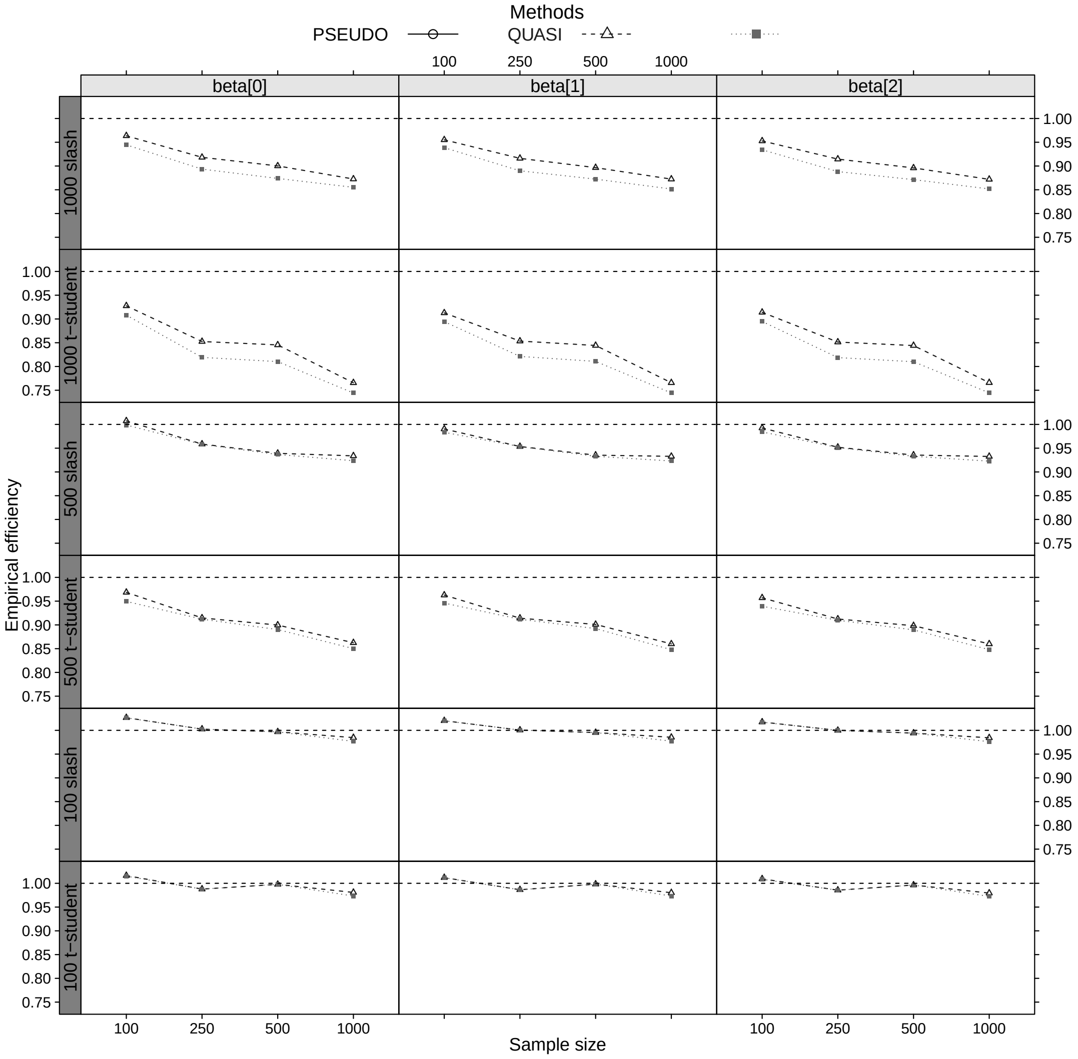

Figure S3.: Empirical efficiency for regression parameters by estimation methods (quasi-likelihood (QMLE), pseudo-likelihood (PMLE) and maximum likelihood (MLE)), sample size and simulation scenarios.

regression models.

The empirical efficiency presented values close to 1 for the small variation simulation scenarios, however, when the amount of variation increases both QMLE and PMLE loss efficiency. The loss were around 10% and 20% for the medium and large variation scenarios, respectively. The results are worse for large samples. The PMLE presents efficiency slightly closer to the nominal level than the QMLE.

2. Data analyses

In this section we shall present two extra illustrative examples of Tweedie regression models. The data that are analysed and the programs that were used to analyse them can be obtained from: omitted for double-blind reviewing

2.1. *Income dynamics in Australia*

We consider some aspects of a cross-section study on earnings of 595 individuals for the year 1982 in Australia. The data set is available in the package `AER` [3] for the statistical software `R`. The response variable `wage` is known to be highly-right skewed. The data set has 12 covariates: `experience` years of full-time work experience; `weeks` weeks worked; `occupation` factor two levels (white-collar, blue-collar); `industry` factor two levels (no;yes) indicating if the individual work in a manufacturing industry; `south` factor two levels (no;yes) indicating if the individuals resides in the south; `smsa` factor two levels (no;yes) indicating if the individual resides in a standard metropolitan area; `gender` factor indicating gender (male, female); `union` factor two levels (no, yes) indicating if the individual's wage set by a union contract; `ethnicity` factor indicating ethnicity, African-American (afam) or not (other). The main goal of the investigation was to assess the effect of the covariates on the wage. We fitted the Tweedie regression model with linear predictor composed by all covariates by using the three estimation methods. Table S1 shows the estimates and standard errors for the regression, dispersion and power parameters.

Supplementary Table S1.: Regression, dispersion and power parameter estimates and standard errors (SE) by estimation methods for the income dynamics data.

	Estimation methods					
Parameter	MLE		QMLE		PMLE	
	Estimate	SE	Estimate	SE	Estimate	SE
`Intercept`	5.8580	0.1723	5.8480	0.1813	5.9137	0.1859
`experience`	0.0056	0.0013	0.0056	0.0014	0.0068	0.0013
`weeks`	0.0034	0.0026	0.0035	0.0028	0.0041	0.0030
`occupation`	-0.1870	0.0365	-0.1893	0.0362	-0.1977	0.0352
`industry`	0.0716	0.0293	0.0731	0.0302	0.0229	0.0322
`south`	-0.0375	0.0305	-0.0363	0.0320	-0.0104	0.0341
`smsa`	0.1644	0.0293	0.1658	0.0297	0.1456	0.0312
`married`	0.1172	0.0478	0.1218	0.0523	0.0902	0.0538
`gender`	-0.3389	0.0570	-0.3346	0.0567	-0.4039	0.0562
`union`	0.1265	0.0314	0.1331	0.0298	0.0839	0.0293
`education`	0.0577	0.0065	0.0578	0.0069	0.0543	0.0074
`ethnicity`	-0.1793	0.0506	-0.1772	0.0510	-0.1466	0.0484
δ	-5.9848	1.1117	-6.8587	2.0409	-7.1317	1.8857
p	2.5354	0.1605	2.6656	0.2979	2.7012	0.2735

The results in Table S1 show that the MLE and QMLE approaches strongly agree in terms of estimates and standard errors for the regression coefficients. The PMLE approach presents estimates slightly different from the MLE and QMLE approaches. Regarding the dispersion parameters, although the slightly difference in terms of estimates and standard errors, the confidence intervals from the QMLE and PMLE approaches contain the MLE estimates.

Concerning the effect of the covariates the MLE and QMLE approaches agree that the covariates `weeks` and `south` are non-significant. On the other hand, the PMLE approach also indicated that the covariates `industry` and `married` are non-significant. Regarding the other covariates the three approaches agree that they are significant.

In order to compare the fit of Tweedie regression model with more standard approaches, we also fitted the Gaussian, gamma and inverse Gaussian regression models for the income dynamics data set. The maximized values of the log-likelihood function were -4437.51, -4318.08 and -4316.52 for the Gaussian, gamma and inverse Gaussian models, respectively. Furthermore, the maximized value of the log-likelihood function for the Tweedie regression model was -4312.39, which in turn shows the better fit of the Tweedie regression model, as expected. In terms of computational time for this data set, the QMLE approach was 45 and 0.15 times faster than the MLE and PMLE approaches, respectively.

2.2. *Gain in weight of rats*

The third example concerns to a standard Gaussian regression model. The goal of this example is to show that the quasi- and pseudo-likelihood approaches can estimate values of the power parameter between 0 and 1, where the maximum likelihood estimator does not exist. We used the `weightgain` data set available in the `HSAUR` package [1]. This data set corresponds to an experiment to study the gain in weight of rats fed on four different diets, distinguished by the amount of protein (low and high) and by source of protein (beef and cereal). The data set has 40 observations.

We fitted the Gaussian, gamma, inverse Gaussian and Tweedie regression models for the `weightgain` data set. The linear predictor was composed of the two main covariates `source` and `type` along with the interaction term, for all models. The values of the maximized log-likelihood were -162.84, -164.21, -165.36 and -163.50 for the Gaussian, gamma, inverse Gaussian and Tweedie models, respectively. These results showed that the Gaussian distribution provides the best fit for this data set, judging by the maximized log-likelihood value. In that case, the MLE method is not able to indicate the best fit. It is due to the non-trivial restriction on the power parameter space. Thus, we fitted the model using the approaches QMLE and PMLE. Table S2 presents the estimates and standard errors for the regression, dispersion and power parameters, obtained by MLE, QMLE and PMLE approaches.

Supplementary Table S2.: Regression, dispersion and power parameter estimates and standard errors (SE) by estimation methods for the gain in weight of rats data.

	Estimation methods					
Parameter	MLE		QMLE		PMLE	
	Estimate	SE	Estimate	SE	Estimate	SE
`Intercept`	4.5891	0.0504	4.6051	0.0454	4.6050	0.0453
`source`	−0.1263	0.0734	−0.1519	0.0693	−0.1517	0.06867
`type`	−0.2235	0.0750	−0.2331	0.0694	−0.2337	0.06922
`source:type`	0.1827	0.1069	0.2096	0.1036	0.2108	0.1026
δ	0.6323	8.1352	3.3614	8.7203	3.3355	9.0088
p	1.0590	1.8400	0.4350	1.9484	0.4408	2.0129

The results in Table S2 show that the three approaches strongly agree in terms of estimates and standard errors for the regression coefficients. The value of the power parameter was estimated smaller than 1 by the QMLE and PMLE approaches, as expected, since the Gaussian distribution provides the best fit for this data. On the other hand, the maximum likelihood method estimated the power parameter close to 1 the border of the parameter space, in that case a non-optimum model. All approaches presented large standard errors for the power and dispersion parameters. In terms of computation time, for this application the PMLE approach was 94 and 0.15 times faster than the MLE and QMLE approaches, respectively.

3. Extra figures

References

[1] Everitt, B. S. and Hothorn, T. [2015]. *HSAUR: A Handbook of Statistical Analyses Using R (1st Edition).* R package version 1.3-7.

[2] Huber, P. J. and Ronchetti, E. M. [2009]. *Robust Statistics*, John Wiley & Sons, Inc., London.

[3] Kleiber, C. and Zeileis, A. [2008]. *Applied Econometrics with R*, Springer-Verlag, New York.

[4] Osorio, F. [2016]. *heavy: Robust estimation using heavy-tailed distributions.* R package version 0.3.
URL: *http://cran.r-project.org/package=heavy*

[5] Rigby, R. A. and Stasinopoulos, D. M. [2005]. Generalized additive models for location, scale and shape,(with discussion), *Applied Statistics* **54**: 507–554.

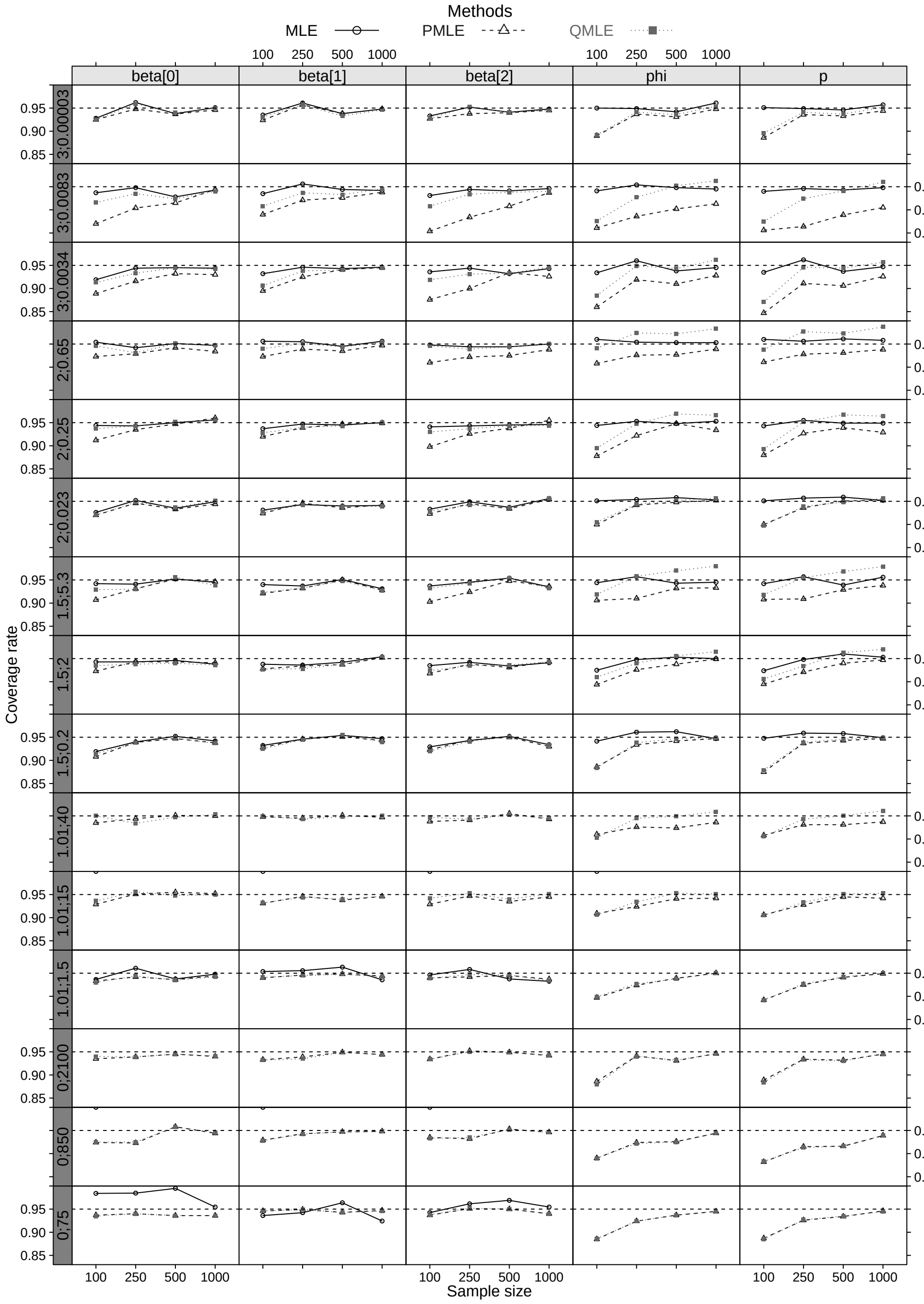

Figure S4.: Coverage rate for each parameter $(\beta_0, \beta_1, \beta_2, \phi, p)$ by estimation methods (maximum likelihood (MLE), pseudo-likelihood (PMLE) and quasi-likelihood (QMLE)), sample size and different values of the power and dispersion parameters $(p; \phi)$.

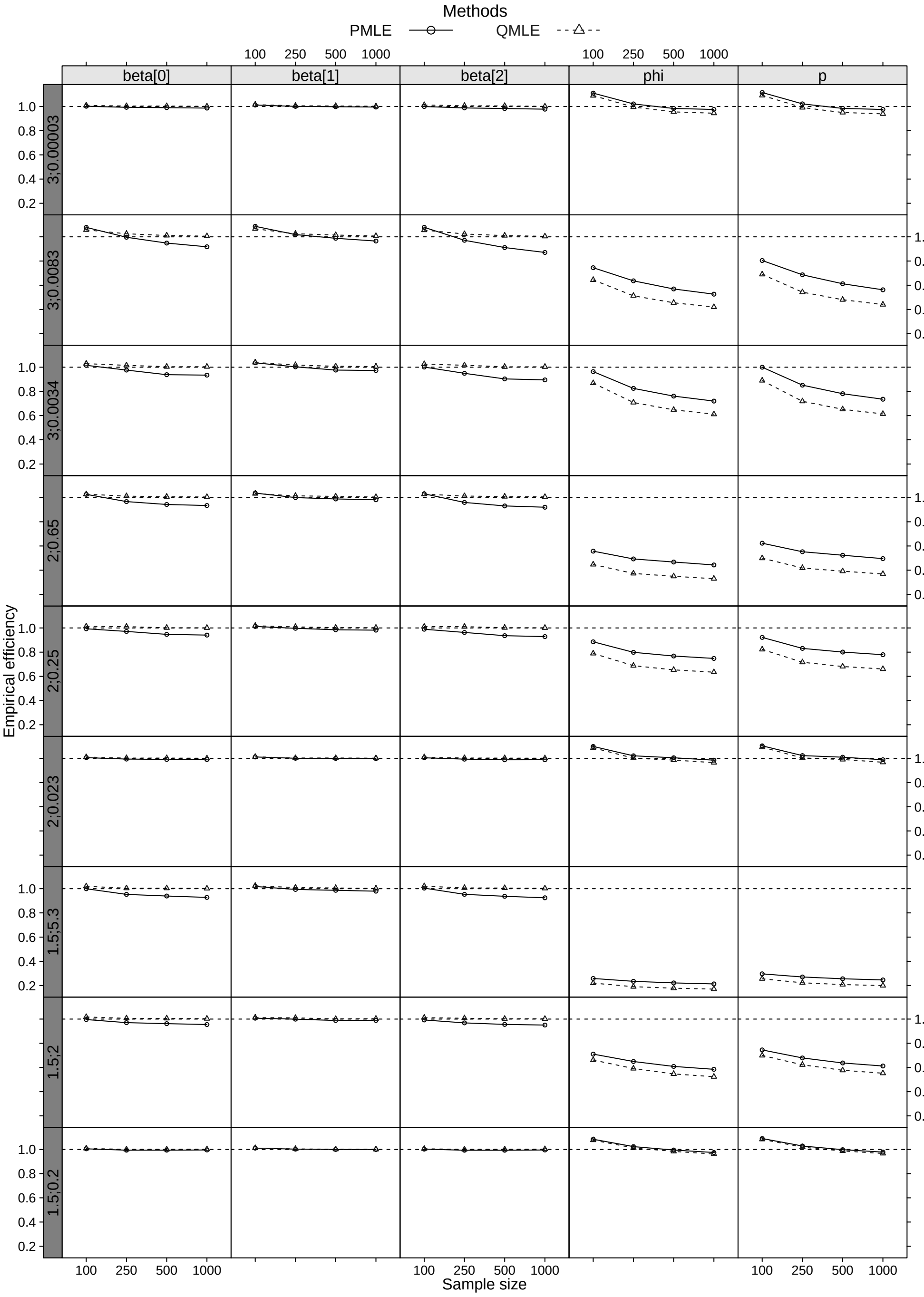

Figure S5.: Empirical efficiency for each parameter $(\beta_0, \beta_1, \beta_2, \phi, p)$ by estimation methods (maximum likelihood (MLE), pseudo-likelihood (PMLE) and quasi-likelihood (QMLE)), sample size and different values of the power and dispersion parameters $(p; \phi)$.