Robust regression approach to the analysis of cost data

Petrinco M[*a], Barbati G[b], Pagano E[c], Gregori D[b]

[a] Department of Public Health and Microbiology and Department of Statistics and Applied Maths "Diego de Castro", University of Turin, Italy

[b] Department of Public Health and Microbiology, University of Turin, Italy

[c] Unit of Cancer Epidemiology, University of Turin, CERMS and CPO-Piemonte, Italy

* Corresponding author: Michele Petrinco, Phd student

Department of Public Health and Microbiology, University of Turin

Via Santena 7 – 10126 Turin, Italy

Tel: +39 0116334572; Mobile: +39 3405197308; Fax: +39 0116334664

Email address: petrinco@econ.unito.it

ABSTRACT

The population mean cost of patients with a certain pathology is the parameter of interest for allocating health resources. It generally depends upon a number of covariates and the presence of outliers yields difficult its estimation. Recent research in parametric robust estimation propose estimators for a class of regression models with asymmetric (or symmetric) error distribution: the 'Truncated Maximum Likelihood' (TML) estimators (Marazzi and Yohai, 2004). The TML procedure assumes that the error term belongs to a location-scale family distribution, like Gaussian or Log-Weibull, and a TML-estimate is computed in three steps. First, an initial high breakdown point but inefficient robust estimate is computed; second, observations that are unlikely under the estimated model are rejected; third, the maximum likelihood estimate is computed with the retained observations.

In the present work, we compared results obtained by this parametric robust procedure with respect to the GLM (Generalized Linear Model) Gamma with log link, both in a simulation study and in a cardiovascular trial. The simulation experiment has been performed both at the nominal model and in presence of outliers.

In presence of outliers the robust procedure outperforms the Gamma estimator both in the simulation scenario and in the real data application.

INTRODUCTION

Cost associated with a certain pathology could be modelled as a positive random variable with asymmetric distribution. Often the population mean cost is the parameter of interest for allocating resources in the health care system, and it depends upon a number of covariates. Unfortunately, cost data may contain outliers and the mean is a difficult parameter to be estimated well: the sample mean, which is the natural estimate, is very non robust (Marazzi and Ruffieux, 1999). In the real applications, often a few atypical observations drastically change the mean estimate and the common tests of means (e.g. t-test and its variants) lead to a different decision when these outliers are removed from the data set (Marazzi and Barbati, 2002).

One of the main approaches to regression with asymmetric errors are Generalized Linear Models (GLM) (McCullagh and Nelder, 1989), that allow modelling of the mean with the help of covariates using maximum likelihood estimation, that is very sensitive to outliers. Recent research in parametric robust estimation have proposed the use of high efficiency and high breakdown point estimators suitable for a class of regression models with asymmetric (or symmetric) error distribution: the 'Truncated Maximum Likelihood' estimators (TML, Marazzi and Yohai, 2004). Within this framework a robust parametric mean is defined as the mean of the estimated model and it can be interpreted as a robust estimate of the population mean after removal of the extreme observations.

In the present work, we compared results obtained by applying a Gamma GLM with respect to the TML estimator both on simulations and on a real dataset, characterized by the presence of an heavy tail in the cost distribution: the COSTAMI data, an international randomized trial on costs of treatment of the uncomplicated myocardial infarction (AMI).

METHODS

The GLM Gamma takes the form: $\mathbf{C}=\mathbf{X}\,\beta + \varepsilon$, where **C** is the outcome, i.e. in our case the cost vector, **X** is the covariate matrix and **β** are the associated regression coefficients; **ε** is a random error term, distributed as a Gamma with a mean of zero and variance σ^2. In the version of generalized linear model framework that we consider, we assume that E(**C**/**X**) exhibits an exponential conditional mean or log-

link relationship: E(**C**/**X**)=exp(**X β**). This model is known to fit reasonably well the positively skewed cost data (Basu et al., 2004).

The TML procedure assumes that in the model: log(**C**)=**X** β + ε, the error term **ε** belongs to a location-scale family distribution, like Gaussian or Log-Weibull. A TML-estimate is computed in three steps. As a first step, an initial high breakdown point but inefficient S-estimate (Rousseeuw and Yohai, 1984) is computed; in a second step, observations that are unlikely under the estimated model are rejected; in a third step, the maximum likelihood estimate is computed with the retained observations.

In the second step of the above procedure, a rejection rule has to be chosen to determine the outliers observations. In the present application we used an adaptive version (ATML-estimator) of this rejection rule, defined as follows: (i) the empirical distribution of the observed likelihoods with respect to the initial estimate is obtained, by using the standardized residuals, and compared with the theoretical one; (ii) observations with the smallest likelihoods are rejected, so that the empirical distribution of the remaining observed likelihoods is stochastically smaller than the theoretical one. No observation is therefore (asymptotically) rejected when the data agree with the model. The ATML-estimator (Adaptive Truncated Maximum Likelihood) is therefore fully efficient at the model; nevertheless, its breakdown point is not smaller than the breakdown point of the initial estimator (e.g. 50%) (Marazzi and Yohai , 2004).

Simulations

The performance of the GLM Gamma versus the ATML-estimator was evaluated with the help of Monte Carlo simulations. Two covariates were considered and observations (x_{1i}, x_{2i}, y_i) were generated according to the nominal model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i \qquad i = 1,..., n$$

$$x_1 \sim Unif(0,1)$$

$$x_2 \sim N(0,1)$$

with $\beta_0 = 0, \beta_1 = \beta_2 = 1$ and the error distribution alternatively set to be standard Gaussian and Log-Weibull with shape and scale equal to one, in line with the

assumption and the study design of (Marazzi and Yohai, 2004). It were passed to the Gamma model the exponential values of the generated y data.

Figure 1 panel (a) shows an example of simulated cost data with Log-Weibull errors. Six different sample sizes (k=20, 50, 200, 500, 1000, 2000) were considered for each distribution; for each of the 1000 Monte Carlo experiments the two models, GLM Gamma and ATML-estimator, were fitted. At each replication of the simulation experiments, the Mean Squared Error (MSE) and the coverage probability (i.e. the proportion of 95% confidence intervals for the estimated coefficients containing the true value of the parameters) of the two regression coefficients β_1, β_2 were computed for the two estimators.

Simulations under outliers contamination

In order to assess the behaviour of the estimators in presence of outliers, the Log-Weibull case was considered, being the more realistic scenario for a cost distribution. We generated contaminated samples where a fraction ε, respectively of 10%, 20% and 30%, of the error distribution was replaced with uniform random values distributed in a range between the maximum of the errors and twice and a half this maximum. Figure 1 panel (b) shows an example of simulated cost data with Log-Weibull errors containing 30% of outliers. At a fixed sample size of 1000, GLM Gamma and ATML-estimator were estimated for one-thousand Monte Carlo experiments. At each replication of the simulation experiments, the coverage probability of the two regression parameters and the Mean Squared Error were computed for the two estimators.

All computations were performed with the software S-Plus version 6.0 with the Robeth and TMLest libraries.

Real data: the COSTAMI Trial

Between January 1998 and August 2000, patients from various participating centres (Italy, Turkey, Spain and Poland) with a recent AMI were recruited and randomly assigned to one of four treatment strategies. Data were collected within two different prospective randomised multi-centre trials based on the same protocol criteria of patients inclusion and exclusion, described elsewhere (Desideri et al, 2003, 2005).

In these trials an early discharge strategy of treatment was compared with three variants of the usual treatment of patients with AMI (Ryan et al, 1999) on a sample of 720 patients with a recent uncomplicated AMI. It is to be noted that in the present study, the factor Strategy has been reduced at two levels: the comparison was done between the early discharge strategy versus the other strategies, taking as reference. Data were collected on patient characteristics, AMI severity measures, prescribed medications and centre characteristics. Total medical costs per patients were measured as the sum of initial hospital costs and follow-up hospital and outpatients costs at 1 year. Costs are expressed in Euro. In the analysis we used an "one-country" cost approach, applying unit cost estimates from one centre to use for resources calculated in each centre. The use of resources was calculated considering direct medical costs of hospitalizations, investigations and interventions and quantified using the mean reimbursement for the diagnosis-related group in the Friuli Venezia Giulia Region, the site of the coordinating centre.

The set of covariates considered for comparing models included: demographic characteristics (sex, age), clinical characteristics of the patients (diabetes, hypertension, previous AMI), the strategy of treatment and an additional set of centre level variables (number of beds in cardiac department, number of hospitalizations and number of MI treated by the hospital). Variables were included in the analysis on the basis of an a priori reason for being associated with treatment costs. They were all inserted in the models and no formal selection regarding covariates has been conducted. Missing values of each variable inserted in the models were excluded from the analysis.

RESULTS

Simulations

Table 1 gives coverage probability for the regression coefficients, Mean Squared Error (MSE) at the nominal Normal and Log-Weibull error distributions for the two estimators. In terms of the coverage probability the Gamma model outperforms the ATML estimator for the Normal case; in the Log-Weibull scenario starting from n=200 the two models show a similar behaviour. The MSE for the Gamma model is sensitive to the sample size, showing a decreasing

behaviour for increasing dimensions; on the contrary, the ATML presents quite low and stable values of the MSE nearly independently from the sample size.
As expected, the Gamma model is globally performing well, having a maximum MSE in estimating the mean cost of about 16% in presence of very small sample sizes (n=20) and presenting a good performance in terms of coverage.

Simulations under outliers contamination

Table 2 gives Coverage probability of the estimated regression coefficients and the average Mean Squared Error of the two estimators for the contaminated Log-Weibull error distribution. For increasing percentages of contamination degree the ATML estimator performs better than the Gamma in terms of coverage probability. The MSEs of the ATML estimator are very close over the different percentages of point contamination. In addition, in all scenarios they result strongly smaller than the MSEs of the Gamma estimator.
An important effect of the outliers is in fact to inflate the classical scale estimates of the ATML estimator.
Figure 2 shows an example of simulated cost data on a log scale with 10% of outliers contamination plotted versus the corresponding predicted costs by the ATML procedure: filled circles denote outliers marked adaptively by the robust procedure and triangles indicate the *x*-range of the "true" outliers added. To be noted the nearly complete detection (93%) of the outliers with respect to the majority of the data at the nominal model (empty circles).

Real data

Table 3 shows the results provided by the GLM Gamma, the ATML estimates (Normal and Log-Weibull errors) and the GLM Gamma applied on the reduced dataset, i.e. by removing the observations marked as outliers by the ATML Log-Weibull procedure. We selected the Log-Weibull adaptive cut-off for its smaller scale estimate and better fit to the data with respect to the Normal model.
In the initial GLM Gamma, the significant covariates on cost are Strategy and Hypertension, this last associated with increasing cost. In the Log-Weibull ATML estimate, Strategy becomes much more significant in lowering cost, Hypertension

is no longer significant, instead Diabetes is marked as a relevant factor for cost increase. Moreover, two variables characterizing the hospital (number of beds in cardiology and MI treated) are indicated as influencing cost. By appling the GLM Gamma on the reduced dataset, the same conclusions of the robust model on the full dataset are drawn.

In table 4, descriptive statistics of the significant covariates and of the costs in the full dataset, in the group of observations retained and in the outliers group are reported. Outliers are clinically characterized by a major incidence of hypertension and diabetes, and come from hospitals with less beds in cardiology and less number of MI treated on average. All these factors contribute significantly to a large difference in median cost between the „majority“ of cases (2892 €) versus the outliers group (15800 €).

Figure 3 shows the cost data distribution of the COSTAMI dataset: it seems evident the presence of the two populations detected by the robust procedure: the majority of observations with costs under 10000 € and the outliers group with costs ranging from 10000 up to about 35000 €.

CONCLUSIONS

In the present paper we compared results obtained by a parametric robust procedure with respect to a classical GLM Gamma model, widely applied in the cost regression analysis, both on simulations and on a real dataset.

In the simulation framework, the robust procedure showed a satisfactory performance for large sample sizes, comparable to the Gamma model; the performance of the ATML estimator does not appear to be highly appreciable when n is small ($n<200$), as also reported in (Marazzi and Yohai, 2004). The ATML estimator showed an outperforming behaviour in presence of outliers, the more realistic situation in the cost analysis.

In the real dataset, the significance of some parameters changed, with consequent important changes in the interpretation of the study results. As it is well known, even a few atypical observations can drastically change the mean estimate producing different decision in the common tests: for example when comparing cost means among different hospitals or over different periods of time this problem could not be ignored. Moreover, by isolating the outliers group in a

parametric adaptive way, i.e. not fixing an a-priori cut-off, that could be highly subjective, but instead following the parametric fit of a model on the sample data, these observations with “extremely high” values for the cost measure could be characterized on the basis of their observed covariates, and this might help in identifying class of patients with higher risk of having a huge impact on the health care expenditure.

A future extension of the present ATML procedure will be in developing the class of TML estimators in the case of gamma distributed random errors; at present the gamma TML is available only for the univariate case. Moreover, a promising development could be to generalize the TML approach to case of censored responses, since, as its is well known, cost censoring is very informative and standard survival techniques cannot be applied.

ACKNOWLEDGMENTS

We would like to thank Prof. Alfio Marazzi for his comments and suggestions.

This work was supported by the Compagnia di San Paolo.

REFERENCES

Desideri A, Fioretti PM, Cortigiani L, et al. Cost of strategies after myocardial infarction (COSTAMI): a multicentre, international, randomized trial for cost-effective discharge after uncomplicated myocardial infarction. Eur Heart J. 2003;24:1630-9.

Desideri A, Fioretti PM, Cortigiani L, et al. Pre-discharge stress echocardiography and exercise ECG for risk stratification after uncomplicated acute myocardial infarction: results of the COSTAMI-II (cost of strategies after myocardial infarction) trial. Heart. 2005;91:146-51.

Basu A, Manning WG, Mullahy J. Comparing alternative models: log vs Cox proportional hazard? Health Econ. 2004;13:749-65.

McCullagh P., Nelder J.A., 1989. Generalized linear models. Second Edition. Chapman and Hall, New York.

Marazzi A, Barbati G. Robust parametric means of asymmetric distributions: estimation and testing. Estadistica. 2002; 54: 47-72.

Marazzi A., Yohai V. J. Adaptively truncated maximum likelihood regression with asymmetric errors. Journal of Statistical Planning and Inference, 2004, Vol 122/1-2, pp. 271-291.

Marazzi A., Ruffieux C. The truncated mean of an asymmetric distribution. Computationla Statistics and Data Analysis. 1999, 32: 79-100.

Ryan TJ, Antman EM, Brooks NH, et al. 1999 update: ACC/AHA Guidelines for the Management of Patients With Acute Myocardial Infarction: Executive Summary and Recommendations: A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee on Management of Acute Myocardial Infarction). Circulation. 1999;100:1016-30.

Rousseeuw, P.J., Yohai V. J. Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis*, J. Franke, W. Hardle and R.D. Martin (Eds). Lecture Notes in Statistics, (1984), 26, Springer, NY, 256-272.

Table 1. Coverages for regression coefficients, Mean Squared Error (MSE) at the nominal Normal and Log-Weibull models (1000 samples for each sample size).

		GLM Gamma			ATML		
		Cov(β_1)	Cov(β_2)	MSE	Cov(β_1)	Cov(β_2)	MSE
Normal	20	0.881	0.896	0.16548	0.728	0.708	0.00574
	50	0.917	0.909	0.01964	0.807	0.813	0.03539
	200	0.936	0.921	0.03358	0.883	0.876	0.01432
	500	0.937	0.953	0.00076	0.897	0.879	0.01694
	1000	0.955	0.961	0.00357	0.905	0.889	0.01279
	2000	0.946	0.948	0.00098	0.897	0.913	0.00882
Log Weibull	20	0.890	0.900	0.15284	0.768	0.764	0.11393
	50	0.928	0.934	0.01754	0.877	0.871	0.07205
	200	0.940	0.933	0.01438	0.933	0.904	0.03893
	500	0.936	0.946	0.00452	0.938	0.933	0.02758
	1000	0.948	0.947	0.00284	0.950	0.941	0.01746
	2000	0.933	0.943	0.00001	0.929	0.941	0.01271

Table 2. Coverage probability for regression coefficients, Mean Squared Error, Scale estimate for the ATML estimator under different percentages of outliers contamination (1000 samples for fixed sample size=1000).

	GLM Gamma			ATML			
	Cov(β_1)	Cov(β_2)	MSE	Cov(β_1)	Cov(β_2)	MSE	*Scale*
$\varepsilon = 10\%$	0.921	0.905	0.02298	0.872	0.921	0.00767	1.09
$\varepsilon = 20\%$	0.798	0.847	0.05991	0.894	0.906	0.00326	1.30
$\varepsilon = 30\%$	0.873	0.913	0.03286	0.937	0.951	0.00757	1.66

Table 3. Coefficient estimates, standard errors and t-values provided by the GLM Gamma, the ATML-estimates (Normal and LogWeibull errors) and the GLM Gamma on the outliers-free dataset.

	GLM Gamma			*Normal*			*LogWeibull*			*GLM Gamma (outliers excluded)*		
	β	st.err.	t	β	st.err.	t	β	st.err.	t	β	st.err.	t
Intercept	9.188	0.309	29.68*	8.790	0.234	37.63*	8.990	0.277	32.39*	8.83	0.231	38.23*
Strategy	-0.115	0.043	-2.68*	-0.201	0.060	-3.36*	-0.402	0.068	-5.93*	-0.361	0.060	-6.03*
Gender (M vs F)	-0.036	0.114	-0.32	-0.051	0.079	-0.65	-0.096	0.084	-1.14	-0.082	0.078	-1.06
Age	-0.003	0.004	-0.79	-0.000	0.003	-0.09	0.000	0.003	0.11	0.000	0.003	0.07
Hypertension	0.181	0.084	2.15*	0.051	0.059	0.87	-0.018	0.068	-0.26	-0.022	0.058	-0.38
Previous AMI	-0.058	0.165	-0.35	-0.056	0.116	-0.48	0.140	0.102	1.37	0.151	0.106	1.42
Diabetes	0.181	0.115	1.57	0.112	0.080	1.41	0.198	0.083	2.37*	0.198	0.080	2.47*
Beds in Cardiology	-0.001	0.003	-0.33	0.006	0.002	3.03*	0.006	0.002	3.56*	0.006	0.002	3.35*
Hospitalizations	0.000	0.000	1.05	0.000	0.000	-0.03	0.000	0.000	1.30	0.000	0.000	1.02
N° of MI treated	-0.000	0.001	-0.78	-0.001	0.000	-3.43*	-0.001	0.000	-2.98*	-0.001	0.000	-3.35*
Scale estimate			0.924			0.736			0.565			0.527

* p-value<0.05

Table 4. Descriptive statistics of the significant covariates and of the cost in the full dataset, in the group of observations retained and in the outliers group. Results are number (percentage) and for continuous variables: median [1st; 3rd quartile] and mean (standard deviation).

	Full data (n=635)	Observations Retained (n=456)	Observations excluded (n=179)
Strategy n (%)	233 (36.69%)	163 (35.75%)	70 (39.11%)
Hypertension n (%)	282 (44.41%)	191 (41.88%)	91 (50.84%)
Diabetes n (%)	97 (15.28%)	68 (14.91%)	29 (16.20%)
Beds in Cardiology	21 [14 ; 26] 25.15 (17.75)	24 [16 ; 26] 26.13 (18.54)	21 [10 ; 26] 22.66 (15.33)
N° of MI treated	183 [112 ; 300] 213.94 (111.21)	183 [112 ; 302] 218.49 (112.46)	183 [112 ; 300] 202.36 (107.39)
Cost (€)	4200 [2511.5;11708.9] 8492 (8681.9)	2892 [2279.6 ; 4672.6] 4058.09 (2706.9)	15800.7 [12378.3; 27488] 19787.33 (8430.7)

Figure 1. Panel (a): simulated cost data with Log-Weibull errors; panel (b): simulated cost data with Log-Weibull errors and 10% of outliers contamination; some extreme data are truncated for graphical reasons.

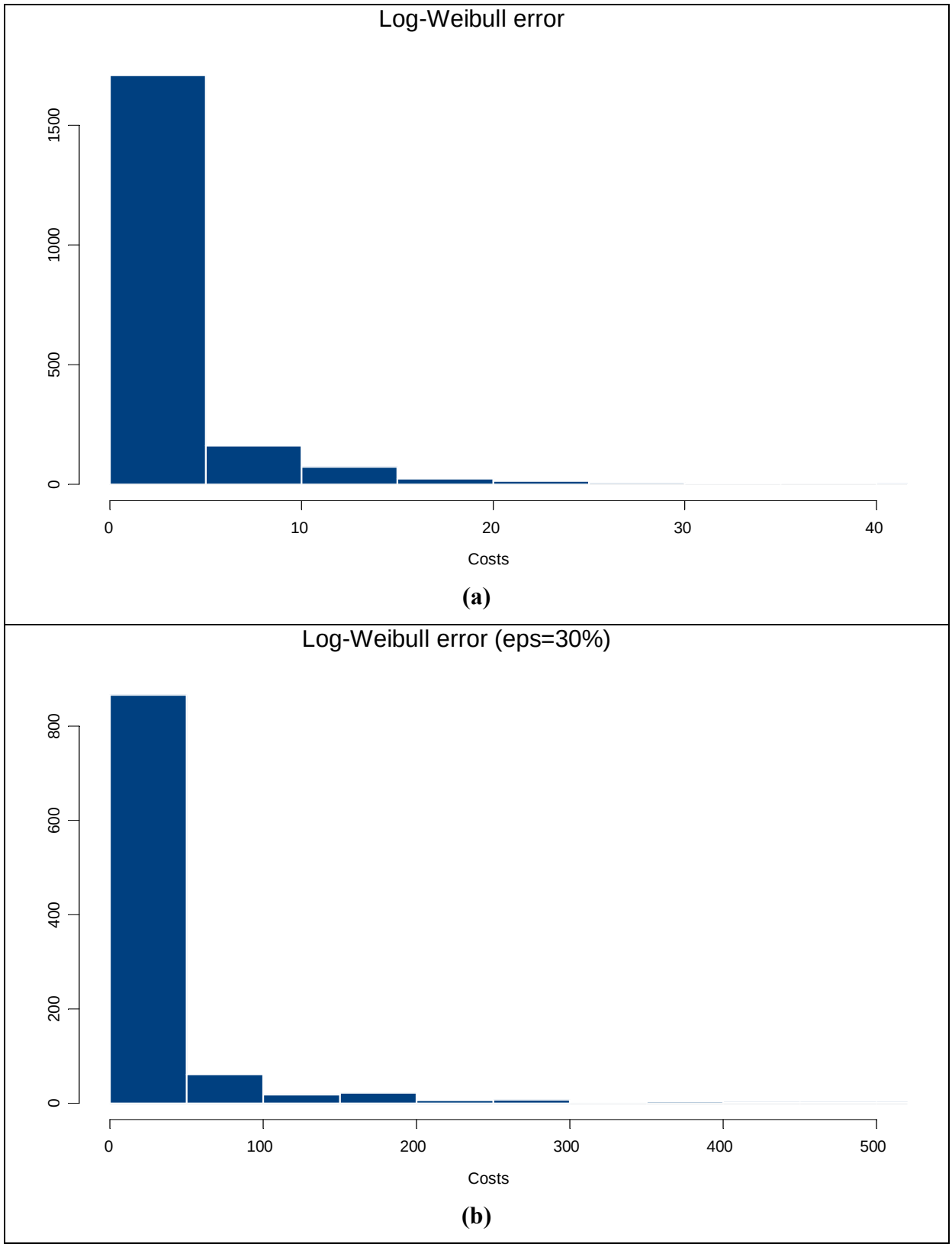

Figure 2. Simulated log(cost) data with Log-Weibull errors and 10% of outliers contamination *versus* predicted log(cost) by the ATML procedure. Filled circles denote outliers marked by the robust procedure; triangles indicate abscissa of the "true" outliers added.

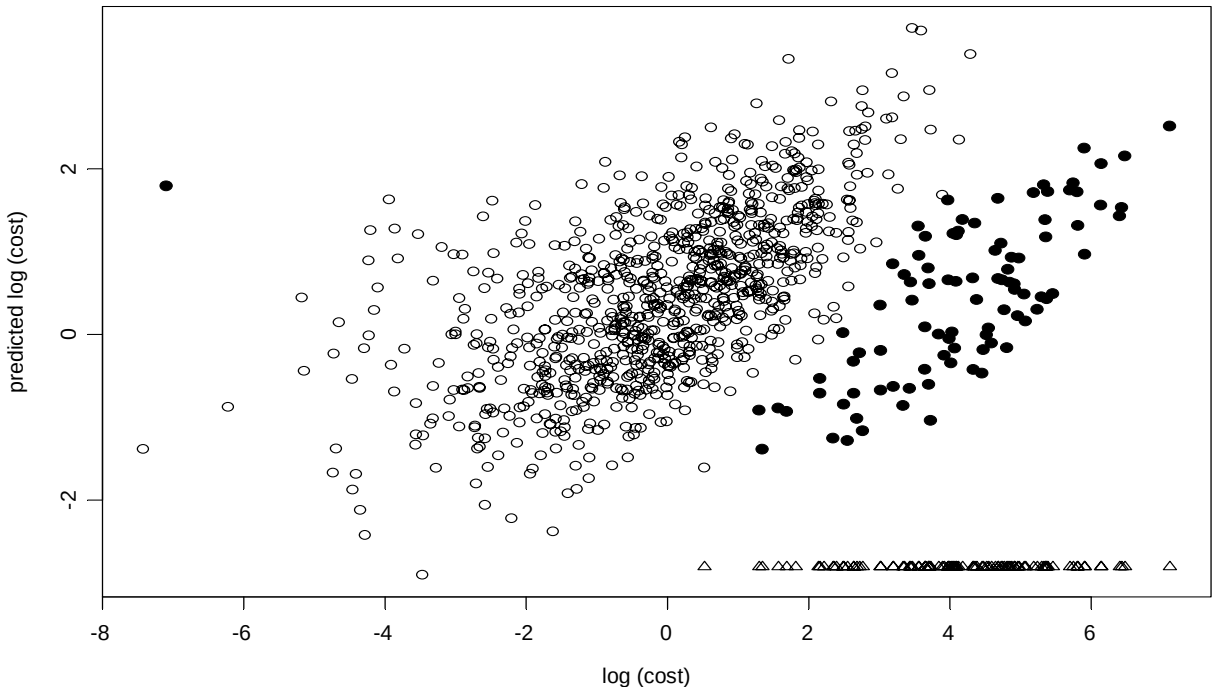

Figure 3. Cost data of the COSTAMI trial.