A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures

Eva Cantoni and Elvezio Ronchetti

No 2004.03

Cahiers du département d’économétrie
Faculté des sciences économiques et sociales
Université de Genève

Mai 2004

Département d’économétrie
Université de Genève, 40 Boulevard du Pont-d’Arve, CH -1211 Genève 4
http://www.unige.ch/ses/metri/
A robust approach for skewed and heavy-tailed outcomes in the analysis of health care expenditures

Eva Cantoni and Elvezio Ronchetti
Department of Econometrics
40, Bd Du Pont d’Arve
University of Geneva
CH - 1211 Geneva 4, Switzerland

May 2004

Abstract
In this paper robust statistical procedures are presented for the analysis of skewed and heavy-tailed outcomes as they typically occur in health care data. The new estimators and test statistics are extensions of classical maximum likelihood techniques for generalized linear models. In contrast to their classical counterparts, the new robust techniques show lower variability and excellent efficiency properties in the presence of small deviations from the assumed model, i.e. when the underlying distribution of the data lies in a neighborhood of the model. A simulation study, an analysis on real data, and a sensitivity analysis confirm the good theoretical statistical properties of the new techniques.

JEL classification: C10, I10.

KEY WORDS: Deviations from the model; GLM modeling; health econometrics; heavy tails; robust estimation; robust inference.

*This research has been carried out under the IRIS program of the Universities of Lausanne and Geneva and the Swiss Federal Institute of Technology, Lausanne. It has been financially supported by the Conférence Universitaire Suisse and its partners.
1 Introduction

Modeling medical expenses is an important building block in cost management and a large research effort has been put in the analysis of this type of data. Many papers discuss the many different aspects related to modeling such data. It is impossible to give a full and representative list of this extensive literature, which includes, for example, Duan, Manning, Morris, and Newhouse (1983), Goldman, Leibowitz, and Buchanan (1998), Manning, Newhouse, Duan, Keeler, and Leibowitz (1987) and many others. The importance of the issue – and its policy implications – makes health economists and other empirical researchers even more aware of the importance of a careful statistical analysis.

From a statistical point of view, the goal is to estimate \( \mu = E(Y|X) \), where \( Y \) is the response (health care expenditure, length of stay, utilization of health care services, to name a few) and \( X \) is a set of explanatory variables (age, sex, income, out-of-pocket price, health status, etc.). The characteristics of the distribution of \( Y \) are such that standard methodology is often inappropriate. For instance, two main issues arise: (i) the measurements of the outcome are positive (or nonnegative) and highly skewed, which contrast with the Gaussian (or at least symmetric) distributional assumption of many standard statistical techniques and (ii) the thickness of the tail of the distribution is often determined by a small number of heavy users.

A possible fix to the skewness problem (issue (i)) is to transform the data. The merits of this approach have been largely discussed in the literature (see Manning, 1998; Mullahy, 1998; Manning and Mullahy, 2001 and references therein). While the transformed model has the advantage to fit in the setting of standard linear regression – which has a long tradition in health economics – it presents several drawbacks. First of all, the interpretability of the model coefficients is often difficult on a different scale than the original, secondly the quality of the retransformed parameter estimates is typically poor without appropriate corrections (see e.g. the nonparametric smearing estimator of Duan, 1983) and – last but not least – the transformed data will have only an approximate normal distribution (for example, the far-right tail of the transformed data is typically still too long even if one assumes a log-normal distribution\(^1\)). Issue (ii) can be viewed as a particular aspect of the broader robustness issue which arises from the fact that models are at best ideal approximations of the underlying process and deviations from the distributional assumptions are always present in real data; for a general overview on

\(^{1}\)See, for instance, the example of Section 3.5 in Duan, Manning, Morris, and Newhouse (1983).
robust statistics see Huber (1981) and Hampel, Ronchetti, Rousseeuw, and Stahel (1986).

Two recent papers focus on robust estimation (Marazzi and Barbati, 2003 and Marazzi and Yohai, 2004) and develop robust estimates for location-scale models on the log-scale which can be used for typical data on health care expenditures. Their work is based on a truncated maximum likelihood regression where the errors are allowed to have asymmetric distribution (e.g. Weibull).

In this paper we pursue a different approach that addresses jointly the skewness and robustness problem (issues (i) and (ii) above) by building on the unified framework of generalized linear models (GLM, see McCullagh and Nelder, 1989). These models are very attractive to handle a large variety of continuous and discrete data and have already been applied in health economics settings (e.g. Blough, Madden, and Hornbrook, 1999, Manning and Mullahy, 2001 and Gilleskie and Mroz, 2004). Because the GLM technique is based on maximum likelihood or quasi-likelihood, it is very sensitive to spurious observations\(^2\). Cantoni and Ronchetti (2001) developed robust versions of estimators and tests for GLM in the case of binomial and Poisson models. Here we consider an extension of their method to other GLM settings, for example the Gamma family. This approach is attractive because it enjoys some interesting advantages over the existing approaches mentioned above. First, the target value \(\mu\) is modeled directly making inference straightforward and avoiding the need of (re-)transformation. Moreover, it enables to go beyond the location-scale family considered in the previously published robust literature and allows some flexibility through the choice of the link function (e.g. logarithmic, inverse) and of the distribution of \(Y\) through its expectation-variance relationship. Finally, a class of test statistics for the comparison of nested models naturally comes along for variable selection. An additional diagnostic feature of our robust approach is the automatic identification of the outlying observations of the process.

Health care expenditure data often come with an excess of zeros, see Mullahy (1997). In these cases, a popular approach is the well-known two-part model, where the mass at zero is modeled separately\(^3\). The approach introduced above and described in detail below concentrates on the estimation of the determinants of the level of \(y_i | y_i > 0\), the so-called Part 2 of the two-part model. This is because our work was motivated by the example in Section 5 where we only observe positive responses. If the data at hand comes with ze-

\(^2\)In fact it was noticed by Manning and Mullahy (2001) that “GLM models can yield very imprecise estimates if the log-scale error is heavy tailed”.

\(^3\)For a discussion on the appropriate use of the two-part model we refer to Jones (2000, Sec. 4).
ros, the binary responses of Part 1 (occurrence or non-occurrence of medical expenses) can be modelled robustly with a binary regression (e.g. logistic), as treated in detail in Cantoni and Ronchetti (2001). An alternative approach would consider specific distributions that model directly the excess of zero, either via the likelihood of an hurdle model or via a zero-inflated distribution4. This approach can be robustified and is subject of ongoing research.

The paper is organized as follows. In Section 2 we briefly introduce the GLM methodology. Section 3 is devoted to a short introduction of the robust approach and to the definition of our estimation and variable selection procedure. In Section 4 the benefits of our technique are confirmed and further supported by a simulation study, whereas in Section 5 we present a study on real data that motivated our work.

2 GLM modeling

We consider the modeling framework of GLM where the response variable $Y_i$, for $i = 1, \ldots, n$, is drawn from a distribution belonging to the exponential family, such that $E[Y_i|x_i] = \mu_i$ and $V[Y_i|x_i] = v(\mu_i)$ for $i = 1, \ldots, n$ and

$$g(\mu_i) = \eta_i = x_i^T \beta \text{ or equivalently } \mu_i = E(Y_i|x_i) = g^{-1}(x_i^T \beta) = g^{-1}(\eta_i), \quad (1)$$

for $i = 1, \ldots, n$, where $\beta \in \mathbb{R}^p$ is the vector of parameters, $x_i \in \mathbb{R}^p$ is a set of explanatory variables, and $g(.)$ is the link function.

Two elements define model (1): the link function, which can be for example logarithmic (or logit or probit), and the mean-variance relationship. In particular, if $v(\mu_i)$ is constant we obtain a non-linear homoscedastic regression model. Models with $v(\mu_i)$ proportional to $\mu_i$ define Poisson-type distributions, possibly over-dispersed. Finally, if $v(\mu_i)$ is proportional to $\mu_i^2$ we obtain the Gamma, the homoscedastic log-normal and the Weibull distributions.

Although the methodology developed here can be applied to the entire class of GLM, the application and simulation in this paper will concentrate on a Gamma model with log-link and variance structure defined by $v(\mu_i) = \mu_i^2/\nu$. It has been reported by several authors (e.g. Blough, Madden, and Hornbrook, 1999, Gilleskie and Mroz, 2004) that this characteristic (the variance proportional to the squared mean) is observed for many health care

4Both these approaches are discussed in Mullahy (1986). Note that they imply overdispersion, but they also express unobserved heterogeneity, see the discussion in Mullahy (1997).
expenditures data. Moreover, these models can be seen as issued from a multiplicative model \( y_i = \exp(x_i^T \beta) \cdot u_i \), where the error term \( u_i \) has constant variance. More specifically we consider a parametrisation of the Gamma density function such that one parameter identifies \( \mu_i \), namely

\[
f_{\mu_i, \nu}(y_i) = \frac{\nu}{\mu_i} \cdot \exp\left(-\frac{\nu y_i}{\mu_i}\right) \cdot \left(\frac{\nu y_i}{\mu_i}\right)^{\nu-1} \frac{\Gamma(\nu)}{\mu_i},
\]

(2)

see also McCullagh and Nelder (1989, p. 30). In this case \( E(Y_i) = \mu_i \) and \( V(Y_i) = \mu_i^2/\nu \).

3 Robust approach

As mentioned in the Introduction, health data often show heavy-tailed distributions, which may be due to the presence of a few heavy users. These points highly affect the estimation and inference of the parameters of the model. The basic idea of robust statistics is to consider the distribution of the data as coming from a neighborhood of the postulated model. Then, robust estimates and test statistics are constructed such that the estimated parameters are consistent at the postulated model and stable in a neighborhood of it. This stability is achieved at the price of a slight loss of efficiency at the model. This can be viewed as an insurance premium one is willing to pay to protect against biases and losses of efficiency due to deviations from the assumed model.

An important mathematical tool that measures the robustness of an estimator is the influence function (Hampel, 1974). For a sample \( z = (z_1, \ldots, z_n) \) it is defined by

\[
IF(z; T, F) = \lim_{\epsilon \to 0} \left( \frac{T(F_\epsilon) - T(F)}{\epsilon} \right),
\]

(3)

where \( T(F) \) is a functional that defines the estimator \( T(F^{(n)}) \), \( F^{(n)} \) is the empirical distribution function, \( F_\epsilon = (1 - \epsilon)F + \epsilon \Delta_z \), and \( \Delta_z \) is a distribution that puts all its mass at \( z \). The influence function measures the effect on the estimate of an infinitesimal contamination at the point \( z \), standardized by the amount of contamination. The maximal marginal effect of an observation \( z \) on \( T \) is approximately \( \epsilon \cdot IF(z; T, F) \). Therefore a bounded influence function is a desirable robustness property for an estimator (see Hampel, Ronchetti, Rousseeuw, and Stahel, 1986 for details). For instance, the maximum likelihood estimator of a Gamma generalized linear model has an influence function proportional to the score function, that is, proportional...
to
\[
\frac{\partial \log(f_{\mu,\nu}(y_i))}{\partial \beta} = \frac{\partial \log(f_{\mu,\nu}(y_i))}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \beta} = \nu \frac{(y_i - \mu_i)}{\mu_i^2} \cdot \frac{1}{g'(x_i^T \beta)} \cdot x_i
\]
which is neither bounded with respect to \(y_i\), nor with respect to \(x_i\). This explain the non-robustness properties of this estimator. As we shall see, the estimator proposed in the next Section has a bounded influence function, therefore ensuring stability in the presence of deviations from the Gamma model defined above.

### 3.1 Robust estimating equations

To address robustness (in the sense of local stability, as measured by the influence function), Cantoni and Ronchetti (2001) suggested to estimate the parameter \(\beta\) via M-estimation (Huber, 1981), that is through a set of estimating equations of the form \(\sum_{i=1}^{n} \Psi(y_i, \beta, \nu) = 0\). They suggested to obtain a (Mallows quasi-likelihood) estimator of the regression parameter \(\beta\) of model (1) by solving
\[
\sum_{i=1}^{n} \left[ \psi(r_i) w(x_i) \frac{1}{v^{1/2}(\mu_i)} \mu_i' - a(\beta) \right] = 0, \tag{4}
\]
where \(r_i = (y_i - \mu_i)/v^{1/2}(\mu_i)\) are the Pearson residuals and \(\mu_i' = \partial \mu_i / \partial \beta\). The correction term \(a(\beta) = \frac{1}{n} \sum_{i=1}^{n} E[\psi(r_i)] w(x_i) \frac{1}{v^{1/2}(\mu_i)} \mu_i'\) ensures Fisher consistency with respect to the mean parameter \(\mu\) at the model.

The form defined in (4) is suggested by the classical estimating equations\(^5\), which are recovered if \(\psi\) is the identity function and \(w(x_i) \equiv 1\), in which case \(a(\beta) = 0\). Other choices of \(\psi\) can be introduced to control large deviations in the \(y\)-space, whereas leverage points are downweighted by the weights \(w(x_i)\). For example these weights can be a function of the diagonal elements of the hat matrix \(H = X(X^T X)^{-1} X^T\) (e.g. \(w(x_i) = \sqrt{1 - H_{ii}}\) or proportional to the inverse of the Mahalanobis distances, see Cantoni and Ronchetti (2001) for further details. A common choice for \(\psi\) to ensure robustness is the so-called Huber’s function defined by \(\psi_c(r) = r \cdot \min(1, c/|r|)\), see Panel (a) of Figure 1. This function is the identity between \(-c\) and \(c\), whereas values of \(r\) larger than \(c\) in absolute value are replaced by \(c \cdot \text{sign}(r)\). Therefore, the contribution of an observation \(y_i\) to the estimating equations (4) is preserved as in the classical case if its residual \(r_i\) is not too large, and reduced otherwise. The constant \(c\) allows one to tune the robustness-efficiency compromise. From a practical point of view, values of \(c\) between 1 and 2 typically guarantee robustness with a reasonable level of efficiency.

\(^5\)see, for example, Manning and Mullahy (2001, p. 466).
Figure 1: Huber’s $\psi_c(r)$ function and Huber’s weights $\tilde{w}(r) = \psi_c(r)/r$.

One can take advantage of the fact that the robust technique can provide automatically a reliable diagnostic measure for the outlying observations by looking at the weights computed in the robust fitting procedure. In fact, the set of estimating equation (4) can be rewritten as

$$\sum_{i=1}^{n} \left[ \tilde{w}(r_i) w(x_i) \frac{1}{\nu^{1/2}(\mu_i)} \mu_i' - a(\beta) \right] = 0,$$

where $\tilde{w}(r) = \psi(r)/r$.

Therefore $\tilde{w}$ and $w$ will give information on how each observation is handled. If the Huber’s $\psi_c$ function is used, then the weights $\tilde{w}(r)$ are plotted in Panel (b) of Figure 1.

One could argue that a similar effect could be obtained by performing diagnostic to identify outlying observations on the basis of a classical analysis and then remove the unusual data points from the sample. This is not a wise approach because a masking effect can occur, where a single large outlier may mask others. This means that the distorted data appear to be the norm rather than the exception. This is due to the fact that classical estimates are affected by outlying points and tend to be pulled in direction of outlying points.

The set of estimating equations (4) does not take into account that $\nu$ has also to be estimated. To do so, one notices that $\text{Var}((Y_i - \mu_i)/\mu_i) = \nu_i$, and therefore any robust estimator of the variance of $(Y_i - \mu_i)/\mu_i$ can be used. If the variance is estimated by the classical (non robust) estimator, we obtain the estimator for $\nu$ used in the GLM framework, that is $\hat{\nu} = 1/n \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2/\hat{\mu}_i^2$. Alternatively, many robust estimators of variance are available in the literature on robust statistic. We choose a simple M-estimator
(Huber’s Proposal 2), which solves

\[ \sum_{i=1}^{n} \chi_{c}\left( \frac{y_i - \mu_i}{\mu_i/\sqrt{\nu}} \right) = 0, \]  

(6)

where \( \chi_c(u) = \psi^2_c(u) - \theta \), and \( \theta = E(\psi^2_c(u)) \) is a constant that ensures Fisher consistency for the estimation of \( \nu \) (see Hampel, Ronchetti, Rousseeuw, and Stahel, 1986, p. 234).

The distributional and robustness properties of the proposed estimator of the regression parameters can be derived. From standard results on M-estimators, we know that the influence function of the estimator defined by the set of equations (4) at a point \((x, y)\) is given by

\[
IF((x, y); T, F_\beta) = M^{-1}(\psi, F_\beta) \left[ \psi\left( \frac{y - \mu}{\nu^{1/2}(\mu)} \right) w(x) \frac{1}{\nu^{1/2}(\mu)} \mu' - a(\beta) \right],
\]

(7)

where \( M(\psi, F_\beta) = \frac{1}{n} X^T B X \), \( b_i = E[\psi'_c(r_i)] \frac{\sigma_i}{\sigma_i^2} \log h(y_i|x_i, \mu_i)](\frac{1}{\nu^{1/2}(\mu_i)} w(x_i)(\frac{\sigma_i}{\sigma_i^2})^2 \]

are the elements of the diagonal matrix \( B \), and \( h(\cdot) \) is the conditional density or probability of \( y_i|x_i \).

The influence function is bounded with respect to \( y \) for a bounded choice of \( \psi \), and the effect of outliers in the design is controlled with appropriate weights \( w(x) \).

Moreover, under quite general conditions, it can be shown (see Cantoni and Ronchetti, 2001) that the asymptotic distribution of \( \sqrt{n}(\hat{\beta} - \beta) \), where \( \hat{\beta} \) is the solution of (4), is normal with expectation 0 and variance equal to

\[
M^{-1}(\psi, F_\beta) Q(\psi, F_\beta) M^{-1}(\psi, F_\beta),
\]

(8)

with \( A \) a diagonal matrix with elements \( a_i = E[\psi'_c(r_i)]^2 w^2(x_i) \frac{1}{\nu^{1/2}(\mu_i)}(\frac{\sigma_i}{\sigma_i^2})^2. \)

This asymptotic result still holds if a \( \sqrt{n} \)-consistent estimator for \( \nu \) is plugged-in in the estimating equations (4).

The expectation terms appearing in \( a(\beta) \), \( b_i \) and \( a_i \) have to be computed explicitly at the model \( F_\beta \). This can be done for several model distributions including binomial and Poisson (see Cantoni and Ronchetti, 2001) and Gamma (see the Appendix). For other distributions, these terms can be at least approximated numerically.

### 3.2 Robust variable selection

The approach outlined in Section 3.1 has an important added value in that it provides a class of robust test statistics for variable selection by comparison of two nested models. It is well-known that such a global strategy is
more reliable than simply looking at univariate t-test-like statistics in the full model; see for instance Cantoni, Flemming, and Ronchetti (2004). In fact, the estimating equations (4) can be seen as the derivatives with respect to \( \beta \) of the robust quasi-likelihood function \( \sum_{i=1}^{n} Q_M(y_i, \mu_i) \), where

\[
Q_M(y_i, \mu_i) = \int_{s}^{\mu_i} \phi(y_i, t)w(x_i)dt - \frac{1}{n} \sum_{j=1}^{n} \int_{\tilde{t}}^{\mu_j} E[\phi(y_j, t)w(x_j)]dt,
\]

where \( \phi(y_i, t) = \psi\left( (y_i - t)/v^{1/2}(t) \right)/v^{1/2}(t) \), \( \tilde{s} \) such that \( \phi(y_i, \tilde{s}) = 0 \), and \( \tilde{t} \) such that \( E[\phi(y_i, \tilde{t})] = 0 \). Therefore to compare a model \( \mathcal{M}_p \) with \( p \) variables (corresponding to a parameter \( \beta = (\beta_1, \ldots, \beta_p) \)) to a nested model \( \mathcal{M}_{p-q} \) with only \( (p-q) \) variables \( (\beta = (\beta_1, \ldots, \beta_{p-q}, 0, \ldots, 0)) \), a test statistic can be constructed based on twice the difference of the quasi-likelihood functions

\[
\Lambda_{QM} = 2 \left[ \sum_{i=1}^{n} Q_M(y_i, \tilde{\mu}_i) - \sum_{i=1}^{n} Q_M(y_i, \tilde{\mu}_i) \right],
\]

where \( \tilde{\mu}_i \) and \( \tilde{\mu}_i \) are the estimators obtained under models \( \mathcal{M}_p \) and \( \mathcal{M}_{p-q} \) respectively\(^7\).

Under the null hypothesis that \( H_0 : \beta_{p-q+1} = \ldots = \beta_p = 0 \) and under quite general conditions, \( \Lambda_{QM} \) is asymptotically distributed as \( \sum_{i=1}^{q} d_i N_i^2 \), where \( N_1, \ldots, N_q \) are independent standard normal variables, \( d_1, \ldots, d_q \) are the \( q \) positive eigenvalues of the matrix \( Q(\psi, F_\beta)(M^{-1}(\psi, F_\beta) - \hat{M}^+(\psi, F_\beta)) \), and \( \hat{M}^+(\psi, F_\beta) \) is such that \( \hat{M}^+(\psi, F_\beta)_{11} = M(\psi, F_\beta)_{11}^{-1} \) and \( \hat{M}^+(\psi, F_\beta)_{12} = 0 \), \( \hat{M}^+(\psi, F_\beta)_{21} = 0 \), \( \hat{M}^+(\psi, F_\beta)_{22} = 0 \) (see Proposition 1 in Cantoni and Ronchetti, 2001). Notice that in our case (Gamma model) the second set of integrals in (9) can be computed explicitly because \( E[\phi(y_j, t)w(x_j)] \) is proportional to \( 1/t \), and \( E[\psi_i(r_i)] \) is independent of \( \mu_i \). The statistic \( \Lambda_{QM} \) is a generalisation of the classical GLM quasi-deviance statistic (Wedderburn, 1974 and Blough, Madden, and Hornbrook, 1999), that can be obtained with an identity function \( \psi \) and \( w(x_i) \equiv 1 \).

By means of general results in Cantoni and Ronchetti (2001), the robustness properties of \( \Lambda_{QM} \) can be formally assessed: the asymptotic level and power under small deviations from the model are stable as long as an estimator of \( \beta \) with bounded influence function is used. The S-PLUS code for the robust approach is available from http://www.unige.ch/ses/metri/cantoni/.

\(^6\)Often \( \phi(0) = 0 \), therefore the choice \( \tilde{s} = \tilde{t} = y_i \) fulfils these conditions.

\(^7\)Note that \( \Lambda_{QM} \) is independent of \( \tilde{s} \) and \( \tilde{t} \).
4 Simulation results

We conduct a small simulation study to compare the classical and our new robust approach. We generated data from a Gamma model with log-link. We assumed that $\nu = 1$ and $\mu_i = g^{-1}(x_i^T \beta)$, where $\beta = (1, 0.2, 0.2, 0.2)^T$ and $x_i = (1, x_{i1}, \ldots, x_{i4})$, with $x_{i1} \sim Bin(1, 0.5)$, $x_{i2}$ is categorical (3 levels with probabilities of 0.5, 0.35 and 0.15 respectively), $x_{i3}$ and $x_{i4} \sim N(0, 1)$.

A thousand samples of size 1000 are generated from a Gamma model (see (2)) with log-link. A thousand of corresponding contaminated samples are obtained by multiplying by 10 5% of randomly chosen responses.

This design has been chosen to mimic a variety of situation arising in practice. For instance, the binary independent variable could represent the gender of an individual, the categorical variable could represent health status (or race or marital status, for example) and the normal distributed variable could represent the (standardized) age or (standardized) educational level (years of completed schooling).
Figure 3: Regression parameters estimation and their standard errors for contaminated data.

For the two classes of data (contaminated and non-contaminated), we first look at the quality of the estimated parameters by both a classical GLM and our robust technique (with $\psi = \psi_c$, $c = 1.5$ and $w(x_i) \equiv 1$).

The simulation results are displayed in Figure 2-4. The estimated regression parameters and their standard error for non-contaminated data (that is, at the model) of Figure 2 appear to be in line with the true values for both the classical and the robust technique. The estimated standard errors of the robust technique are slightly larger than their classical counterparts as theoretically expected due to the small loss of efficiency incurred. The results for the contaminated set of data (Figure 3) are quite different. In fact, the intercept coefficient is not well estimated, even more so with classical GLM. Moreover, the estimated coefficients for the classical technique are not biased but exhibit a large (spurious) variability, and their standard errors are overestimated. This would impact the inference of the classical analysis by hiding significant effects. The large variability observed in the classical
estimates is the consequence of the bad estimation of the scale parameter, as it appears in Figure 4. The classical technique is highly affected by a small fraction (5%) of contaminated observations of the data and overestimates the variability of the majority of the sample data. Note that if no contamination is present, the robust and classical estimators of $\nu$ perform similarly.

5 An example on Swiss data

In this section, we consider a sample of 100 patients hospitalized at the Centre Hospitalier Universitaire Vaudois in Lausanne (Switzerland) during 1999 for “medical back problems” (APDRG 243). The outcome is the cost of stay (in Swiss francs) and the explanatory variables are: length of stay (LOS, in days), admission type (ADM: 0=planned, 1=emergency), insurance type (INS: 0=regular, 1=private), age in years (AGE), sex (SEX: 0=female, 1=male) and discharge destination (DEST: 1=home, 0=another health institution).

The observed cost of stay in the sample varies between 1’584.20 and 42’117.92 Swiss Francs, the median being 11’125.98. The median age is 56.5 years (the youngest patient is 16 years old and the oldest is 93 years old). The length of stay ranges from 2 to 64 days. 60 individuals out of the 100 of the sample were admitted in emergency and only 9 patients had private insurance. Both sexes are well represented in the sample with 53 men and 47 women. After being treated, 82 patients went home directly.
Table 1: Coefficient estimates and standard errors from a classical and a robust analysis.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Classical</th>
<th>Robust</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>7.2338 0.1469</td>
<td>7.2523 0.1049</td>
</tr>
<tr>
<td>log(LOS)</td>
<td>0.8222 0.0280</td>
<td>0.8391 0.0200</td>
</tr>
<tr>
<td>ADM</td>
<td>0.2136 0.0500</td>
<td>0.2221 0.0357</td>
</tr>
<tr>
<td>INS</td>
<td>0.0933 0.0791</td>
<td>0.0093 0.0565</td>
</tr>
<tr>
<td>AGE</td>
<td>-0.0005 0.0013</td>
<td>-0.0010 0.0009</td>
</tr>
<tr>
<td>SEX</td>
<td>0.0951 0.0500</td>
<td>0.0727 0.0357</td>
</tr>
<tr>
<td>DEST</td>
<td>-0.1043 0.0693</td>
<td>-0.1230 0.0495</td>
</tr>
</tbody>
</table>

scale: 0.0496  scale: 0.0243

5.1 Fit of the model

We report the estimated parameters and their standard errors in Table 1. The first two columns give the classical analysis, whereas the second set of columns reports the results with the robust estimation via (4), where we used a Huber’s ψ function with $c = 1.5$ and $w(x_i) \equiv 1$. If in addition weights $w(x_i) = \sqrt{1 - H_{ii}}$ on the design space are used, similar results are obtained (not shown here).

Only small differences appear on the values of the estimated coefficients between the classical and the robust analysis except for INS, where there is a difference by a factor of 10 (it is not a typo!). There are at the contrary major discrepancies between the estimated standard errors of the two techniques, the ones based on the robust approach being much smaller. This is in line with the conclusions of the simulations study of Section 4. It is mainly due to the fact that the scale estimate for the classical analysis is twice as large as the one from the robust analysis. The conclusions from both analyses are quite different: if no doubt arises on the significance of the Intercept, log(LOS) and ADM on both analysis, the robust analysis would suggest a significant effect also for DEST, and less clearly for SEX. In view of the results of the simulation study in Section 4, the robust analysis has to be considered more reliable.

To identify the observations exhibiting a different pattern than the majority of the data, we can look at the weights $\tilde{w}$. When fitting the full model to the dataset at hand, we have five observations with a weight less or equal than 0.5, namely $\tilde{w}_{14} = 0.23$, $\tilde{w}_{21} = 0.50$, $\tilde{w}_{28} = 0.24$, $\tilde{w}_{44} = 0.42$ and $\tilde{w}_{63} = 0.32$. The particular behaviour of these observations can for example...
be highlighted in the pairwise plot of the cost of stay ($Y$) against $\log(LOS)$ (Figure 5): the pattern of the downweighted observations is different from the pattern of the majority of the data. Note that, although surprising at first sight, the far right point in Figure 5 received full weight because the model is such that it allows for variability increasing with $\mu_i^2$ and therefore with $x$.

### 5.2 Sensitivity analysis for variable selection

This example can also serve the purpose of illustrating how the $p$-values of classical tests are sensitive to outliers, whereas the robust tests are more stable. We consider the model as in 5.1 including all the available variables and test whether the variable SEX is significant in the model. To do so, we let $y_{21}$ span the range of all the values of the sample (about $1'500 - 45'000$) on a grid of 100 points (see Figure 5). For each point of the grid, we compute the classical and the robust $p$-values, that is the $p$-values obtained with the
Figure 6: p-values for testing whether the variable SEX is significant in the model of Table 1 when letting $y_{21}$ range between 1'500 and 45’000.

test statistics (10) with a Huber’s function with $c = \infty$ (reproducing the classical deviance approach) and $c = 1.5$ respectively.

The results are displayed in Figure 6. The difference of behavior between the two methods is striking, even more so if one thinks that only one point out of 100 is causing it. The p-value associated to the classical test statistics ranges from 4.4% to 21.9%. On the other hand, the p-value of the robust test statistics is much more stable and varies only between 4 and 8.4%. It provides a consistent message of near significance for the SEX variable which is based on the structure of the overwhelming majority of the data and is not affected by a single data point.

Acknowledgements

The authors thank A. Marazzi and A. Holly for helpful discussions, and A. Marazzi for providing the dataset analysed in Section 5.

Appendix

In this Appendix we provide explicit expression for $E[\psi(r_i)]$, $E[\psi(r_i)^2]$ and $E[\psi(r_i) \frac{\partial}{\partial \mu_i} \log h(y_i|x_i, \mu_i)]$ when $\psi = \psi_c$ is the Huber’s function with tuning constant $c$ (see Figure 1) and when $Y_i$ is issued from a Gamma distribution with parameters $\mu_i$ and $\nu$ as defined by (2).

We first show that the variable $R_i = (Y_i - \mu_i)/\nu^{1/2}(\mu_i)$ has a distribution
independent of $\mu_i$. The density function of $R_i$ is given by

$$f_\nu(r_i) = \frac{\nu^{\nu/2} \exp(\sqrt{\nu} (\sqrt{\nu} + r_i)) (\sqrt{\nu} + r_i)^{\nu-1}}{\Gamma(\nu)}, \quad r_i > -\sqrt{\nu},$$

which is in fact a Gamma density of the form (2) with $\mu_i = \sqrt{\nu}$, but with shifted origin to $-\sqrt{\nu}$. Let us also define

$$G(t, \kappa) = \exp(-\sqrt{\nu} (\sqrt{\nu} + t)) (\sqrt{\nu} + t)^\kappa \mathbb{I}_{(t > -\sqrt{\nu})}$$

where $\mathbb{I}_{(t > -\sqrt{\nu})} = 1$ if $t > -\sqrt{\nu}$ and 0 otherwise.

We then have

$$E\left[\psi_c(Y_i - \mu_i)\right] = E\left[\psi_c(R_i)\right] = \int_{-\sqrt{\nu}}^\infty \psi_c(r_i) f_\nu(r_i) \mathbb{I}_{(R_i > -\sqrt{\nu})} dr_i$$

$$= \nu c \left( P(R_i > c) - P(R_i < -c) \right) + \int_{-c}^c r_i f_\nu(r_i) \mathbb{I}_{(R_i > -\sqrt{\nu})} dr_i. \quad (12)$$

The integral in (12) can be computed by

$$\int_{-c}^c r_i f_\nu(r_i) \mathbb{I}_{(R_i > -\sqrt{\nu})} dr_i =$$

$$= \int_{-c}^c (\sqrt{\nu} + r_i) f_\nu(r_i) \mathbb{I}_{(R_i > -\sqrt{\nu})} dr_i - \sqrt{\nu} P(-c < R_i < c) =$$

$$= \frac{\nu^{(\nu-1)/2}}{\Gamma(\nu)} \left[ G(-c, \nu) - G(c, \nu) \right],$$

where integration by parts has been used in the last step.

Similarly, we obtain:

$$E\left[\psi_c^2(Y_i - \mu_i)\right] = E\left[\psi_c^2(R_i)\right] = \int_{-\infty}^\infty \psi_c^2(r_i) f_\nu(r_i) \mathbb{I}_{(R_i > -\sqrt{\nu})} dr_i$$

$$= \nu c^2 \left( P(R_i < -c) + P(R_i > c) \right) + P(-c < R_i < c)$$

$$+ \frac{\nu^{(\nu-1)/2}}{\Gamma(\nu)} \left[ G(-c, \nu + 1) - G(c, \nu + 1) \right]$$

$$+ \frac{\nu^{\nu/2}}{\Gamma(\nu)} \left( \frac{\nu + 1}{\nu} - 2 \right) \left[ G(-c, \nu) - G(c, \nu) \right].$$

For the computation of the third term, we first notice that $\frac{\partial}{\partial \mu_i} \log f_{\mu_i, \nu}(Y_i) = (Y_i - \mu_i)/\mu_i^2 = \sqrt{\nu} R_i / \mu_i$. This term will depend on $\mu_i$. We then use the
same reasoning as above to compute
\[
E[\psi_c(R_i) \frac{\partial}{\partial \mu_i} \log f_{\mu_i,Y_i}(Y_i)] = \sqrt{\nu} E[\psi_c(R_i)R_i] \\
= \frac{\nu^{\nu/2}}{\mu_i \Gamma(\nu)} \left[ G(-c, \nu) + G(c, \nu) \right] + \frac{\nu^{\nu/2}}{\mu_i \Gamma(\nu)} P(-c < R_i < c) \\
+ \frac{\nu^{\nu/2}}{\mu_i \Gamma(\nu)} \left[ G(-c, \nu + 1) - G(c, \nu + 1) \right] \\
+ \frac{\nu^{(\nu+1)/2}}{\mu_i \Gamma(\nu)} \left( \frac{\nu + 1}{\nu} - 2 \right) \left[ G(-c, \nu) - G(c, \nu) \right].
\]

References


