Outlier Identification Rules for Generalized Linear Models

Sonja Kuhnt and Jörg Pawlitschko

Department of Statistics, University of Dortmund

44221 Dortmund, Germany

Abstract

Observations which seem to deviate strongly from the main part of the data may occur in every statistical analysis. These observations, usually labelled as outliers, may cause completely misleading results when using standard methods and may also contain information about special events or dependencies. Therefore it is of interest to identify them. We discuss outliers in situations where a generalized linear model is assumed as null-model for the regular data and introduce rules for their identification. For the special cases of a loglinear Poisson model and a logistic regression model some one-step identifiers based on robust and non-robust estimators are proposed and compared.

1 Introduction

In the statistical analysis of data one often is confronted with observations that "appear to be inconsistent with the remainder of that set of data" (Barnett and Lewis, 1994) or, more generally, that "are far away $[\cdots]$ from the pattern set by the majority of the data" (Hampel et al., 1986). Such observations are usually called "outliers". They may have a high impact on the statistical analysis and can cause completely misleading results when

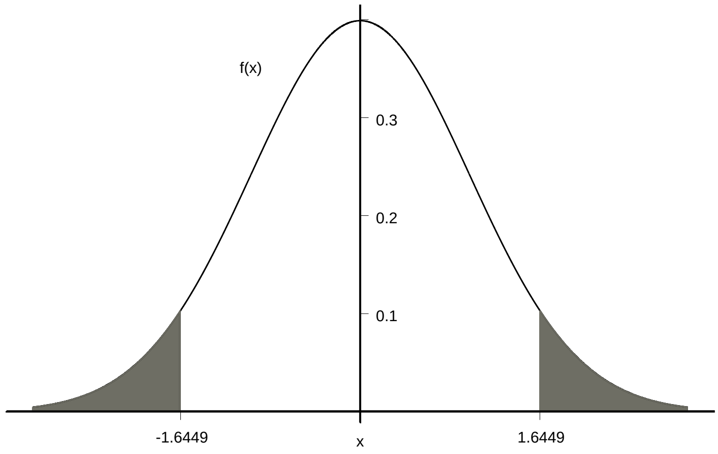

Figure 1: 0.05-outlier region of the $\mathcal{N}(0,1)$-distribution

using standard procedures. Sometimes, however, these outliers themselves provide the most interesting aspect of the data, for instance an unexpected long lifetime observed in a clinical trial or a surprisingly high insurance claim. Although the problem how to identify and to handle outliers has been subject of numerous investigations, there is no general accepted formal definition of outlyingness. Most authors, however, agree in that the notion “outlier” is only meaningful in relation to a hypothesized statistical model for the “good” data, the so-called null model. We treat outliers in the sense of Davies and Gather (1993) who define outliers in terms of their position relative to the null model. E.g. for any normal distribution $\mathcal{N}(\mu, \sigma^2)$ and any α, $0 < \alpha < 1$, the corresponding α-outlier region is defined by

$$out(\alpha, \mu, \sigma^2) \;=\; \left\{x\colon\; |x-\mu| > \sigma\, z_{1-\alpha/2}\right\}$$

where $z_{1-\alpha/2}$ denotes the $(1-\alpha/2)$-quantile of the standard normal distribution. Any number x is called an α-outlier with respect to $\mathcal{N}(\mu, \sigma^2)$ if

$x \in out(\alpha, \mu, \sigma^2)$. Figure 1 shows the 0.05-outlier region of the $\mathcal{N}(0, 1)$ distribution. Any number with an absolute value larger than 1.96 is a 0.05-outlier.

We will extend this general approach to outlyingness to null models for structured data situations such as regression models and contingency tables which are summarized in the broad class of generalized linear models (GLM). This unifying family of models has been introduced by Nelder and Wedderburn (1972) and has had a major influence on statistical modelling in a number of modern applications. GLM essentially extend the classical linear model in two ways: data are not necessarily assumed to be normally distributed and the mean is not necessarily modelled as a linear combination of certain covariates but some function of the mean is. Generalized linear models are introduced in more detail in Section 2. A formal definition of outliers for GLM is given in Section 3 together with methods aiming at the identification of outliers in observed samples. Section 4 provides some examples. Possible areas for further research are discussed in Section 5.

2 Generalized linear models

Consider the situation where it is of interest to explain a univariate response variable by a set of p fixed or stochastic covariates. Let $(Y_1, X_1), \ldots, (Y_n, X_n)$, with $X_i = (X_{i1}, \ldots, X_{ip})'$, be a sample of n observations. A generalized linear model (GLM) is characterized by two assumptions:

- *Distributional Assumption:* For each Y_i, $i = 1, \ldots, n$, the conditional distribution of Y_i given $X_i = x_i$ belongs to an exponential family with expectation $E(Y_i|X_i = x_i) = \mu_i$ and variance $Var(Y_i|X_i = x_i) = \phi V(\mu_i)$, where V is a known variance function and $\phi > 0$ a common dispersion parameter not depending on i.

- *Structural Assumption:* There exists a so-called link function g, which is a known one-to-one, sufficiently smooth function, and a parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)' \in \mathbb{R}^p$ such that for each i, $i = 1, \ldots, n$, the expectation μ_i is related to a linear predictor via

$$g(E(Y_i|X_i = x_i)) \; = \; \sum_{j=1}^{p} x_{ij}\, \beta_j.$$

The distributional assumption includes for example the families of normal, inverse Gaussian, gamma, Poisson and binomial distributions. The common linear regression model is part of this model class; in this case the responses have a normal distribution, g is the identity function, $\phi = \sigma^2$, and $V(\mu) = 1$.

3 Outliers and identification rules

A key element of GLM is the fact that different vectors of covariates may give different distributions for the corresponding responses. It is indeed this assumption of non-identical distributions which causes the new problem in identifying outliers. We look at the distributions P_i of the responses given the covariates, $\{Y_i|X_i = x_i, i = 1, \ldots, n\}$ and start by defining α_i-outlier regions for the individual conditional distributions. This definition can be derived from a more general definition of outlier regions given in Gather et al. (2003).

Let $\mathcal{P}$ be an exponential family such that $P_i \in \mathcal{P}$ has density f_i with respect to a dominating measure and has (known) support $supp(P_i)$. Given $\alpha_i \in (0, 1)$ the α_i-outlier region of $P_i \in \mathcal{P}$ is then defined as

$$out(\alpha_i, P_i) \; = \; \{x \in supp(P_i) \colon \; f_i(x) < K(\alpha_i)\} \tag{1}$$

where

$$K(\alpha_i) \; = \; \sup\big\{K > 0 \colon \; P_i\big(\{x \colon \; f_i(x) < K\}\big) \; \leq \; \alpha_i\big\}.$$

With $inl(\alpha_i, P_i) = supp(P_i) \setminus out(\alpha_i, P_i)$ we define the corresponding α_i-inlier region of P_i. Each point $x \in out(\alpha_i, P_i)$ is called an α_i-outlier relative to P_i and each $x \in inl(\alpha_i, P_i)$ an α_i-inlier. This definition formalizes the element of "unlikeliness" that is associated with the more informal definitions of an outlier cited in the introduction. Furthermore, this definition requires no further properties of a point x to be classified as outlier with respect to P_i than being contained in $out(\alpha_i, P_i)$. Especially outliers need not to come from a special outlier-generating mechanism as it is usually assumed in the literature.

Let now $\{y_i | x_i, i = 1, \ldots, n\}$ be a sample that under the null model is assumed to come i.i.d. from a certain GLM. An observed response y_i is then identified as α_i-outlier if it lies in the α_i-outlier region of the corresponding conditional distribution. The levels α_i should be chosen such that under the null model the probability of the occurrence of any outlier in the whole sample does not exceed a given $\tilde{\alpha}$. If equal values of the α_i, $i = 1, \ldots, n$, are desired, a natural choice depending on the sample size is given by

$$\alpha_i \; = \; 1 - (1 - \tilde{\alpha})^{1/n}. \tag{2}$$

The task of identifying all outliers in a sample $(\mathbf{y}_n | \mathbf{x}_n) = \{y_i | x_i, i = 1, \ldots, n\}$ can now be described as the task to find all those y_i which are located in the corresponding outlier region $out(\alpha_i, P_i)$.

Roughly spoken, there are two important types of outlier identification rules, namely rules that proceed in one step and rules that operate stepwise. In the following we focus on the first type of rules. For a GLM, a so-called simultaneous or one-step outlier identifier essentially consists in a set of empirical versions $OI_i(\alpha_i, \mathbf{y}_n | \mathbf{x}_n)$, $i = 1, \ldots, n$, of the α-outlier regions $out(\alpha_i, P_i)$, $i = 1, \ldots, n$. Each point located in $OI_i(\alpha_i, \mathbf{y}_n | \mathbf{x}_n)$ then is classified as α_i-outlier with respect to the corresponding P_i. The main problem is that the P_i are only partially known. Since the different distributions of the responses are only caused by the different values of the covariates, the P_i share the same unknown characteristics, namely the parameter vector $\boldsymbol{\beta}$ and the dispersion parameter ϕ.

To make the performance of different outlier identifiers comparable it is useful to standardize them in an appropriate way. Davies and Gather (1993) suggest two approaches in the i.i.d. case which can be transferred to the more complex setting of a GLM as well. In this case, the first standardization consists in the requirement that under the null model H_0 one has

$$P_{H_0}\big(Y_i \notin OI_i(\alpha_i, \mathbf{Y}_n|\mathbf{X}_n),\ i = 1, \ldots, n\big) \ \geq\ 1 - \gamma \tag{3}$$

for some $\gamma > 0$ which is often chosen equal to $\tilde{\alpha}$. Their second suggestion leads to the requirement that under the null model one has

$$P_{H_0}\big(OI_i(\alpha_i, \mathbf{Y}_n|\mathbf{X}_n) \subset out(\alpha_i, P_i),\ i = 1, \ldots, n\big) \ \geq\ 1 - \gamma \tag{4}$$

with γ chosen as in (3). Since both approaches inevitably lead to the laborious task of deriving (or simulating) a large number of normalizing constants we suggest to work without such a type of standardization and to estimate the regions $out(\alpha_i, P_i)$ directly. If $\tilde{\alpha}$ is chosen reasonably small this approach leads to identification rules which are not susceptible to identify too much regular observations as outliers. For estimating the true outlier regions one needs estimators of P_i, $i = 1, \ldots, n$, and these are obtained by plugging estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ and $\hat{\phi}$ of ϕ into the corresponding densities f_i. The classical estimator in GLM is the Maximum Likelihood (ML) estimator which, however, has the disadvantage of being not robust in most cases. For other data situations (see e.g. Davies and Gather, 1993, Becker and Gather, 1999) it has been shown that reliable outlier identification rules should be based on robust estimators of the model parameters. Especially, outlier identifiers that are constructed with non-robust estimators are prone to the effects of masking and swamping. Masking occurs if an identification rule fails to identify some outlier although the sample contains two or more apparently outlying observations (which then "mask" themselves). Swamping occurs if some apparent outlier(s) in the sample cause the identification rule to classify a regular observation as outlier as well. These findings lead us to recommend the use of robust estimators for the construction of outlier identifiers in GLM as well. We present two examples in the next section.

4 Examples

4.1 Loglinear Poisson models

As a first illustration of outlier identification in GLM, we look at the problem of identifying outlying cells in contingency tables. We concentrate on a 7×8 table from Yick and Lee (1998) containing student enrolment figures from seven community schools in Australia for eight different periods of the year, see Table 1.

93	96	99	99	147	144	87	87
138	141	141	201	189	153	135	114
42	45	42	48	54	48	45	45
63	63	72	66	78	78	82	63
60	60	54	51	51	45	39	36
174	165	156	156	153	150	156	159
78	69	84	78	54	66	78	78

Table 1: Student enrolments data (Yick and Lee, 1998)

The assumed model is that of independence between the row and column classification. The 56 cell counts y_i, $i \in \{1, \ldots, 56\}$, are taken to be outcomes of independent Poisson distributed random variables with individual expectations $E(Y_i) = \mu_i = \exp(x_i'\boldsymbol{\beta})$, where we have the logarithm as link function. The x_i, $i \in \{1, \ldots, 56\}$, are defined by the independence assumption and consist only of entries $-1, 0, 1$ if effect coding is used (Fahrmeir and Tutz, 2001). In case of the Poisson distribution it is not possible to give a simple expression for the outlier region, which is always an upper tail region or the union of an upper and a lower tail region. However, it can easily be derived using the definition. Every α_i-outlier region of a Poisson distribution $Poi(\widehat{\mu}_i)$ based on an estimate $\widehat{\mu}_i$ can be seen as an "empirical version" of the outlier region $out(\alpha_i, Poi(\mu_i))$, as discussed in Section 3. A one-step outlier identification rule can then be defined by identifying all cell counts lying in the corresponding region $out(\alpha_i, Poi(\widehat{\mu}_i))$ as α_i-outliers. With $\tilde{\alpha} = 0.1$ the choice of the individual levels according to (2) leads to $\alpha_i = 1 - (1 - 0.1)^{\frac{1}{56}} = 0.00188$.

The classical estimator for contingency tables is the ML-estimator. Some robust alternatives have been proposed in the last years, including estimates based on the median polish method (Mosteller and Parunak, 1985), L_1-estimates (Hubert, 1997), minimum Hellinger distances (Kuhnt, 2000), least median of chi-squares residuals and least median of weighted squared residuals (Shane and Simonoff, 2001).

105.20	103.74	105.20	113.48	117.86	111.05	100.98	94.49
149.65	147.57	149.65	161.43	167.67	157.97	143.65	134.41
45.56	44.93	45.56	49.15	51.05	48.09	43.73	40.92
69.76	68.79	69.76	75.25	78.16	73.64	66.96	62.66
48.90	48.22	48.90	52.74	54.78	51.61	46.93	43.92
156.69	154.51	156.69	169.02	175.55	165.40	150.40	140.73
72.23	71.23	72.23	77.92	**80.93**	76.25	69.33	64.88

Table 2: Maximum likelihood estimates

In case of ML-estimates (Table 2) only observation $y_{53} = 54$ lies in the 0.00118-outlier region of the distribution given by the estimate, $out(0.00118, Poi(80.93)) = \mathbb{N} \setminus \{55, \ldots, 110\}$, and is thereby identified as outlier.

94.34	95.42	99.53	98.47	**109.55**	**101.67**	95.32	88.48
138.70	140.26	146.33	**144.77**	161.06	149.47	140.13	130.09
44.54	45.05	46.99	46.49	51.72	48.00	45.00	41.77
67.17	67.94	70.87	70.11	78.00	72.39	67.86	63.00
46.59	47.13	49.16	48.63	54.10	50.21	47.07	43.70
152.80	154.54	161.20	159.48	177.43	164.66	154.37	143.31
76.38	77.25	80.58	79.72	**88.69**	82.31	77.17	71.63

Table 3: Median polish estimates

We also use median polish estimates, which are the means of the results of two sweeps of median polish on the logarithm of the cell counts once starting with the rows and once with the columns. Using these estimates, see Table 3, the four observations y_5, y_6, y_{12} and y_{53} are identified as 0.00188-outliers. Yick and Lee (1998) obtain the same set of outliers or, depending on the outlier identification procedure used, the set $\{y_5, y_6, y_{12}, y_{13}\}$, which they can explain from subject knowledge. Observations y_5 and y_6 are collected during a period in which a group of transient seasonal fruit picker families have moved near to this school, thus inflating the school's enrolments. y_{12} might show an unexpected high value due to a significant number of people moving into the area for an aboriginal funeral procession, which lasted around three month. Yick and Lee suggest that due to this funeral actually y_{13} might be the outlying observation and y_{53} is judged discordant due to swamping.

4.2 Logistic regression

Consider the case that the responses are binomially distributed according to $Y_i | X_i = x_i \sim Bin(m_i, p_i)$, $i = 1, \ldots, n$. We suppose that $p_i = 1/(1 + \exp(-x_i'\boldsymbol{\beta}))$ for some parameter vector $\boldsymbol{\beta}$ that is, we have a logistic regression

x_i Concentration (g / 100 cc)	m_i number of exposed insects	y_i number of killed insects
0.10	47	8
0.15	53	14
0.20	55	24
0.30	52	32
0.50	46	38
0.70	54	50
0.95	52	50

Table 4: Data for the toxicity example

model with grouped data: the link function from the structural assumption of the GLM is chosen as the logit function. The corresponding outlier regions can essentially be derived as in the Poisson case. Again, for the construction of a reliable one-step outlier identifier we need a robust estimator of $\boldsymbol{\beta}$. For this purpose we may e.g. use the Least Median of Weighted Squares (LMWS) or Least Trimmed Weighted Squares (LTWS) estimators as proposed in Christmann (2001). As an example look at the data in Table 4 which are taken from Myers et al. (2002). These data report the result from a toxicity experiment which has been conducted to investigate the effect of different doses of nicotine on the common fruit fly.

A reasonable model for this set of data is a logistic regression model with

$$p_i \;=\; \frac{1}{1+\exp(-(\beta_0+\beta_1\,\ln x_i))}. \tag{5}$$

For the data in Table 4 the ML- and LMWS-estimators of the model parameters are $\hat{\beta}_0^{ML} = 3.1236$, $\hat{\beta}_1^{ML} = 2.1279$, and $\hat{\beta}_0^{LMWS} = 3.3478$, $\hat{\beta}_1^{LMWS} = 2.3047$, respectively. Suppose now that the number of killed insects for the concentration of 0.70 g / 100 cc has been wrongly reported as 5 instead of 50. Then the parameter estimators are given by $\hat{\beta}_0^{ML} = 0.9278$, $\hat{\beta}_1^{ML} = 0.8964$, and $\hat{\beta}_0^{LMWS} = 3.3368$, $\hat{\beta}_1^{LMWS} = 2.2989$, respectively. For $\tilde{\alpha} = 0.01$ we now

x_i Concentration (g / 100 cc)	$inl(\alpha_i, Bin(m_i, \hat{p}_i^{ML}))$	$inl(\alpha_i, Bin(m_i, \hat{p}_i^{LMWS}))$
0.10	$\{3, \ldots, 21\}$	$\{0, \ldots, 14\}$
0.15	$\{7, \ldots, 28\}$	$\{5, \ldots, 25\}$
0.20	$\{10, \ldots, 32\}$	$\{12, \ldots, 34\}$
0.30	$\{13, \ldots, 35\}$	$\{22, \ldots, 43\}$
0.50	$\{16, \ldots, 37\}$	$\{31, \ldots, 45\}$
0.70	$\{24, \ldots, 46\}$	$\{43, \ldots, 54\}$
0.95	$\{26, \ldots, 46\}$	$\{45, \ldots, 52\}$

Table 5: Estimated inlier regions for the toxicity example

estimate the α_i-inlier regions for the distributions of the number of killed insects, where $\alpha_i = 0.00143$, $i = 1, \ldots, 7$, is determined according to condition (2). These estimated inlier regions are contained in Table 5. Here $\hat{p}_i^{ML}$ ($\hat{p}_i^{LMWS}$) denotes the plug-in estimator of p_i when inserting the ML- (the LMWS-) estimator of β_0 and β_1 into the right hand side of (5). Again, an observation is identified as α_i-outlier if it is not contained in the corresponding estimated α_i-inlier region.

Note that in this example both rules detect the wrongly reported number of killed insects at 0.70 g / 100 cc correctly as 0.0143-outlier. However, the rule based on the ML-estimator also identifies the numbers at 0.50 and 0.95 g / 100 cc as outlying. This is an example of the swamping effect that clearly demonstrates the unreliability of outlier identification rules based on non-robust methods.

5 Outlook

In this paper we have only discussed outliers with respect to the conditional distribution of the responses given the covariates which are treated as if they are fixed. This is the appropriate approach e.g. for the loglinear Poisson model of Section 4.1 where the x_i reflect the structure of a certain contingency table. However, in many cases the covariates are actually random. Hence it makes sense to consider outlier regions for the distribution of the covariates and the joint distribution of responses and covariates as well. For a GLM with normal distribution of the responses and the identity as link function (i.e. a linear regression model with normal errors) the distribution of the covariates is often assumed to be multivariate normal. In this case the joint distribution of responses and covariates is a multivariate normal distribution as well. Hence the outlier identifiers discussed e.g. in Becker and Gather (1999) can be applied for this setting. For other distributions of the responses, especially in the discrete case, it will be more complicated to model the joint distribution of responses and covariates and hence to derive the corresponding outlier regions. This should be the task of further research.

Acknowledgement

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475) is gratefully acknowledged.

References

BARNETT, V. and LEWIS, T. (1994): *Outliers in Statistical Data.* 3^{rd} ed., Wiley, New York.

BECKER, C. and GATHER, U. (1999): The Masking Breakdown Point of Multivariate Outlier Identification Rules. *Journal of the American Statistical Association, 94, 947–955.*

CHRISTMANN, A. (2001): Robust Estimation in Generalized Linear Models. In: J. Kunert, G. Trenkler (Eds.) *Mathematical Statistics with Applications in Biometry: Festschrift in Honour of Siegfried Schach.* Eul-Verlag, Lohmar, 215–230.

DAVIES, P.L. and GATHER, U. (1993): The Identification of Outliers. *Journal of the American Statistical Association, 88, 782–792.*

HAMPEL, F.R., RONCHETTI, E.M., ROUSSEEUW, P.J. and STAHEL, W.A. (1986): *Robust Statistics – The Approach Based on Influence Functions.* Wiley, New York.

HUBERT, M. (1997): The Breakdown Value of the L_1 Estimator in Contingency Tables. *Statistics and Probability Letters, 33, 419–425.*

FAHRMEIR, L. and TUTZ, G. (2001): *Multivariate Statistical Modelling Based on Generalized Linear Models.* 2^{rd} ed., Springer Verlag, New York.

GATHER, U., KUHNT, S. and PAWLITSCHKO, J. (2003): Concepts of Outlyingness for Various Data Structures. In: J.C. Misra (Ed.): *Industrial Mathematics and Statistics.* Narosa Publishing House, New Dehli, to appear.

KUHNT, S. (2000): *Ausreißeridentifikation im Loglinearen Poissonmodell für Kontingenztafeln unter Einbeziehung robuster Schätzer.* Dissertation, Department of Statistics, University of Dortmund, Germany.

MOSTELLER, F. and PARUNAK, A. (1985): Identifying Extreme Cells in a Sizable Contingency Table: Probabilistic and Exploratory Approaches. In: Hoaglin, D.C., Mosteller, F. and Tukey, J.W. (Eds.): *Exploring Data Tables, Trends and Shapes.* Wiley, New York, 189–224.

MYERS, R.H., MONTGOMERY, D.C. and VINING, G.C. (2002): *Generalized Linear Models.* Wiley, New York.

NELDER, J.A. and WEDDERBURN, R.W.M. (1972): Generalized Linear Models. *Journal of the Royal Statistical Society A, 134, 370–384.*

SHANE, K.V. and SIMONOFF, J.S. (2001): A Robust Approach to Categorical Data Analysis. *Journal of Computational and Graphical Statistics, 10, 135–157.*

YICK, J.S. and LEE, A.H. (1998): Unmasking Outliers in Two-Way Contingency Tables. *Computational Statistics* & *Data Analysis, 29, 69–79.*