A NOTE ON CONTAMINATION MODELS AND OUTLIERS

Jürgen Wellmann

GSF–National Research Center for Environment and Health

Institute of epidemiology, D–85764 Neuherberg

Ursula Gather

University of Dortmund, Department of Statistics

D–44221 Dortmund

Key Words: slippage model; univariate data; expected number of outliers

ABSTRACT

In order to describe or generate so-called outliers in univariate statistical data, contamination models are often used. These models assume that k out of n independent random variables are shifted or multiplicated by some constant, whereas the other observations still come i.i.d. from some common target distribution. Of course, these contaminants do not necessarily stick out as the extremes in the sample. Moreover, it is the amount and magnitude of 'contamination' which determines the number of obvious outliers. Using the concept of Davies and Gather (1993) to formalize the outlier notion we quantify the amount of contamination needed to produce a prespecified expected number of 'genuine' outliers. In particular, we demonstrate that for samples of moderate size from a normal target distribution a rather large shift of the contaminants is necessary to yield a certain expected number of outliers. Such an insight is of interest when designing simulation studies where outliers should occur as well as in theoretical investigations on outliers.

CONTAMINATION MODELS

The problem of outliers in statistical data has been treated for a long time, and several attempts have been made to give the vague term outlier a precise meaning. One approach consists in formulating distribution models which should describe the occurence of spurious data better than some null model. Such models may be called outlier generating models. To give a definition we start with a null model of i.i.d. normally distributed random variables, where the joint distribution underlying the observations $x_1, \ldots, x_n$ is a product distribution

$$\bigotimes_{i=1}^{n} N(\mu, \sigma^2), \quad -\infty < \mu < \infty, \sigma^2 > 0. \tag{1}$$

Many outlier generating models now assume that only the majority of the data comes i.i.d. from the target distribution. Those data may be called regular observations. The remainder are the so called contaminated observations, i.e. observations from a different source.

We look at an often used special case (cf. David (1979)) of a model introduced by Ferguson (1961). Let the mapping $\pi : \{1, \ldots, n\} \to \{0, 1\}$ indicate whether the i-th observation comes from the target distribution ($\pi(i) = 0$) or not ($\pi(i) = 1$), $i = 1, \ldots, n$. Let k denote the number of contaminants, that means that π maps exactly k of the indices onto 1. The function π is treated as an unknown parameter in this mean slippage model with the following joint distribution for the $x_1, \ldots, x_n$:

$$\begin{aligned} \bigotimes_{i=1}^{n} G_{\pi(i)}, \qquad & G_0 = N(\mu, \sigma^2), G_1 = N(\mu + a, \sigma^2), \\ & -\infty < \mu, a < \infty, \sigma^2 > 0, \pi \in \Pi, \end{aligned} \tag{2}$$

where Π denotes the set of all functions π of the above type. Similar models with fixed π have been used in simulation studies by several authors, e.g. Jain, Pingel, and Davidson (1982), Rosner (1983) and Marashinghe (1985).

OUTLIER REGIONS

Models of the above kind explain spurios data reasonably well. In a setting of analytical chemical measurements for example, one may think of an unwanted substance which happened to contaminate the sample under consideration and supplies an additional amount of the substance which is measured. Nevertheless there will be hardly any chance to validate model assumptions like a mean slippage model. Moreover, Barnett and Lewis (1994) "... define an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data". Thus, outliers are supposed to be something surprising and outlier generating models should therefore capture this element of surprise, i.e. they should be able to produce sufficiently aberrant values.

Note that the contaminants in the above model need not even stick out as extremes in the dataset. But in univariate, unimodally distributed data, only the extremes can raise suspicion. As a consequence, rules for the identification of outliers are tailored to detect sufficiently aberrant extremes and at least in order to compare and validate these rules we have to formalize what we mean by 'sufficiently aberrant'. We will use the concept of Davies and Gather (1993) who define an α-outlier-region for a normal distribution $N(\mu, \sigma^2)$ as

$$\text{out}(\alpha, \mu, \sigma^2) = \{x : |x - \mu| > \sigma z(\alpha)\}, \tag{3}$$

for some given α, $0 < \alpha < 1$, where $z(\alpha)$ is the $(1 - \alpha/2)$-quantile of the standard normal distribution. Any observation in the outlier-region $\text{out}(\alpha, \mu, \sigma^2)$ is called α-outlier, irrespective of the random mechanism that has placed it into this region.

Note that for some $N(\mu, \sigma^2)$-random variable X the probability to be located in $\text{out}(\alpha, \mu, \sigma^2)$ equals α. For a set of n independent random variables $X_1, \ldots, X_n$ we can adjust α according to n by taking $\alpha_n = 1 - (1 - \alpha)^{1/n}$, where $\alpha \in (0, 1)$ is a small, given number, such that we have

$$\Pr\Big(X_i \in \text{out}(\alpha_n, \mu, \sigma^2) \text{ for at least one } i = 1, \ldots, n\Big) = \alpha.$$

We use $\alpha = 0.05$ throughout this paper which gives values of $z(\alpha_n)$ equal to 2.800, 3.016, 3.283, 3.474, and 4.050 for $n = 10, 20, 50, 100$, and $1\,000$ respectively.

CONTAMINANTS IN THE OUTLIER REGION

An outlier generating model like (2) will of course only produce a sufficient number of α-outliers when k and a are large enough. In order to investigate this in more detail we will consider the random number M of α-outliers, produced by some outlier generating models.

Considering the mixture model as an outlier generating model, Tukey (1960) already investigated expected numbers of outliers in a similar sense. He took a mixture model with n independent observations, where $n - K$ are regular observations from $N(\mu, \sigma^2)$ and the remaining K come from $N(\mu, (b\sigma)^2)$, $b \geq 1$. The number K is a binomially distributed random variable with parameters n and λ. This is equivalent to a model of n i.i.d. observations from a mixture distribution

$$(1 - \lambda)N(\mu, \sigma^2) + \lambda N(\mu, (b\sigma)^2). \tag{4}$$

Therefore, under model (4), M is binomially distributed with parameters n and p, where p denotes the probability of one observation to fall into $\text{out}(\alpha, \mu, \sigma^2)$, i.e. $p = (1 - \lambda)\alpha + 2\lambda(1 - \Phi(z(\alpha)/b))$, where Φ denotes the c.d.f. of the standard normal distribution. The numbers of regular or contaminated observations in $\text{out}(\alpha, \mu, \sigma^2)$ under model (4) are therefore binomially distributed with parameters n and $p_0 = (1 - \lambda)\alpha$ or $p_1 = 2\lambda(1 - \Phi(z(\alpha/b))$ respectively.

Tukey considered a mixture distribution with $\lambda = 0.01$ and $b = 3$. Despite the threefold increase of the standard deviation of the contaminants this mixture distribution is close to the normal distribution as Tukey has demonstrated by comparing some quantiles of both distributions. Consequently the fraction of outliers produced by this mixture distribution is low. Tukey was interested in the expected number of observations in the tail region outside $\mu \pm 2.512\sigma$, which corresponds to the 0.006- or 0.994-quantile of the normal distribution and therefore to the 0.012-outlier region in the sense of Davies and Gather (1993). We extended Tukey's example and found that for more extreme outlier regions, which are attained by using α_n as above instead of $\alpha = 0.012$, the fraction p of α_n-outliers is smaller than the fraction λ of contaminants (see table I). Moreover, Tukey noted the disturbing fact that the expected number of contaminated observations in the 0.012-outlier region, np_1 in our notation, is considerably lower than the expected number of regular observations, i.e. than np_0. For his example we obtain $p_0 = 0.011\,880$ and $p_1 = 0.004\,024$. Only for large n which yields quite extreme outlier regions we get values of p_0 lower than p_1 (see table I).

TABLE I

Probabilities $p = p_0 + p_1$ for an observation from a mixture distribution with $\lambda = 0.01$ and $b = 3$ to be placed in $\mathrm{out}(\alpha_n, \mu, \sigma^2)$

n	α_n	$z(\alpha_n)$	p_0	p_1	p
1	0.050 00	1.960	0.049 500	0.005 135	0.054 635
Tukey's example		2.512	0.011 880	0.004 024	0.015 904
10	0.005 12	2.800	0.005 065	0.003 507	0.008 572
20	0.002 56	3.016	0.002 536	0.003 147	0.005 683
50	0.001 03	3.283	0.001 015	0.002 737	0.003 752
100	0.000 51	3.474	0.000 508	0.002 469	0.002 976
1 000	0.000 05	4.050	0.000 051	0.001 771	0.001 821

In the following we will not consider the mixture model any longer because of its deficiency of being a model of i.i.d. variables which is anyway not very appropriate to describe outliers as exceptional observations. Therefore, we concentrate on the Ferguson-model (2). In order to compute the distribution of the number of α-outliers under model (2), let I denote a subset of $\{1, \ldots, n\}$ of size $m, 0 \leq m \leq n$. Suppose that j of the X_i with $i \in I$ are contaminated, $j \leq \min\{m, k\}$, and $m - j$ are regular observations. The number $m - j$ of regular observations in this subset cannot exceed the number $n - k$ of regular observations in the whole sample, i.e. we have $j \geq m - n + k$. For fixed m and j there are $\binom{k}{j}\binom{n-k}{m-j}$ distinct subsets of size m with j contaminants. Under model (2), the probability that exactly those X_i with $i \in I$ are α-outliers equals

$$\begin{aligned} &(\Phi(-z(\alpha) - a) + 1 - \Phi(z(\alpha) - a))^j \quad \times \quad \alpha^{m-j} \\ \times \quad &(\Phi(z(\alpha) - a) - \Phi(-z(\alpha) - a))^{k-j} \quad \times \quad (1 - \alpha)^{n-k-m+j}. \end{aligned} \tag{5}$$

Denote this probability by $\tau_j(\alpha, a),\ j = \max\{0, m - n + k\}, \ldots, \min\{m, k\}$. Summarizing over all 'admissible' sets of size m we get for the distribution of the random variable M

$$\Pr(M = m) = \sum_{j=\max\{0,m-n+k\}}^{\min\{m,k\}} \binom{k}{j}\binom{n-k}{m-j}\tau_j(\alpha, a), \qquad m = 0, \ldots, n.$$

In order to compare those distributions for various values of n, k and α, we may ask now, as some type of a reverse question to Tukey's, how large the shift parameter a must be to ensure a sufficient number of α-outliers. More precisely, we determine the shift parameter a which achieves that the expected number of α-outliers is equal to the number k of assumed contaminants in model (2), i.e.

$$E(M) = E\left(M(\alpha, a)\right) = k. \tag{6}$$

Our results are listed in the table below for $n = 10, 20, 50, 100$, and some k, together with the standard deviation of M at the corresponding value of a.

TABLE II

Values of a which fulfill (6) under model (2) with k contaminants and the standard deviation of $M(\alpha, a)$ with this a

$n = 10$	$\alpha = .050$	$z = 1.96$	$\alpha = .012$	$z = 2.51$	$\alpha_n = .0051$	$z = 2.80$
k	a	std(M)	a	std(M)	a	std(M)
1	2.086	0.822	3.749	0.451	4.484	0.300
2	2.802	0.837	4.177	0.432	4.844	0.284
3	3.152	0.801	4.423	0.406	5.059	0.266
4	3.399	0.750	4.609	0.377	5.224	0.247
$n = 20$	$\alpha = .050$	$z = 1.96$	$\alpha = .012$	$z = 2.51$	$\alpha_n = .0026$	$z = 3.02$
k	a	std(M)	a	std(M)	a	std(M)
1	0.000	0.975	3.258	0.633	4.674	0.308
2	2.086	1.162	3.749	0.637	5.010	0.302
3	2.533	1.190	4.003	0.626	5.199	0.294
4	2.802	1.183	4.177	0.610	5.333	0.285
7	3.283	1.099	4.521	0.554	5.609	0.258
9	3.506	1.019	4.691	0.510	5.750	0.237
$n = 50$	$\alpha = .050$	$z = 1.96$	$\alpha = .012$	$z = 2.51$	$\alpha_n = .0010$	$z = 3.28$
k	a	std(M)	a	std(M)	a	std(M)
1			2.290	0.907	4.926	0.313
2			3.071	0.990	5.250	0.312
3	1.173	1.656	3.398	1.008	5.426	0.309
4	1.771	1.778	3.601	1.011	5.547	0.306
7	2.464	1.879	3.961	0.994	5.778	0.296
9	2.706	1.879	4.113	0.975	5.883	0.290
24	3.566	1.570	4.738	0.785	6.342	0.231
$n = 100$	$\alpha = .050$	$z = 1.96$	$\alpha = .012$	$z = 2.51$	$\alpha_n = .0005$	$z = 3.47$
k	a	std(M)	a	std(M)	a	std(M)
1					5.111	0.315
2			2.290	1.283	5.432	0.315
3			2.797	1.365	5.604	0.314
4			3.071	1.399	5.721	0.313
7	1.535	2.445	3.509	1.429	5.941	0.308
9	1.946	2.564	3.681	1.428	6.037	0.305
24	2.961	2.609	4.287	1.334	6.417	0.279
49	3.585	2.200	4.754	1.100	6.746	0.229

All computations refer to model (2) with $\mu = 0$ and $\sigma^2 = 1$. We consider α-outliers with $\alpha = 0.05$ and, in analogy to Tukey (1960), $\alpha = 0.012$, as well as $\alpha_n = 1 - (1 - 0.05)^{1/n}$. When the expected number of α-outliers is higher than k even without any contamination the corresponding fields in the table are left blank.

DISCUSSION

Contamination models may produce some outliers but they lack in yielding enough outliers as extreme values, since contaminants can be realized even just in the bulk of the data. In the present paper we have investigated the relation between a contamination model which is most popular when carrying out simulation studies with outliers and outlier regions in the sense of Davies and Gather (1993).

As a result we show in table II that the shift-parameter a in model (2) has to be remarkably large in order to supply the desired expected number of outliers the sense of (6). This is true especially for the α_n-outlier-case, even with small k, and can be seen for other situations as well. A quite large shift-parameter, at least 4, say, is even necessary to get a reasonable number of (less extreme) α-outliers as can be seen from our table since the variation of M is large at smaller values of a. This is remarkable since then for example the 97.5% quantile of the regular observations (1.96) is still lower than the 2.5% quantile of the shifted observations (2.04).

As a side issue, table II shows that for larger numbers k of contaminants a larger shift parameter a is needed to obtain a sufficient number of outliers. This means that it is not sufficient to increase the number of contaminants k in order to increase the number of expected outliers in the same size, which can be interpreted as a deficiency of the slippage model to generate large numbers of outliers.

Tukey showed that slight contamination in the mixture model hardly produces outliers, either. In a similar sense we have demonstrated that a considerable amount of contamination is necessary to make the contaminants in a Ferguson-model show up as outliers.

These findings are helpful for the design of simulation studies with outliers and they may be of help in other theoretical considerations on outlier models as well.

ACKNOWLEDGEMENTS

Work of the first author on this topic was partly supported by DFG Grants DA 237/1-2 and GA 338/2-2 while he was assistant scientist at the University of Dortmund, Department of Statistics.

BIBLIOGRAPHY

Barnett, V. and Lewis, T. (1994). *Outliers in Statistical Data*, Third Edition, John Wiley & Sons, Chichester, New York.

David, H.A. (1979). "Robust Estimation in the presence of outliers", *Robustness in Statistics,* (R.L. Launer and G.N. Wilkinson, Eds.). Academic Press, New York.

Davies, P.L. and Gather, U. (1993). "The identification of multiple outliers", invited Paper with discussion and reply, *Journal of the American Statistical Association*, 88, Theory and Methods, 782–792.

Ferguson, T.S. (1961). "On the rejection of outliers", *Proceedings of the Fourth Berkeley Symposium on Mathematical Statictics and Probability*, 1, 253–287.

Jain, R.B. Pingel, L.A. and Davidson, J.L. (1982). "A unified approach for estimation and detection of outliers", *Communications in Statistics*, A11, 2953–2976.

Marashinghe, M.G. (1985). "A Multistage Procedure for Detecting Several Outliers in Linear Regression", *Technometrics*, 27, 395–399.

Rosner, B. (1983). "Percentage Points for a Generalized ESD Many-Outlier Procedure", *Technometrics*, 25, 165–172.

Tukey, J.W. (1960). "A survey of sampling from contaminated distributions", *Contributions to Probability and Statistics*, (I. Olkin, Ed.). University Press, Stanford, California.