

**A Note on an Alternative Outlier Model**

by

**R.D. Cook, N. Holschuh, S. Weisberg**

**University of Minnesota**

**Technical Report No. 373**

**May 1979**

### Abstract

This paper examines modeling a single outlier in the normal theory fixed effects linear model as arising from an unknown observation with inflated variance. The maximum likelihood estimates are characterized in terms of standard least squares statistics. The estimated position of the outlier does not necessarily agree with the estimated position under the usual mean slippage outlier model, and an example where they differ is presented. A sufficient and common condition for agreement is given.

### Keywords

Outliers, Linear Models, Maximum Likelihood Estimation, Mean Slippage Models, Variance Slippage Models.

### Acknowledgement

This work was supported in part by grant 1-R01-GM25587 from the National Institute of General Medical Science, U.S. Department of Health, Education and Welfare. Part of the work was completed while the first author was a Hartley Visiting Fellow, University of Southampton, England.

## 1. Introduction

A common approach to modeling outliers in the fixed effects linear model is to assume that outliers result from slippages in the expected values of contaminated observations (For review, see Barnett and Lewis, 1978). In the case of a possible single outlier, this leads to considering the Studentized residuals, since, under normality and the possibility of each observation being the outlier, the maximum likelihood estimate for the position of the possible outlier corresponds to the case with the largest absolute Studentized residual. Also, the likelihood ratio test statistic for the presence of an outlier in this model is a monotonic function of the largest absolute Studentized residual (Srikantan, 1961; Tietjen, Moore, and Beckman, 1973; Ellenberg, 1976). Intuitively, one might expect similar results to be true whenever the effect of a single outlier is modeled identically for all observations, regardless of values of the independent variables or fixed effects. We shall see that this is not necessarily so.

We consider here a single outlier model that assumes an outlier arises from an error term with an inflated variance. In section 2 the model is defined and shown to be, after a parameter transformation, a special case of Harville's general linear model (Harville, 1977). The joint maximum likelihood estimates (MLE's) are characterized in section 3, and an example where the estimated outlier does not correspond to the observation with the largest absolute Studentized residual is presented in section 4. In the final section several issues pertinent to the modeling of outliers are discussed in terms of this model and the mean slippage model.

Our intent here is not to propose this model as a common replacement for the mean slippage model, but rather, to examine the performance of the

mean slippage model under a plausible alternative.

## 2. The Model

The  $n$ -dimensional observation vector  $y$  is assumed to have the representation

$$y = X\beta + z ,$$

where  $X$  is a known  $n \times p$  ( $n > p$ ) full rank matrix,  $\beta$  is an unknown  $p$ -dimensional parameter vector, and  $z$  is an  $n$ -dimensional random vector of errors. Furthermore,  $z$  has the representation

$$z = \pi e ,$$

where  $e$  follows an  $n$ -dimensional normal distribution with mean zero and covariance matrix  $\sigma^2 \begin{bmatrix} w & 0 \\ 0 & I_{n-1} \end{bmatrix}$  with  $\sigma^2 > 0$ ,  $w \geq 1$ , both unknown; and  $\pi$  is an unknown  $n \times n$  permutation matrix from the set  $\Pi$  composed of  $n$ -dimensional permutation matrices  $\pi_i$ ,  $i = 1, \dots, n$  where  $\pi_i$  permutes only the first and  $i$ th coordinates. Hence, the parameter space  $\Theta$  is

$$\Theta = \{ \theta = (\beta, \sigma^2, w, \pi) \mid \beta \in R^p, \sigma^2 > 0, w \geq 1, \pi \in \Pi \} .$$

This permits the possible presence of a single outlier with inflated variance in the usual normal theory linear model.

It is instructive to observe that under the 1-1 parameter transformation  $g: (\sigma^2, w) \rightarrow (\tau, \eta)$  given by  $g(\sigma^2, w) = (\sigma^2, \sigma^2(w-1))$ ,  $y$  has the representation

$$y = X\beta + \pi e_1 u + v , \quad \text{where } e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{n \times 1} ,$$

$u$  is a univariate normal with mean 0 and variance  $\eta$  ( $\geq 0$ ), and  $v$  is an  $n$ -dimensional normal vector, independent of  $u$ , with mean 0 and covariance matrix  $\pi I_n$ . The problems and properties of maximum likelihood estimation in general linear models of this form are examined in Harville's 1977 review paper. For generalizations of the variance inflation model it may be preferable to use this parameterization. In the present case it is convenient to derive the MLE's for the original parameterization.

### 3. Maximum Likelihood Estimation

To maximize the likelihood function, first fix the value of the permutation matrix  $\pi$  and maximize over  $\theta$  in terms of  $\beta$ ,  $\sigma^2$ , and  $w$ . Repeating for each of the  $n$  possible permutation matrices gives  $n$  values of the likelihood function, the largest being the maximum of the likelihood function. The permutation matrix associated with this largest value is the MLE of  $\pi$  and the values of  $\beta$ ,  $\sigma^2$ , and  $w$  that produce this value are their MLE's. Therefore, we determine the permissible values of  $\beta$ ,  $\sigma^2$ , and  $w$  that maximize the likelihood function when  $\pi$  is assumed fixed at  $\pi_1$ . Except for an additive constant the log-likelihood function is

$$l_1(\theta; y) = -\frac{n}{2} \log(\sigma^2) - \frac{1}{2} \log(w) - \frac{1}{2\sigma^2} (y - X\beta)' (\pi_1 W)^{-1} (y - X\beta)$$

where  $W = \begin{bmatrix} w & 0 \\ 0 & I_{n-1} \end{bmatrix}$ .

We must assume that  $n > p+1$ , else the log-likelihood is unbounded.

Now for  $w$  fixed and positive, standard normal theory results have

$l_1(\theta; y)$  maximized at  $(\beta, \sigma^2) = (\hat{\beta}_1(w), \hat{\sigma}_1^2(w))$ , where

$$\hat{\beta}_1(w) = (X'(\pi_1 W)^{-1}X)^{-1}X'(\pi_1 W)^{-1}y,$$

$$\hat{\sigma}_1^2(w) = \frac{1}{n}(y'(\pi_1 W)^{-1}y - y'(\pi_1 W)^{-1}X\hat{\beta}_1(w)).$$

It is useful to express  $\hat{\beta}_1(w)$  and  $\hat{\sigma}_1^2(w)$  in terms of their MLE's (denoted  $\tilde{\beta}$  and  $\tilde{\sigma}^2$ ) under the model with no outliers. Letting  $X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}$  with

$x_j$ ,  $j=1, \dots, n$  being  $p$ -dimensional vectors,  $v_i = x_i'(X'X)^{-1}x_i$ ,  $\tilde{y}_i = x_i'\tilde{\beta}$ ; we have  $\tilde{\beta} = (X'X)^{-1}X'y$ ,  $\tilde{\sigma}^2 = \frac{1}{n}(y'y - y'X\tilde{\beta})$ , and, assuming  $1-v_i > 0$ ,  $t_i^2 = (y_i - \tilde{y}_i)^2 / (\tilde{\sigma}^2 n(n-p)^{-1}(1-v_i))$ , the square of the usual  $i$ th Studentized residual. Then

$$\hat{\beta}_1(w) = \tilde{\beta} - (y_i - \tilde{y}_i) \left( \frac{w-1}{w-(w-1)v_i} \right) (X'X)^{-1}x_i,$$

$$\hat{\sigma}_1^2(w) = \tilde{\sigma}^2 \left( 1 - \left( \frac{w-1}{w+v_i(1-v_i)} \right)^{-1} \frac{t_i^2}{n-p} \right).$$

When  $1-v_i=0$ ,  $\hat{\sigma}_1^2(w) = \tilde{\sigma}^2$  and  $\hat{\beta}_1(w) = \tilde{\beta}$ .

We note that  $\hat{\beta}_1(1) = \tilde{\beta}$ ,  $\hat{\sigma}_1^2(1) = \tilde{\sigma}^2$  and, if  $1-v_i > 0$ ,

$$\lim_{w \rightarrow +\infty} \hat{\beta}_1(w) = \tilde{\beta} - (y_i - \tilde{y}_i)(1-v_i)^{-1}(X'X)^{-1}x_i,$$

the usual MLE of  $\beta$  if the  $i$ th observation is ignored, and

$$\lim_{w \rightarrow +\infty} \hat{\sigma}_1^2(w) = \tilde{\sigma}^2 (1 - t_i^2 (n-p)^{-1}) = (n-1)n^{-1} \tilde{\sigma}_{(i)}^2,$$

where  $\tilde{\sigma}_{(i)}^2$  is the usual MLE of  $\sigma^2$  ignoring the  $i$ th observation. Now  $\ell_1(\theta; y)$  evaluated at  $(\beta, \sigma^2) = (\hat{\beta}_1(w), \hat{\sigma}_1^2(w))$  is, except for an additive constant, proportional to

$$h_1(w) = -n \log(\hat{\sigma}_1^2(w)) - \log(w).$$

It only remains to find the value of  $w$ , say  $\hat{w}$ , over the range  $[1, +\infty)$  that maximizes  $h_1(w)$ . Then  $(\beta, \sigma^2, w) = (\hat{\beta}_1(\hat{w}), \hat{\sigma}_1^2(\hat{w}), \hat{w})$  maximizes  $\ell_1(\theta; y)$ .

The two special cases when  $v_1=0$  or  $1$  are handled separately. If  $v_1=1$ ,  $\hat{\sigma}_1^2(w)=\tilde{\sigma}^2$  so  $\hat{w}=1$ . If  $v_1=0$ , which occurs only when  $x_1=0^*$ , straightforward differentiation of  $h_1(w)$  gives  $\hat{w}=\max\{1, (n-1)y_1^2/(n\tilde{\sigma}^2-y_1^2)\}$ .

To determine  $\hat{w}$  when  $0 < v_1 < 1$ , we need the following directly verifiable facts:

$$(1) \lim_{w \rightarrow +\infty} \hat{\sigma}_1^2(w) = (n-1)n^{-1} \tilde{\sigma}_{(1)}^2 > 0 \text{ (since } n-1 > p \text{);}$$

$$(2) \lim_{w \rightarrow 0} \hat{\sigma}_1^2(w) = \tilde{\sigma}^2 \left( 1 + \left( \frac{1-v_1}{v_1} \right) \frac{t_1^2}{n-p} \right) > 0;$$

$$(3) \frac{d}{dw} (h_1(w)) = \frac{\binom{n}{n-p} \left( \frac{t_1^2}{1-v_1} \right)}{\left( w + \frac{v_1}{1-v_1} \right) \left( w + \frac{v_1}{1-v_1} - (w-1) \frac{t_1^2}{n-p} \right)} - \frac{1}{w};$$

(4) when they exist the real roots to the equation  $\frac{d}{dw} (h_1(w)) = 0$  are

$$\frac{t_1^2(n+2v_1-1) - 2v_1(n-p) \pm \left( t_1^2 (4nv_1 (t_1^{2-n+p}) + (n-1)^2 t_1^2) \right)^{1/2}}{2(1-v_1) (n-p-t_1^2)}$$

From (1) and (2) it is clear that  $\lim_{w \rightarrow 0} h_1(w) = +\infty$  and  $\lim_{w \rightarrow +\infty} h_1(w) = -\infty$ .

Thus,  $\hat{w}$  is equal to one or else it is the largest root in (4), assuming it exists and is greater than one. In the situation where the largest root in (4) is a candidate, two cases can occur. If  $\frac{d}{dw} (h_1(w))|_{w=1} > 0$ , real roots necessarily exist and the largest maximizes  $h_1(w)$ ,  $w \in [1, +\infty)$ .

From (3), it follows that this occurs if, and only if,  $(y_1 - \tilde{y}_1)^2 / \tilde{\sigma}^2 > 1$ .

In contrast, when  $v_1=0$  this is a necessary and sufficient condition for  $\hat{w}$  to be greater than one. If  $\frac{d}{dw} (h_1(w))|_{w=1} \leq 0$ ,

either  $w=1$  or the largest root maximizes  $h_1(w)$  and one must evaluate  $h_1(w)$  at both points.

The above discussion characterizes the values of  $(\beta, \sigma^2, w)$  that maximize  $\ell_1(\theta; y)$ . Following this procedure for all  $\pi_1 \in \Pi$  determines the MLE's of  $(\beta, \sigma^2, w, \pi)$ . It would be convenient if a simple statistic,

\* This can occur only if an intercept term is not included in the model.

such as the largest absolute Studentized residual in the mean slippage single outlier model, produced the MLE of  $\pi$ . This is true only in a special, but common, situation. It suffices to consider two possible permutation matrices since the MLE of  $\pi$  could be determined by pairwise comparisons of maximized values of  $l_1(\theta; y)$  or, equivalently,  $h_1(w)$ . Hence, we wish to know conditions for  $\sup_{w \geq 1} h_1(w) \geq \sup_{w \geq 1} h_j(w)$ ,  $i \neq j$ ,

or, equivalently, for  $\inf_{w \geq 1} (\hat{\sigma}_i^2(w)w^{1/n}) \leq \inf_{w \geq 1} (\hat{\sigma}_j^2(w)w^{1/n})$ . A sufficient condition is  $\hat{\sigma}_i^2(w) \leq \hat{\sigma}_j^2(w)$  for all  $w \geq 1$ . From the above expression for  $\hat{\sigma}_k^2(w)$ ,

$$\hat{\sigma}_i^2(w) - \hat{\sigma}_j^2(w) = \left[ \frac{\tilde{\sigma}^2(w-1)(n-p)^{-1}}{\left(w + \frac{v_i}{1-v_i}\right) \left(w + \frac{v_j}{1-v_j}\right)} \right] \left[ w(t_j^2 - t_i^2) + \left(\frac{v_i}{1-v_i}\right)t_j^2 - \left(\frac{v_j}{1-v_j}\right)t_i^2 \right].$$

Thus  $t_i^2 \geq t_j^2$  and  $v_i(1-v_i)^{-1}t_j^2 - v_j(1-v_j)^{-1}t_i^2 \leq t_i^2 - t_j^2$ , imply  $\hat{\sigma}_i^2(w) - \hat{\sigma}_j^2(w) \leq 0$  for  $w \in [1, +\infty)$ , and if either inequality is strict,  $\hat{\sigma}_i^2(w) - \hat{\sigma}_j^2(w) < 0$  for  $w \in (1, +\infty)$ . The second inequality is equivalent to  $(y_i - \tilde{y}_i)^2 \geq (y_j - \tilde{y}_j)^2$ . Hence, if observation  $i$  has both a larger absolute Studentized residual and larger absolute residual than observation  $j$ , then  $\sup_{w \geq 1} h_i(w) \geq \sup_{w \geq 1} h_j(w)$ .

If  $\sup_{w \geq 1} h_i(w)$  occurs at  $w=1$ ,  $\sup_{w \geq 1} h_i(w) = -n \log(\tilde{\sigma}^2)$  which does not depend on  $i$ . So, if one observation has  $h_i(w)$  maximized for  $w \in (1, +\infty)$ , it must be true that the MLE of  $w$  is greater than one. But, since

$$\sum_{i=1}^n \frac{(y_i - \tilde{y}_i)^2}{\tilde{\sigma}^2} = n, \quad \max_{1 \leq i \leq n} \frac{(y_i - \tilde{y}_i)^2}{\tilde{\sigma}^2} > 1,$$

with probability one, implying that  $\sup_{w \geq 1} h_i(w)$  occurs for  $w \in (1, +\infty)$

for some  $i$ . Consequently, the MLE of  $w$  is always greater than one and, in the situation where the observation with the largest absolute Studentized residual also has the largest absolute residual from the ordinary least squares fit, the corresponding permutation matrix is the MLE of  $\pi$ .

In the following section we produce an example where the observation with the largest absolute Studentized residual is not estimated to be the possible outlier. In fact one can construct examples where the estimated outlier corresponds to neither the largest absolute Studentized residual nor the largest absolute residual.

#### 4. An Example

This data involves an examination of a less costly method for measuring the thickness of non-magnetic coatings of galvanized zinc on iron and steel (referenced in Freeman, 1942). The dependent  $y$ -values are thickness measurements ( $10^{-5}$  in.) on 11 pieces of coated iron and steel made by the less costly, but untested, non-destructive magnetic method. The independent  $x$ -values are the 11 accurate measurements from the standard, destructive stripping method. As an informative first step, one might fit a simple linear regression model for the expectation of  $y$ , i.e.,

$$E(y_i) = \alpha + \gamma x_i.$$

We explore this data assuming the single outlier model discussed above. For the sake of this example the  $y$ -value for the 9th observation was changed from 250 to 248.4.

The relevant calculations for each observation are presented below. The residuals come from the ordinary least squares fit and "t-test" refers to the square root of the F-test (with appropriate sign attached) for testing whether an observation has a slippage of its mean, assuming it is the only such possibly contaminated observation.  $\hat{w}$  is the value

that satisfies  $h_1(\hat{w}) = \sup_{w>1} h_1(w)$ .

Table 1. Data for example<sup>a</sup>

Observation	X	Y	Residual	Studentized Residual	t-test	$\hat{w}$	$h_1(\hat{w})$
1	116	105	-.68	-.05	-.04	1	-58.63
2	132	120	.18	.01	.01	1	-58.63
3	104	85	-10.07	-.69	-.66	1	-58.63
4	139	121	-5.01	-.34	-.32	1	-58.63
5	114	115	11.09	.75	.73	1	-58.63
6	129	127	9.83	.66	.64	1	-58.63
7	720	630	-9.67	-1.06	-1.07	1	-58.63
8	174	155	-1.95	-.13	-.12	1	-58.63
9	312	248.4	-30.56	-2.03	-2.60	9.15	-55.18
10	338	310	8.05	.54	.52	1	-58.63
11	465	443	28.77	2.04	2.63	10.37	-55.21

<sup>a</sup>The MLE's for  $(\alpha, \gamma, \sigma^2, w, \pi)$ , say  $(\hat{\alpha}, \hat{\gamma}, \hat{\sigma}^2, \hat{w}, \hat{\pi})$ , are:

$$\hat{\alpha} = 4.62, \hat{\gamma} = .89, \hat{\sigma}^2 = 123.32, \hat{w} = 9.15, \hat{\pi} = \pi_9.$$

For comparison, the MLE's for  $(\alpha, \gamma, \sigma^2)$  assuming no outliers are:

$$\tilde{\alpha} = 3.12, \tilde{\gamma} = .88, \tilde{\sigma}^2 = 206.42.$$

We see that observation 11 has the largest absolute Studentized residual, but observation 9 with the largest absolute residual is the estimated outlier. With the original y-value of 250 for the 9th observation, the ordering in both sets of absolute residuals for these observations is the same but observation 11 is the estimated outlier. Ignoring multiple testing considerations, it appears on the basis of the t-test values both observations 9 and 11 are suspect.

### 5. Comments

Several relevant issues in the handling of outliers are illustrated by comparisons of the mean slippage single outlier model and the variance inflation (slippage) single outlier model discussed here.

As demonstrated by the example, the estimated position of the possible outlier is not necessarily the same for both models. This is further evidence of the well recognized fact that determination of an outlier can depend critically on the form of the outlier model assumed to be operating in the data. For example, in the single sample situation an outlier in the usual scale may not be viewed as such if the same model is fit in the log scale. On the other hand, we reiterate that under these two models the estimated position of the outlier will be the same when the largest absolute Studentized residual corresponds to the largest absolute residual. This will occur in all balanced designed experiments where all residuals have the same variance.

Interestingly, the two models, to some degree, represent two different approaches to the treatment of outliers (see Barnett and Lewis, 1978). One approach is to eliminate an estimated outlier from an analysis, usually after a test, and analyze the remainder of the data with the uncontaminated model. This is essentially the effect of the mean slippage model for a single outlier. The MLE's for  $\beta$  and  $\sigma^2$  are the same (except for a divisor of  $n$  rather than  $n-1$  in the estimate of  $\sigma^2$ ) as those removing the estimated outlier and fitting the remainder of the data with the usual normal theory linear model. An alternative approach is the specific accommodation of the outlier in the model, as in the variance inflation model. Once the MLE's of  $w$  and  $\pi$  are found, the MLE's of  $\beta$  and  $\sigma^2$  are obtained by a weighted least squares fit where the only observation without unit weight is the estimated outlier, which has weight  $\hat{w}^{-\frac{1}{2}}$ . We note that modeling a single outlier as arising from both a mean and variance slippage is an overparameterization leading to the same maximum likelihood estimates as the mean slippage model.

We found the MLE of  $w$  is always greater than one. An important problem is the performance of the MLE's when, in fact, there are no outliers, i.e.,  $w = 1$ . This requires some knowledge of the small sample distribution properties of the MLE's. Unfortunately, this distribution problem appears to be intractable. In contrast we do have some distribution properties for the mean slippage single outlier model. Under the assumption of no outliers a monotonic transformation of each absolute Studentized residual follows an  $F(1, n-p-1)$  distribution. Consequently we can, at least, apply Bonferonni bounds for the largest of the absolute Studentized residuals, which corresponds to the likelihood ratio test statistic for testing the hypothesis of no outlier. For the variance inflation model, no simple procedure is available.

Finally, generalizing this variance inflation model is conceptually easy but computationally it poses problems. The immediate generalization is that possibly  $k$  out of the  $n$  observations have variances  $w\sigma^2$ . The method for maximizing the likelihood is the same but now the roots of  $\frac{d}{dw}(h(w)) = 0$  are roots of a polynomial of degree  $2k + 2$  ( $k \geq 2$ ). Solving this polynomial for all  $\binom{n}{k}$  possible outlier positions could be unfeasible.

## References

- Barnett, V. and T. Lewis (1978). Outliers in Statistical Data, John Wiley and Sons, New York.
- Ellenberg, J.H. (1976). "Testing for a Single Outlier from a General Linear Regression". Biometrics, 32, 637-645.
- Freeman, H.A. (1942). Industrial Statistics, John Wiley and Sons, New York.
- Harville, D.A. (1977). "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems" (with discussion). J. Amer. Statist. Ass., 72, 320-340.
- Srikantan, K.S. (1961). "Testing for the Single Outlier in a Regression Model." Sankhyā, A, 23, 251-260.
- Tietjen, G.L.; Moore, R.H.; and R.J. Beckman (1973). "Testing for a Single Outlier in Simple Linear Regression". Technometrics 15, 717-721.