L_1-*Statistical Procedures and Related Topics*
IMS Lecture Notes – Monograph Series (1997) Volume 31

Properties of L^1 residuals

Stephan Morgenthaler

Swiss Federal Institute of Technology, Lausanne, Switzerland

Abstract: The paper discusses the behavior of residuals from least-absolute-deviations (or L^1) fits of linear models. Particular emphasis is given to data arising by way of designed experiments. The paper argues that the L^1 method of fitting such models should be discouraged. The method is inefficient when compared to other robust methods while not being any simpler to compute. The residuals obtained by L^1 fitting exhibit several weaknesses. First of all they are ambiguous in the sense that there are a multitude of L^1 fits, sometimes quite far apart. Second, typical algorithms produce as many exact zero residuals as there are contrasts fitted. As a result, the non zero residuals do not give an accurate reflection of the errors that occurred during the experimental runs.

Key words: Least absolute deviations, factorial designs, outliers detection, uniqueness of fit.

AMS subject classification: 62K15, 62J05.

1 Introduction

Let $y = X\theta + \varepsilon$ be a linear model with uncorrelated, centered, and homoskedastic errors $\varepsilon_1, \ldots, \varepsilon_n$. As indicated, we take n to be the number of observations, whereas p denotes the dimension of the regression parameter θ. The least-squares residuals are $r = (I - H)y$, where $\hat{y} = Hy = X(X^TX)^{-1}X^Ty$ is the least-squares fit. If the error distribution has two moments, it follows that $E(r) = X\theta - HX\theta = 0$, and $Var(r) = \sigma^2(I - H)$. The use of such residuals for outlier detection and other diagnostic purposes has been explored in great detail in the statistical literature (see for example Belsley, Kuh and Welsch, 1980; Cook and Weisberg, 1982).

Residuals from an L^1 fit are not so easily described. Throughout this article, we denote by $\tilde{\theta}$ a parameter fit obtained by minimizing the least-

absolute-deviations and designate by e the corresponding residuals $e = y - X\tilde{\theta}$. Because $\mathbb{R}^\kappa$ equipped with the L^1 norm is only a weakly convex normed linear space, the best approximation to y of the form $X\tilde{\theta}$ is in general not unique. In fact, the best approximations form a convex set in the p-dimensional column space of X. It is widely-known that among the best approximations there is always one, for which at least p of the components of e are exactly equal to zero. The algorithms based on linear programming techniques always identify one of these solutions, because they correspond to extremal points of the linear programming problem.

For Gaussian errors, the least-squares fit $\hat{\theta}$ is fully efficient, whereas the least-absolute-deviation fit $\tilde{\theta}$ reaches an asymptotic efficiency of $2/\pi = 63.7\%$. In balanced factorial models the element h_{ij} of H, proportional to the covariance of the least-squares fit $\hat{y}_i$ and $\hat{y}_j$, depends in a simple way on the factor settings at runs i and j. In the case of a two-way ANOVA with factors f_1 and f_2, for example, there are four cases, distinguished by the comparison of (f_{1i}, f_{2i}) and (f_{1j}, f_{2j}). In particular, the diagonal elements h_{ii} are all equal to p/n, where p is the dimension of the column space of X. In the following, we restrict our discussion to this case of a balanced factorial model.

The asymptotic behavior of the residuals is for L^1 and L^2 the same, as long as p/n tends to zero with increasing n. The residuals are asymptotically equivalent to a sample from the error distribution. Asymptotic considerations are, however, of minor interest when one discusses properties of residuals. The case $p \approx n$ is of much more practical concern.

2 Identifying a small flock of outliers

Least-squares residuals have a tendency to behave much like a sample from a Gaussian distribution. Stem-and-leave plots or normal plots do often not reveal anything of interest. This is due to the dependence imposed on the residuals by the requirement that $r^T X = 0$. For that reason, glaring error structures will be lost or not faithfully translated into residual structures. An example of this sort concerns the presence of a few **outliers** among the measurement errors. To illustrate what happens, suppose the residual r has a fixed size, e.g. $r^T r = 1$ and we seek to maximize $w^T r$ for a fixed vector of component weights w. The solution to this constrained optimisation problem yields a maximal value of $w^T w - w^T H w$ achieved, when $r \propto (I - H)w$. Thus, if we maximize a single component of a (unit) least-squares residual r, the largest possible value is $\sqrt{1 - h_{ii}} = \sqrt{(n-p)/n}$. If we maximize the sum of two components, the largest value for the sum of the ith and jth residuals is $\sqrt{2 - h_{ii} - h_{jj} - 2h_{ij}}$, etc.

Example 1 *A simple illustration of these facts can be given by using the* 2^2 *main effects model. If all observations are zero except one, which is equal to* c*, the residuals are equal to* $\pm c/4$*. This residual vector has* L^2*-norm equal to* $c^2/4$ *and* $c = 2$ *normalizes it. The largest possible component of a unit residual is, therefore, equal to* $1/2 = \sqrt{(4-3)/4}$*, which means that the largest percentage of the* L^2*-norm of a residual that resides on a single component is 25%.*

In a 3×3 main effects model, the corresponding number is $2/3 = \sqrt{(9-5)/9}$, which implies that the largest percentage of the L^2-norm residing in a single residuals is equal to 44.4%. These examples illustrate the fact that the ability of a design to show a single outlier by way of a large individual residual depends on the ratio p/n.

The general formula given above can be used to judge the ability of a given design to point out in a single experiment two outliers by two large residuals. It is immediately clear that this capacity is dependent on the positions (on the indices) of the outlying observations, since h_{ij} depends on i and on j.

The picture is maybe clarified, if we pose the question differently. Given a residual vector r, what maximal percentage of its L^2-norm $r^T r$ can be explained by 1, or 2, or 3, etc. components. Let $I = \{i_1, \ldots, i_m\} \subset \{1, \ldots, n\}$ denote a set of m indices. It turns out that the answer to the above query is equal to $1 - \lambda_{\min}(I)$, where $\lambda_{\min}(I)$ denotes the smallest eigenvalue of the minor H_I of the hat matrix determined by the intersection of the rows and columns from I. This is easy to show and we leave it to the reader to check the statement.

Example 2 *When the number* m *of residuals we wish to check is equal to 1, the minor* H_I *is equal to the scalar* $h_{i_1,i_1} = p/n$ *and* $1 - \lambda_{\min}(I) = 1 - p/n$*, which is a result we already knew. If we pass to* $m = 2$ *for the* 2^2 *design, all minors of dimension 2 have a minimal eigenvalue of* $1/2$*. The largest percentage of the total norm explained by two components, i.e., by half the components, is equal to 50%. The situation in the* 3×3 *case is different. The smallest eigenvalue of 2-dimensional minors is either 1/3 or 4/9, depending on the position of the pair within the 3 by 3 table. The maximal percentage of the total norm that can be explained by 2 of the 9 residuals is, therefore, 66.7%. Since with a single component, one can at most explain a percentage of 44.4%, this is a bit disappointing. Evidently, two outliers will result in two residuals that stick out much less than was the case with a single outlier. This kind of behavior is typical for least-squares fits.*

What is the answer to the same question in the case of least-absolute-

deviations residuals e? For values of m smaller than the **exact-fit-point** of the L^1 method, the residuals can be completely concentrated on any of set of m components. The exact-fit-point N_{ef} of an equivariant fitter can be defined as the largest number of non zero obserservations that one can add in any position to the vector of observations $y = 0$ without changing the fit from $\tilde{y} = 0$. In the best of situations, this point is equal to or close to $n/4 - 1$ for the least-absolute-deviations regression method (for details, see Ellis and Morgenthaler, 1992).

A thorough discussion of the break down and oulier resistance problem in designed experiments is given in Müller (1995). If we wish to be able to fit all contrasts in a given model, the crucial quantity is the maximal number of experimental runs that by themselves are not enough to determine a fit of the model. In the 3×3 main effects design this number is equal to six, which is bigger than the five dimensions of the parameter space. Any equivariant fitter breaks down, as soon as a majority of the $3 = 9 - 6$ remaining observations are faulty.

For $m > N_{\text{ef}}$, depending on the position of the outliers, different outcomes are possible.

Example 3 *In a 3×3 design, $N_{\text{ef}} = 1$. For $m = 2$ and $m = 3$, the following tables show some of the possiblities.*

a	b	0
0	0	0
0	0	0

a	0	0
0	b	0
0	0	0

a	0	0
0	b	0
0	0	c

In the left-most case, the residual e has in general two non zero components. They can be in the same position as the outliers a and b – – this happens when they are of opposite sign – – or they can be spread to other positions, one at the third position of the first line, the other indicating the more important outlier among a and b. In the middle case, the non zero residuals are confined to the two positions where a and b are observed as long as they have the same sign. Otherwise, the L^1 fit is not unique and a second and third non zero residual can pop up in the first and the last line. If a is -300 and b is 290, the residual table can be as disparate as the follwing two examples:

-300	0	0
0	290	0
0	0	0

-10	0	290
0	0	0
0	-290	0

Since the vector

1	*0*	*1*
0	−1	*0*
0	−1	*0*

lies in the column space of the design matrix X, one can smoothly transform between these two residual vectors without changing the L^1 norm. In the right-most case, the situation is even more complex. If the outliers a, b and c are of equal sign, the residual matrix faithfully reflects this structure. If they are of unequal signs, surprising things can happen. The observed table

300	*0*	*0*
0	*290*	*0*
0	*0*	−250

leads to a unique L^1 fit with residual table equal to

50	−250	*0*
−250	*40*	*0*
0	*0*	*0*

.

In this case, there exists an additive fit explaining the data with (merely) two sizable residuals. When $m = 3$, we are beyond the range, where we can generally expect to distinguish outliers from additive structure in 3×3 tables. Most robust procedures prefer the fit found by the L^1 method over the fit $\tilde{y} = 0$. But one can, of course, imagine procedures that are able to identify any additive structure as long as it is exactly adhered to by a majority (LMS, Rousseeuw and Leroy, 1987). Such a procedure could not distinguish between the two fits which both have at least 5 of the 9 residuals equal to zero.

3 Maximal residuals

The size of the largest residual will often be taken as an indication, whether faulty runs occured during a designed experiment. As we saw in the last section, when only a small number – less than the exact-fit-point of the runs are faulty, and stick out very clearly, then the L^1 fit will produce residuals that can safely be used to identify the faulty runs. Beyond this number of grossly wrong observations and when the outlyingness is less clear cut, the L^1 method is not successful. We also noted in the last section that the L^1 fit has two drawbacks. Firstly, **it does** in general **not produce a unique answer**. This may at first sight seem not to be a concern, but in the case of designed experiments, multiplicity of possible answers is very

common and the set of L^1 solutions can be very varied. Secondly, the solutions found with the known algorithms produce at least p exact zeros among the residuals. Such least-absolute-deviation fits are typically not very appealing when one analyzes them in detail. The greed for exact zeros tends to make the non zero residuals "too" big.

There are several simple arguments which allow us to estimate the average inflation factor that we should expect when passing from least-squares residuals to least-absolute-deviations residuals. First, the L^1 criterion will be about the same between the two solutions, i.e. $\sum_{i=1}^n |e_i| \leq \sum_{i=1}^n |r_i|$ with rough equality for large values of n/p. If this were so and if we further imagine that the exact zeroes are created by randomly selecting the cells, then the fact that the first sum contains p exact zeros makes $|e_i|$ on average $n/(n-p)$ times larger than $|r_i|$. This is equivalent to imagining that the non-zero L^1 residuals are constructed from a L^2 residual to which $p/(n-p)$ parts of p/n L^2 residuals are added. Both of these arguments over-estimate the inflation factor. A final check can also be made on the level of the variance. Suppose that a random selection of $(n-p)/n$ of the L^2 residuals were multiplied by the above inflation factor, i.e., $n/(n-p)$, whereas the others were put equal to zero. Under such a process, the variance of the non zero L^1 residuals would be equal to $(n/(n-p))^2$ times the variance of the L^2 residuals. If we take the rough equality of the L^2 criterion as a guide, we would expect the variation in e to be the same as the variation in r. Since the components in e contain a mixture of p/n exact zeros with zero variance and $(n-p)/n$ non zeros, the variation of the non zero residuals is expected to be $n/(n-p)$ times bigger than the variation in r. This leads to an inflation factor of $\sqrt{(n-p)/n}$, but this time one underestimates the true size. Both factors tend to 1, as p/n tends to zero and both are true some of the time. Typically, when we have only a few degrees of freedom for the error, then the factor $n/(n-p)$ is correct. This is the case for example, when we fit a 2^k design up to the $(k-1)$-factor interactions. But, it is also roughly true for a $2 \times k$ factorial design, where the number of degrees of freedom for the error is $k/2$ and thus arbitrarily large. The last example shows, that it is also the design itself that has an influence on the behavior of the L^1 method.

In the 3×3 design, the inflation factor for the size of residuals is between $\sqrt{9/4} = 1.5$ and $9/4 = 2.25$. Figure 1 illustrates what really happens for four simple designs. Note that an innocent interpretation of the L^1 residuals e would quite often lead to the conclusion that outlying experimental runs were present, because of the large maximal residual size. The simple minded adjustment given above works reasonably well.

Figure 2 repeats the experiment explained in Figure 1, but this time

with errors from a contaminated Gaussian with $1/n$ contamination having a five-fold standard deviation.

Among the 300 simulated experiments, roughly 60% resulted in the correct identification of the faulty run in the sense that the largest sized residual was indeed associated with the contaminated run. This is true for both methods of fitting. However, in the least-squares case, the largest residual does not stick out clearly when compared to the next largest one.

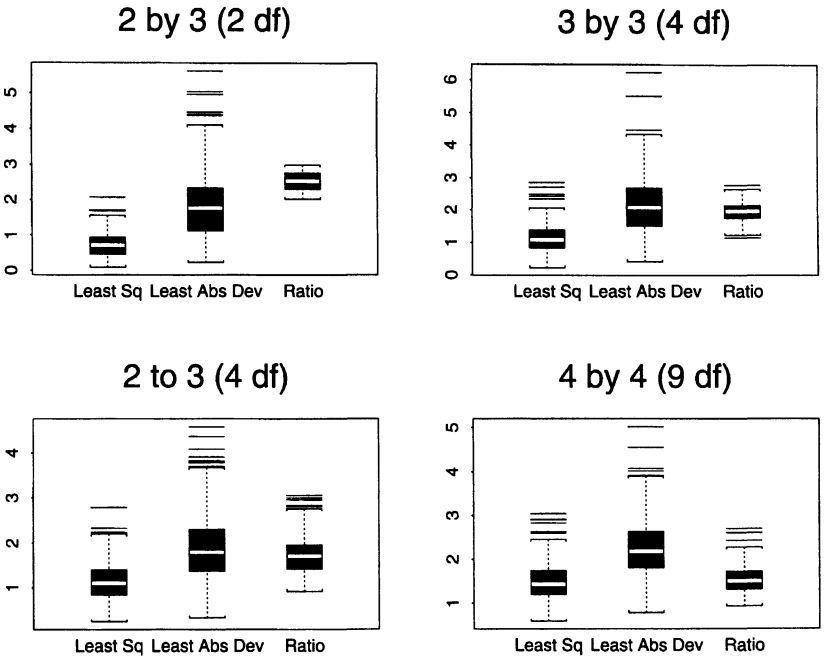

Figure 1: *The figure shows the behavior of the maximal residual size in four different design. The expected inflation factors are (1.73, 3.00) for the* 2×3*, (1.50, 2.25) for the* 3×3*, (1.41, 2.00) for the* 2^3 *main effects and (1.33, 1.78) for the* 4×4*. The average inflation factor of the maximal residual observed in 300 replications are 2.54, 1.93, 1.61 and 1.51. The ratios between the maximal residual sizes computed for each replication is also shown in the plots. These ratios are surprisingly stable.*

Compared to the least-squares residuals, the L^1 residuals do fairly well. But, when challenged by a standard robust estimator, they do worse. If

one uses the simulations shown in Figure 2, but replaces the least-squares residuals by robust residuals based on Tukey's biweight with $6\times$ *MAD*, both produce about an equally large maximal size. However, the ratio of largest to second largest is usually more important for the robust fit, which, therefore, results usually in a clearer picture. This is due to the higher degree of smoothness of such estimators when compared to the L^1 fitter. This also results in an improved relative efficiency.

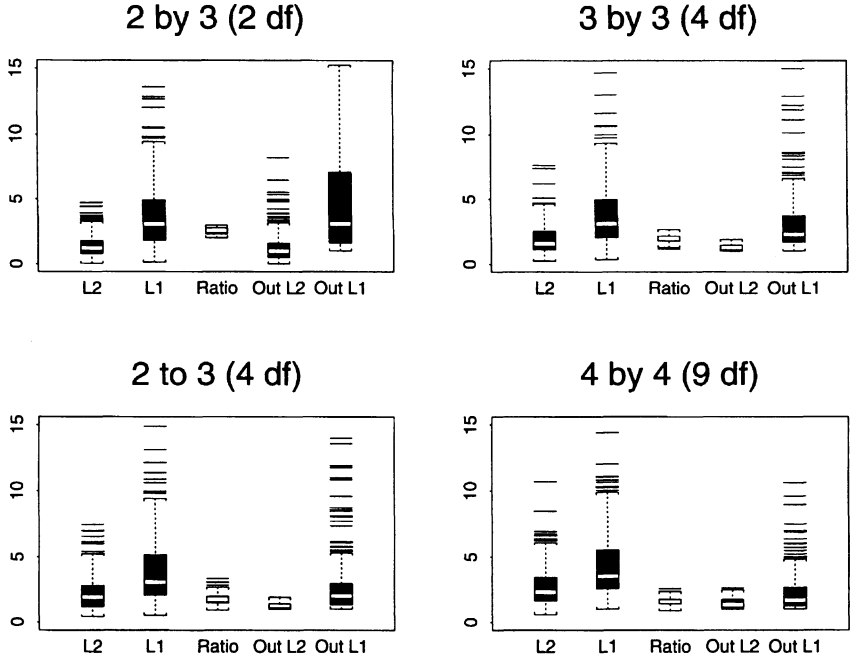

Figure 2: *The figure shows the behavior of the maximal residual size in four different design for an error distribution which is Gaussian in all runs except one. In the exceptional one the variance of the Gaussian error is 25 times larger. The average inflation factor observed in 300 replications are 2.61, 1.97, 1.75 and 1.58 and thus remarkably close to the ones observed for Gaussian data. The similarity with the Gaussian case is a bit surprising, since the L^1 method is supposed to be able to identify more clearly the faulty run. In order to check this, each figure contains a boxplot of the ratio of the largest sized residual to the next largest one – the corresponding boxplots are labelled "Out L1" and "Out L2".*

4 Non uniqueness of the L^1 fit

The ambiguity of the L^1 fit is another problem that the user of this method should be aware of. In the case of the 2×2 main effects model, for example, the common L^1 algorithms will result in a residual e consisting of three exact zeroes and one non-zero residual whose size is exactly four times as large as the size of the L^2 residuals. The run in which the single non zero residual is placed, is completely arbitrary. All intermediate fits preserve the L^1 norm. The same problems occurs in sufficiently balanced replicated 2×2 designs such as the one presented in Sheather and McKean (1992, Example 2, p. 153). In their example, the L^1 solution is not unique. In fact, there is a 1-dimensional family of fits, which contains in the center a solution close to the L^2 fit.

In the 2×3 design, the L^1 fit results in general in 2 non zero residuals, which are placed in two different columns. The placement of the these two residuals is arbitrary to the extent that we can make them switch rows. All intermediate solutions preserve the L^1 norm. In the 3×3 design, things become more complicated. In general, there are four non zero residuals. The values and placements of the non zero cells are usually not unambiguously determined. If the pattern of the non zero cells – indicated by $*$ – is of the form

$*$	$*$	0
$*$	$*$	0
0	0	0

,

then the fit is unique. If we have, however, a pattern of the form

$*$	0	0
0	$*$	$*$
0	$*$	0

,

there is in general a 2-dimensional set of L^1 solutions.

If we want to recommend the use of L^1 fitting, it seems important to me to produce an algorithm which enumerates all extremal points of the polygone of L^1 fits instead of simply picking one, somewhat at random. This will allow the user to judge, inhowfar the criterion is really identifying outlying points or whether it simply produces large residuals by artificially zeroing others.

5 Estimating error variation

The undesirable features of the L^1 residuals that we have discussed above will, of course, have an effect on their ability to predict the error variability.

The p exact zeroes among the residuals are a property of the design, the fitter and the algorithm. They contain no information about the error distribution. It is, therefore, quite natural to compute the variance of the non zero L^1 residuals as an indication of the variance σ^2 of the errors. Consequently, consider an L^1 solution with at least p zero residuals and at most $n-p$ non zero residuals. It is evident that the sum of squares of the non zero residuals over-estimates the error variation and the question is by how much. To answer this question, suppose we tried to reconstruct the L^2 residuals r on the basis of the L^1 residuals e. If $p = n-1$, then e would contain a single non-zero residual, which we would evenly distribute over the n $|r_i|$. These reconstructed L^2 residuals would then have a sum of squares and, since there is a single degree of freedom, a mean square equal to $\sum_{i=1}^n e_i^2/n$. Now, suppose we have several non zero L^1 residuals, which are all of equal size $|e_i| = R$. Evenly distributing them over all n observations, leads to a size of $|r_i| = R(n-p)/n$. The mean square of these reconstructed r_i is equal to $(n-p)/n^2 \sum_{i=1}^n R^2 = \sum_{i=1}^n e_i^2/n$, i.e., the same formula as before. One proposal for estimating the error standard deviation from an L^1 fit would, therefore, be the following:

$$u = \sqrt{\sum_{i=1}^n e_i^2/n}\,.$$

Because of the non-uniqueness, $\sum_{i=1}^n |e_i|$, which is uniquely determined, is in some sense a more appropriate basis for estimating σ, the standard deviation of the error. In this case, the over-size of the L^1 residuals does not play any role either. If we re-size them and in some way reconstruct pseudo-L^2 residuals, their sum of absolute values would remain the same. How would one estimate σ based on a set of L^2 residuals r? Since the marginal distribution of each r_i has expectation zero and variance $\sigma^2(1-h_{ii}) = \sigma^2(n-p)/n$, we have – for Gaussian errors – $E(|r_i|) = \sqrt{2/\pi}\sqrt{(n-p)/n}\sigma$. The statistic

$$\frac{1}{n}\sum_{i=1}^n |r_i| \frac{\pi}{2}\sqrt{n/(n-p)} = \frac{\pi/2}{\sqrt{n(n-p)}} \sum_{i=1}^n |r_i|$$

is, therefore, an unbiased estimate of σ. In replacing $\sum_{i=1}^n |r_i|$ by $\sum_{i=1}^n |e_i|$, we obtain an estimate that underestimates σ, but should still be a useful indicator:

$$v = \frac{\pi/2}{\sqrt{n(n-p)}} \sum_{i=1}^n |e_i|\,.$$

Figure 3 shows with various plots, how the two estimates behave for Gaussian errors.

A similar behavior is found for other error distributions with a finite second moment.

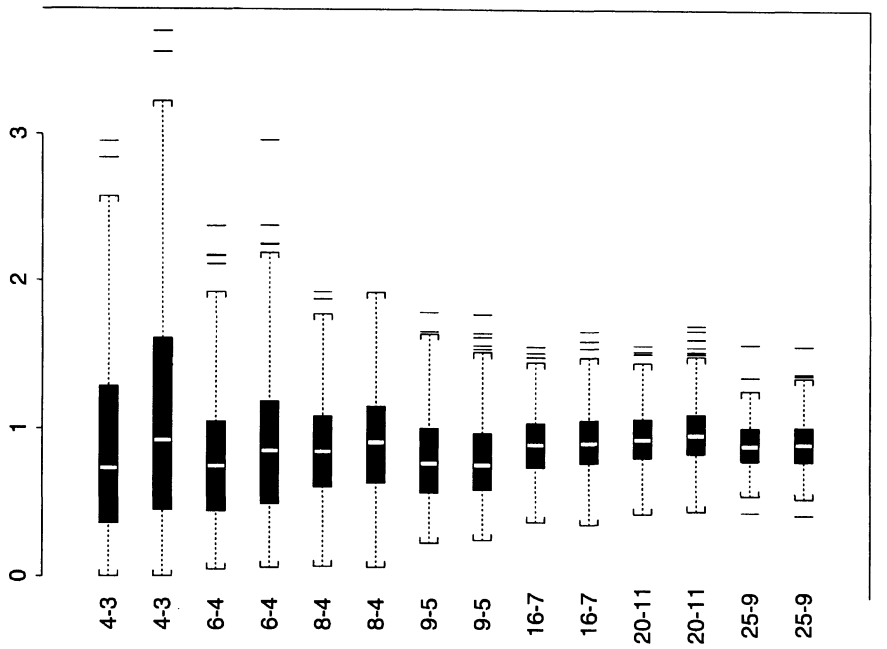

Figure 3: *The boxplots show the behavior of the two estimates u and v for various balanced designs and Gaussian errors. The labels indicate for each boxplot the number of observations n and the dimension of the parameter space p. Each design is represented with two boxplots, the first one for u, the second one for v. The true value of σ is equal to 1. The average values over 300 replications are: 0.811, 0.778, 0.843, 0.819, 0.877, 0.968, 0.885 for u and 1.02, 0.853, 0.888, 0.803, 0.897, 0.984, 0.887 for v. Both estimates tend to underestimate σ.*

6 Conclusions

The L^1 method has several drawbacks and in particular leads to some undesirable features built into the residuals. In my opinion it is not a suitable method for fitting of ANOVA data the following reasons:

(1) The computation of the L^1 fit is not as easy as the computation of the

L^2 fit. It is comparable in difficulty to robust fits.

(2) The L^1 residuals have some idiosyncrasies that should be known to the user of this method. Ignorance will lead to wrong interpretations. They cannot in a straighforward manner replace L^2 residuals.

(3) The non-uniqueness of the L^1 fit in ANOVA problems is the rule rather than the exception. We lack easily available algorithms which exhibit the whole solution set.

(4) The resistance of the L^1 fit to outliers is not good enough. One can do better with competing methods that are computationally about equivalent.

References

[1] Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). *Regression Diagnostics.* New York: Wiley.

[2] Cook, R.D. and Weisberg, S. (1982). *Residuals and Influence in Regression.* London: Chapman and Hall.

[3] Ellis, S.P. and Morgenthaler, S. (1992). Leverage and breakdown in L_1 regression. *J. Am. Statist. Assoc.* **87**, 143–148.

[4] Müller, Ch. (1995). Outlier robust inference for planned experiments. Habil. Schrift, Fachbereich Mathematik, Freie Universität Berlin.

[5] Rousseeuw, P.J. and Leroy, A.M. (1987). *Robust Regression & Outlier Detection.* New York: Wiley.

[6] Sheather, S.J. and McKean J.W. (1992). The interpretation of residuals based on L_1 estimation. In *L_1-Statistical Analysis and Related Methods*, Ed. Y. Dodge. Amsterdam: North-Holland.