The Conditional Breakdown Properties of Least Absolute Value Local Polynomial Estimators

Avi Giloni
Sy Syms School of Business
Yeshiva University
500 West 185th Street
New York, NY 10033

Jeffrey S. Simonoff
Leonard N. Stern School of Business
New York University
44 West 4th Street
New York, NY 10012

December 7, 2003

Abstract

Nonparametric regression techniques provide an effective way of identifying and examining structure in regression data. The standard approaches to nonparametric regression, such as local polynomial and smoothing spline estimators, are sensitive to unusual observations, and alternatives designed to be resistant to such observations have been proposed as a solution. Unfortunately, there has been little examination of the resistance properties of these proposed estimators. In this paper we examine the breakdown properties of local polynomial estimation based on least absolute values, rather than least squares. We show that the breakdown at any evaluation point depends on the observed distribution of observations and the kernel weight function used, and make recommendations regarding choice of kernel based on two different breakdown measures. The results suggest that the breakdown point at an evaluation point provides a useful summary of the resistance of the regression estimator to unusual observations.

Key words: Nonparametric regression; robust estimation.

1 Introduction

Nonparametric regression techniques have been shown in recent years to be very effective at identifying and estimating structure in regression data, without requiring restrictive assumptions on the form of the relationship between the target and predicting variables. Many different approaches to this problem have been suggested; see Simonoff (1996), chapter 5, for discussion of many of the possibilities. In this paper we focus on local polynomial estimation based on a single predictor variable. Let $\{x_i, y_i\}, i = 1, \ldots, n$, be the data set at hand. The underlying model assumed for these data is

$$y_i = \mu(x_i) + \varepsilon_i,$$

with ε_i having zero median and $V(\varepsilon|X = x) = \sigma^2(x)$ not necessarily constant. The goal is to estimate $\mu(x)$, the conditional median of y given x (if the errors are symmetric this also corresponds to the conditional mean, assuming the mean exists).

Local polynomial estimation proceeds by fitting a polynomial locally over a small neighborhood centered at any evaluation point x, based on weighted least squares. The estimator $\hat{\mu}(x)$ is then the intercept term $\hat{\beta}_0$ from the weighted least squares regression. The bandwidth h controls the amount of smoothness of $\hat{\mu}(x)$, and can be fixed for all values of x, or locally varied (based on nearest neighbor distance, for example) to allow different levels of smoothing at different locations. Kernel regression corresponds to $p = 0$, and is known to have inferior performance compared to taking $p \geq 1$ (in terms of bias in the boundary region, for example). Assuming a given amount of smoothness of $\mu(\cdot)$, it can be shown that certain local polynomial estimators, combined with appropriate choice of h, can achieve the best possible asymptotic rate of convergence of the estimator to the true curve $\mu(\cdot)$.

As is the case for any estimator based on least squares, local polynomial estimation is susceptible to the effects of observations with unusual response values (outliers). If an observed y_i is sufficiently far from the bulk of observed responses for nearby values of x, $\hat{\mu}(x)$ will be drawn towards the unusual response and away from the majority of the points. This has led to the proposal of the use of other criteria to fit local polynomials that downweight these observations with large residuals. These include *Lowess* (Cleveland, 1979), its successor *loess* (Cleveland and Devlin, 1988), and local versions of M–estimators (Tsybakov, 1986; Fan, Hu, and Truong, 1994; Welsh, 1994). In order to determine these estimators, an iterative process is utilized that typically begins with a least squares based initial estimate (we are not aware of any implementation of such estimators that is not based on a least squares initial estimate). However, since the original residuals are based on a least squares criterion, these robust alternatives still can be sensitive to outliers.

Figure 1 illustrates the problem. The data are from a radioimmunoassay calibration study, and relate counts of radioactivity to the concentration of the dosage of the hormone TSH, in micro units per ml of incubator mixture (Tiede and Pagano, 1979). There is a roughly hyperbolic relationship between counts and concentration, with one clear outlier at $(20, 4478)$. Figure 1 gives both nonrobust and robust loess estimates (based on a local linear model) for these data, based on a nearest neighbor bandwidth covering 65% of the data. As can be seen, both loess estimates are affected by the outlier. While the nonrobust estimate (solid line) is drawn towards the outlier, the robust estimate (dotted line) is driven away from it, resulting in a spurious dip below the bulk of the points. This dip is not a function of choice of the bandwidth, as bandwidths from the smallest possible value

(36% of the data) to one leading to clear oversmoothing (90% of the data) all yield estimates exhibiting it.

True robustness requires an estimator that is not based (even indirectly) on a least squares criterion. Wang and Scott (1994) investigated the least absolute values (LAV) version of local polynomial estimation. They showed that the estimator is the solution to a linear program, and derived asymptotic theory under specific conditions. See also Chaudhuri (1991) and Yu and Jones (1998), who looked at the general case of estimating regression quantiles (LAV corresponds to the median). We note that although LAV regression is an M–estimator, since it can be solved via linear programming, it does not require an iterative process as is the case for ordinary M–estimators.

A characteristic of all of this work is that while the asymptotic properties of the methods have been investigated, the robustness properties have not. Thus, while a primary justification of these methods is their supposed resistance to unusual observations, there are no results that actually quantify this resistance. In this paper one measure of resistance, the breakdown, is used to evaluate robustness. The breakdown of an estimator is the smallest fraction of outliers that can force the estimator to arbitrary values, and is thus a measure of the resistance of the estimator to unusual values. More specifically, the breakdown point of an estimator τ is defined to be the maximum bias that can be caused by replacing any m of the original data points by arbitrary values (Donoho and Huber, 1983). An estimator that is not at all resistant to outliers, such as one based on least squares, thus has breakdown $\frac{1}{n}$. In this paper we propose and investigate a locally varying (conditional on the evaluation point) notion of breakdown that is appropriate for local polynomial estimation. By adapting breakdown results from linear least absolute values regression estimation, we derive the robustness properties of local LAV polynomial estimators. In the next section we propose and discuss the derivation of the breakdown values. Section 3 provides specific examples of conditional breakdown. We investigate its dependence on the local distribution of predictor values and the choice of kernel, and reanalyze the data of Figure 1. Section 4 concludes the paper with discussion of potential future work.

2 Determining the Conditional Breakdown

Since the local polynomial regression estimate $\hat{\mu}(\cdot)$ is implemented by solving many local regression problems, each centered at an evaluation point x, its breakdown properties are defined on a local level as well. We restrict ourselves to kernel functions $K(\cdot)$ that are positive on a bounded interval (typically $[-1, 1]$). When we refer to the conditional breakdown, we are reflecting that, unlike for parametric models, the breakdown value changes depending on the evaluation point x. Several key points illuminate how the notion of conditional breakdown at a point x can be defined.

The first point to recognize is that since the local polynomial estimate is based on a weighted regression, the breakdown of $\hat{\mu}(x)$ is simply the breakdown of a weighted version of the linear regression method being used, whether that is least squares or least absolute values estimation.

We also must recognize that if the evaluation point becomes unbounded (i.e., $|x| \to \infty$), there is no sensible way to define breakdown (or any robustness properties) in the neighborhood of that x. The reason for this is that, unlike in the case of a parametric function μ, it isn't meaningful to talk about the "true" $\mu(x)$ when $x \to \pm\infty$, since μ is only defined by local smoothness ($\mu(\infty)$ is not well–defined). For this reason, we will only treat breakdown at an evaluation point x for bounded x.

Consider now the use of a bandwidth h that is not a function of the local design (a constant bandwidth is an obvious example of this, but h also can vary in ways that do not depend on the observations x_i). In this case, contamination in the predictor variable is no longer relevant, since any value of x_i that goes to $\pm\infty$ eventually has zero weight in the local regression; that is, only observations local to x can have an effect on $\hat{\mu}(x)$. We thus can describe robustness and breakdown in this case by considering the finite sample breakdown point of some regression estimator τ with contamination restricted to the dependent variable.

The situation when using a bandwidth that varies as a function of the design is more complicated. Consider the most common bandwidth choice of this type, the nearest neighbor bandwidth chosen at x to yield a fixed proportion s of observations with nonzero weights (the closest observations to x). If $1 - s$ is greater than the proportion of observations with $|x_i| \to \infty$, then once again contamination in the predictor variable is not relevant, since eventually these x_i's will no longer be in the neighborhood of x and will have zero weight. On the other hand, if $1 - s$ is less than or equal to the proportion of observations with predictor contamination, at least one contaminated observation will have nonzero weight. In this case we can appeal to known breakdown results for LAV regression when there is contamination in the predictor. That is, the breakdown at x of local LAV regression is $\frac{1}{n}$ (the smallest possible value, indicating no robustness). For these reasons, throughout the rest of this paper we refer to the finite sample breakdown point with contamination restricted to the dependent variable simply as the finite sample breakdown point.

The breakdown properties of local LAV–regression estimators can be determined by determining the breakdown properties of each subproblem, which is a weighted LAV regression problem involving the observations for which the weights are positive. The finite sample breakdown point of LAV regression (with contamination restricted to the dependent variable) has been studied previously. Note that in the case of local LAV regression (as opposed to the traditional LAV regression), we are only concerned with the intercept term, i.e., $\hat{\beta}_0$, but this is not an important distinction, since the finite sample breakdown point of $\hat{\beta}_0$ of (weighted) LAV regression is the same as the finite sample breakdown point of (weighted) LAV regression for all of the parameters.

The weighted LAV regression problem with positive, finite weights w_i can be formulated and solved as a linear program with an objective function consisting of the sum of the absolute weighted residuals. Equivalently, the objective function can be taken to be the same as in the case of unweighted LAV regression, changing the data by setting $\widetilde{y}_i = w_i y_i$ and setting the ith set of predictor values to $\widetilde{\mathbf{x}}_i' = (w_i \quad w_i x_i)$. Thus, to calculate the finite sample breakdown of weighted LAV regression with design matrix $\mathbf{X}$ one just needs to calculate the finite sample breakdown of LAV regression with design matrix $\widetilde{\mathbf{X}}$.

We utilize the approaches of Giloni and Padberg (2003), Giloni, Sengupta, and Simonoff (2003), and Mizera and Müller (2001) to calculate this finite sample breakdown. Giloni and Padberg (2003) show that the finite sample breakdown of LAV regression can be solved by mixed integer programming techniques. Giloni, Sengupta, and Simonoff (2003) and Mizera and Müller (2001) provide an algorithm for calculating the finite sample breakdown point of LAV regression that is very efficient when the number of predictor variables is small. We use this methodology to calculate the finite sample breakdown locally for local LAV regression in the next section.

3 Local breakdown and its relationship to kernel choice

In this section we investigate more closely the conditional breakdown properties of the local LAV linear estimators. Since the breakdown is based on a set of weighted LAV regressions, it depends at any evaluation point on both the local distribution of predictor values and the kernel used. While the local distribution of predictors is typically beyond the control of the data analyst, the choice of kernel is not, leaving open the possibility that it might be chosen in such a way as to make the estimator as robust as possible.

The properties of the local LAV linear estimator at an evaluation point depend on the bandwidth used, as that determines the set of observations within the local regression. This suggests that the bandwidth could be chosen so as to maximize robustness (in some sense), but this is a mistaken conclusion. Wang and Scott (1994) derived the bandwidth that minimized the asymptotic average mean squared error of $\hat{\mu}$, showing that it satisfies

$$h_{opt} = \left(\frac{36}{f(0)^2 \int_0^1 \mu''(x)^2 dx} \right)^{1/5} n^{-1/5}, \tag{1}$$

where f is the density of the errors (taking x to be uniform on $[0,1]$ and assuming constant variance for the errors). Thus, the optimal choice of h depends on the curvature of μ and the density of ε, and cannot be set arbitrarily so as to ensure robustness.

Equation (1) assumes use of a uniform kernel, so if a different kernel is used, the bandwidth must be adjusted. Wang and Scott (1994) showed that the equivalent bandwidth when using a kernel K_2 rather than one K_1 satisfies

$$h_{opt}(K_2) = h_{opt}(K_1)[V(K_2)/V(K_1)]^{1/2},$$

where $V(K)$ is the variance of the kernel, $\int x^2K(x)dx$. Table 1 lists the kernels we examine, which include most of the ones used in practice. The interpretation of the table is that, for example, if the bandwidth h yields an appropriate amount of smoothing when using a uniform kernel, the bandwidth $1.291h$ is the appropriate choice when using a quadratic kernel. Thus, any comparisons of robustness across kernels corrects for this scale effect by using equivalent bandwidths. Although (1) is based on a uniform design, the nonuniform case is similar, in that the design only appears as a constant multiplier for the bandwidth, and does not depend on the kernel (Yu and Jones, 1998). Thus, the multipliers in Table 1 are appropriate for any design.

We evaluate the robustness of a particular kernel choice at any evaluation point in two ways. First, we use the breakdown value, the smallest number of observations that can force the estimator to arbitrary values. Note that when comparing kernels we do not wish to use the breakdown point (the proportion of observations in the span of the kernel that can force the estimator to arbitrary values) because the number of observations in the span depends on the appropriate multiplier for the bandwidth for the chosen kernel. Say, for example, that the bandwidth used at evaluation point x using a uniform kernel includes $n_u(x)$ observations, with breakdown point $\alpha_u(x)$. Then, the smallest number of observations that could possibly break down the estimate at x using the uniform kernel is $\lceil n_u(x)\alpha_u(x)\rceil$, where $\lceil\cdot\rceil$ represents the smallest integer greater than or equal to the value. On the other hand, if a quadratic kernel was used, the bandwidth would be 29.1% larger at x, yielding $n_q(x)$ observations in the span of the kernel, with $n_q(x)$ probably larger than $n_u(x)$. The smallest number of observations that could possibly break down the estimate at x using the quadratic kernel is $\lceil n_q(x)\alpha_q(x)\rceil$, where $\alpha_q(x)$ is the breakdown point at x when using the quadratic kernel. The choice of kernel is up to the data analyst, so the preferred choice on the basis of breakdown would be the one with larger value of $n(x)\alpha(x)$ (the breakdown value), not larger $\alpha(x)$ (the breakdown point). This argument also shows why breakdown value is not sufficient to describe resistance in the nonparametric regression context. Since the breakdown value is an increasing function of the number of observations in the span of the kernel, kernels with larger equivalent bandwidths (such as the triweight) have an advantage over kernels with smaller equivalent bandwidths (such as the uniform) in terms of breakdown value.

For this reason, we examine a second measure of breakdown. For a given kernel, say there are $n(x)$ observations in the span of the kernel at evaluation point x, and the breakdown value at that point is $b(x)$. The estimator cannot break down at x if the number of outliers within the span of the kernel is less than $b(x)$, so the probability that the estimator will not break down at x is

$$P(\text{Estimator cannot break down at } x) = \sum_{j=0}^{b(x)-1} P(j \text{ of the observations in the span are outliers}).$$

Say there are k outliers in the sample, and they are spread randomly over the observations in the sample. Then the probability that j of the observations in the span of the kernel are outliers is hypergeometric,

$$P(j \text{ of the observations in the span are outliers}) = \frac{\binom{k}{j}\binom{n-k}{n(x)-j}}{\binom{n}{n(x)}},$$

with $0 \leq j \leq k$. Note that if $k < b(x)$, the estimator cannot possibly break down at x, but as k gets larger, the probability of having too many outliers in the span of the kernel increases, decreasing the probability that the estimator cannot break down. Note also that a smaller bandwidth makes it more likely that the estimator cannot break down, since there are fewer observations in the span of the kernel, implying an advantage for kernels with smaller equivalent bandwidths. Thus, these two measures quantify a tradeoff between choosing kernels using smaller bandwidths and those using larger bandwidths.

We will focus here on nearest neighbor–type bandwidths, rather than fixed–width bandwidths. The reason for this is that fixed–width bandwidths add the complication of including different numbers of observations within the span of the kernel for different evaluation points if the predictor variable design is not uniform (and even if it is uniform at the boundaries). In what follows the ith predictor value satisfies $x_i = G^{-1}[i/(n+1)]$, where $G(\cdot)$ is either the uniform [0, 1], standard Gaussian, or exponential (with mean one) cumulative distribution function (that is, the design density is consistent with either a uniform, Gaussian, or exponential pattern, yielding what might be considered typical design patterns), with $n = 100$. The breakdown measures are determined at an equally–spaced grid of 1000 values.

Figure 2 gives breakdown values for the different kernels for a uniformly distributed design. The bandwidth is taken so that 20% of the observations are covered by the uniform kernel (recall that for other kernels the equivalent number based on Table 1 is used). In this figure, and all following ones, the uniform kernel is represented by a solid line, the quadratic kernel by a dotted line, and the triweight kernel by a dashed line. We have omitted the biweight and tricube kernels from this figure (and most of the following figures) to make them clearer; generally speaking, the properties of these two kernels are similar to those of the triweight kernel. The uniform kernel consistently has the poorest breakdown. The other kernels have similar breakdown values, with the biweight and triweight alternating between values of 7 and 8 outliers for most evaluation points.

It is difficult to separate the curves for the different kernels in the figure, so Table 2 gives the average breakdown values (averaged over the 1000 evaluation points) for each of the kernels. It is clear that the uniform kernel is a decidedly inferior choice in this case, while the breakdown properties of the other kernels are similar.

Not surprisingly, breakdown values are more dependent on the evaluation point when the design is not uniform. Even though the estimate is based on the same number of observations at each evaluation point (since it is uses a nearest–neighbor bandwidth), breakdown is higher in the region where observations are denser. This is related to the connection between breakdown and leverage for the LAV regression estimator. The breakdown point of the estimator drops in the presence of leverage points. Toward the edges of the design, the observations fall asymmetrically, making the ones farthest towards the edges leverage points in the local regressions. If an outlier falls at one of those locations, it is more likely to break down the estimate.

Despite this difference from the uniform design situation, the general breakdown patterns are similar (Figure 3 and Table 2). Once again the uniform is a distinctly inferior choice, while the differences between the other kernels are relatively small. The breakdown values are generally lower than those for the uniform design, in keeping with the effects of leverage noted earlier. The situation for an exponential design (Figure 4 and Table 2) is consistent with the findings for uniform and Gaussian designs. The breakdown is highest in the densest observation region, and the uniform kernel is decidedly inferior to the other kernels.

These results would seem to imply that any kernel other than the uniform kernel is a reasonable choice, with the triweight kernel (slightly) better than the others, but this ignores the preference the breakdown value measure gives to using a larger bandwidth discussed earlier. Figures 5–8 summarize the results of analysis based on the probability that no breakdown can occur (which gives preference to using smaller bandwidths). Figure 5 refers to the uniform design case. When there are only 5 outliers, the probability of no breakdown is virtually 1, since this is below the breakdown value. Differences between the kernels become evident when there are 15 or 25 outliers in the data. The uniform kernel is a relatively strong performer now, but the quadratic kernel is best (this is clearest for 25 outliers). The triweight kernel, which had the highest breakdown values, has the lowest probability of not breaking down as the number of outliers increases. These patterns can be seen more clearly in the top plot of Figure 6. These curves give the values of the probability of no breakdown for each of 1 to 40 outliers, averaged over all of the evaluation points. Under the uniform design, the quadratic kernel is clearly best, with the uniform kernel following behind.

Figure 7 gives the probability of no breakdown for the Gaussian design. In this case the probability of no breakdown decreases markedly near the edges for some kernels even when there are only 5 outliers. The uniform kernel is particularly strong near the edges, with the quadratic kernel best in the middle of the design region. This translates into overall strong performance of these two kernels when averaging over all design points (the middle plot of Figure 6). These patterns carry over to the exponential design (Figure 8 and the bottom plot of Figure 6). Only the uniform kernel keeps the probability of no breakdown relatively high over the entire design, but the quadratic kernel is either the best or second–best choice over the entire region.

The properties of the two breakdown measures together imply a reasonable approach to kernel choice. Since all of the kernels other than the uniform had similar breakdown values, and the uniform and quadratic kernel have the highest probabilities of not breaking down, the quadratic (Epanechnikov) kernel is the best choice from a robustness point of view when fitting an LAV local linear regression. This provides a nice counterpoint to the well–known optimality (in terms of mean squared error) of this kernel for least squares local polynomial estimation, but now based on a breakdown argument for a robust estimator.

We briefly investigate the properties of local LAV linear estimation for a much larger sample size ($n = 1000$). The top plot of Figure 9 gives breakdown values by evaluation point for a uniform design for the three different kernels, again based on a nearest neighbor bandwidth that covers 20% of the observations for the uniform kernel. The patterns are similar to those for $n = 100$, except that the differences between kernels are more pronounced. The triweight kernel comes through as a particularly strong performer, with a breakdown value close to 90 in the center of the data range, with the quadratic and uniform kernels trailing behind. Thus, for the large sample, the advantage in terms of breakdown of a larger equivalent bandwidth is more noticeably more pronounced than for the smaller sample.

As expected, however, this larger bandwidth has a detrimental effect on the probability of not breaking down. The middle plot in Figure 9 gives the probability of no breakdown by evaluation point when there are 260 outliers. The quadratic kernel is a much stronger performer here, having a consistently high probability over the entire data range. The triweight kernel does surprisingly well in the center of the data range, but does poorly near the edges. The usefulness of the quadratic kernel is reinforced in the bottom plot of Figure 9, which averages the probability of no breakdown over all evaluation points for up to 400 outliers. The larger-bandwidth triweight kernel has a clear advantage over the uniform kernel up to roughly

250 outliers, where the uniform becomes noticeably better. The quadratic kernel, on the other hand, is competitive over the entire range of outliers, and thus provides a good choice when the number of outliers is unknown.

Figure 10 examines again the regression relationship for the calibration data. The dotted line in the figure is the local LAV linear estimate based on a 55% nearest neighbor bandwidth and quadratic kernel. The 55% span is roughly equivalent to the 65% span used in loess (which is based on the tricube kernel). The results are very similar for spans between 50% and 75%. The robustness of the estimate is obvious, as it is unaffected by the outlier.

The roughness of the estimate is worth further comment. Local LAV estimates are inherently “jumpy,” but this property is pronounced in this case, because there are only seven distinct predictor design points. In situations where there are more distinct data points the roughness of the estimate is much less noticeable; see, for example, Figure 4 of Wang and Scott (1994) and Figures 3 and 5a of Yu and Jones (1998). A simple correction for the jumpiness of the estimate is to input the estimated regression curve into an ordinary local least squares estimate, thereby smoothing it out. An example of this is given as the solid line in Figure 10. This is a local linear (least squares) estimate derived from the local LAV estimate. The estimate preserves the robustness of the underlying LAV estimate, while exhibiting an intuitively appealing smooth form. Yu and Jones (1998) also noted the benefits of post-smoothing the local LAV estimator, and proposed a “double kernel” method to do this.

4 Conclusion

In this paper we have discussed and examined the robustness properties of local linear estimation based on least absolute values. We have found that the quadratic (Epanechnikov) kernel is a good choice for this estimator, as it provides strong protection in terms of both high breakdown value and high probability of avoiding breakdown for different predictor distributions, particularly for smaller sample sizes. In contrast to other proposed estimators for robust nonparametric regression, local least absolute values polynomial regression has both verifiable robustness properties and known asymptotic convergence properties.

Practical application of these methods requires guidance on bandwidth choice. Wang and Scott (1994) proposed using a robust version of cross–validation for this, and Yu and Jones (1998) suggested modifying a plug–in least squares–based bandwidth. Considering the strong performance of a corrected version of AIC for bandwidth selection found in Hurvich, Simonoff, and Tsai (1998) for nonparametric regression based on least squares, adaptation of the corresponding criterion for LAV regression in Hurvich and Tsai (1990) to the nonparametric regression context seems an interesting potential choice.

An appealing possibility for improving the breakdown of local polynomial estimators would seem to be the use of a more robust criterion function than least absolute values, such as least median of squares or least trimmed squares (Rousseeuw, 1984). While this is relatively straightforward to implement, its properties are very unclear. In particular, unless the asymptotic squared error properties can be derived, it is not possible to compare kernels, since the notion of equivalent bandwidths is not available.

We have restricted ourselves to univariate nonparametric regression in this paper, but many problems involve multiple predictors. The robustness and estimation properties of the local polynomial LAV estimators in that context, including application to additive models (Hastie and Tibshirani, 1990), is an important problem, since outliers are as problematic in this case as in univariate regression. Finally, the theoretical properties of post–estimation smoothing (to reduce the jumpiness in the estimate) are an open, and interesting, question.

Acknowledgments

We thank Ferdie Wang for helpful discussion of this material. The comments of two anonymous referees helped improve the presentation.

References

Chaudhuri, P. (1991), "Nonparametric Estimates of Regression Quantiles and Their Local Bahadur Representation," *Annals of Statistics*, **19**, 760–777.

Cleveland, W.S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, **74**, 829–836.

Cleveland, W.S. and Devlin, S.J. (1988), "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting," *Journal of the American Statistical Association*, **83**, 596–61 0.

Donoho, D.L., and Huber P.J. (1983), "The Notion of Breakdown Point," in *A Festschrift for Erich Lehmann*, eds. P. Bickel, K. Doksum, and J.L. Hodges, Wadsworth, Belmont, CA, 157–184.

Fan, J., Hu, T.–C., and Truong, Y.K. (1994), "Robust Non–Parametric Function Estimation," *Scandinavian Journal of Statistics*, **21**, 433–446.

Giloni, A., and Padberg, M. (2003), "The Finite Sample Breakdown Point for ℓ_1–Regression," to appear in *SIAM Journal on Optimization*.

Giloni, A., Sengupta, B., and Simonoff, J.S. (2003), "Mathematical Programming Methods for Improving Robustness of Regression Estimators," working paper.

Hastie, T.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*, Chapman and Hall, London.

Hurvich, C.M., Simonoff, J.S., and Tsai, C.–L. (1998), "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," *Journal of the Royal Statistical Society, Ser. B*, **60**, 271–293.

Hurvich, C.M. and Tsai, C.–L. (1990), "Model Selection for Least Absolute Deviations Regression in Small Samples," *Statistics and Probability Letters*, **9**, 259–265.

Mizera I. and Müller C.H. (2001), "The Influence of the Design on the Breakdown Point of ℓ_1-type M-estimators," in: A. Atkinson, P. Hackl, and W. Müller, eds., *MODA6 — Advances in Model-Oriented Design and Analysis*, Physica-Verlag, Heidelberg, Germany, 193–200.

Rousseeuw, P.J. (1984), "Least Median of Squares Regression," *Journal of the American Statistical Association*, **79**, 871–880.

Simonoff, J.S. (1996), *Smoothing Methods in Statistics*, Springer–Verlag, New York.

Tiede, J.J. and Pagano, M. (1979), "The Application of Robust Calibration to Radioimmunoassay," *Biometrics*, **35**, 567–574.

Tsybakov, A.B. (1986), "Robust Reconstruction of Functions by the Local–approximation method," *Problems of Information Transmission*, **22**, 133–146.

Wang, F.T. and Scott, D.W. (1994), "The L_1 Method for Robust Nonparametric Regression," *Journal of the American Statistical Association*, **89**, 65–76.

Welsh, A.H. (1994), "Robust Estimation of Smooth Regression and Spread Functions and Their Derivatives," *Statistica Sinica*, **6**, 347–366.

Yu, K. and Jones, M.C. (1998), "Local Linear Quantile Regression," *Journal of the American Statistical Association*, **93**, 228–237.

Kernel	*Formula*	*Variance*	*Multiplier*
Uniform	$\frac{1}{2}$	$\frac{1}{3}$	1.000
Quadratic	$\frac{3}{4}(1-x^2)$	$\frac{1}{5}$	1.291
Biweight	$\frac{15}{16}(1-x^2)^2$	$\frac{1}{7}$	1.528
Triweight	$\frac{35}{32}(1-x^2)^3$	$\frac{1}{9}$	1.732
Tricube	$\frac{70}{81}(1-\lvert x\rvert^3)^3$	.1440329	1.521

Table 1: Multipliers to give equivalent bandwidths for different kernels.

	Design		
Kernel	Uniform	Gaussian	Exponential
Uniform	5.00	4.58	4.43
Quadratic	6.75	5.63	5.29
Biweight	6.99	5.70	5.53
Triweight	7.06	5.72	5.66
Tricube	6.80	5.59	5.40

Table 2: Average breakdown values over 1000 equally-spaced evaluation points for three different designs. The sample size $n = 100$, and the bandwidth is chosen so that it covers 20 observations for the uniform kernel.

Figure 1: Loess estimates for calibration data. The solid line is the ordinary (nonrobust) version of the estimate, while the dotted line is the robust version.

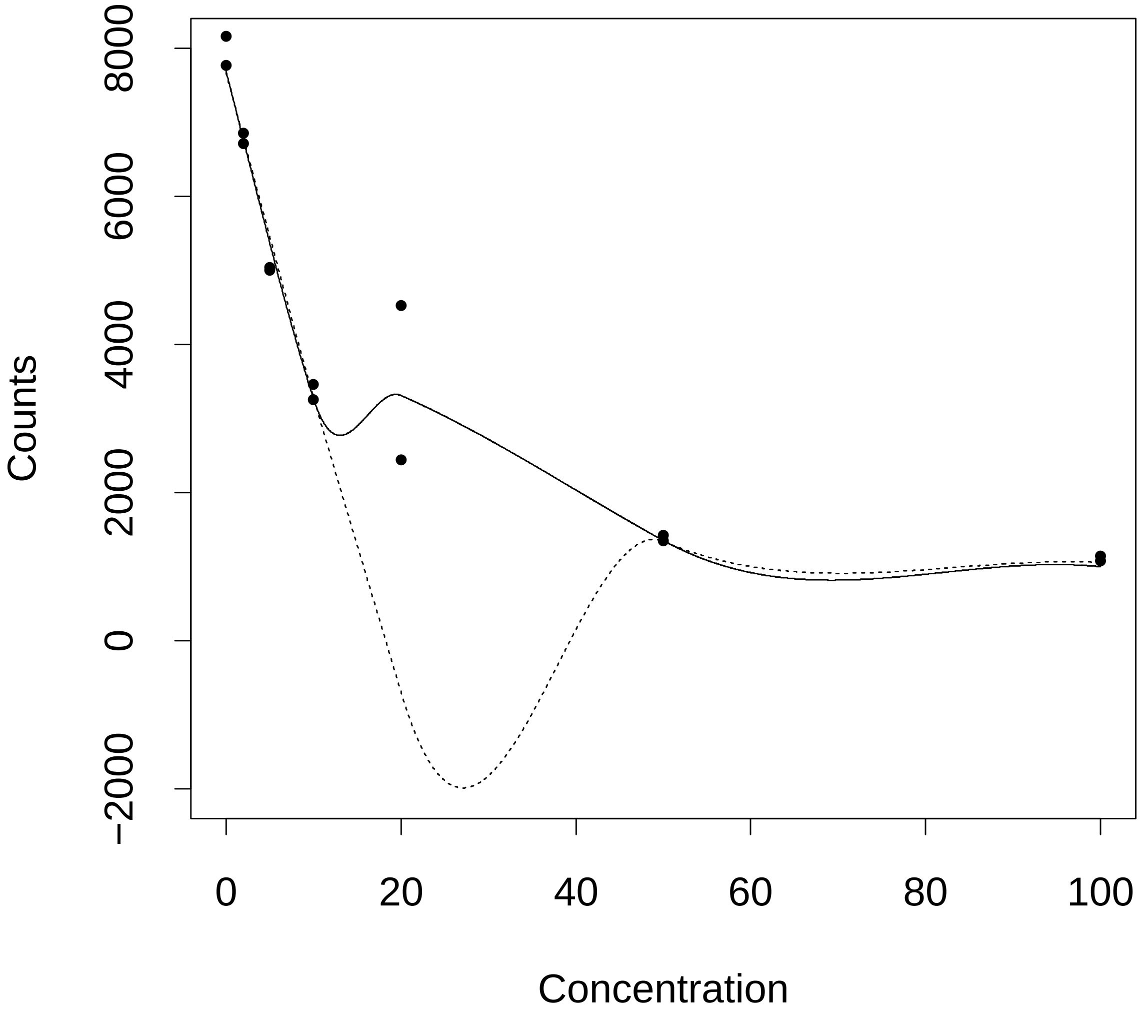


Figure 2: Conditional breakdown values of different kernels for uniform design based on nearest neighbor bandwidth covering 20% of the observations for the uniform kernel ($n = 100$). The solid line refers to the uniform kernel, the dotted line to the quadratic kernel, and the dashed line to the triweight kernel.

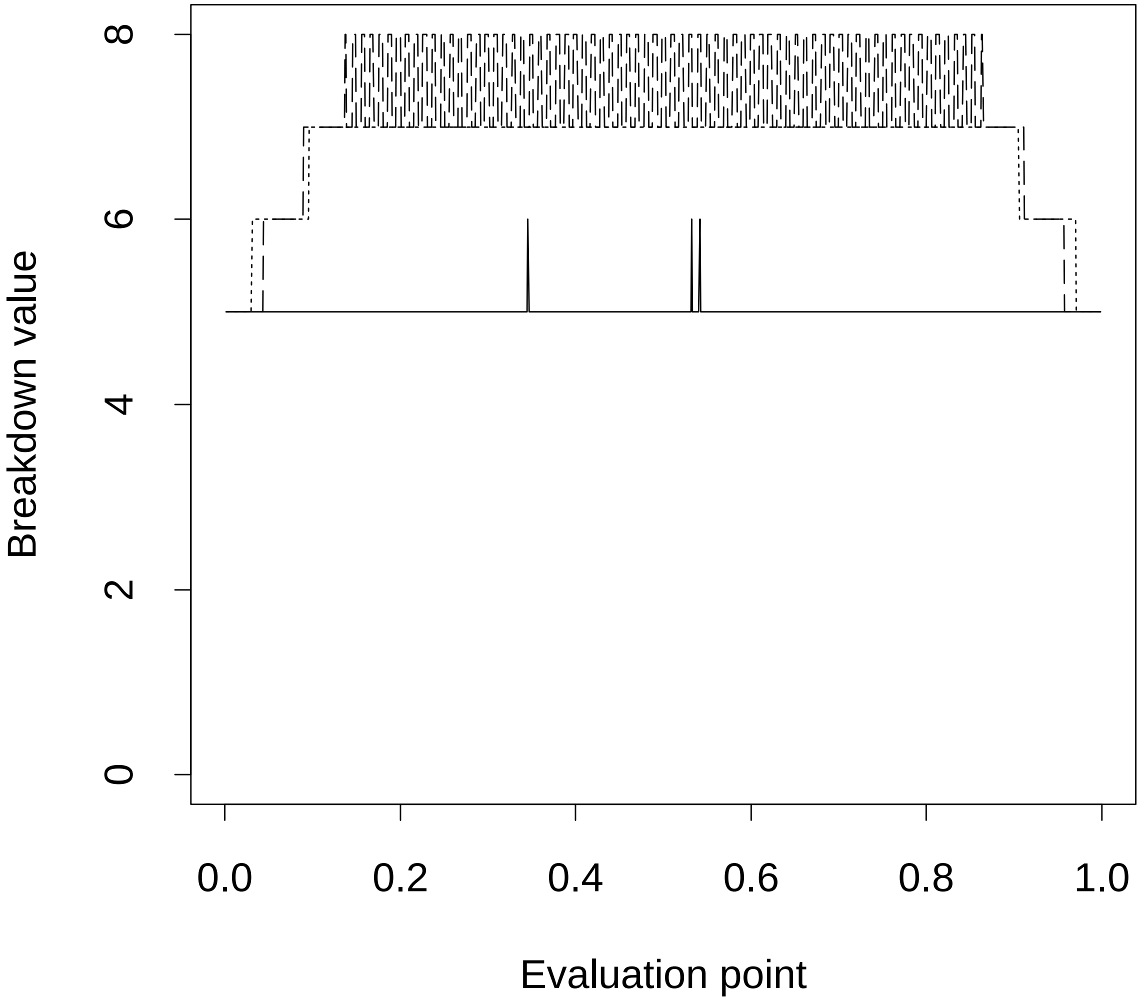


Figure 3: Conditional breakdown values of different kernels for Gaussian design based on nearest neighbor bandwidth covering 20% of the observations for the uniform kernel ($n = 100$). The solid line refers to the uniform kernel, the dotted line to the quadratic kernel, and the dashed line to the triweight kernel.

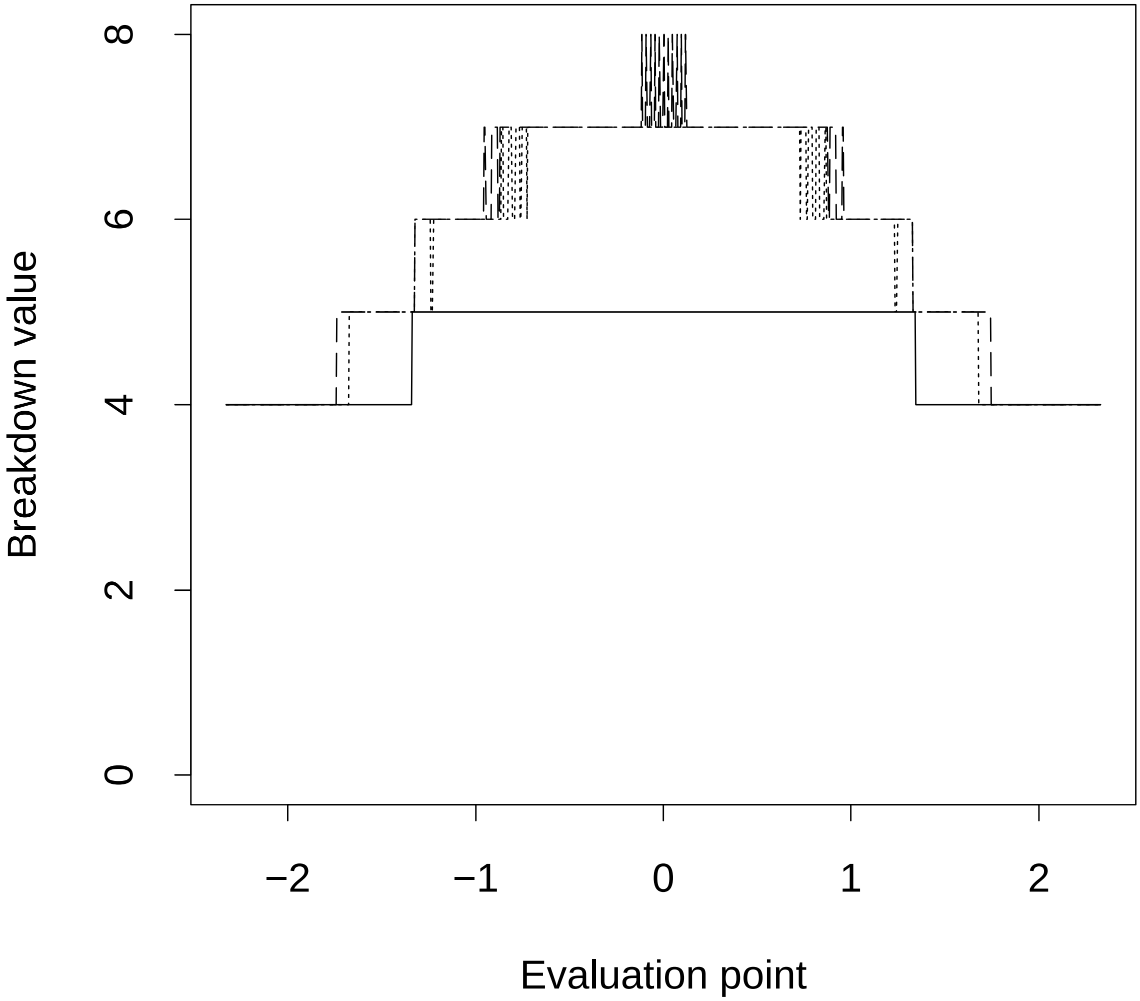


Figure 4: Conditional breakdown values of different kernels for exponential design based on nearest neighbor bandwidth covering 20% of the observations for the uniform kernel ($n = 100$). The solid line refers to the uniform kernel, the dotted line to the quadratic kernel, and the dashed line to the triweight kernel.

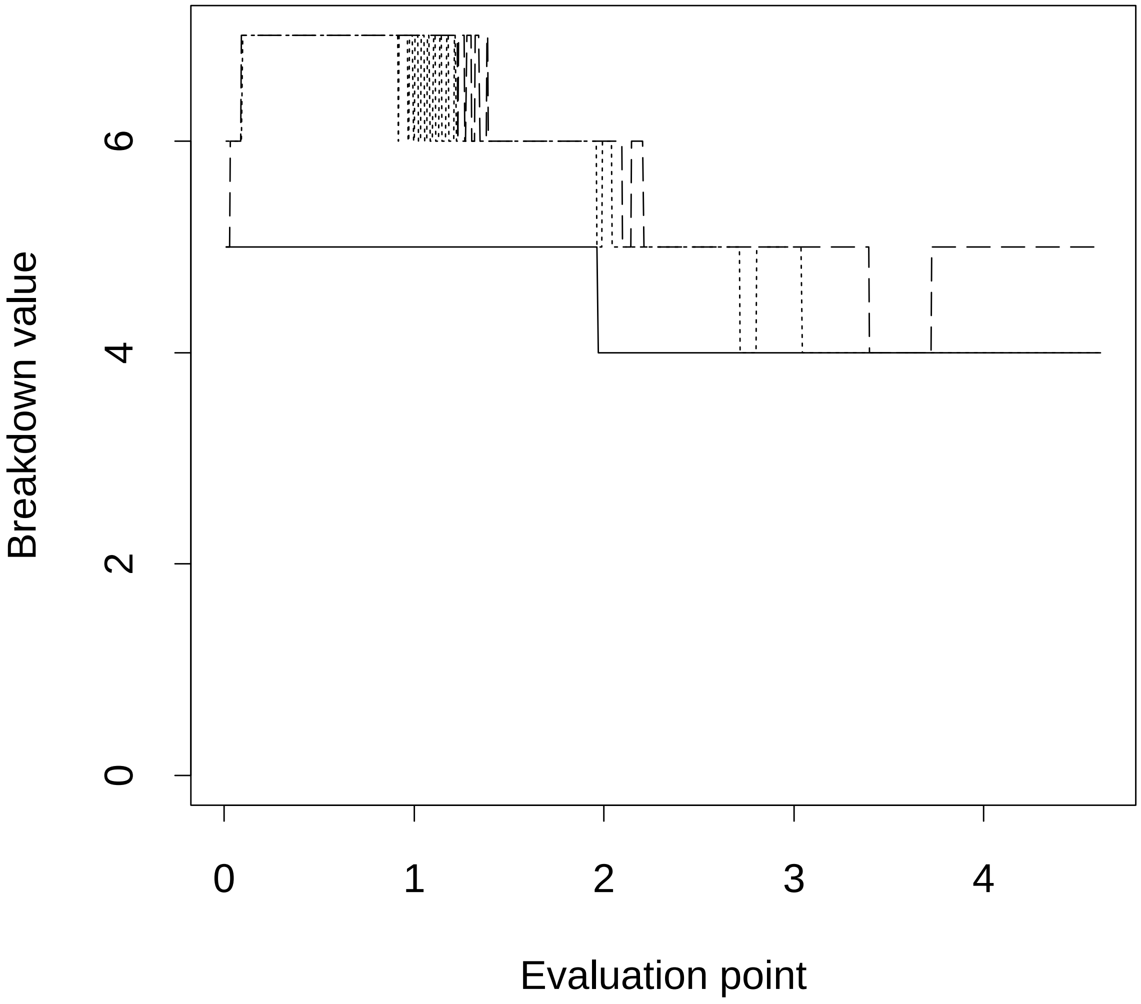
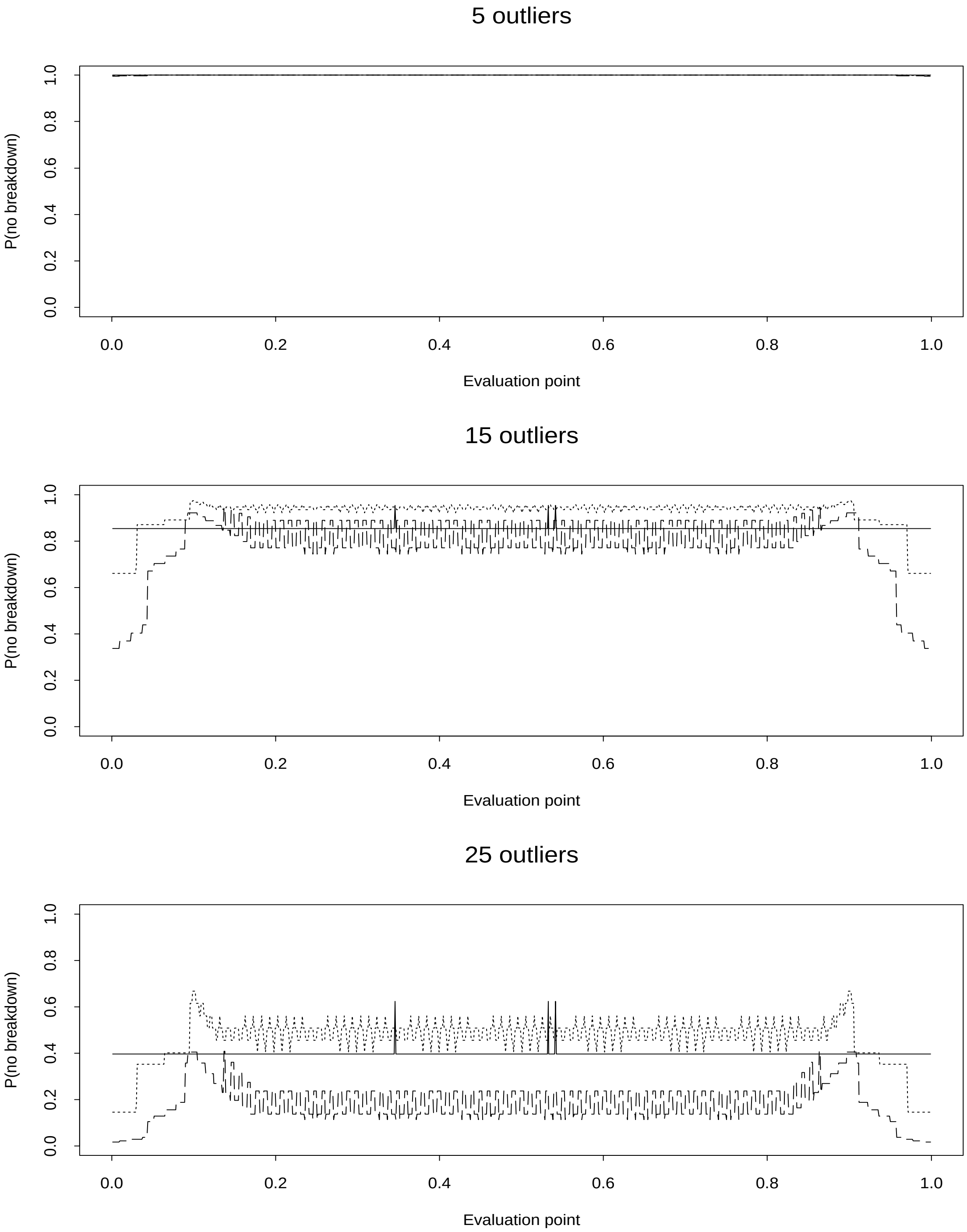


Figure 5: Probability of no breakdown of different kernels for uniform design based on nearest neighbor bandwidth covering 20% of the observations for the uniform kernel ($n = 100$). The solid line refers to the uniform kernel, the dotted line to the quadratic kernel, and the dashed line to the triweight kernel. The top plot refers to 5 outliers in the data, the middle plot to 15 outliers, and the bottom plot to 25 outliers.

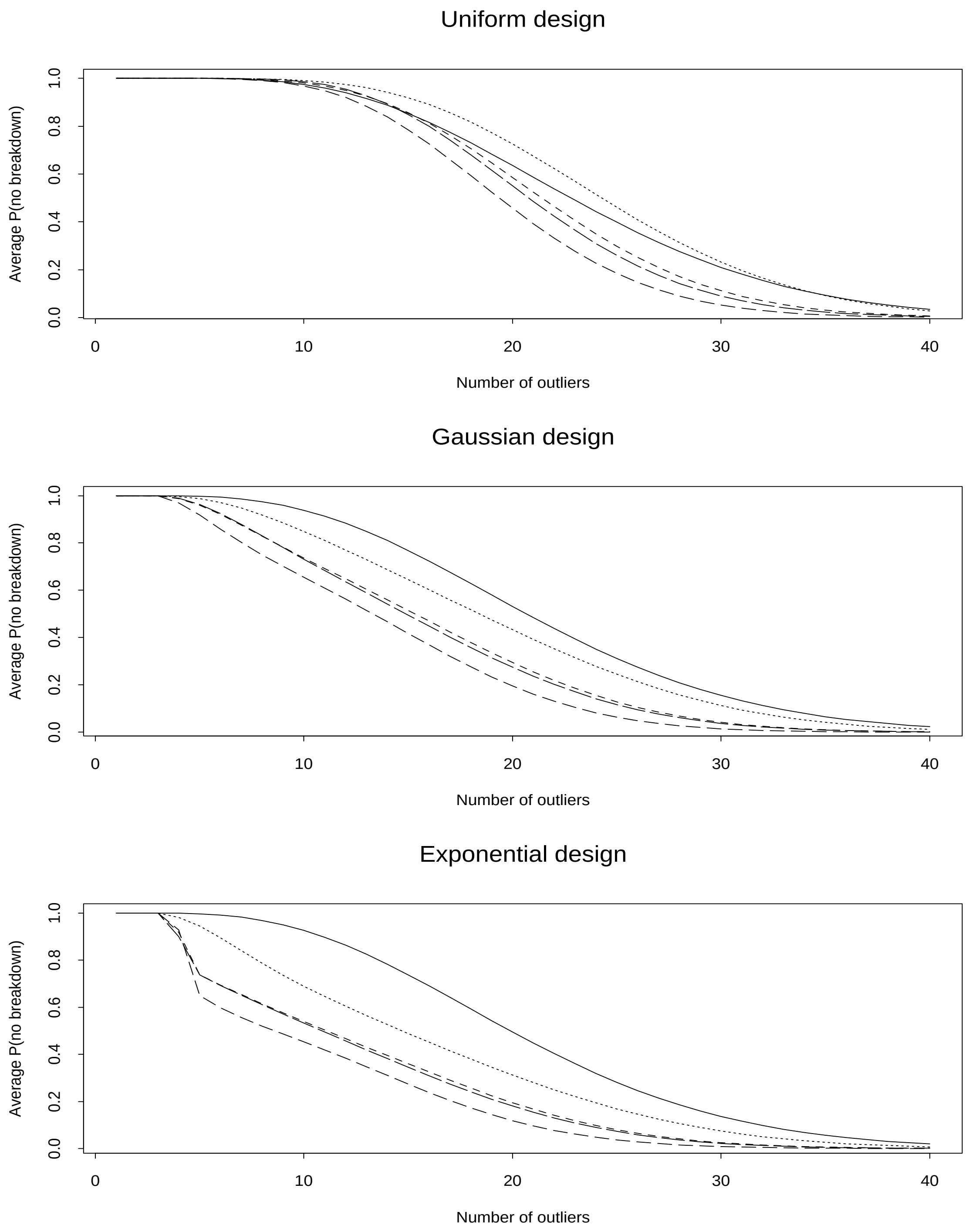


5 outliers

15 outliers

25 outliers

Figure 6: Average probability (over evaluation points) of no breakdown for different numbers of outliers of different kernels for uniform design based on nearest neighbor bandwidth covering 20% of the observations for the uniform kernel ($n = 100$). The solid line refers to the uniform kernel, the dotted line to the quadratic kernel, the short–dashed line to the biweight kernel, the medium–dashed line to the triweight kernel, and the long–dashed line to the tricube kernel. The top plot refers to the uniform design, the middle plot to the Gaussian design, and the bottom plot to the exponential design.



Uniform design

Gaussian design

Exponential design

Figure 7: Probability of no breakdown of different kernels for Gaussian design based on nearest neighbor bandwidth covering 20% of the observations for the uniform kernel ($n = 100$). The solid line refers to the uniform kernel, the dotted line to the quadratic kernel, and the dashed line to the triweight kernel. The top plot refers to 5 outliers in the data, the middle plot to 15 outliers, and the bottom plot to 25 outliers.

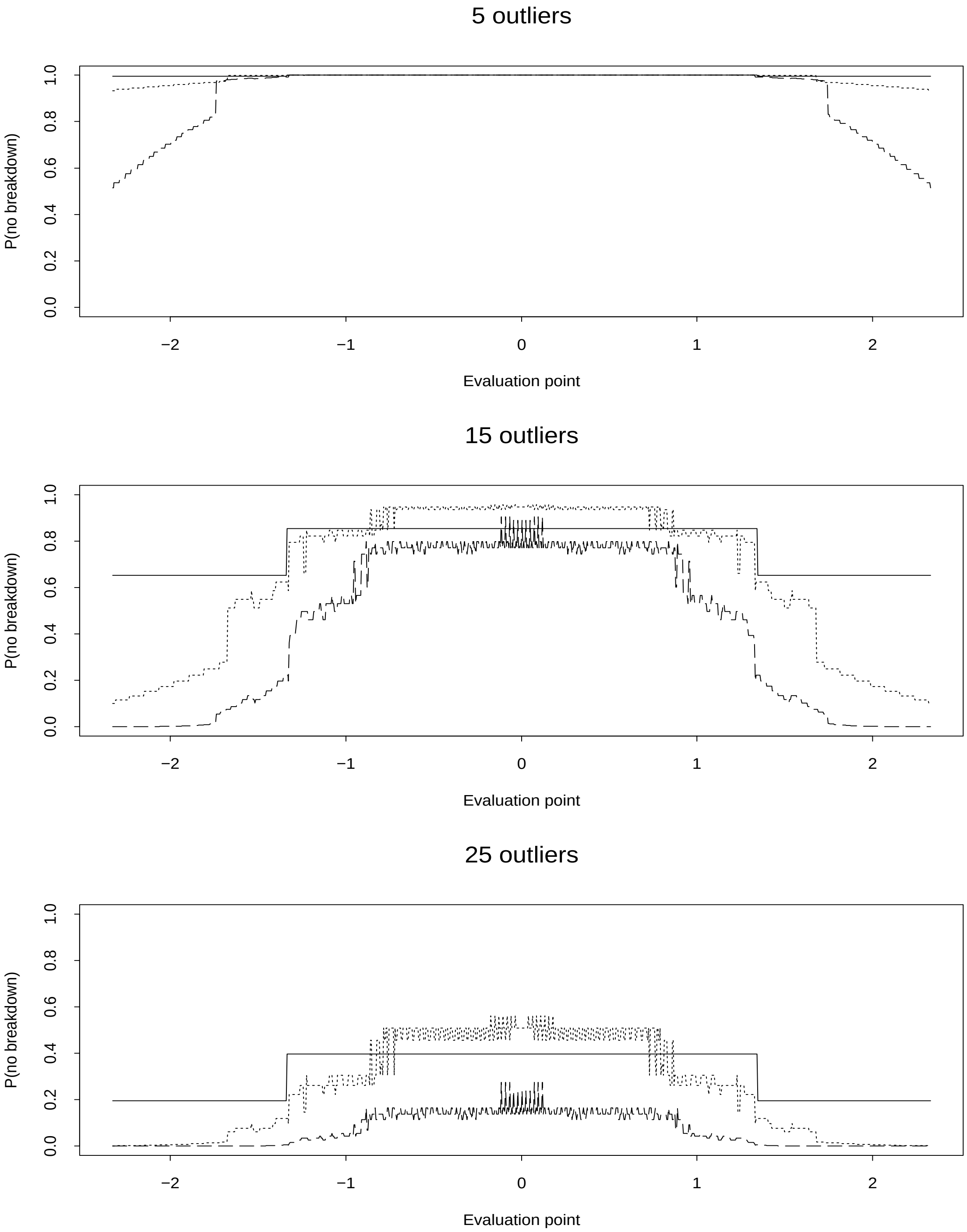
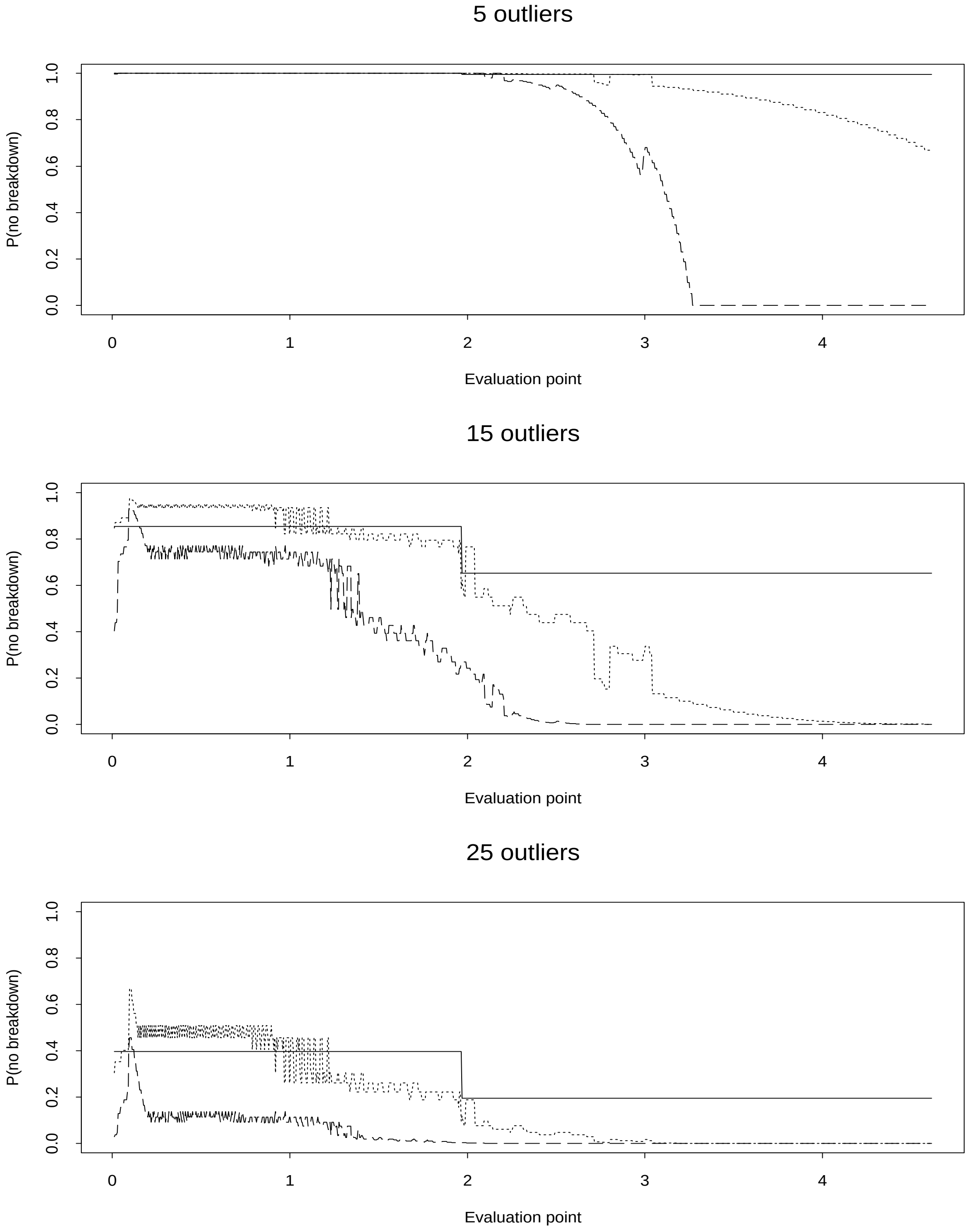


Figure 8: Probability of no breakdown of different kernels for exponential design based on nearest neighbor bandwidth covering 20% of the observations for the uniform kernel ($n = 100$). The solid line refers to the uniform kernel, the dotted line to the quadratic kernel, and the dashed line to the triweight kernel. The top plot refers to 5 outliers in the data, the middle plot to 15 outliers, and the bottom plot to 25 outliers.


5 outliers

15 outliers

25 outliers

Figure 9: Properties of local least absolute values linear estimator using different kernels for a uniform design when $n = 1000$, based on nearest neighbor bandwidth covering 20% of the observations for the uniform kernel. The solid line refers to the uniform kernel, the dotted line to the quadratic kernel, and the dashed line to the triweight kernel. The top plot gives conditional breakdown values, the middle plot gives the probability of no breakdown of when there are 260 outliers, and the bottom plot gives the average probability (over evaluation points) of no breakdown for different numbers of outliers.

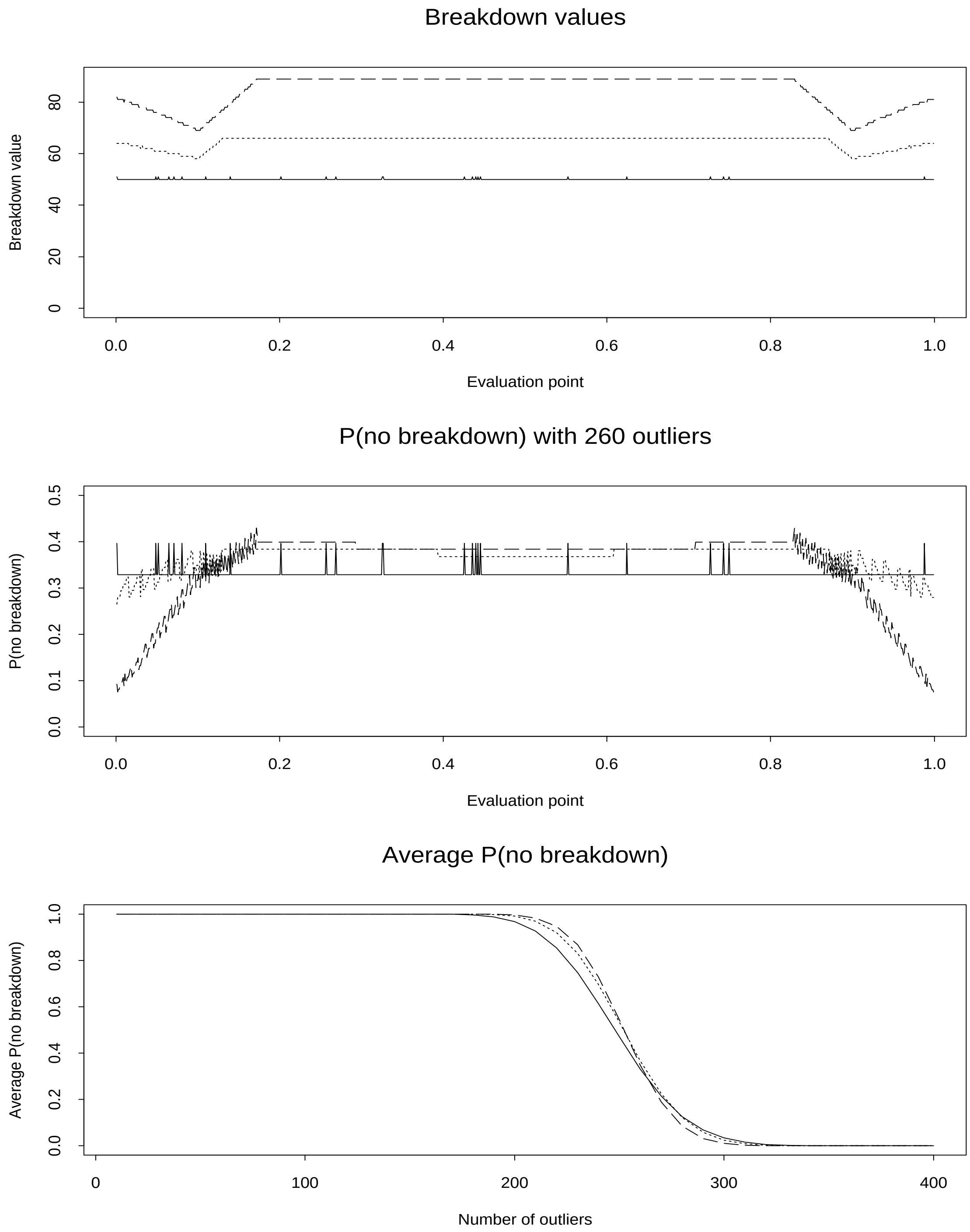
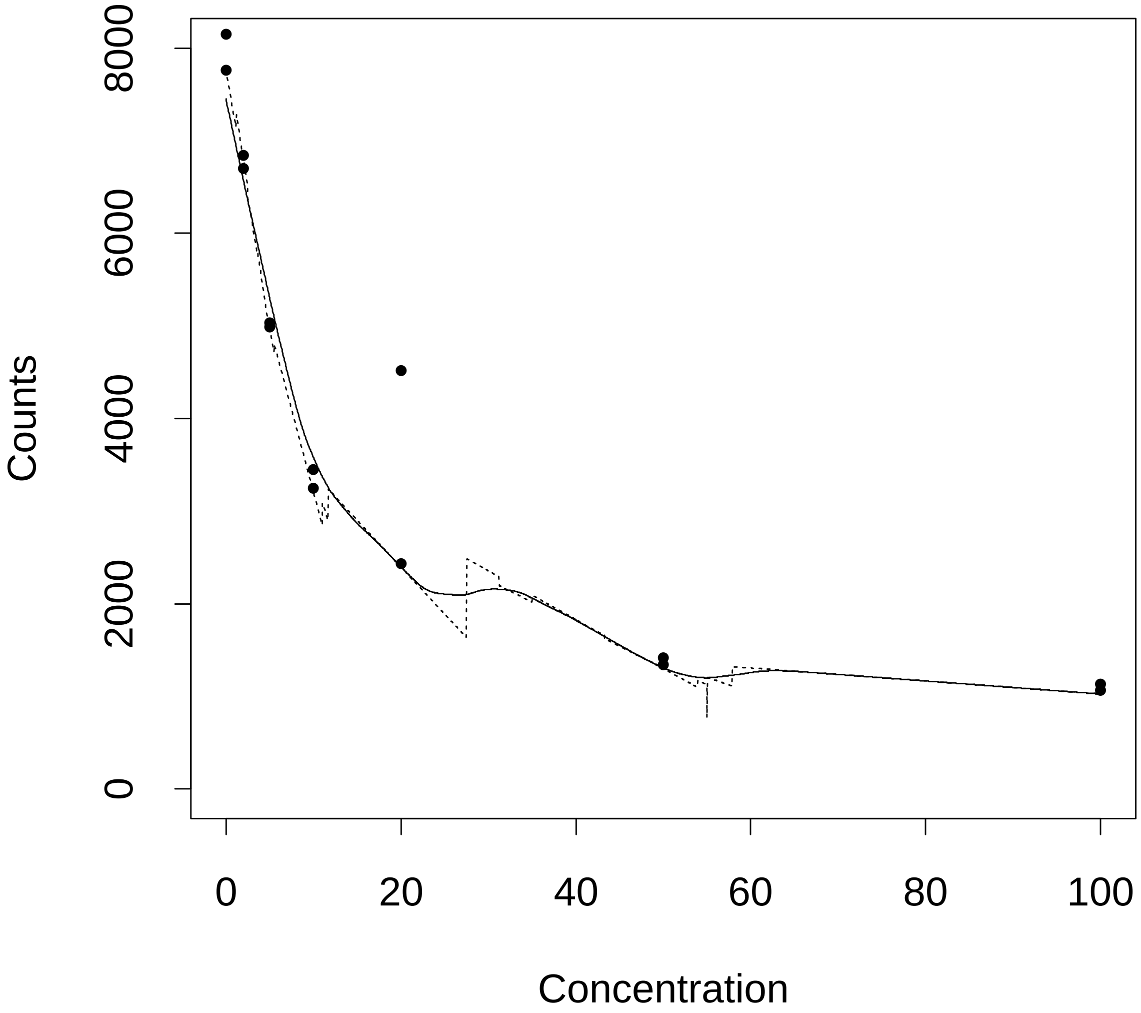


Figure 10: Local least absolute values linear estimates for calibration data. The dotted line is the estimate, while the solid line is a version that has been smoothed.


Counts
Concentration