5-1-2005

Inferences About Regression Interactions Via A Robust Smoother With An Application To Cannabis Problems

Rand R. Wilcox
University of Southern California, rwilcox@usc.edu

Mitchell Earleywine
University of Southern California, Los Angeles

Inferences About Regression Interactions Via A Robust Smoother With An Application To Cannabis Problems

Rand R. Wilcox   Mitchell Earleywine
Department of Psychology
University of Southern California, Los Angeles

A flexible approach to testing the hypothesis of no regression interaction is to test the hypothesis that a generalized additive model provides a good fit to the data, where the components are some type of robust smoother. A practical concern, however, is that there are no published results on how well this approach controls the probability of a Type I error. Simulation results, reported here, indicate that an appropriate choice for the span of the smoother is required so that the actual probability of a Type I error is reasonably close to the nominal level. The technique is illustrated with data dealing with cannabis problems where the usual regression model for interactions provides a poor fit to the data.

Key words: Robust smoothers, curvature, interactions

Introduction

A combination of extant regression methods provides a very flexible and robust approach to detecting and modeling regression interactions. In particular, both curvature and nonnormality are allowed. The main goal in this paper is to report results on the small-sample properties of this approach when a particular robust smoother is used to approximate the regression surface. The main result is that in order to control the probability of a Type I error, an appropriate choice for the span must be used which is a function of the sample size. However, before addressing this issue, we provide a motivating example for considering smoothers when investigating interactions.

A well-known approach to detecting and modeling regression interactions is to assume that for a sample of n vectors of observations,

Rand R. Wilcox (rwilcox@usc.edu) is a Professor of Psychology at the University of Southern California, Los Angeles. M. Earleywine is an Associate Professor at the University of Southern California, Los Angeles.

$$(Y_i, X_{i1}, X_{i2}),$$
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i, \quad (1)$$

$i=1,...,n$, where ε is independent of X_{i1} and X_{i2}, $E(\varepsilon) = 0$. The hypothesis of no interaction corresponds to

$$H_0 : \beta_3 = 0.$$

This approach appears to have been first suggested by Saunders (1956). A practical issue is whether this approach is flexible enough to detect and to model an interaction if one exists. We consider data collected by the second author to illustrate that at least in some situations, a more flexible model is required. The data deal with cannabis problems among adult males. Responses from $n=296$ males were obtained where the two regressors were the participants' use of cannabis (X_1) and consumption of alcohol (X_2). The dependent measure (Y) reflected cannabis dependence as measured by the number of DSM-IV symptoms reported. An issue of interest was determining whether the amount of alcohol consumed alters the association between Y and the amount of cannabis used, and there is the issue of

understanding how the association changes if an interaction exists.

Using a method derived by Stute, González-Manteiga and Presedo-Quindimil (1998), it is possible to test the hypothesis that the model given by equation (1) provides a good fit to the data. If, for example,

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2}^2 + \varepsilon_i,$$

then there is an interaction, but the family of regression equations given by (1) is inappropriate. The Stute et al. method can be applied using the S-PLUS or R function lintest in Wilcox (2003). Estimating the unknown parameters via least squares, this hypothesis is rejected at the .05 level. A criticism is that when testing the hypothesis that (1) is an appropriate model for the data, and when using the ordinary least squares estimator when estimating the unknown parameters, the probability of a Type I error might not be controlled (Wilcox, 2003).

Replacing the least squares estimator with various robust estimators corrects this problem. Here, using the robust M-estimator derived by Coakley and Hettmansperger (1993), or using a generalization of the Theil-Sen estimator to multiple predictors (see Wilcox, 2005), again the hypothesis is rejected. Moreover, the R (or S-PLUS) function pmodchk in Wilcox (2005) provides a graphical check of how well the model given by (1) fits the data when a least squares estimate of the parameters is used, versus a more flexible fit based on what is called a running interval smoother, and a poor fit based on (1) is indicated. Robust variations give similar results. So, at least in this case, an alternative and more flexible approach to testing the hypothesis of no interaction seems necessary.

To provide more motivation for a more flexible approach when modeling interactions, note that equation (1) implies a nonlinear association between Y versus X_1 and X_2. A concern, however, is that a nonlinear association does not necessarily imply an interaction. If, for example, X_1, X_2 and ε are independent and have standard normal distributions, and if $Y = X_1 + X_2^2 + \varepsilon$, the probability of rejecting

H_0: $\beta_3 = 0$ is .18 when testing at the .05 level with a sample size of twenty. Of course, in this case, standard diagnostics can be used to detect the curvature, but experience with smoothers suggest that dealing with curvature is not always straightforward.

Suppose instead $Y = X_1 + X_1^2 + |X_2| + \varepsilon$, so there is no interaction even though there is a nonlinear association. Then with a sample size of fifty, and when testing at the .05 level, the probability of rejecting H_0: $\beta_3 = 0$ is .30. In contrast, using the more flexible method described here, the probability of rejecting the hypothesis of no interaction is .042.

If we ignore the result that (1) is an inadequate model for the cannabis data and simply test H_0: $\beta_3 = 0$ (using least squares in conjunction with a conventional T test), or if we test H_0: $\beta_3=0$ using a more robust hypothesis testing method derived for the least squares estimator that is based on a modified percentile bootstrap method (Wilcox, 2003), or when using various robust estimators (such as an M-estimator with Schweppe weights or when using the Coakley-Hettmansperger estimator), we reject. But an issue is whether we reject because there is indeed an interaction, or because the model provides an inadequate representation of the data. And another concern is that by using an invalid model, an interaction might be masked.

A more general and more flexible approach when investigating interactions is to test the hypothesis that there exists some functions f_1 and f_2 such that

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + \varepsilon. \qquad (2)$$

Equation (2) is called a generalized additive model, a general discussion of which can be found in Hastie and Tibshirani (1990). A special case is where $f_1(X_1) = \beta_1 X_1$, $f_2(X_2) = \beta_2 X_2$, but (2) allows situations where the regression surface is not necessarily a plane, even when there is no interaction. If the model represented by (2) is true, then there is no interaction in the following sense. Pick any two values for X_2,

say 6 and 8. Then no interaction means that the regression line between Y and X_1, given that $X_2 = 6$, is parallel to the regression line between Y and X_1, given that $X_2 = 8$.

For completeness, Barry (1993) derived a method for testing the hypothesis of no interaction assuming an ANOVA-type decomposition where

$$Y = \beta_0 + f_1(X_1) + f_2(X_2) + f_3(X_1, X_2) + \varepsilon,$$

in which case the hypothesis of no interaction is

$$H_0 : f_3(X_1, X_2) \equiv 0.$$

Barry (1993) used a Bayesian approach assuming that the (conditional) mean of Y is to be estimated and that prior distributions for f_1, f_2 and f_3 can be specified. The goal in this article is to investigate the small-sample properties of a non-Bayesian method where the mean is replaced by some robust measure of location (cf. Samarov, 1993).

Methodology

There are, in fact, many approaches that might be used that are based on combinations of existing statistical techniques. The problem is finding a combination of methods that controls the probability of a Type I error in simulations even when the sample size is relatively small. One possibility is to use some extension of the method in Dette (1999), this was considered, but in simulations no variation was found that performed well in terms of controlling the probability of a Type I error.

Only one method was found that performs well in simulations; it is based on a combination of methods summarized in Wilcox (2005). The approach is outlined here, and the computational details are relegated to Appendices A and B. Briefly, the method begins by fitting the model given by (2) using the so-called backfitting algorithm (Hastie & Tibshirani, 1990) in conjunction with a what is called a running interval smoother. Generally, smoothers are methods for approximating

regression lines without forcing them to have a particular shape such as a straight line. As with most smoothers, the running interval smoother is based in part on something called a span, κ, which plays a role when determining whether the value X is close to a particular value of X_1 (or X_2). Details are provided in Appendix A.

There are many ways of fitting the model given by (2). Here, the focus is on a method where the goal is to estimate a robust measure of location associated with Y, given (X_1, X_2), because of the many known advantages such measures have (e.g., Hampel, Ronchetti, Rousseeuw & Stahel, 1986; Huber, 1981; Staudte & Sheather, 1990; Wilcox, 2003, 2005). Primarily for convenience, the focus is on a 20% trimmed mean, but various robust M-estimators are certainly a possibility. The advantages associated with robust measures of location include an enhanced ability to control the probability of a Type I error in situations where methods based on means are known to fail, and substantial gains in power, over methods based on means, even under slight departures from normality. (Comments about using the mean, in conjunction with the proposed method, are made in the final section of this paper.) Here, the main reason for not using a robust M-estimator (with say, Huber's Ψ), is that this estimator requires division by the median absolute deviation (MAD) statistic, and in some situations considered here, when the sample size is small, MAD is zero.

The running interval smoother provides a predicted value for Y, given (X_{i1}, X_{i2}), say $\hat{Y}_i$; see Appendix A. Next, compute the residuals $r_i = Y_i - \hat{Y}_i$. If the model given by (2) is true, meaning that there is no interaction, then the regression surface when predicting r, given (X_1, X_2), should be a horizontal plane. The hypothesis that this regression surface is indeed a horizontal plane can be tested using the method derived by Stute et al. (1998). The details can be found in Appendix B.

Results

Simulations were conducted as a partial check on the ability of the method, just outlined, to control the probability of a Type I error. Values for X_1, X_2 and ε were generated from four types of distributions: normal, symmetric and heavy-tailed, asymmetric and light-tailed, and asymmetric and heavy-tailed. For non-normal distributions, observations were generated from a g-and-h distribution which is described in Appendix C. The goal was to check on how the method performs under normality, plus what would seem like extreme departures from normality, with the idea that if good performance is obtained under extreme departures from normality, the method should perform reasonably well with data encountered in practice. The correlation between X_1 and X_2 was taken to be either ρ=0 or ρ=.5.

Initial simulation results revealed that the actual probability of a Type I error, when testing at the .05 level, is sensitive to the span, κ. (Härdle & Mammen, 1993, report a similar result for a method somewhat related to the problem at hand.) If the span is too large, the actual Type I error probability can drop well below the nominal level. When testing at the .05 level, simulations were used to approximate a reasonable choice for κ. Here, the span corresponding to the sample sizes 20, 30, 50, 80 and 150 are taken to be .4, .36, .18, .15 and .09, respectively. It is suggested that when 20≤*n*≤150, interpolation based on these values be used, and for *n*>150 use a span equal to .09. For *n*>150 and sufficiently large, perhaps the actual Type I error probability is well below the nominal level, but exactly how the span should be modified when *n*>150 is an issue that is in need of further investigation.

Table 1 contains $\hat{\alpha}$, the estimated probability of making a Type I error when testing at the .05 level. *n*=20, and when *Y*=ε or $Y = X_1 + X_2^2 + \varepsilon$. (The *g* and *h* values are explained in Appendix C.) Simulations were also run when $Y = X_1 + X_2 + \varepsilon$, the results were very similar to the case *Y*=ε, so for brevity they are not reported. No situation was found

where the estimated probability of a Type I error exceeded the nominal .05 level. The main difficulty is that when marginal distributions have a skewed, heavy-tailed distribution, ρ=.5, and there is curvature, the estimated probability of a Type I error dropped below .01. This situation corresponds to what would seem like an extreme departure from normality as indicated in Appendix C.

An Illustration

Returning to the cannabis data described in the introduction, the hypothesis of no interaction is rejected at the .05 level when testing the model given by (2). (The test statistic described in Appendix B is *D*=3.37 and the .05 critical value is 1.79.) To provide some overall sense of the association, Figure 1 shows an approximation of the regression surface based on a smooth derived by Cleveland and Devlin (1988) called loess. (Using the robust smooth in Wilcox, 2003, section 14.2.3, gives similar results when the span is set to 1.2.) Note the nonlinear appearance of the surface. Also, there appears to be little or no association over some regions of the X_1 and X_2 values.

Figure 2 shows the plot based on X_1 and X_2 versus the residuals corresponding to the generalized additive model given by (2). This plot should be a horizontal plane if there is no interaction. As is evident, the surface appears to be nonlinear, at least to some degree

To further explore the nature of the interaction, first it is noted that the quartiles associated with X_2 (alcohol use) are -0.732, -0.352 and 0.332. The left panel of Figure 3 shows three smooths between *Y* and X_1; they are the smooths between *Y* and X_1 given that

$$X_2 = -0.73, \quad X_2 = -0.352 \quad \text{and} \quad X_3 = 0.332.$$

(These smooths were created using a slight generalization of the kernel regression estimator in Fan, 1993; see R or S-PLUS function kercon in Wilcox, 2005, Ch. 11.)

Table 1: Estimated probability of a Type I error, $n=20$.

g	h	g	h	$Y=\varepsilon$		$Y = X_1 + X_2^2 + \varepsilon$	
				$\rho=0$	$\rho=.5$	$\rho=0$	$\rho=.5$
0.0	0.0	0.0	0.0	.033	.034	.047	.035
		0.0	0.5	.039	.034	.026	.031
		0.5	0.0	.045	.043	.045	.034
		0.5	0.5	.037	.035	.035	.032
0.0	0.5	0.0	0.0	.031	.032	.019	.015
		0.0	0.5	.032	.024	.020	.012
		0.5	0.0	.033	.031	.016	.013
		0.5	0.5	.029	.024	.023	.013
0.5	0.0	0.0	0.0	.029	.022	.036	.022
		0.0	0.5	.031	.020	.032	.014
		0.5	0.0	.040	.039	.037	.028
		0.5	0.5	.029	.027	.025	.020
0.5	0.5	0.0	0.0	.028	.024	.024	.003
		0.0	0.5	.026	.017	.015	.003
		0.5	0.0	.035	.029	.014	.006
		0.5	0.5	.020	.015	.015	.007

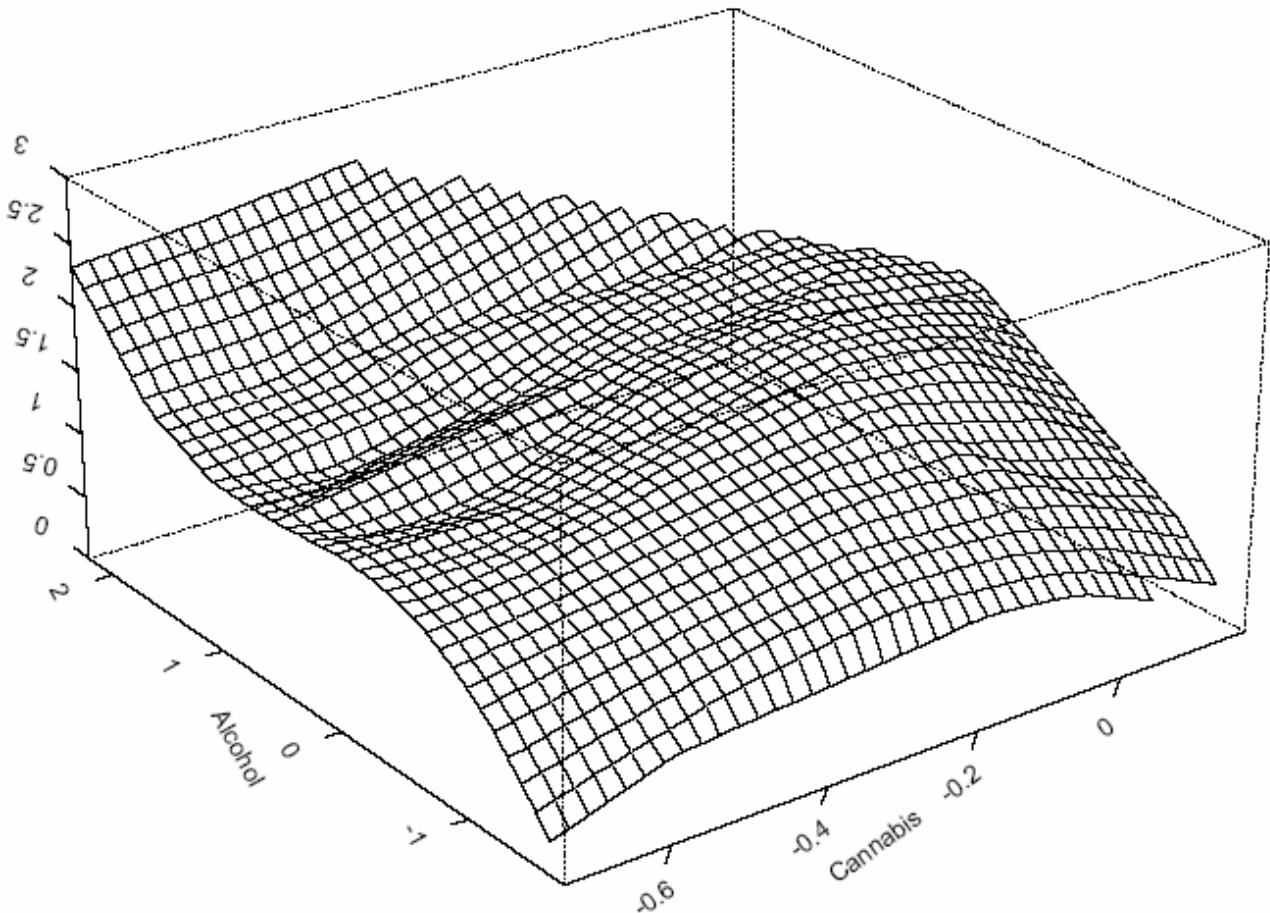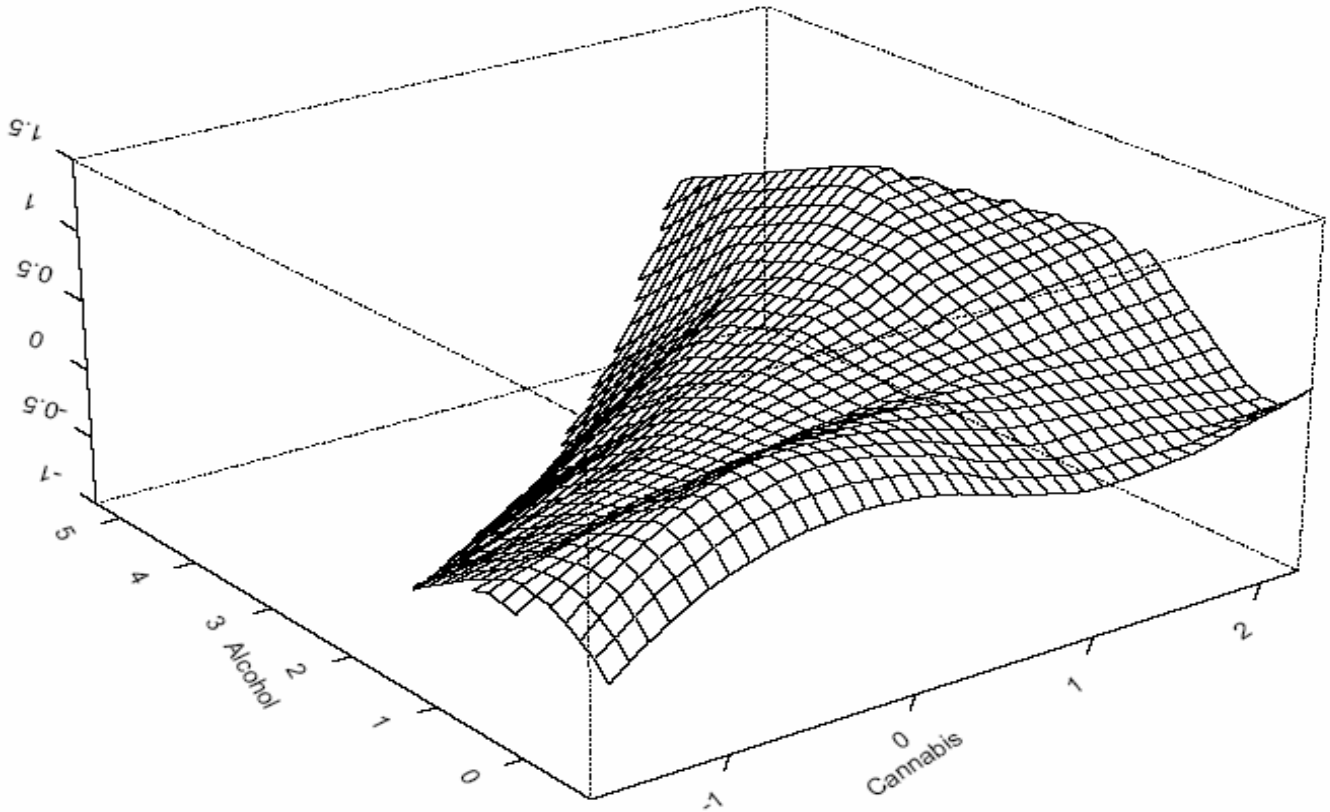Figure 1: An approximation of the regression surface based on the smoother loess.

Figure 2: A smooth of the residuals stemming from the generalized additive model versus the two predictors.

When there is no interaction, all three regression lines should be approximately parallel which is not the case. The regression lines corresponding to $X_2 = -0.73$ and -0.352 are reasonably parallel, and they are approximately horizontal suggesting that there is little association between Y and X_1 for these special cases.

But for $X_3 = 0.332$, the association changes, particularly in the right portion of Figure 1 where the association becomes more positive. If the data are split into two groups according to whether X_{i2} is less than the median of the values $X_{12},...,X_{n2}$, -0.352, and then create a smooth between Y and X_1, the result is shown in right panel of Figure 3.

Figure 3: Some smooths used to investigate interactions.

Conclusion
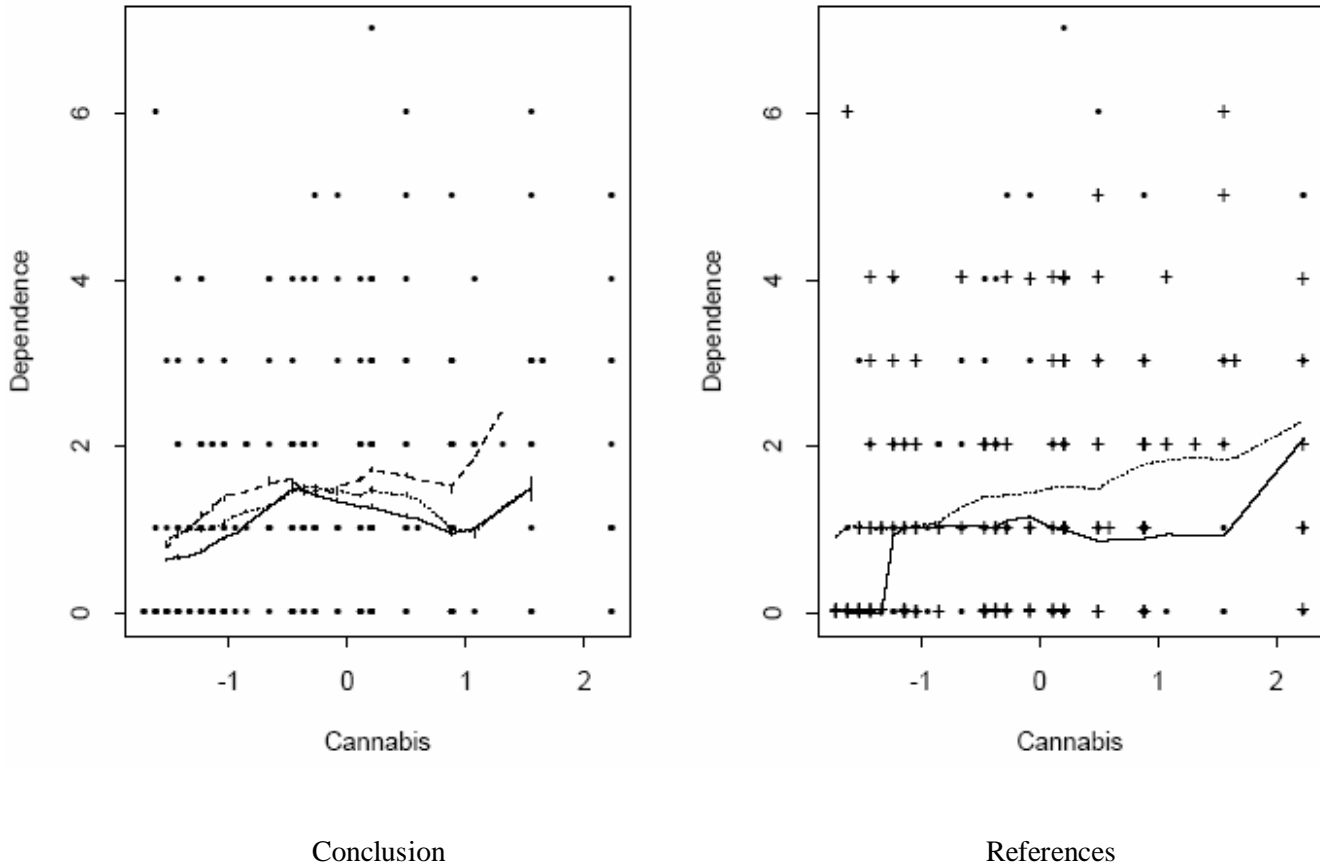
In principle, the method in this article can be used with any measure of location. It is noted, however, that if the 20% trimmed mean is replaced by the sample mean, poor and unstable control over the probability of a Type I error results.

Finally, all of the methods used in this paper are easily applied using the S-PLUS or R functions in Wilcox (2005). (These functions can be downloaded as described in chapter 1.) Information about S-PLUS can be obtained from www.insightful.com, and R is a freeware variant of S-PLUS that can be downloaded from www.R-project.org. For convenience, the relevant functions for the problem at hand have been combined into a single function called adtest. If, for example, the *X* values are stored in an S-PLUS matrix x, and the *Y* values are stored in y, the command adtest(x,y) tests the hypothesis that the model given by (2) is true.

References

Barry, D. (1993). Testing for additivity of a regression function. *Annals of Statistics, 21*, 235-254.

Cleveland, W. S., & Devlin, S. J. (1988). Locally-weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association, 83*, 596-610.

Coakley, C. W., & Hettmansperger, T. P. (1993). A bounded influence, high breakdown, efficient regression estimator. *Journal of the American Statistical Association, 88*, 872-880.

Dette, H. (1999). A consistent test for the functional form of a regression based on a difference of variances estimator. *Annals of Statistics, 27*, 1012-1040.

Fan, J. (1993). Local linear smoothers and their minimax efficiencies. *The Annals of Statistics, 21*, 196-216.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust Statistics*. New York: Wiley.

Hardle, W. (1990). *Applied Nonparametric Regression.* Cambridge: Cambridge University Press.

Hardle, W., & Mammen, E. (1993). Comparing non-parametric versus parametric regression fits. *Annals of Statistics, 21*, 1926-1947.

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized Additive Models*. New York: Chapman and Hall.

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. Hoaglin, F. Mosteller & J. Tukey (Eds.) *Exploring Data Tables Trends and Shapes*. New York: Wiley.

Huber, P. J. (1981). *Robust Statistics*. New York: Wiley.

Rosenberger, J. L., & Gasko, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Underststanding Robust and Exploratory Data Analysis.* (pp. 297-336). New York: Wiley.

Samarov, A. M. (1993). Exploring regression structure using nonparametric functional estimation. *Journal of the American Statistical Association, 88*, 836-847.

Saunders, D. R. (1956). Moderator variables in prediction. *Educational and Psychological Measurement, 16*, 209-222.

Stute, W., Gonzlez-Manteiga, W. G. & Presedo-Quindimil, M. P. (1998). Bootstrap approximations in model checks for regression. *Journal of the American Statistical Association, 93*, 141-149.

Wilcox, R. R. (2003). *Applying Contemporary Statistical Techniques*. San Diego, CA: Academic Press.

Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*, 2nd Edition. San Diego, CA: Academic Press.

Appendix A

We begin by describing how to compute a 20% trimmed mean based on a sample of m observations. Put the observations in ascending order yielding $W_{(1)} \leq \cdots \leq W_{(m)}$. Let $\ell = [.2m]$, where $[.2m]$ means to round $.2m$ down to the nearest integer. Then the 20% trimmed mean is

$$\overline{X}_t = \frac{1}{n-2\ell} \sum_{t=\ell+1}^{n-\ell} W_{(i)}.$$

In terms of efficiency (achieving a small standard error relative to the usual sample mean), 20% trimming performs very well under normality but continues to perform well in situations where the sample mean performs poorly (e.g., Rosenberger & Gasko, 1983).

Now, we describe the running interval smoother in the one-predictor case. Consider a random sample $(X_1, Y_1), ..., (X_n, Y_n)$ and let κ be some constant that is chosen in a manner to be described. The constant κ is called the span. The median absolute deviation (MAD), based on $X_1, ..., X_n$, is the median of the n values $|X_1 - M|, ..., |X_n - M|$, where M is the usual median. Let MADN=MAD/.6745. Under normality, MADN estimates σ, the standard deviation. Then the point X is said to be close to X_i if

$$|X_i - X| \leq \kappa \times MADN.$$

Thus, for normal distributions, X is close to X_i if X is within κ standard deviations of X_i. Then $\hat{Y}_i$ is the 20% trimmed mean of the Y_j values for which X_j is close to X_i. In exploratory work, a good choice for the span is often $\kappa = .8$ or 1, but for the situation at hand an alternative choice is needed.

Virtually any smoother, including the one used here, can be extended to the generalized additive model given by (2) using the backfitting algorithm in Hastie and Tibshirani (1990). Set $k=0$ and let f_j^0 be some

initial estimate of f_j ($j=1$, 2). Here, $f_j^0(X_j) = S_j(Y \mid X_j)$, where $S_j(Y \mid X_j)$ is the running interval smooth based on the jth predictor, ignoring the other predictor under investigation. Next, iterate as follows.

1. Increment k.
2. Let

$$f_1^k(X_1) = S_1(Y - f_2^{k-1}(X_2) \mid X_1)$$

and

$$f_2^k(X_2) = S_2(Y - f_1^{k-1}(X_1) \mid X_2).$$

3. Repeat steps 1 and 2 until convergence.

Finally, estimate β_0 with the 20% trimmed mean of the values $Y_i - \sum f_j^k(Y_i \mid X_{ij})$, $i=1,...,n$. The computations are performed by R or S-PLUS function adrun in Wilcox (2005).

Appendix B

This appendix describes the method for testing the hypothesis of no interaction. Fit the generalized additive model as described in Appendix A yielding $\hat{Y}_i$, and let $r_i = Y_i - \hat{Y}_i$, $i=1,...,n$. The goal is to test the hypothesis that the regression surface, when predicting the residuals, given (X_{i1}, X_{i2}), is a horizontal plane. This is done using the wild bootstrap method derived by Stute, González-Manteiga and Presedo-Quindimil (1998). Let $\bar{r}_t$ be the 20% trimmed mean based on the residuals $r_1,...,r_n$. Fix j and set $I_i = 1$ if simultaneously $X_{i1} \leq X_{j1}$ and $X_{i2} \leq X_{j2}$, otherwise $I_i = 0$.

Let

$$
\begin{aligned}
R_j &= \frac{1}{\sqrt{n}} \sum I_i(r_i - \bar{r}_t) \\
&= \frac{1}{\sqrt{n}} \sum I_i v_i, \qquad (3)
\end{aligned}
$$

where $\qquad v_i = r_i - \bar{r}_t.$

The test statistic is the maximum absolute value of all the R_j values. That is, the test statistic is

$$D = max \mid R_j \mid. \qquad (4)$$

An appropriate critical value is estimated with the wild bootstrap method as follows. Generate $U_1,...,U_n$ from a uniform distribution and set

$$V_i = \sqrt{12}(U_i - .5),$$

$$v_i^* = v_i V_i,$$

and

$$r_i^* = \bar{r}_t + v_i^*.$$

Then based on the n pairs of points (X_1, X_2, r_1^*), ..., (X_n, X_n, r_n^*), compute the test statistic as described in the previous paragraph and label it D^*. Repeat this process B times and label the resulting (bootstrap) test statistics $D_1^*,...,D_B^*$. Here, $B=500$ is used. Finally, put these B values in ascending order yielding $D_{(1)}^* \leq \cdots \leq D_{(B)}^*$. Then the critical value is $D_{(u)}^*$, where $u=(1-\alpha)B$ rounded to the nearest integer. That is, reject if

$$D \geq D_{(u)}^*.$$

Appendix C

Details regarding the simulations are as follows. Observations were generated where the marginal distributions have a g-and-h distribution (Hoaglin, 1985) which includes the normal distribution as a special case. More precisely, observations Z_{ij}, (i=1,...,n; j=1, 2) were initially generated from a multivariate normal distribution having correlation ρ, then the marginal distributions were transformed to

$$
X_{ij} = \begin{cases}
\dfrac{exp(gZ_{ij})-1}{g}\, exp(hZ_{ij}^2 / 2), & \text{if } g > 0 \\[2ex]
Z exp(hZ_{ij}^2 / 2), & \text{if } g = 0
\end{cases}
$$

where g and h are parameters that determine the third and fourth moments. The four (marginal) g-and-h distributions examined were the standard normal (g=h=0), a symmetric heavy-tailed distribution (g=0, h=.5), an asymmetric distribution with relatively light tails (g=.5, h=0), and an asymmetric distribution with heavy tails (g=h=.5). Here, two choices for ρ were considered: 0 and .5.

Table 2 shows the theoretical skewness (κ_1) and kurtosis (κ_2) for each distribution considered. When g>0 and h>1/k, $E(X^k)$ is not defined and the corresponding entry in Table 2 is left blank. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985). Some of these distributions might appear to represent extreme departures from normality, but the idea is that if a method performs reasonably well in these cases, this helps support the notion that they will perform well under conditions found in practice.

Table 2: Some properties of the g-and-h distribution.

g	h	κ_1	κ_2
0.0	0.0	0.00	3.0
0.0	0.5	0.00	---
0.5	0.0	1.75	8.9
0.5	0.5	---	---