DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 14 | Issue 1 Article 4

5-1-2015

Estimating the Strength of an Association Based on a Robust Smoother

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

Recommended Citation
Wilcox, Rand (2015) "Estimating the Strength of an Association Based on a Robust Smoother," *Journal of Modern Applied Statistical Methods*: Vol. 14 : Iss. 1 , Article 4.
DOI: 10.22237/jmasm/1430452980
Available at: http://digitalcommons.wayne.edu/jmasm/vol14/iss1/4

Journal of Modern Applied Statistical Methods
May 2015, Vol. 14, No. 1, 2-11.

ISSN 1538 − 9472

Invited Article
Estimating the Strength of an Association Based on a Robust Smoother

Rand Wilcox
University of Southern California
Los Angeles, CA

It is known that the more obvious parametric approaches to fitting a regression line to data are often not flexible enough to provide an adequate approximation of the true regression line. Many nonparametric regression estimators, often called smoothers, have been derived that are aimed at dealing with this problem. The paper deals with the issue of estimating the strength of an association based on the fit obtained by a robust smoother. A simple approach, already known, is to estimate explanatory power in a fairly obvious manner. This approach has been found to perform reasonably well when using the smoother LOESS. But when using a running interval, which provides a simple way of using any robust measure of location, the method performs poorly, even with a reasonably large sample size. The paper suggests an alternative estimation method that performs much better in simulations.

Keywords: Running interval smoother, explanatory power, cross-validation, Well Elderly 2 Study

Introduction

Consider a situation where the conditional measure of location of some random variable *Y*, given *X*, is given by

$$M\left(Y \mid X\right) = g\left(X\right) \tag{1}$$

where $g(X)$ some unknown function. As is evident, a common strategy is to assume $g(X) = \beta_0 + \beta_1 X$, where β_0 and β_1 are unknown parameters that are typically estimated using ordinary least squares (OLS) regression with the goal of estimating the conditional mean of *Y* given *X*. There are, however, well known

Dr. Wilcox is Professor of Psychology at the University of Southern California. Email him at rwilcox@usc.edu.

concerns with this approach. First, it is often the case that assuming a straight regression line is unsatisfactory, which has led to the derivation of many nonparametric regression estimators, often called smoothers (e.g., Efromovich, 1999; Eubank, 1999; Fan & Gijbels, 1996; Fox, 2001; Green & Silverman, 1993; Györfi, et al., 2002; Härdle, 1990; Hastie & Tibshirani, 1990). Of course, some parametric model might be used to deal with any curvature, but often the more obvious strategies (e.g., include a quadratic term) are not flexible enough in terms of giving a reasonably accurate approximation of the true regression line.

Another concern with least squares regression, as well as the bulk of the smoothers that have been derived, is that they are designed to estimate the conditional mean of *Y*, one concern being that the population mean is not robust in the general sense summarized, for example, by Hampel et al., (1986), Huber and Ronchetti (2009), Staudte and Sheather (1990). (The population mean has an unbounded influence function and its breakdown point is zero.) A related concern is that even a single outlier can highly influence the sample mean, which in turn can give a distorted view of the typical value of *Y* given *X*. Cleveland (1979) derived a smoother (generally known as LOESS) aimed at estimating the conditional mean of *Y* and suggested how it might be modified to handle outliers among the dependent variable. Another robust approach is the running interval smoother in Wilcox (2012). It is more flexible than LOESS in the sense that virtually any robust measure of location can be used. For example, it is easily applied when the goal is to estimate the conditional median, trimmed mean or M-estimator of *Y*. It also can be used to estimate any quantile of interest.

A fundamental goal is estimating the strength of an association given a fit to data. An approach when using any smoother is to use some robust version of explanatory power (e.g., Wilcox, 2012). Explanatory power is

$$\xi^2 = \frac{\tau^2\left(\hat{Y}\right)}{\tau^2\left(Y\right)'}$$

where τ^2 is some measure of variation and $\hat{Y}$ is the predicted value of *Y* based on some fit to the data. The square root of explanatory power is called the explanatory strength of the association. To put ξ^2 in perspective, if $\hat{Y}$ is based on the OLS regression line and τ^2 is taken to be the usual variance, ξ^2 reduces to R^2, the usual coefficient of determination.

Estimating explanatory power would seem to be straightforward. Given a random sample (X_i, Y_i), $i = 1, \cdots n$, let $\hat{Y}_i$ be the predicted value of Y given that $X = X_i$. Let $\hat{\tau}^2\left(\hat{Y}\right)$ be an estimate of $\tau^2(\hat{Y})$ based on $\hat{Y}_1, \cdots, \hat{Y}_n$ and let $\hat{\tau}^2\left(Y\right)$ be an estimate of $\tau^2(Y)$ based on $Y_1, \cdots, Y_n$. The an estimate of explanatory power is simply

$$\hat{\xi}^2 = \frac{\hat{\tau}^2\left(\hat{Y}\right)}{\hat{\tau}^2\left(Y\right)} \qquad (2)$$

This approach seems to perform reasonably well when using LOESS, but when using the running interval smoother, it performs poorly: it can be severely biased (Wilcox, 2008). The goal in this paper is to suggest another estimation method that gives substantially better results.

The next section describes the details of the proposed estimation method. The following section reports simulation results comparing the new estimator to the estimator studied in Wilcox (2008). The final section illustrates the new method using data from the Well Elderly 2 study.

The Proposed Method

The measure of location used here is a 20% trimmed mean. For $Y_1, \cdots, Y_n$ the sample 20% trimmed mean is

$$\frac{1}{n-2g}\sum_{i=g+1}^{n-g} Y_{(i)}$$

where $g = .2n$ rounded down to the nearest integer and $Y_{(1)} \leq \cdots \leq Y_{(n)}$ are the values $Y_1, \cdots, Y_n$ written in ascending order. The 20% trimmed mean has nearly the same efficiency as the mean under normality, but it continues to have high efficiency, relative to the usual sample mean, when sampling from heavy-tailed distributions.

The measure of variation that is used is the 20% Winsorized variance. For $i = 1, \cdots, g$, let $W_i = Y_{(g+1)}$. For $i = g + 1, \cdots, n - g$, let $W_i = Y_{(i)}$ and for $i = n - g + 1, \cdots, n$ let $W_i = Y_{n-g}$. Then the Winsorized variance is just the usual sample variance based on the Winsorized values $W_1, \cdots, W_n$.

The running-interval smoother is applied as follows. For some constant *f*, declare *x* to be close to X_i if

$$\left|X_i - x\right| \leq f \times MADN$$

where $MADN = MAD/.6745$, *MAD* is the median of the values.

$|X_1 - M|, \cdots, |X_n - M|$ and *M* is the usual sample median of the X_i values. Let $N(X_i) = \{\ j{:}|X_j - X_i| \leq f \times MADN\ \}$. That is, $N(X_i)$ indexes the set of all X_j values that are close to X_i. Then $M(Y \mid X_i)$ is taken to be some measure of location based on all Y_j values such that $j \in N\left(X_i\right)$ and here, a 20% trimmed mean is used. It appears that often a good choice for the span, *f*, is $f = 1$ (e.g., Wilcox, 2012) and this value is used here.

Method M1

Letting $\hat{Y}_i = M\left(Y \mid X_i\right)$ based on the running interval smoother just described, method M1 consists of simply computing (2) using the Winsorized variance.

Method M2

Method M2 differs from method M1 in two fundamental ways. First, $\hat{Y}_i$ is based on a leave-one-out cross-validation approach in conjunction with the running interval smoother. That is, $\hat{Y}_i$ in method M1 is replaced by $\breve{Y}_i = M\left(Y \mid X_i\right)$, which is based on $(X_1,Y_1), \cdots, (X_n,Y_n)$, ignoring the point (X_i,Y_i) rather than using all *n* points. For notational convenience, let T_i be the trimmed mean of $Y_1, \cdots, Y_n$, excluding Y_i. The other difference, compared to method M1, is that the estimate of explanatory power is replaced by

$$\breve{\xi}^2 = \frac{\tau^2\left(T_1,\cdots,T_n\right) - \tau^2\left(\breve{Y}_1,\cdots,\breve{Y}_n\right)}{\tau^2\left(T_1,\cdots,T_n\right)} \tag{3}$$

Note that (3) mimics a standard way of writing the coefficient of determination. That is, it reflects the proportion of variation accounted for by the dependent variable and the fit obtained by the running interval smoother.

Simulation Results

Simulations were used to compare the bias and mean squared error of methods M1 and M2 when estimating ξ. For the first set of simulations data were generated from the model $Y = \frac{1}{\sqrt{3}} X + e$. The true value of ξ^2 was determined by noting that $\xi^2 = \tau_x^2 / \left(\tau_x^2 + \tau_e^2\right)$, in which case the explanatory strength of the association is $\xi = .5$. The sample size is taken to be 50. Both X and e were taken to have one of four g-and-h distributions, which contain the standard normal distribution as a special case. More precisely, if Z has a standard normal distribution, then

$$\begin{aligned} W &= \frac{\exp(gZ)-1}{g} \exp\left(h\frac{Z^2}{2}\right), \text{if } g > 0 \\ &= Z\exp\left(h\frac{Z^2}{2}\right), \text{ if } g = 0 \end{aligned}$$

has a g-and-h distribution where g and h are parameters that determine the first four moments. The four distributions used here were the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution ($h = 0.2$, $g = 0.0$), an asymmetric distribution with relatively light tails ($h = 0.0$, $g = 0.2$), and an asymmetric distribution with heavy tails ($g = h = 0.2$). Table 1 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. More properties of the g-and-h distribution are summarized by Hoaglin (1985).

Table 1. Some properties of the g-and-h distribution

g	h	κ_1	κ_2
0.0	0.0	0.00	3.00
0.0	0.2	0.00	21.46
0.2	0.0	0.61	3.68
0.2	0.2	2.81	155.98

Let $\hat{\xi}_1$ and $\hat{\xi}_2$ be the estimates of ξ based on methods M1 and M2, respectively. Bias was measured with $E\left(\hat{\xi}_j - \xi\right)$, $j = 1, 2$. To add perspective, bias also was measured with the median difference. The accuracy of the estimators was also measured with mean squared error, $E\left(\hat{\xi}_j - \xi\right)^2$, as well as the median squared error.

Table 2 shows the estimated bias when $n = 100$ and $Y = \beta X + e$ for three choices of the slope: 0, .5 and 1. As can be seen, generally M2 is less biased, and in various situations substantially so despite the reasonably large sample size. Note that the bias associated with M1 can be quite severe, the estimates being approximately −.2 in some cases.

Table 2. Estimated mean bias and median bias, $Y = \beta X + e$, $n = 100$

g	h	β	*mean bias*		*median bias*	
			M1	***M2***	***M1***	***M2***
0.0	0.0	0.0	.110	.081	.101	.000
0.0	0.2	0.0	.115	.078	.104	.000
0.2	0.0	0.0	.110	.085	.101	.000
0.2	0.2	0.0	.115	.082	.105	.000
0.0	0.0	0.5	-.140	-.099	-.139	-.065
0.0	0.2	0.5	-.178	-.072	-.178	-.035
0.2	0.0	0.5	-.144	-.108	-.142	-.070
0.2	0.2	0.5	-.179	-.081	-.138	-.045
0.0	0.0	1.0	-.132	-.074	-.129	-.057
0.0	0.2	1.0	-.197	-.059	-.197	-.039
0.2	0.0	1.0	-.139	-.077	-.134	-.057
0.2	0.2	1.0	-.201	-.064	-.200	-.047

Table 3 reports the estimated squared error. Method M2 does not dominate. But M1 never offers a striking advantage, while in some situations M2 is substantially better.

Tables 4 and 5 report the estimated bias and squared error loss when $Y = .5X^2 + e$. In terms of bias, the advantage of M2 over M1 is even more striking compared to the results in Table 2. Also, in terms of both the mean and median squared error, all indications are that M2 performs better than M1.

Table 3. Estimated mean squared error (MSE) and median squared error (MEDSE), *Y* = *βX* + *e*, *n* = 100

g	h	β	*MSE*		*MEDSE*	
			M1	*M2*	*M1*	*M2*
0.0	0.0	0.0	.016	.021	.010	.000
0.0	0.2	0.0	.017	.021	.011	.000
0.2	0.0	0.0	.016	.023	.010	.000
0.2	0.2	0.0	.018	.022	.011	.000
0.0	0.0	0.5	.019	.044	.009	.011
0.0	0.2	0.5	.030	.038	.018	.011
0.2	0.0	0.5	.020	.048	.010	.012
0.2	0.2	0.5	.031	.040	.019	.011
0.0	0.0	0.7	.024	.018	.017	.005
0.0	0.2	0.7	.047	.017	.039	.004
0.2	0.0	0.7	.026	.019	.018	.005
0.2	0.2	0.7	.049	.019	.040	.005

Table 4. Estimated mean bias and median bias, *Y* = .5X^2 + *e*, *n* = 100

g	h	*mean bias*		*median bias*	
		M1	*M2*	*M1*	*M2*
0.0	0.0	-.201	-.085	-.208	-.050
0.0	0.2	-.182	-.015	-.191	.025
0.2	0.0	-.203	-.067	-.210	-.036
0.2	0.2	-.182	.004	-.190	.043

Table 5. Estimated mean squared error (MSE) and median squared error (MEDSE), *Y* = .5X^2 + *e*, *n* = 100

g	h	*MSE*		*MEDSE*	
		M1	*M2*	*M1*	*M2*
0.0	0.0	.045	.037	.043	.013
0.0	0.2	.039	.036	.036	.025
0.2	0.0	.046	.036	.044	.012
0.2	0.2	.040	.035	.036	.016

An Illustration

The Well Elderly 2 study (Clark et al., 2012; Jackson et al., 2009) was generally concerned with assessing the efficacy of an intervention strategy aimed at improving the physical and emotional health of older adults. One goal was to determine the association between the cortisol awakening response (CAR) and a measure of depressive symptoms after intervention. CAR is defined to be the change in cortisol concentration that occurs during the first hour after waking from sleep. Extant studies (e.g., Clow et al., 2004; Chida & Steptoe, 2009) indicate that various forms of stress are associated with the CAR.

Simply using Pearson's correlation yields $r = .07$, which is not significant at the .05 level when using Student's t test ($p = .22$). There are outliers suggesting the use of some robust generalization of Pearson's correlation. The skipped correlation in Wilcox (2012, section 9.4.3) is estimated to be .07. Kendall's tau and Spearman's rho are .038 and .057, respectively. So all of these correlation coefficients fail to detect any association and suggest that any association that might exist is relatively weak. However, a test of the hypothesis that the regression line is straight (using the method in Wilcox, 2012, section 11.6.1) is significant ($p < .001$). Based on method M1, the strength of the association is estimated to be .12 compared to .31 using method M2.

Concluding Remarks

It is not being suggested that better-known correlation coefficients should be abandoned in favor of method M2. If, for example, a correct parametric model has been specified, under normality Pearson's correlation provides a more accurate estimate of the true association in terms of both bias and mean squared error. A difficulty is that no single estimator dominates and the optimal estimator depends in part on the true nature of the association, which of course is unknown. If, for example, a smoother suggests that the regression line is reasonably straight, and if outliers do not appear to be a serious issue, Pearson's correlation seems reasonable. But it can be difficult determining whether some specified parametric model is sufficiently accurate to justify using something other than method M2. In the illustration, for example, the hypothesis of a straight line was rejected. But even if this hypothesis is not rejected, there is the issue of whether the test of the hypothesis that the regression line is straight has enough power to justify assuming a straight line when estimating the strength of the association. Strategies for deciding which estimator to use, or how to resolve any discrepancies among the estimators that are used, are in need of further study.

The running interval smoother can be used when there are two or more independent variables. A few simulations were run with two independent variables yielding results similar to those reported in Tables 2 and 3. But a more extensive investigation is in order.

References

Chida, Y. & Steptoe, A. (2009). Cortisol awakening response and psychosocial factors: A systematic review and meta-analysis. *Biological Psychology, 80*, 265-278.

Clark, F., Jackson, J., Carlson, M., Chou, C.-P., Cherry, B. J., Jordan-Marsh, M., … Azen, S. P. (2012). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2 Randomised Controlled Trial. *Journal of Epidemiology and Community Health, 66*, 782-790. doi:10.1136/jech.2009.099754

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association, 74*, 829-836.

Clow, A., Thorn, L., Evans, P. & Hucklebridge, F. (2004). The awakening cortisol response: Methodological issues and significance. *Stress, 7*, 29-37.

Efromovich, S. (1999). *Nonparametric curve estimation: Methods theory and applications*. New York: Springer-Verlag.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. New York: Marcel Dekker.

Fan, J. & Gijbels, I. (1996). *Local polynomial modeling and its applications*. Boca Raton, FL: CRC Press.

Fox, J. (2001). *Multiple and generalized nonparametric regression*. Thousand Oaks, CA: Sage.

Green, P. J. & Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: A roughness penalty approach*. Boca Raton, FL: CRC Press.

Györfi, L., Kohler, M., Krzyzk, A. & Walk, H. (2002). *A distribution-free theory of nonparametric regression*. New York: Springer Verlag.

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

Hardle, W. (1990). Applied nonparametric regression. *Econometric Society Monographs No. 19*, Cambridge, UK: Cambridge University Press

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 distributions. In D. Hoaglin, F. Mosteller and J. Tukey (Eds.) *Exploring Data Tables, Trends, and Shapes.* (pp. 461-515). New York: Wiley.

Huber P and Ronchetti E. M. (2009). *Robust statistics.* 2nd Ed. New York: Wiley.

Jackson, J., Mandel, D., Blanchard, J., Carlson, M., Cherry, B., Azen, S., … Clark, F. (2009). Confronting challenges in intervention research with ethnically diverse older adults: the USC Well Elderly II trial. *Clinical Trials, 6*, 90-101.

Staudte, R. G. & Sheather S. J. (1990). *Robust estimation and testing*. Wiley: New York.

Wilcox, R. R. (2008). Estimating explanatory power in a simple regression model via smoothers. *Journal of Modern Applied Statistical Methods*, *7*(2), 368-375. http://digitalcommons.wayne.edu/jmasm/vol7/iss2/2/

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. New York: Elsevier.