11-1-2008

Estimating Explanatory Power in a Simple Regression Model Via Smoothers

Rand R. Wilcox

University of Southern California, rwilcox@usc.edu

Follow this and additional works at: http://digitalcommons.wayne.edu/jmasm

Part of the Applied Statistics Commons, Social and Behavioral Sciences Commons, and the Statistical Theory Commons

Recommended Citation

INVITED ARTICLES
Estimating Explanatory Power in a Simple Regression Model Via Smoothers

Rand R. Wilcox
University of Southern California

Consider the regression model $Y = \gamma(X) + \varepsilon$, where $\gamma(X)$ is some conditional measure of location associated with Y, given X. Let $\hat{Y}$ be some estimate of Y, given X, and let $\tau^2(Y)$ be some measure of variation. Explanatory power is $\eta^2 = \tau^2(\hat{Y}) / \tau^2(Y)$. When $\gamma(X) = \beta_0 + \beta_1 X$ and $\tau^2(Y)$ is the variance of Y, $\eta^2 = \rho^2$, where ρ is Pearson's correlation. The small-sample properties of some methods for estimating a robust analog of explanatory power via smoothers is investigated. The robust version of a smoother proposed by Cleveland is found to be best in most cases.

Key words: strength of association, smothers, effect size, robust methods and nonparametric regression.

Introduction

Consider the simple, nonparametric regression model

$$Y = \gamma(X) + \varepsilon, \quad (1)$$

where X and ε are independent random variables, and $\gamma(X)$ is some unknown function that represents some conditional measure of location associated with Y given X. A fundamental goal is measuring the strength of the association between Y and X. Certainly the best-known approach is to assume

$$\gamma(X) = \beta_0 + \beta_1 X,$$

estimate β_0 and β_1 via ordinary least squares, and then use ρ^2, where ρ is Pearson's correlation. It is well known that Pearson's correlation is not robust (e.g., Wilcox, 2005) and can yield a highly misleading sense about the strength of the association among the bulk of the points. Yet another concern is the assumption that the regression line is straight. Situations are encountered where this assumption seems to be a reasonable approximation of reality, but experience with nonparametric regression methods (e.g. Efromovich, 1999; Eubank, 1999;

Rand R. Wilcox (rwilcox@usc.edu) is Professor of Psychology. He is the author of seven textbooks on statistics, the most recent of which is *Basic Statistics: Understanding Conventional Methods and Modern Insights* (2009, New York, Oxford University Press)

Fan & Gijbels, 1996; Fox, 2001; Green & Silverman, 1993; Gyofri et al., 2002; Hardle, 1990; Hastie & Tibshirani, 1990), sometimes called smoothers, suggest that it is common to encounter situations where this is not the case.

Let $\hat{Y}$ be some estimate of Y given X, and let $\tau^2(Y)$ be some measure of variation associated with the marginal distribution of Y. Then a general approach to measuring the strength of the association between Y and X, called explanatory power, is

$$\eta^2 = \frac{\tau^2(\hat{Y})}{\tau^2(Y)} \quad (2)$$

(e.g., Doksum, Blyth, Bradlow, Meng, & Zhao, 1994; Wilcox, 2003, p. 506). If it is assumed that the conditional distribution of Y given X has the form

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

where $E(\varepsilon) = 0$, and if τ^2 is taken to be the usual variance, $\eta^2 = \rho^2$. It is well-known, however, that the usual variance and Pearson's correlation are not robust. Roughly, small changes in any distribution can substantially alter ρ resulting in a potentially misleading sense about the strength of the association among the bulk of the points. In particular, slight departures from normality can be a practical concern when interpreting ρ.

A simple method for robustifying η^2 is to take τ^2 to be some robust measure of variation. Many such measures have been proposed, comparisons of which are reported by Lax (1985). Based on efficiency, Lax concludes that two so-called A-estimators are best, one of which corresponds to the percentage bend midvariance that was studied by Shoemaker and Hettmansperger (1982). It can be designed to have a reasonably high breakdown point, its efficiency compares well to the usual sample variance under normality, and its standard error can be substantially smaller than the standard error of the sample variance when sampling from a heavy-tailed distribution. For these

reasons it is used here, but this is not to suggest that all other measures of variation have no practical value for the problem at hand.

In addition to many robust measures of variation, there are many nonparametric regression methods that might be used when trying to deal with curvature. Here, no attempt is made to examine all such methods when estimating explanatory power, but rather to consider a few methods that appear to deserve serious consideration, with the goal of finding one method that performs well over a fairly broad range of situations when the sample size is small. In particular, three estimates of η^2 are considered that are based on three nonparametric regression estimators: the robust version of the method in Cleveland (1979), a particular version of a kernel regression estimator derived by Fan (1993), and the running interval smoother (e.g., Wilcox, 2003, section 11.4.4). Consideration was given to a variation of the running interval smoother based on bootstrap bagging (e.g., Buhlmann & Yu, 2002), but it performed rather poorly in the simulations reported here, so further details are omitted.

To add perspective, some results are included assuming

$$\gamma(X) = \beta_0 + \beta_1 X$$

with β_0 and β_1 estimated using the robust method derived by the Theil (1950) and Sen (1968) as well as the ordinary least squares estimator. Of course, when there is curvature, any method that assumes

$$\gamma(X) = \beta_0 + \beta_1 X$$

has the potential to perform poorly. The issue here is how much is sacrificed when a nonparametric estimate of the regression line is used and the regression line is indeed straight. As is well known, there are many robust alternatives to the Theil-Sen estimator that have excellent theoretical properties. The Theil-Sen estimator is used because, in terms of efficiency, it seems to perform about as well as the ordinary least squares (OLS) estimator when the error term has a normal distribution, and it continues

to perform well in situations where OLS performs poorly (e.g., Wilcox, 2005). If the regression line is straight, perhaps there is some practical advantage to using some other robust estimator, but this issue is not addressed here. The primary goal is to consider methods that can be used when curvature might exist. Although not considered here, another well-known approach to nonparametric regression is based on what are called splines, and so for completeness, some comments seem in order. Some informal comparisons with other smoothers suggest that sometimes splines are not quite as satisfactory as other methods (Hardle, 1990; Wilcox, 2005). For this reason, they are not considered, but in fairness, it seems that an extensive formal comparison with the regression methods used here has not been made.

An attempt could be made to fit a parametric model in a manner that takes into account curvature, but simulating this process is difficult. The results reported here suggest that, even when fitting a correct parametric model, little is gained relative to method C, which is described below.

Methodology

The Percentage Bend Midvariance

The objective now is to summarize how the percentage bend midvariance measure of dispersion is computed. For a recent summary of how this measure of dispersion compares to other robust measures of variation, see Wilcox (2005, section 3.12). Let $X_1, \ldots, X_n$ be a random sample. For some β satisfying $0 < \beta < .5$, compute $(1-\beta)n+.5$, round the result to the nearest integer, and label the result m. The choice $\beta = .1$ results in good efficiency under normality, but a relatively low breakdown point. That is, with $\beta = .1$, only 10% of the observations have to be changed to destroy this measure of dispersion. Accordingly, $\beta = .2$ is used. Let $W_i = |X_i - M|, i = 1, \ldots, n$, and let $W_{(1)} \le \ldots \le W_{(n)}$ be the W_i values written in ascending order. Set $\hat{\omega}_\beta = W_{(m)}$, and

$$U_i = \frac{X_i - M}{\hat{\omega}_\beta}.$$

Let $a_i = 1$ if $U_i < 1$; otherwise $a_i = 0$. The estimated percentage bend midvariance is

$$\tau^2 = \frac{n\omega_\beta^2 \sum \psi^2(U_i)}{(\sum a_i)^2}, \quad (3)$$

where $\psi(x) = \max[-1, \min(1, x)]$.

Fan's Kernel Regression Estimator

The first of the nonparametric regression methods considered here stems from Fan (1993). $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample of n points. Let $K(u)$ be the Epanechnikov kernel given by

$$K(u) = \frac{3}{4}(1 - \frac{1}{5}u^2)/\sqrt{5}$$

If $|u| < \sqrt{5}$; otherwise $K(u) = 0$. Let $h = \min(s, IQR/1.34)$, where s is the standard deviation of the X values and IQR is the interquartile range. Bjerve and Doksum (1993) take $h = s$, but it is well known that a robust measure of variation, such as the interquartile range, can have practical value when using a kernel density estimator (e.g., Silverman, 1986).

There is the issue of how to estimate IQR. Many quantile estimators have been proposed, comparisons of which were made by Parrish (1990) as well as Dielman, Lowry, and Pfaffenberger (1994). Here the interquartile range is estimated via the so-called ideal fourths (Frigge, Hoaglin, & Iglewicz, 1989). Perhaps some alternative quantile estimator offers a practical advantage for the problem at hand, but this goes beyond the scope of this paper.

To be more precise, the ideal fourths are computed as follows. Let $X_{(1)} \le \ldots \le X_{(n)}$ be the observations written in ascending order. Estimates of the lower quartile typically have the form

$$q_1 = (1-\ell)X_{(j)} + \ell X_{(j+1)}$$

The ideal fourths are computed by taking j to be the integer portion of (n/4)+(5/12) and

$$\ell = \frac{n}{4} + \frac{5}{12} - j$$

The estimate of the upper quartile is taken to be $q_2 = (1-\ell)X_k + \ell X_{(k-1)}$ where k=n-j+1, in which case the interquartile range is estimated with $IQR = q_2 - q_1$. Let $m(x) = E(Y | X = x)$. Then $m(x)$ is estimated with $\hat{m}(x) = b_0 + b_1 x$, where b_0 and b_1 are determined via weighted least squares with weights $w_i = K((X_i - x)/h)$. This will be called method F.

Cleveland's Method

To outline Cleveland's method, for any x, let $\delta_i = |X_i - x|$. Sort the δ_i values and retain the κn pairs of points that have the smallest δ_i values, where κ is some number between 0 and 1 and is called the span. Let

$$Q_i = \frac{|x - X_i|}{\delta_m}, \text{ and if } 0 \le Q_i < 1, \text{ set}$$

$w_i = (1 - Q_i^3)^3$, otherwise $w_i = 0$. Next, use weighted least squares to estimate m(x) using w_i as weights.

Cleveland (1979) also discussed a robustified version of this method, which is used here. In effect, extreme Y values get little or no weight, and so they have little or no impact on the smooth. (An outline of these additional computations can also be found in Hardle, 1990, p. 192.) Both R and S-PLUS provide access to a function, called lowess, which performs the computations, and the R version was used in the simulations reported here using the default value κ =.75. This will be called method C.

The Running-Interval Smoother

Finally, the so-called running interval smoother was considered. For some constant f, declare x to be close to X_i if

$$|X_i - x| \le f \times MADN,$$

where MADN=MAD/.6745, MAD is the median of the values $|X_1 - M|, \ldots, |X_n - M|$, and M is the usual sample median of the X_i values. Let

$N(X_i) = \{j :| X_j - X_i | \le f \times MADN\}$. That is, $N(X_i)$ indexes the set of all X_j values that are close to X_i. Then m(X_i) is taken to be some measure of location based on all Y_j values such that $j \in N(X_i)$. Here, a 20% trimmed mean is used. It 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. It appears that often a good choice for the span, f, is f=1 (e.g., Wilcox, 2005) and this value is used here. However, results in the next section indicate that this choice can be relatively ineffective for the problem at hand; a smaller value for f seems to be desirable, at least with small sample sizes. But even now, all indications are that Cleveland's method gives superior results. This will be called method R.

The Theil-Sen Estimator

This section reviews how the Theil-Sen estimator is computed. Let X_i and X_j be any two X values such that $X_i > X_j$. Denote the slope corresponding to the two points (X_i, Y_i) and (X_j, Y_j) by b_{1ij}. The median of all such slopes is the Theil-Sen estimate of β_1 and is labeled b_{1ts}. The intercept is estimated with $b_{0ts} = M_y - b_{1ts}M_x$, where M_y and M_x are the sample medians corresponding to the Y and X values, respectively. Estimation of explanatory power via the Theil-Sen estimator will be called method TS.

Estimating Explanatory Power

Based on the regression estimators just described, explanatory power is estimated in an obvious way. For each X_i, compute $\hat{Y}_i$, the estimate of Y given that $X = X_i$. Then explanatory power is estimated with

$$\hat{\eta}^2 = \frac{\hat{\tau}^2(\hat{Y})}{\hat{\tau}^2(Y)},$$

where $\hat{\tau}^2(Y)$ indicates the estimated percentage bend midvariance based on $Y_1, \ldots, Y_n$.

Results

Simulations were used to the check the small sample properties of the methods just described Here, two types of regression lines are considered: $Y = X + \varepsilon$ and $Y = X^2 + \varepsilon$. In both cases, bias was found to be an important issue, as will be seen. It is noted that additional simulations were run with $Y = \varepsilon$, in which case $\eta^2 = 0$, again bias is an issue, but for brevity, no additional details are given. For $Y = X^2 + \varepsilon$, no results are reported when using OLS and method TS, since they are based on the assumption that $Y = \beta_0 + \beta_1 X + \varepsilon$ and are clearly unsatisfactory when in fact $Y = X^2 + \varepsilon$. Both X and ε are assumed to have one of four g-and-h distributions (Hoaglin, 1985), which contains the standard normal distribution as a special case. If Z has a standard normal distribution, and if $g > 0$, then

$$W = \frac{\exp(gZ) - 1}{g} \exp(hZ^2 / 2)$$

has a g-and-h distribution where g and h are parameters that determine the first four moments. If $g = 0$, this last equation is taken to be $W = Z \exp(hZ^2 / 2)$. The four distributions were the standard normal ($g = h = 0$), a symmetric heavy-tailed distribution (h=.5, g=0), an asymmetric distribution with relatively light tails (h=0, g=.5), and an asymmetric distribution with heavy tails (g=h=.5). Table 1 shows the theoretical skewness (κ_1) and kurtosis (κ_2) for each distribution considered. When h=.5, the fourth moment is not defined and the value for κ_2 is left blank. Additional 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	0.0	3.0
0.0	0.5	0.00	---
0.5	0.0	0.61	9.7
0.5	0.5	2.81	---

There remains the problem of determining the population value of η^2 when and ε have some specified distribution. First consider the case $Y = X + \varepsilon$, where both X and ε are assumed to have one of four g-and-h distributions previously described. Then the correct estimate of Y is $\hat{Y} = X$, in which case $\tau^2(\hat{Y}) = \tau^2(X)$, which was determined by randomly sampling n=100,000 observations from the distribution under consideration. As for $\tau^2(Y)$, the following process was used. First generate 5000 values for both ε and X, which yields 5000 values for Y. Computing τ based on these 5000 values yields an estimate of τ. Here, this process was repeated 5000 times, and the average of the resulting τ values is taken to be the population value of $\tau^2(Y)$. And of course, having determined both $\tau^2(\hat{Y})$ and $\tau^2(Y)$, η^2 is taken be $\tau^2(\hat{Y}) / \tau^2(Y)$. As for the case $Y = X^2 + \varepsilon$, the same process was used. For $Y = X + \varepsilon$, the values of η^2 were found to be .499, .409, .338, .314 corresponding to (g,h)=(0,0), (.5,0), (0,.5) and (.5,.5), respectively. As for $Y = X^2 + \varepsilon$, the values were found to be .323, .242, .365 and .330.

Each replication in the simulations consisted of generating n values for X, another n values for ε, computing $Y = X + \varepsilon$ or $Y = X^2 + \varepsilon$, and then applying the estimators described in the previous section. Two sample sizes were considered: n=30 and 100. Here, X and ε have the same g-and-h distribution.

Table 2: Estimated bias

g	h	TS	C	F	R	OLS
			$Y = X + \varepsilon$			
0.0	0.0	.017	.007	-.005	-.109	.019
0.5	0.0	.028	.040	.021	-.052	.094
0.0	0.5	.042	.047	.396	-.015	.158
0.5	0.5	.045	.050	.313	.013	.207
			$Y = X^2 + \varepsilon$			
0.0	0.0	---	.022	.009	-.112	---
0.5	0.0	---	.086	.021	-.019	---
0.0	0.5	---	.084	-.013	-.003	---
0.5	0.5	---	.121	.047	.077	---

Table 3: Estimated squared standard error

g	h	TS	C	F	R	OLS
			$Y = X + \varepsilon$			
0.0	0.0	.031	.034	.037	.035	.038
0.5	0.0	.029	.035	.051	.051	.062
0.0	0.5	.035	.039	83.875	.047	.178
0.5	0.5	.034	.040	6.490	.063	2.452
			$Y = X^2 + \varepsilon$			
0.0	0.0	---	.035	.033	.038	---
0.5	0.0	---	.074	.052	.076	---
0.0	0.5	---	.142	.559	.135	---
0.5	0.5	---	.159	1.018	.343	---

This process was repeated 1000 times yielding 1000 estimates of η^2, say $\eta_1^2, \ldots, \eta_{1000}^2$. Bias was estimate with

$$\frac{1}{1000} \sum (\hat{\eta}_i^2 - \eta^2)$$

and the squared standard error of η^2 was estimated with

$$\frac{1}{999} \sum (\eta_i^2 - \bar{\eta}^2)^2 ,$$

where $\bar{\eta}^2 = \sum \eta_i^2 / 1000$. The results are summarized in Tables 2 and 3 for the case n=30.

First consider bias. Method F performs well when the regression line is straight and when both X and ε have symmetric distributions. But when the distributions are skewed, bias can be severe, suggesting that method F be eliminated from consideration. Method R performs reasonably well, except under normality where it performs poorly. Increasing n to 100, it still performs poorly, in terms of bias, for this special case. Only method C has relatively low bias, and it competes well with OLS and method TS, even when the regression line is straight. However, when there is curvature, now the bias of method C is rather high compared to method F. Again, method R is found to be unsatisfactory under normality.

As for the squared standard error of the estimators, Table 3 indicates that method F can be relatively disastrous when the regression line is straight and sampling is from skewed distributions. And for heavy-tailed distributions, OLS does not perform well compared to methods C and R. Method R competes reasonably well with method C, but there are obvious exceptions. Generally, method C performed best among the situations considered.

To provide some sense of how method C improves when $Y = X^2 + \varepsilon$, as n gets large, some additional simulations were run with n=100 for the cases (g, h)=(0.0, 0.5) and (0.5, 0.5). Now the bias of method C was estimated to be .088 and .080, respectively. So for the skewed, heavy-tailed distribution considered here, the reduction in bias is substantial, but for the skewed, light-tailed distribution the amount of bias remains about the same. Method F has small bias for these situations, but its squared standard error is relatively high. Method R has about the same amount of bias as method C and a smaller standard error, but because it performs poorly in other situations, it would seem that it should be used with caution.

Conclusion

One limitation of the results reported here is that, when using a smoother, the span was chosen to be a fixed constant that is often used as the default value. Checks made when using method R indicate that a smaller span can improve its performance considerably. However, it remains unknown how best to adjust the span when estimating explanatory power, and even for the adjustments considered here (f=.7 and .5), it was found that method C remains a bit more satisfactory in most situations.

Although method C offers protection against the deleterious effects of outliers among the Y values, it is known that a sufficient number of outliers can negatively affect its performance relative to method R (Wilcox, 2005).

This was one of the main reasons for considering method R and it might explain why method C can be unsatisfactory when there is curvature and when dealing with extremely heavy-tailed distributions. Perhaps in most practical situations this is not an issue, but the extent to which this is true is difficult to determine.

When the usual variance is used, rather than the percentage bend midvariance, results in Doksum and Samarov (1995) suggest estimating explanatory power with r^2, the square of Pearson's correlation, rather than with the ratio of the variances of $\hat{Y}$ and Y. An analog of this approach is to use the percentage bend correlation (Wilcox, 2005, p. 391). Consideration was given to this approach, but it proved to be unsatisfactory in the simulations described here.

Perhaps the most surprising result is that there is little or no advantage to fitting a straight line to the data, versus using something like method C, when in fact the regression line is straight and when using the percentage bend variance. Consequently, method C is recommended for general use.

References

Bjerve, S. & Doksum, K. (1993). Correlation curves: Measures of association as functions of covariate values. *Annals of Statistics*, *21*, 890-902.

Buhlmann, P. & Yu, B. (2002). Analyzing bagging. *Annals of Statistics*, *30*, 927-961.

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

Davison, A. C., Hinkley, D. V. & Young, G. A. (2003). Recent developments in bootstrap methodology. *Statistical Science, 18*, 141-157.

Dielman, T., Lowry, C., & Pfaffenberger, R. (1994). A comparison of quantile estimators. *Communications in Statistics--Simulation and Computation, 23*, 355-371.

Doksum, K., Blyth, S., Bradlow, E., Meng, X.-L., Zhao, H. (1994). Correlation curves as local measures of variance explained by regression. *Journal of the American Statistical Association*, *89*, 571-582.

Doksum, K. A. & Samarov, A. (1995). Nonparametric estimation of global functionals and a measure of the explanatory power of covariates in regression. *Annals of Statistics, 23*, 1443--1473.

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

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

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

Frigge, M., Hoaglin, D. C., & Iglewicz, B. (1989). Some implementations of the boxplot. *American Statistician, 43*, 50-54.

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

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

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 & J. Tukey (Eds.), *Exploring data tables, trends and shapes*. New York: Wiley.

Lax, D. A. (1985). Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions. *Journal of the American Statistical Association*, *80*, 736-741.

Parrish, R. S. (1990). Comparison of quantile estimators in normal sampling. *Biometrics*, *46*, 247-257.

Sen, P. K. (1968). Estimate of the regression coefficient based on Kendall's tau. *Journal of the American Statistical Association*, *63*, 1379-1389.

Shoemaker, L. H. & Hettmansperger, T. P. (1982) Robust estimates and tests for the one- and two-sample scale models. *Biometrika, 69*, 47-54.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. New York: Chapman and Hall.

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Indagationes Mathematicae, 12*, 85-91.

Wilcox, R. R. (2003). *Applying contemporary statistical techniques*. New York: Academic Press.

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. New York: Academic Press.