Journal of Modern Applied Statistical Methods

Volume 16 | Issue 2 Article 3

12-1-2017

The Regression Smoother LOWESS: A Confidence Band That Allows Heteroscedasticity And Has Some Specified Simultaneous Probability Coverage

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

Recommended Citation
Wilcox, R. (2017). The Regression Smoother LOWESS: A Confidence Band That Allows Heteroscedasticity And Has Some Specified Simultaneous Probability Coverage. Journal of Modern Applied Statistical Methods, 16(2), 29-38. doi: 10.22237/jmasm/1509494580

Journal of Modern Applied Statistical Methods
November 2017, Vol. 16, No. 2, 29-38.
doi: 10.22237/jmasm/1509494580

ISSN 1538 − 9472

The Regression Smoother LOWESS: A Confidence Band That Allows Heteroscedasticity And Has Some Specified Simultaneous Probability Coverage

Rand Wilcox
University of Southern California
Los Angeles, CA

Many nonparametric regression estimators (smoothers) have been proposed that provide a more flexible method for estimating the true regression line compared to using some of the more obvious parametric models. A basic goal when using any smoother is computing a confidence band for the true regression line. Let $M(Y|X)$ be some conditional measure of location associated with the random variable Y, given X and let x be some specific value of the covariate. When using the LOWESS estimator, an extant method that assumes homoscedasticity can be used to compute a confidence interval for $M(Y|X = x)$. A trivial way of computing a confidence band is to compute confidence intervals for K covariate values, each having probability coverage $1 - \alpha$. But an obvious concern is that the simultaneous probability coverage can be substantially smaller than $1 - \alpha$. A method is suggested for dealing with this issue that allows heteroscedasticity and simultaneously performs better than the Bonferroni method or the Studentized maximum modulus distribution.

Keywords: nonparametric regression, confidence band, heteroscedasticity

Introduction

Let $M(Y|X)$ be some conditional measure of location associated with the random variable Y, given X. Nonparametric regression estimators provide an approach to estimating $M(Y|X)$ that deal with curvature in a flexible manner beyond the more obvious parametric models that might be used. Numerous nonparametric regression estimators have been derived and their practical importance is well

Rand Wilcox is a professor at the University of Southern California. His main interests are robust methods. Email him at rwilcox@usc.edu.

established (e.g., Hastie & Tibshirani, 1990; Efromovich, 1999; Eubank, 1999; Fan & Gijbels, 1996; Fox, 2000; Green & Silverman, 1993; Györfi et al., 2002; Härdle, 1990; Wilcox, 2017).

The goal in this paper is to examine methods for computing a confidence interval for $M(Y|X)$ based on the smoother derived by Cleveland (1979), generally known as LOWESS, in manner that allows heteroscedasticity and provides simultaneous probability coverage $1 - \alpha$ for K values of the independent variable, where K is relatively large. From a robustness point of view, LOWESS is important because it includes a method that down-weights outliers among the dependent variable Y.

Let x be some specific value for the independent variable X. Assuming homoscedasticity, a method for computing a confidence interval for $M(Y|X = x)$ has been derived (Cleveland et al., 1992), which has been implemented via the R function `lowess`. So it is a trivial matter to compute a $1 - \alpha$ confidence interval for a collection of values for the covariate, say $x_1, \cdots, x_K$. It is evident, however, that the simultaneous probability coverage will, in general, be substantially smaller than $1 - \alpha$. And there is the added concern that when in fact there is heteroscedasticity, an incorrect estimate of the standard error is being used.

Here, heteroscedasticity is addressed with a bootstrap estimate of the standard error of $\hat{\theta}_k$, where $\hat{\theta}_k$ is the estimate of $M(Y|X = x_k)$ ($k = 1, \cdots, K$) based on LOWESS.

A simple way of achieving simultaneous probability coverage greater than or equal to $1 - \alpha$ is to compute a $1 - \alpha/K$ confidence interval for each of K covariate values of interest. That is, use the Bonferroni method. Another strategy is to use the Studentized maximum modulus distribution. But both of these strategies are too conservative meaning that the actual probability coverage will be substantially larger than the nominal level, in which case the widths of the confidence intervals will be larger than necessary. The strategy here is to find an adjustment for the confidence intervals that achieves simultaneous probability coverage under normality and homoscedasticity, and then study how well the method performs when dealing with non-normality and heteroscedasticity. The method for adjusting the confidence intervals has certain similarities to using a Studentized maximum modulus distribution, but it differs in ways that will be fairly evident.

LOWESS

Consider the basic strategy used by LOWESS. Not all of the many computational details are provided here, which are summarized by Cleveland et al. (1992). The main goal is to provide some sense of how the span is used and determined. The choice for the span turns out to play an important role given the goal of computing confidence intervals having simultaneous probability coverage $1 - \alpha$.

Given the goal of estimating $M(Y|X = x_k)$, let $(X_1,Y_1), \cdots, (X_n,Y_n)$ be a random sample and let $\delta_i = |X_i - x_k|$. Next, sort the δ_i values and retain the pn pairs of points that have the smallest δ_i values, where p is a number between 0 and 1. The value of p represents the proportion of points used to predict Y and is generally referred to as the span. Let δ_m be the maximum value of the δ_i values that are retained. Set

$$Q_i = \frac{\delta_i}{\delta_m}$$

If $0 \leq Q_i < 1$, set $w_i = (1 - Q_i^3)^3$, otherwise set $W_i = 0$. Finally, use weighted least squares to predict Y using $w_1, \cdots, w_n$ as weights. Because the weights change with X, generally a different regression estimate of Y is used when the value of X is altered.

There are refinements beyond the computational steps just described, but for the brevity the many details are omitted. As previously noted, the method includes the ability of down weighting outliers among the independent variable Y. The main point here is that the choice for the span, p, will be found to play a crucial role.

Description of the Proposed Method

Let $(X_1^*, Y_1^*), \cdots, (X_n^*, X_n^*)$ be a bootstrap sample, which is obtained by resampling with replacement n points from $(X_1,Y_1), \cdots, (X_n,Y_n)$. Let $\hat{\theta}_k^*$ be the estimate of $M(Y|X = x_k)$ based on this bootstrap sample. Repeat this process B times yielding $\hat{\theta}_{kb}^*$ ($b = 1, \cdots, B$). From basic principles (e.g., Efron & Tibshirani, 1993), an estimate of the squared standard error of $\hat{\theta}_k$ is

$$s_k^2 = \frac{1}{B-1}\sum\left(\hat{\theta}_{kb}^* - \bar{\theta}_k\right)^2,$$

where $\bar{\theta}_k = \sum \hat{\theta}_{kb}^* / B.$.

Here, two strategies for choosing $x_1, \cdots, x_K$ were considered. The first used $K = n$ values based on all of the observed values of the independent variable X.

The adjusted confidence intervals, based on the basic strategy described momentarily, performed well in simulations for some situations, but not others, so this approach was abandoned.

To describe the second strategy, let M be the usual sample median based on $X_1, \cdots, X_n$, let $L = M - 1.5\,MADN$ and $U = M - 1.5\,MAD$, where MAD is the median of

$$|X_1 - M|, \cdots, |X_n - M|$$

and $MADN$ is MAD divided by 0.6745. To add perspective, it is noted that under normality, $MADN$ estimates the standard deviation. Then $x_1, \cdots, x_K$ are taken to be K values evenly spaced between L and U, inclusive. Here the focus is on $K = 25$.

Now focus on a single value of the independent variable, x_k, and note that for some specified constant θ_0,

$$H_0 : \theta_k = \theta_0 \tag{1}$$

can be tested using the test statistic

$$T_k = \frac{\hat{\theta}_k - \theta_0}{s_k}, \tag{2}$$

where the null distribution is taken to be a Student's T distribution with degrees of freedom as indicated by Cleveland et al. (1992), which is computed by the R function `loess`. Let p_k be the resulting p-value, let

$$p_m = \min(p_1, \cdots, p_K)$$

and let p_α be the α quantile of the distribution of p_m. As is evident, if the null hypothesis given by (1) is rejected if and only if $p_k \leq p_\alpha$, the probability of one or

more Type I errors is α. In terms of confidence intervals, if a $1 - p_\alpha$ confidence interval is computed for each θ_k, the simultaneous probability coverage is $1 - \alpha$.

Simulations are used to estimate p_α when dealing with independent standard normal distributions. More precisely, generate n pairs of points from a bivariate normal distribution having correlation zero, perform the K tests as just described, and determine p_m, the minimum p-value among these K tests. This process is repeated N times yielding say $p_{m1}, \cdots, p_{mN}$, in which case p_α can be estimated with some quantile estimator. Here, the Harrell and Davis (1982) estimator is used with $N = 4{,}000$. For convenience, this method for computing confidence intervals will be called method C henceforth.

Shown in Table 1 are some estimates of p_α when $\alpha = 0.05$ and the sample size n ranges between 50 and 2,000. Note that based on the Bonferroni method with $K = 25$, each of the K tests would be performed at the 0.002 level. If, for example, the Studentized maximum modulus distribution is used with fifty degrees of freedom, in effect p_α is taken to be 0.0022. Generally, using the method described here will result in shorter confidence intervals. Roughly, the reason is that the T_k values are highly correlated, which is taken into account when computing p_α. Also note that as n increases, initially the estimates of p_α decrease and then they increase. The reason for this is unclear.

Table 1. Estimates of p_α based on 4,000 replications

n	p_α
30	0.00360
50	0.00266
70	0.00240
100	0.00288
150	0.00300
200	0.00354
300	0.00387
500	0.00440
1000	0.00408
2000	0.00451

Simulation Results

Simulations were used to study the small-sample properties of method C. Data were generated based on the model

$$Y = X^a + \lambda(X)\epsilon$$

for $a = 0$, 1 and 2. Both X and ϵ were generated from one of four types of distributions: normal, symmetric and heavy-tailed, asymmetric and light-tailed, and asymmetric and heavy-tailed. More precisely, both the error term and the distribution of the independent variable were taken to be one of four g-and-h distributions (Hoaglin, 1985) that contain the standard normal distribution as a special case. If Z has a standard normal distribution and $g > 0$, then

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

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

$$W = Z \exp\left(hZ^2/2\right).$$

The four distributions used here were the standard normal ($g = h = 0.0$), an asymmetric 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 2 shows the skewness (κ_1) and kurtosis (κ_2) for each distribution. Additional properties of the g-and-h distribution are summarized by Hoaglin (1985).

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

g	h	κ_1	κ_2
0.00	0.00	0.00	3.00
0.00	0.20	0.00	21.46
0.20	0.00	0.61	3.68
0.20	0.20	2.81	155.98

Table 3 summarizes the simulation results for method C based on $a = 0$, sample sizes 50 and 100, and when the default value for the span is used, namely $p = 2/3$. Shown are estimates of α when the goal is to achieve simultaneous probability coverage $1 - \alpha = 0.95$. Similar results were obtained with $a = 1$ and $a = 2$. Bradley (1978) has suggested that as a general guide, when computing a

0.95 confidence interval, at a minimum the actual probability coverage should be between 0.925 and 0.975. All of the estimates satisfy this basic criterion.

Table 3. Estimates of α when the goal is to achieve simultaneous probability coverage $1 - \alpha = 0.95$.

g	*h*	*n*	*VP*1	*VP*2	*VP*3
0.0	0.0	50	0.050	0.061	0.047
0.0	0.0	100	0.050	0.051	0.043
0.0	0.2	50	0.032	0.035	0.026
0.0	0.2	100	0.029	0.034	0.026
0.2	0.0	50	0.061	0.056	0.049
0.2	0.0	100	0.053	0.067	0.045
0.2	0.2	50	0.055	0.050	0.037
0.2	0.2	100	0.033	0.065	0.046

Method C continues to perform well with $n = 200$ and $n = 300$. But with $n = 500$ and when sampling from a skewed distribution, it can be unsatisfactory when there is heteroscedasticity. That is, the estimates of α exceed 0.075. Increasing the number of bootstrap samples to 400 improved matters in some cases. But what was more effective was reducing the span. For $n = 500$, reducing the span to $p = 0.5$ yielded estimates less than 0.05 for all situations considered. But under normality and homoscedasticity, the estimate was 0.016.

Illustration

Method C is illustrated using data from the Well Elderly 2 study (Clark et al., 2012) that dealt with an intervention program aimed at improving the physical and emotional wellbeing of older adults. A portion of the study focused on the association between the cortisol awakening response (CAR) and a measure of depressive symptoms based on the Center for Epidemiologic Studies Depressive Scale (CESD). CAR refers to the change in cortisol concentration that occurs 30-60 minutes after waking from sleep. A CESD score greater than 15 is regarded as an indication of mild depression. A score greater than 21 indicates the possibility of major depression.

Figure 1 shows the estimate of the regression line as well as a collection of confidence intervals having simultaneous probability coverage approximately equal to 0.95. (Leverage points were removed.) The horizontal dotted line in Figure 1 corresponds to CESD = 15. So Figure 1 indicates that for CAR values between −0.2 and 1.5, after intervention, a reasonable decision is that the typical

participant does not have any indication of mild depression. Outside this interval, it is unclear the extent to which this is the case.

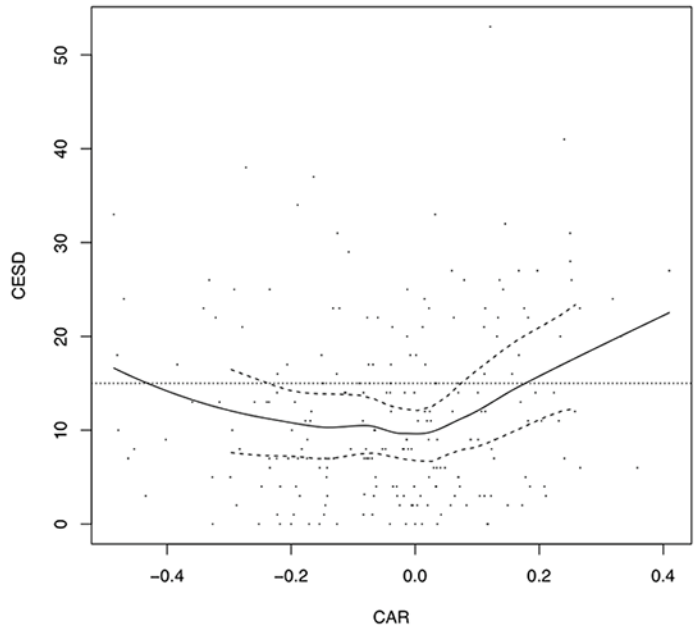

Figure 1. Confidence intervals for the typical CESD score using the CAR as the independent variable. The horizontal dotted line corresponds to CESD = 15. (CESD values greater than 15 are considered an indication of mild depression.)

Conclusion

Method C offers a more satisfactory way of computing confidence intervals compared to the simple approach of computing $1-\alpha$ confidence intervals for each value of the independent variable of interest. The method performed well in simulations, in terms of achieving estimates of α less than 0.075 for $n \leq 500$, provided the span is chosen appropriately. However, there is room for improvement because as the sample size increases, the actual probability coverage becomes increasingly unstable in terms of how the data are generated. Avoiding estimates of α greater than 0.075 can be achieved by choosing the span to be sufficiently small, but at the expense of estimates less than 0.025 when there is normality and homoscedasticity.

A broader issue is whether some variation of method C can be used in conjunction with other smoothers. One of the many smoothers of interest is the running interval smoother (e.g., Wilcox, 2017) because it provides a simple and effective method for dealing with situations where M(Y|X) is any robust measure of location of interest. Preliminary results indicate that an adjustment of the confidence intervals, similar to what was used here, is not straightforward. The details of how best to proceed are under investigation.

Finally, the R function `lplotCI` applies method C and has been added to the library of R functions stored at Dornsife.usc.edu/cf/labs/wilcox/wilcox-faculty-display.cfm in the file Rallfun-v32.

References

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144-152. doi: 10.1111/j.2044-8317.1978.tb00581.x

Clark, F., Jackson, J., Carlson, M., et al. (2012). Effectiveness of a lifestyle intervention in promoting the well-being of independently living older people: results of the Well Elderly 2 Randomise 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*(368), 829-836. doi: 10.2307/2286407

Cleveland, W. S., Grosse, E. & Shyu, W. M. (1992) Local regression models. In J. M. Chambers & T. J. Hastie, Eds. *Statistical Models in S*. New York: Wadsworth & Brooks/Cole (Springer-Verlag), Chapter 8.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. New York: Springer-Verlag. doi: 10.1007/b97679

Efron, B. & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.

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. (2000). *Multiple and Generalized Nonparametric Regression*. Thousands Oaks, CA: Sage. doi: 10.4135/9781412985154

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. doi: 10.1007/b97848

Hall, P. & Sheather, S. J. (1988). On the distribution of a Studentized quantile. *Journal of the Royal Statistical Society, B, 50*, 380-391.

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

Härdle, W. (1990). *Applied Nonparametric Regression (Econometric Society Monographs No. 19)*. Cambridge, UK: Cambridge University Press.

Harrell, F. E. & Davis, C. E. (1982). A new distribution-free quantile estimator. *Biometrika, 69*(3), 635-640. doi: 10.1093/biomet/69.3.635

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

Hettmansperger, T. P. & McKean, J. W. (1998). *Robust Nonparametric Statistical Methods*. London: Arnold.

Hettmansperger, T. P., & Sheather, S. J. (1986). Confidence intervals based on interpolated order statistics. *Statistics & Probability Letters, 4*(2), 75-79. doi: 10.1016/0167-7152(86)90021-0

Hoaglin, D. C. (1985). Summarizing shape numerically: The g-and-h distribution. In D. C. Hoaglin, F. Mosteller & J. W. Tukey (Eds.) *Exploring Data Tables Trends and Shapes*. New York: Wiley, pp. 461-514. doi: 10.1002/9781118150702.ch11

Huber, P. J. & Ronchetti, E. (2009). *Robust Statistics* (2nd Ed.). New York: Wiley. doi: 10.1002/9780470434697

Sheather, S. J. & McKean, J. W. (1987). A comparison of testing and confidence intervals for the median. *Statistics & Probability Letters, 6*(1), 31-36. doi: 10.1016/0167-7152(87)90055-1

Staudte, R. G. & Sheather, S. J. (1990). *Robust Estimation and Testing*. New York: Wiley. doi: 10.1002/9781118165485

Wilcox, R. R. (2017). *Introduction to Robust Estimation and Hypothesis Testing* (4th Ed.). San Diego, CA: Academic Press.