SOME REMARKS ABOUT THE GASSER–SROKA–JENNEN-STEINMETZ VARIANCE ESTIMATOR

BU-1088-MA November, 1992

by

N. S. Altman and C. P. Paulson

SOME REMARKS ABOUT THE GASSER–SROKA–JENNEN-STEINMETZ VARIANCE ESTIMATOR

N. S. Altman and C. P. Paulson

Biometrics Unit, 337 Warren Hall, Cornell University, Ithaca, NY 14853-7801
and
105 N. Sunset Drive, Ithaca, NY 14850

Key Words and Phrases: nonparametric regression; residual variance.

ABSTRACT

This article proposes some simplifications of the residual variance estimator of Gasser, Sroka, and Jennen-Steinmetz (GSJ, 1986) which is often used in conjunction with nonparametric regression. The GSJ estimator is a quadratic form of the data, which depends on the relative spacings of the design points. When the errors are independent, identically distributed Gaussian variables, and the true regression curve is flat, the estimate is distributed as a weighted sum of χ^2 variables. By matching the first two moments, the distribution can be approximated by a χ^2 with degrees of freedom determined by the coefficients of the quadratic form. Computation of the estimated degrees of freedom requires computing the trace of the square of an $n \times n$ matrix, where n is the number of design points. In this article, (n–2)/3 is shown to be a conservative estimate of the approximate degrees of freedom, and (n–2)/2 is shown to be conservative for many designs. In addition, a simplified version of the estimator is shown to be asymptotically equivalent, under many conditions.

1. INTRODUCTION

This article discusses estimation of error variance for the nonparametric regression model

$$y_i = \mu(t_i) + \epsilon_i, \qquad i = 1, \cdots, n$$

where the y_i are the observed data, the t_i are fixed design points, μ is the regression function and the errors ϵ_i are independent and identically distributed with $E(\epsilon) = 0$ and $Var(\epsilon) = \sigma^2$.

A number of nonparametric estimators are available for approximating the regression function, including kernel and nearest neighbor regression (Gasser and Müller,1979), smoothing splines (Silverman, 1985), and local linear regression (Cleveland, 1979; Friedman, 1984).

An estimate of residual variance is often needed for use with the nonparametric regression estimate for a variety of reasons, including selection of smoothing parameters, outlier detection, and construction of confidence and prediction bands. A number of estimators have been suggested based on divided differences of the observed y's (Rice, 1984; GSJ, 1986; Buckley and Eagleson, 1989). The GSJ estimator is often used.

The form of the GSJ estimator is

$$\hat{\sigma}^2 = \frac{1}{n-2}\,\mathbf{y}'\mathrm{D}\mathbf{y}\ ,$$

where the $n \times n$ matrix D is tridiagonal and depends on the spacings of the t's. D is defined explicitly in Section 2. Using an approximation of Box (1954), the distribution of $\hat{\sigma}^2$ can be approximated by a scaled chi-squared distribution with the scaling and degrees of freedom a function of the trace of D^2.

This paper was motivated by the desire to find a simple estimate of residual variance that could be used when introducing nonparametric regression to consulting clients and beginning students in statistics (Altman, 1992). The GSJ estimator is readily computed from the divided difference formula given in Section 2. However, computing the required trace is beyond the skills of most beginning students.

Section 3 shows that for many reasonable designs D may be computed as if the t's were equally spaced. For equally spaced designs, $\mathrm{tr}(D^2) = (n-2)(1+17/18)-1$ for "tr" the matrix trace. Section 4 shows that in general $\mathrm{tr}(D^2)$ is bounded above by $3(n-2)$ and for special designs it is bounded above by $(n-2)(1+17/18)-1$. The use of an upper bound for the actual trace leads to an estimate of degrees of freedom that is somewhat larger than the Box approximation. This results in conservative confidence intervals for the true variance. Normal theory pointwise confidence intervals for the regression function using a Student-t distribution on the estimated degrees of freedom will be narrower than intervals based on the Box approximation, but will be conservative compared to intervals commonly in use which employ critical values from the standard normal distribution (Eubank, 1988, p. 147).

2. DEFINING THE GSJ ESTIMATOR

The GSJ estimator is based on the residual sum of squares computed by local estimation of $\mu(t_i)$ by the line through (t_{i-1}, y_{i-1}) and (t_{i+1}, y_{i+1}). The residual is the difference between y_i and the value of the line at t_i. To be more precise let $\Delta_i = t_{i+1} - t_{i-1}$ for $2 \le i \le n-1$ and define

$$a_i = \begin{cases} \dfrac{t_{i+1}-t_i}{\Delta_i}, & \text{if } \Delta_i \neq 0 , \\ \dfrac{1}{2} , & \text{if } \Delta_i = 0 , \end{cases} \qquad \text{and} \qquad b_i = 1 - a_i .$$

Then the residual is $e_i = a_i y_{i-1} + b_i y_{i+1} - y_i$.

Let $c_i^2 = \left(a_i^2+b_i^2+1\right)^{-1}$. Then $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=2}^{n-1} c_i^2 e_i^2$. Re-expressing $\hat{\sigma}^2$ as $\mathbf{y}'\mathrm{D}\mathbf{y}/(n-2)$, we find that $D = A'C^2A$, where C is the $(n-2) \times (n-2)$ diagonal matrix

$$C_{ii} = c_{i+1}$$

and A is the $(n-2) \times n$ tridiagonal matrix with

$$A_{i,i} = a_{i+1},\ A_{i,i+1} = -1,\ A_{i,i+2} = b_{i+1} .$$

Note that $\mathrm{tr}(D) = n-2$.

GSJ showed that $\hat{\sigma}^2$ is strongly consistent for the true error variance under the following regularity conditions:

A: $\max |t_i - t_{i+1}| = 0(1/n)$.

B: The errors $\{\epsilon_i\}$ are independent and identically distributed, with $E(\epsilon_1) = 0$, $\mathrm{Var}(\epsilon_1) = \sigma^2$ and $E(\epsilon_1^4) < \infty$.

C: The regression function μ is continuous.

In addition, $\hat{\sigma}^2$ is asymptotically normal if

D: $|\mu(t)-\mu(s)| \le M|t-s|^{\alpha}$, with $\alpha > 1/4$.

3. SIMPLIFICATION OF THE GSJ ESTIMATOR

When the design points are equally spaced, $a_i = b_i = 1/2$. In this case, the computations are considerably simpler. Propositions 1 and 2 below show that when the design can be transformed to uniform using a sufficiently smooth monotone function then $\hat{\sigma}^2$ can be computed as if the design were equally spaced. The proofs of Propositions 1 and 2 are clear, and are not given.

Proposition 1: Suppose the design is regular, in the sense that $t_i = h(i/n)$, where h is a strictly monotone, uniformly continuous (respectively, Lipshitz-continuous) function. Then $\mu \circ h$ satisfies C (respectively, D) when μ does, and the conclusions of the theorem of GSJ continue to hold if we take $a_i = b_i = 1/2$.

Proposition 2: Suppose the design converges to a density with strictly monotone, uniformly continuous inverse integral transform, h. Then, for large n, $t_i \approx h(i/n)$, and the conclusions of the theorem of GSJ continue to hold if we take $a_i = b_i = 1/2$.

4. APPROXIMATE DEGREES OF FREEDOM

When the regression function is linear, and the errors are normally distributed, the finite sample distribution of $\hat{\sigma}^2$ is (using the notation of GSJ 1986)

$$\frac{\sigma^2}{(n-2)} \sum_{j=1}^{n-2} \lambda_j \chi_j^2 ,$$

where the χ_j^2 are independent χ^2 variables on 1 degree of freedom and the λ_j are the eigenvalues of D^2 (Box, 1954). By matching the first two moments, the distribution may be approximated by a $\tau^2\chi^2$ distribution on δ degrees of freedom, where $\delta = (n-2)^2/tr(D^2)$ and $\tau^2 = \sigma^2/\delta$. For elementary students, the need to compute $tr(D^2)$ is prohibitive. Other more accurate approximations are given by Solomon and Stephens (1977) and Buckley and Eagleson (1988) but are even more difficult to compute. Remark 1 and Proposition 3 provide some upper bounds on $tr(D^2)$ that can be used to compute, for example, conservative tests and confidence intervals for the estimate of the $\mu(t)$.

Remark 1: When $a_i = \gamma$, for all i,

$$tr(D^2) = (n-2)\{1+\{(1+\gamma^2(1-\gamma)^2\}/\{2(\gamma^2-\gamma+1)^2\}\} - \{1+2\gamma^2(1-\gamma)^2\}/\{2(\gamma^2-\gamma+1)^2\} .$$

In particular, when $\gamma = 1/2$, $tr(D^2)$ has a maximum, $tr(D^2) = (n-2)(1+17/18)-1$. As γ approaches 0 or 1, $tr(D^2)$ approaches a minimum, $tr(D^2) = (n-2)(3/2) - 1/2$. So, for designs of this type, $(n-2)/2 \leq \delta \leq 2(n-2)/3$.

Proposition 3: For any design, $tr(D^2) \leq 25(n-2)/9 + 28/9$. For $n \geq 16$, $tr(D^2) \leq 3(n-2)$, and $\delta \geq (n-2)/3$.

The proof of proposition 3 is in the Appendix.

The three propositions provide a simplification of the estimates of GSJ. For most designs, it is adequate to use weights $a_i = 1/2$. When the design is highly irregular, however, it may be preferable to use the weights suggested by GSJ. In this case, Proposition 3 provides a conservative bound on the degrees of freedom, which is useful for "quick and dirty" computations.

APPENDIX

Proof of Proposition 3:

Let $G = CAA'C$. Then

$$\begin{aligned} \mathrm{tr}(D^2) &= \mathrm{tr}(A'C^2AA'C^2A) \\ &= \mathrm{tr}(CAA'C^2AA'C) \\ &= \mathrm{tr}(G^2) \\ &= \sum_{i=1}^{n-2} \sum_{j=1}^{n-2} G_{ij}^2 \end{aligned}$$

and

$$G_{ij} = \begin{cases} 0\,, & |i-j| > 2\,, \\ 1\,, & i = j\,, \\ -(a_{i+1}+b_i)c_ic_{i+1}\,, & |i-j| = 1\,, \\ a_{i+2}b_ic_ic_{i+2}\,, & |i-j| = 2\,. \end{cases}$$

Since $c_k^2 \le \frac{2}{3}$ for every k, $(c_ic_{i+j})^2 \le \frac{4}{9}$. So

$$G_{ij}^2 \le \tfrac{4}{9}(a_{i+1}+b_i)^2 \text{ for } |i-j| = 1 \quad \text{and} \quad G_{ij}^2 \le \tfrac{4}{9}a_{i+2}^2+b_i^2 \text{ for } |i-j| = 2\,.$$

This gives $\mathrm{tr}(D^2) \le n-2 + \frac{8}{9}\sum_{i=1}^{n-3}(a_{i+1}+b_i)^2 + \frac{8}{9}\sum_{i=1}^{n-4} a_{i+2}^2b_i^2$. We bound the last two terms of this sum as follows. First,

$$\begin{aligned} \sum_{i=1}^{n-3}(a_{i+1}+b_i)^2 &= \sum_{i=2}^{n-2}a_i^2 + \sum_{i=1}^{n-3}b_i^2 + 2\sum_{i=1}^{n-3}a_{i+1}b_i \\ &\le \sum_{i=1}^{n-2}(a_i+b_i)^2 + \sum_{i=1}^{n-3}a_{i+1}b_i + \sum_{i=1}^{n-4}a_{i+2}b_{i+1} + a_2b_1 \\ &\le n-2 + \sum_{i=1}^{n-3}a_{i+1}b_i + \sum_{i=1}^{n-4}a_{i+2}b_{i+1} + a_2b_1\,. \end{aligned}$$

Then $\quad \sum_{i=1}^{n-4}a_{i+2}^2\,b_i^2 \le \sum_{i=1}^{n-4}a_{i+2}b_i\,, \qquad$ since $0 \le a_{i+2} \le 1$ and $0 \le b_i \le 1$.

Let $$U = \sum_{i=1}^{n-3} a_{i+1}b_i + \sum_{i=1}^{n-4} a_{i+2}b_{i+1} + \sum_{i=1}^{n-4} a_{i+2}b_i \text{ and set}$$

$$V = \sum_{i=1}^{n-3} a_i b_{i+1} + \sum_{i=1}^{n-4} a_{i+1}b_{i+2} + \sum_{i=1}^{n-4} a_i b_{i+2} .$$

Use of telescoping sums shows that

$$U-V = -3a_1 - 2a_2 + a_{n-3} + 3a_{n-2} + a_{n-1} + a_n ,$$

while Lemma 1 (below) gives

$$U+V \le 2(n-4) + a_{n-2} + 2a_{n-1} + 2a_n - 2a_{n-3}a_{n-2} ,$$

so that

$$U \le n-4 + \tfrac{1}{2}(-3a_1-2a_2+a_{n-3}+4a_{n-2}+3a_{n-1}+3a_n-2a_{n-3}a_{n-2}) .$$

This gives

$$\begin{aligned} tr(D^2) &\le n-2 + \tfrac{8}{9}(n-2) + \tfrac{8}{9}a_2b_1 + \tfrac{8}{9}U \\ &\le \frac{25(n-2)}{9} + \frac{28}{9} . \end{aligned}$$

For $n \ge 16$,

$$tr(D^2) \le 3(n-2) .$$

Lemma 1: If $0 \le x \le 1$, $0 \le y \le 1$ and $0 \le z \le 1$, then

$$x + y + z - xy - xz - yz \le 1 .$$

Proof: The inequality $(1-x)(1-y)(1-z) \ge 0$ implies $-(x+y+z-xy-xz-yz) \ge xyx-1$ and, hence, that $x + y + z - xy - xz - yz \le 1$.

BIBLIOGRAPHY

Altman, N.S. (1992) "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *Amer. Statist.* **46**, 175-185.

Box, G.E.P. (1954) "Some theorems on quadratic forms applied in the study of analysis of variance problems, I: effect of inequality of variance in the one-way classification," *Ann. Math. Statist.* **25**, 290-302.

Buckley, M.J. and Eagleson, G.K. (1988) "An approximation to the distribution of quadratic forms in normal random variables," *Austral. J. Statist.* **30A**, 150-159.

Buckley, M.J. and Eagleson, G.K. (1989) "A graphical method for estimating the residual variance in nonparametric regression," *Biometrika* **76**, 203-210.

Cleveland, W.S. (1979) "Robust locally weighted regression and smoothing scatterplots," *J. Amer. Statist. Assoc.* **74**, 829-836.

Eubank, R.L. (1988) *Spline Smoothing and Nonparametric Regression,* Marcel Dekker, Inc., New York.

Friedman, J.H. (1984) "A Variable Span Smoother," LCS Technical Report 5, Stanford, Department of Statistics.

Gasser, T. and Müller, H.G. (1979) "Kernel estimation of regression functions," *Smoothing Techniques for Curve Estimation 23-67* Lecture Notes in Math. 757. Springer-Verlag, Berlin.

Gasser, T., L. Sroka and C. Jennen-Steinmetz (1986) "Residual variance and residual pattern in nonlinear regression," *Biometrika*, **73**, 625-33.

Solomon, H. and Stephens, M. (1977) "Distribution of a weighted sum of chi-square variables," *J. Amer. Statist. Assoc.* **72**, 881-885.

Stone, C.J. (1977) "Consistent Nonparametric Regression," *Ann. Statist.* **5**, 595-645.