Coplots, Nonparametric Regression, and Conditionally Parametric Fits

William S. Cleveland
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974
wsc@research.att.com

February 1, 1993

Abstract

A nonparametric regression surface is *conditionally parametric* in a proper subset of the factors if, given the values of the factors not in the subset, the surface is a member of a parametric class as a function of the subset. Conditionally parametric surfaces can provide a parsimonious fit in the common situation where the effects due to some factors are significant but limited. One approach to nonparametric regression is local regression, specifically the local fitting of linear and quadratic polynomials of the factors. Recent work of Fan [10, 11, 13, 12] and of Hastie and Loader [17], has shown that local regression is superior to kernel estimation and modified kernel estimation, two methods that have had extensive theoretical investigation, but that work poorly in practice. Local regression fitting can be modified in a simple way to produce conditionally parametric fits. The conditioning plot, or coplot, is a graphical method that is particularly helpful for carrying out regression studies, in particular, for determining factors that can be taken to be conditionally parametric.

1 Introduction

The International Symposium in Hong Kong

The Symposium on Multivariate Analysis was held at Hong Kong Baptist College in March 1992. One extraordinary aspect was its organization. The myriad tasks of the conference were carried out with by 78 undergraduates and 4 graduate students in maths and stats. What was so striking was the enormous efficiency, enthusiasm, and pride with which they ran the conference. Energy abounded. Figure 1 is a quantile plot — the ith order statistic graphed against $(i - 0.5)/78$ — of the time in hours spent by the undergraduates. The mere existence of the data speaks to the fastidiousness

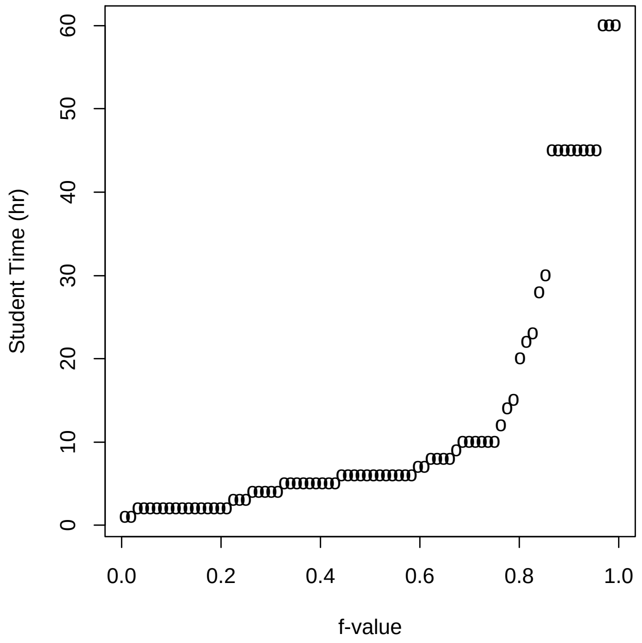

Figure 1: Quantile plot of times spent by undergraduates organizing and running the Symposium.

of the operation. For a citizen of the United States, even a well-traveled one, this was a revelation. Sometimes the march of history — often thought to move to the loud beat of political leaders' machinations — steps to other things, seemly small, but the real drivers of revolutions. In this case it is a quest for knowledge, a penchant for perfection, constancy of purpose, and teamwork. Those in the West have felt their effects acutely as drastically shifting national economies. The Symposium provided the opportunity to see the drivers first hand.

Tutorial

My contribution to the Symposium was a tutorial to a group that — with the exception of one person, a good friend who will go nameless — deserve commendation for staying awake on the first day despite widespread jet lag. The topic, graphical methods for data analysis and computer intensive methods, was based on a forthcoming book, *Visualizing Data* [5]. This paper cannot be a documentation of the tutorial because it makes no sense to rewrite sections of the book. Instead I will discuss an idea used in the book but not treated in depth — conditionally parametric fits. On route to this I will briefly discuss coplots, which are treated in great detail in the book, and some issues of nonparametric regression that have arisen recently.

Contents of the Paper

The paper begins with a data set: measurements of three variables from an industrial experiment [3]. One variable is a response and the other two are factors. The data are presented on a *coplot*, a graphical method that is particularly useful for diagnosing models used in regression analysis. This graphical method is the first topic of the paper.

Many multivariate data sets — for example, the data from the industrial experiment mentioned above — are better fitted by nonparametric regression surfaces than parametric ones because the latter lack the flexibility to track all but very simple patterns. One method of nonparametric regression is *local fitting*. The method is an old idea first used by time series analysts to smooth their data [20]. It was brought to more general regression studies in the 1970s [4, 25]. Making this computer-intensive method a practical one required that a number of computational methods be developed [8], but this work proved successful, and local regression is now regularly used in regression studies [9]. There are a multitude of other nonparametric regression methods. One is splines [22, 24, 26]. But for two or more factors, splines have resisted solutions to nasty computational problems, and at this writing are still an n^3 operation. Another is kernel estimates [21, 27] and their adjusted versions [16]. But kernel estimates have the disadvantage that they do not work. All it takes is a serious effort to use them in practice to find out this out. New and exciting theoretical results of Fan [10, 11, 13, 12], of Hastie and Loader [17] show why this is the case and why local fitting of linear and quadratic polynomials does far better. The second topic in this paper is nonparametric regression, in particular, what works and what does not.

Nonparametric regression surfaces can consume a large number of degrees of freedom in following the patterns of complex data. One way to conserve degrees of freedom is to fit a special class of nonparametric regression surfaces called *conditionally parametric fits.* This class is the third topic of the paper.

Finally, an appendix provides information about a package of public-domain subroutines, written in Fortran and C, that carry out local fitting, including conditionally parametric fits.

2 Coplots

Figure 2 is a scatterplot matrix that graphs the data from the industrial experiment mentioned earlier [3]. A single cylinder engine was run and three variables were measured. The response, which will be denoted by NO_x, is the concentration of NO plus the concentration of NO_2 in the engine exhaust, normalized by the amount of work of the engine. The units are μg of NO_x per joule. One factor is the equivalence ratio, E, at which the engine was run. E is a measure of the richness of the air and fuel mixture. As E increases, the amount of fuel increases. The second factor is C the compression ratio of the engine. There were 88 runs of the experiment, so the data consist of 88 measurements of three variables. The scatterplot of NO_x against E shows a strong nonlinear dependence with a peak between 0.8 and 1.0. The scatterplot of NO_x against C shows no apparent dependence; however, we should not at this point draw any firm conclusion since it is possible that a dependence is being masked by the strong effect of E. The scatterplot of C and E shows that the values of the two variables are nearly uncorrelated and that C takes on one of five values.

These data have been ill treated in the past. In the original analysis, log NO_x was modeled [3]. This makes the variance of the errors decrease with the mean of the response. A fourth degree polynomial of E and C was fitted to the data by stepwise regress. Such high-order polynomial fitting is at best

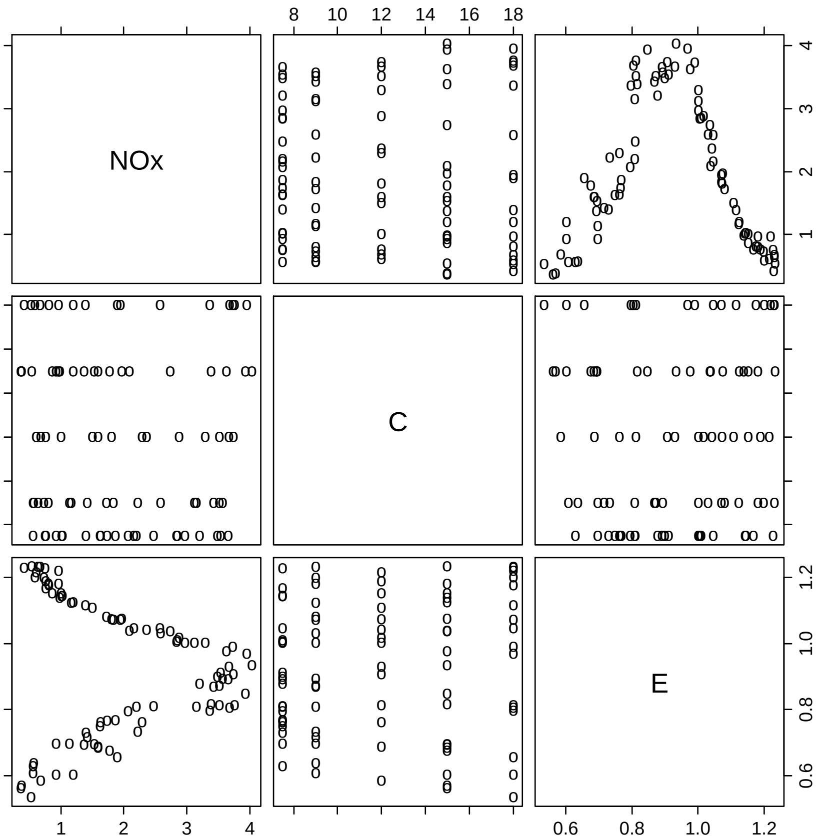

Figure 2: Scatterplot matrix of NO_x, C, and E.

a dubious practice. Another analysis of these data ignored an interaction between C and E that cannot be removed by transformation was ignored [23].

Figure 3 is a coplot of $\mathrm{NO_x}$ against C given E [5]. The dependence panels are the 3×3 array, and the given panel is at the top. On each dependence panel, $\mathrm{NO_x}$ is graphed against C for those observations whose values of E lie in an interval, and a smooth curve has been added using the nonparametric regression procedure loess, which will be described in the next section. Thus each panel shows how $\mathrm{NO_x}$ depends on C for E held fixed to an interval. The intervals are shown on the given panel; as we move from left to right through these intervals, we move from left to right and then bottom to top through the dependence panels. The intervals have two properties: approximately the same number of observations lie in each interval and approximately the same number of observations lie in two successive intervals. The data analyst specifies the number of intervals, 9 in Figure 3, and the target fraction of points shared by successive intervals, 1/2 in Figure 3.

Figure 4 is a coplot of $\mathrm{NO_x}$ against E given C. Since C takes on five values, we have simply conditioned on each of these five values.

The coplot in Figure 4 shows that $\mathrm{NO_x}$ has a strong nonlinear dependence on E with a peak value near $E = 0.9$ for each conditioning on C. The coplot of Figure 3 shows that $\mathrm{NO_x}$ depends on C, but less strongly. That is, over the range of values of E and C in the data set, $\mathrm{NO_x}$ undergoes greater change as a function of E for C held fixed than as a function of C for E held fixed. Given E, the surface appears to be linear as a function of C. As the conditioning value of E increases, the slope of the linear dependence on C first increases and then decreases to zero. The complex nonlinear behavior of the data as a function of E given C makes the fitting of a parametric surface an unsatisfactory approach. Nonparametric regression is the appropriate tool here. But the observation of a simple function of C given E is exactly the property that will be exploited by the conditionally parametric fits of section of Section 4.

3 Nonparametric Regression

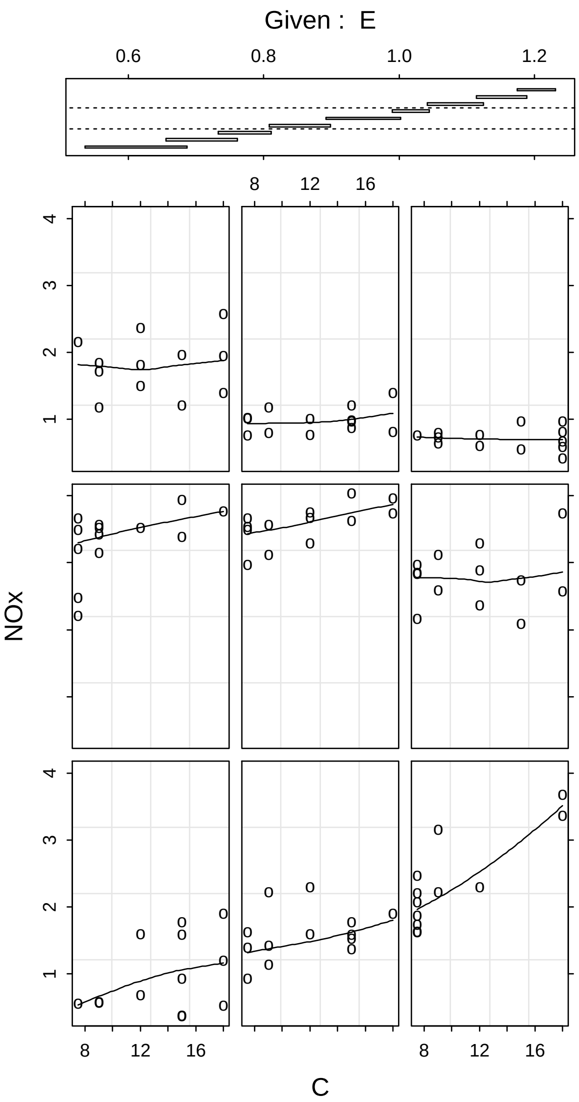

Figure 3: Coplot of $\mathrm{NO_x}$ against C given E.

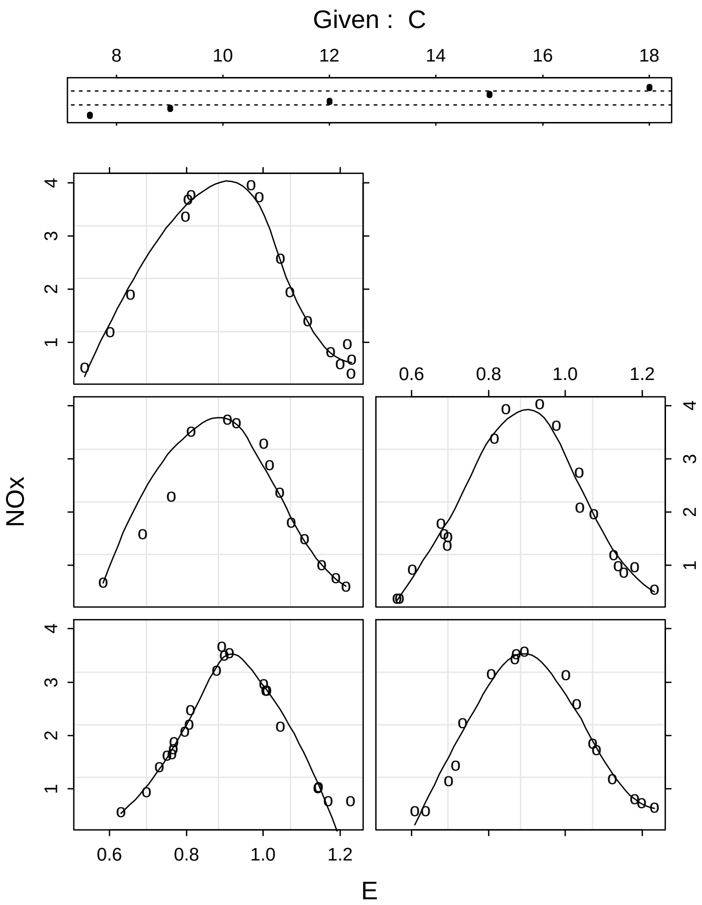

Figure 4: Coplot of NO_x against E given C.

Local Regression Models

Suppose, for each $i = 1$ to n, that y_i is a measurement of the response and x_i is a corresponding vector of measurements of p factors. The model for local regression has the same basic structure as that for parametric regression:

$$y_i = g(x_i) + \varepsilon_i,$$

where g is the regression surface and the ε_i are random errors. In carrying out a local regression, we specify properties of the errors and the regression surface. Here we will suppose that the errors are i.i.d. Gaussian. (We can relax this assumption and suppose only that they are i.i.d. and symmetric.) For each x in the space of the factors, we suppose that in a certain neighborhood of x, the regression surface is well approximated by a linear or quadratic polynomial in the factors. The overall sizes of the neighborhoods are specified by a parameter, α, that will be defined shortly. Size, of course, implies a metric in the space of the factors, which we will also define.

Loess

Loess is one particular method of estimation for local regression models [4, 6, 7]. The following describes the loess computation at one point, x, in the space of the factors.

Let $\Delta_i(x)$ be the Euclidean distance from x to x_i. This provides a metric in the space of the factors, but the x_i do not have to be the raw measurements. Typically, it makes sense to take x_i to be the raw measurements normalized in some way. Here, we normalize by dividing the factors by their 10% trimmed sample standard deviation. Let $\Delta_{(i)}(x)$ be the values of these distances ordered from smallest to largest.

The smoothness of the loess fit depends on the specification of a neighborhood parameter, $\alpha > 0$. Suppose $\alpha \leq 1$. Let q be equal to αn truncated to an integer. Let

$$T(u;t) = \begin{cases} (1-(u/t)^3)^3, & \text{for } 0 \leq u < t \\ 0 & \text{for } u \geq t \end{cases}$$

be the *tricube weight function*. We define a weight for (x_i, y_i) by

$$w_i(x) = T(\Delta_i(x)/\Delta_{(q)}(x)).$$

For $\alpha > 1$, the $w_i(x)$ are defined by the same formula except that $\Delta_{(q)}(x)$ is replaced by $\Delta_{(n)}(x)\alpha^{1/p}$. The $w_i(x)$, which we will call the *neighborhood weights*, decrease or stay constant as x_i increases in distance from x.

If we have specified the surface to be locally well approximated by a linear polynomial—that is, if λ is 1 — then a linear polynomial in the factors is fitted to y_i using weighted least squares with the weights $w_i(x)$; the value of this fitted polynomial at x is the loess fit, $\hat{g}(x)$. If λ is 2, a quadratic is fitted. In fitting quadratics is is also sometimes useful to drop the squares of some or even all of the factors in doing the fitting. An example will be given later.

This loess fitting method applies when the data are Gaussian. Modifications can be made in straightforward ways to accommodate a specification of a symmetric, possibly long-tailed, distribution [2, 4]. In the Gaussian case, sampling distributions of the estimates have been worked out [7], but only very rough approximations exist for the symmetric case [9].

How do we decide between locally linear and locally quadratic fitting? Locally linear fitting is sufficient when the curvature in the surface is relatively gentle. But if there is substantial curvature, in particular, peaks or valleys, locally quadratic fitting almost always does better. A locally quadratic surface that fits such data without substantial lack of fit will typically be smoother than a locally linear surface that fits the data. The reason is that the added flexibility of the quadratic polynomial often allows us to substantially increase the value of α.

The engine data require locally quadratic fitting because there is substantial curvature: a ridge in the surface near $E = 0.9$ that was revealed by the coplots. Diagnostic plotting of residuals shows that with this specification, an α of about 0.25 is the largest the data will tolerate without lack of fit. The resulting fit with $\alpha = 0.25$ has 21.6 equivalent degrees of freedom and $\hat{\sigma} = 0.18$. Diagnostic plotting also shows that the Gaussian assumption used in the fitting is appropriate.

Figure 6 shows a coplot of the surface as a function of E given C. For each of a collection of values of C, shown at the top on the given panel, the surface is graphed as a function of E. In other words, we are graphing slices of the surface using planes perpendicular to the C axis. 99% confidence intervals are shown at selected positions. Figure 5 shows a coplot of the surface as a function of C given E. There is one distressing property — to much variation as function of C, but we will cure this problem shortly.

Why Kernel Smoothers Do Not Work

Kernel smoothers amount to locally constant fitting, that is, locally fitting a zero degree polynomial. In practice, the trade-off between variance

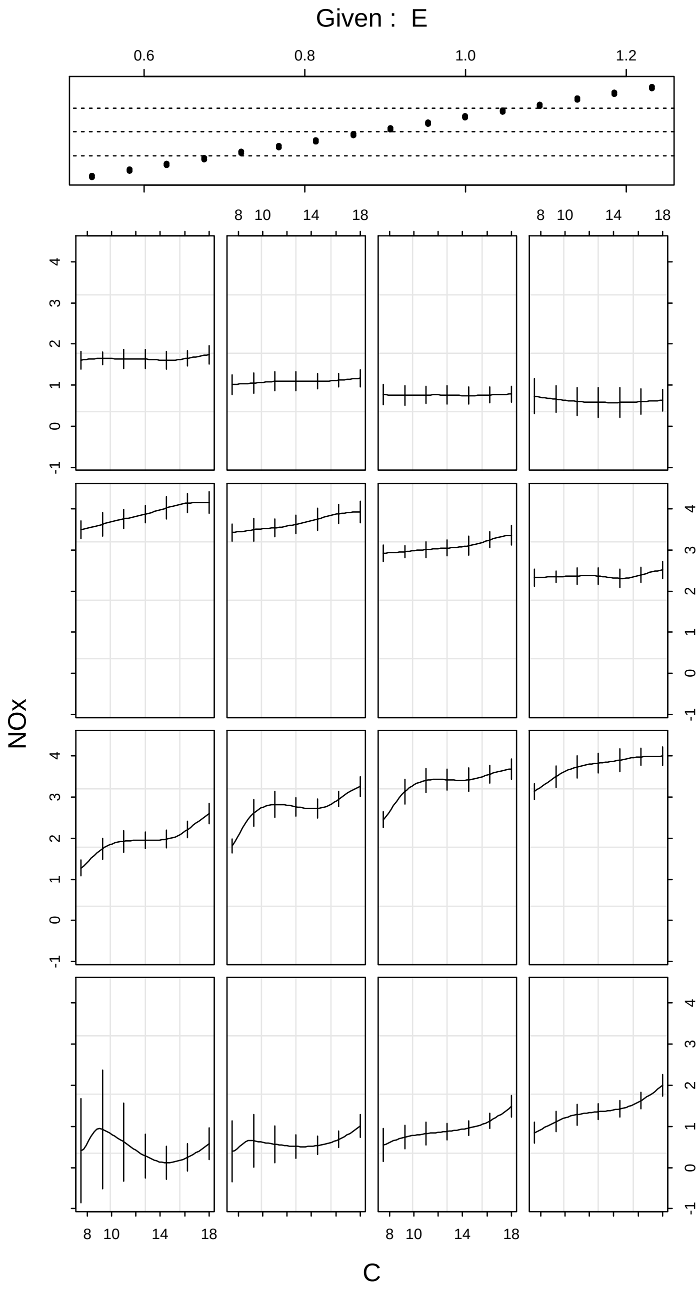

Figure 5: Coplot of the local regression fit with pointwise 99% confidence intervals.

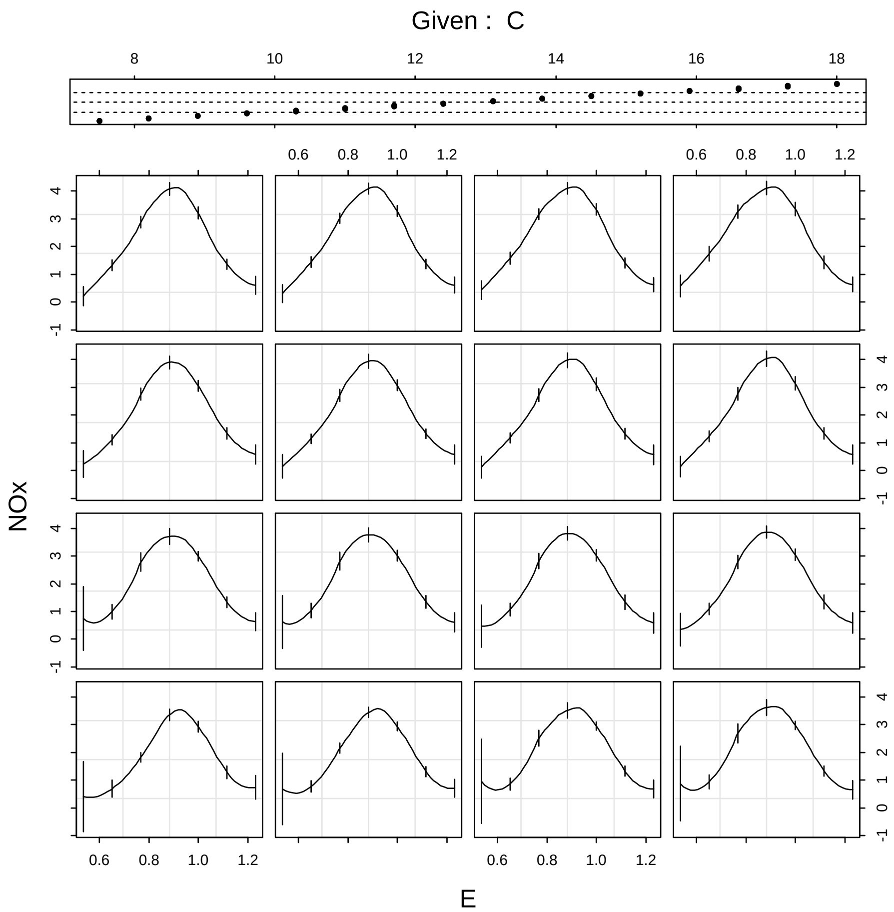

Figure 6: Coplot of the local regression fit with pointwise 99% confidence intervals.

and bias almost never makes it worthwhile to use this crude method of approximation. Locally constant fitting cannot even accommodate a linear effect at the boundaries of the design space of the factors. For the engine data, a kernel estimate would perform wretchedly. In fact, even locally linear fitting does poorly and locally quadratic fitting is needed.

The poor properties of kernel estimates was not fully appreciated by researchers in the mathematics of nonparametric regression until the work of Fan appeared [10, 11, 13, 12]. He shows that locally linear estimates have the same asymptotic variance properties as the standard kernel estimates and the same asymptotic bias properties as modified kernel estimates that correct for bias but inflate variances by 50%. In other words, locally linear estimates have the good variance properties of kernel estimates and the good bias properties of corrected-kernel estimates. More recently, an excellent study by Hastie and Loader [17] — one that balances mathematics, statistical methods, computing, and the processes of data analysis — provides particularly deep insight into this issue.

4 Conditionally Parametric Fits

A nonparametric surface is *conditionally parametric* in a proper subset of the factors if, given the values of the factors not in the subset, the surface is a member of a parametric class as a function of the subset. It makes sense to specify a regression surface to be conditionally parametric in one or more variables if exploration of the data or *a priori* information suggests that the underlying pattern of the data is globally a very smooth function of the variables. Making such a specification when it is valid can result in a more parsimonious fit.

An exceedingly simple modification of loess fitting yields a conditionally parametric surface. We simply ignore the conditionally parametric factors in computing the Euclidean distances that are used in the definition of the neighborhood weights, $w_i(x)$.

The idea of conditionally parametric surfaces and the approach to fitting them with the simple modification of loess were introduced and put to practical use in [9]. The idea has been further discussed and expanded in [19].

Back to the Engine Data

Figure 5, the coplot of the NO_x surface as a function of C given E, has a disturbing property: there is a lot of variation as function of C that is not large compared with the confidence intervals. The fit does not give us the linearity in C given E strongly suggested by the coplots. But we can achieve it by specifying the fit to be conditionally parametric in C. If we do nothing else, it will be conditionally quadratic in C. So we drop C^2 from the local fitting monomials — that is, we fit just C, E, CE, and E^2. The result is a fit conditionally linear in C. The diagnostic plots of residuals now show that we can increase α to 0.5 without introducing lack of fit. The resulting equivalent degrees of freedom is 9.2, a substantial reduction from the 21.6 of the original fit, but $\hat{\sigma} = .18$ remains the same to two decimal places. Thus by specifying a conditionally parametric fit, we have driven down the degrees of freedom without introducing lack of fit. The coplots in Figures 7 and 8 show the resulting fitted surface.

Why Conditionally Parametric Fits Work

A number of methods have been introduced to add structure to nonparametric surfaces to cut back on the degrees of freedom of the fit without sacrificing lack of fit [1, 14, 15, 18]. In any application, this works if the pattern of the data is reasonably well approximated by such a surface. The above methods ultimately rely on the pattern being well approximated by a sum of univariate functions, either of the original predictors or of linear combinations of them, or of transformations of the factor and predictors.

A growing list of practical applications has shown that conditionally parametric fits add structure in ways that is quite useful in practice [19, 9]. This is to be expected because there is a good argument for why we would expect conditionally parametric surfaces to frequently provide a good fit. They represent a natural extension of a phenomenon addressed by a number of regression methods — factors with no effect. A nonparametric regression surface that ignores the uninfluential variables is conditionally parametric in them; the parametric family is a polynomial of degree zero, a constant. If a surface has, instead of no effect, a very limited effect due to these factors compared to other factors, it is reasonable to expect that the effect would be nearly linear or quadratic in these factors given the others. One way to argue for this is to think of Taylor series approximations. Consider the NO_x surface as a function over all imaginable values of C and E. Suppose it is very complex function of both C and E with many peaks and valleys. Taylor series arguments make it clear that C can be varied by a sufficiently

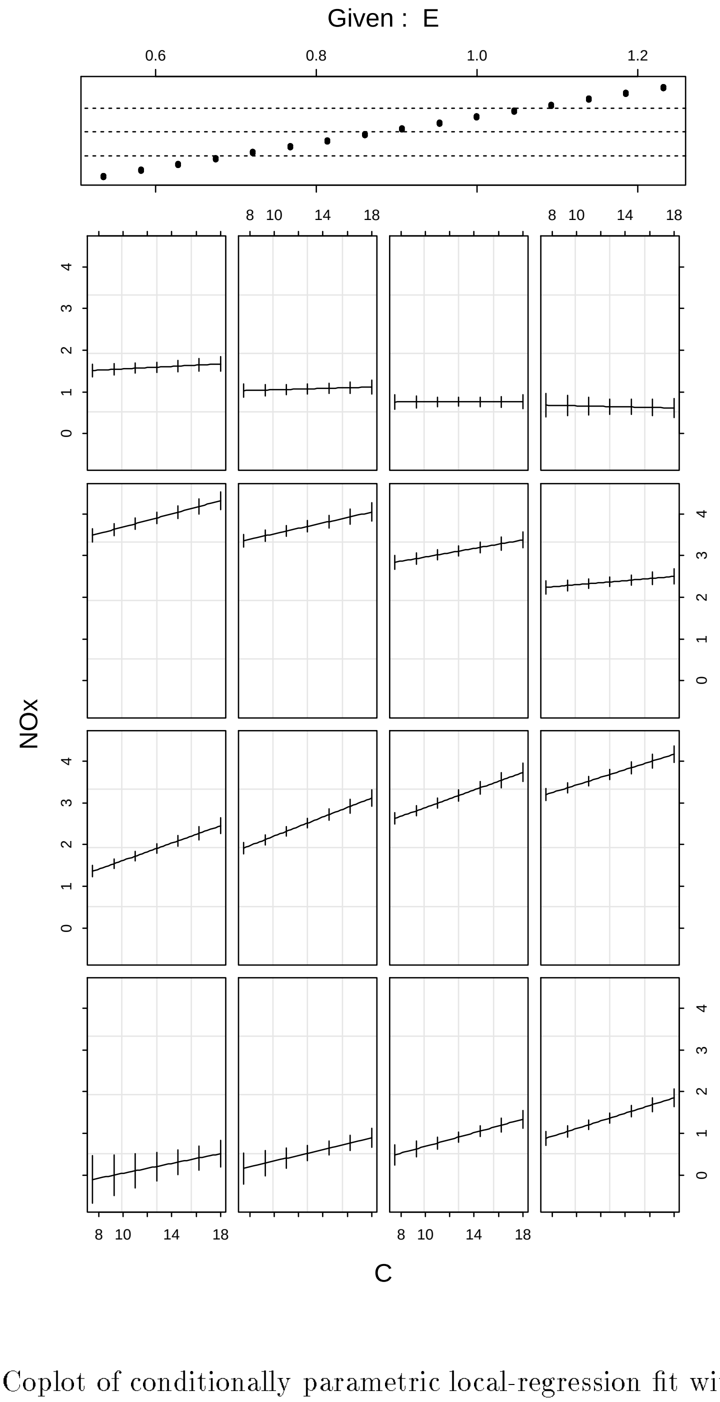

Figure 7: Coplot of conditionally parametric local-regression fit with pointwise 99% confidence intervals.

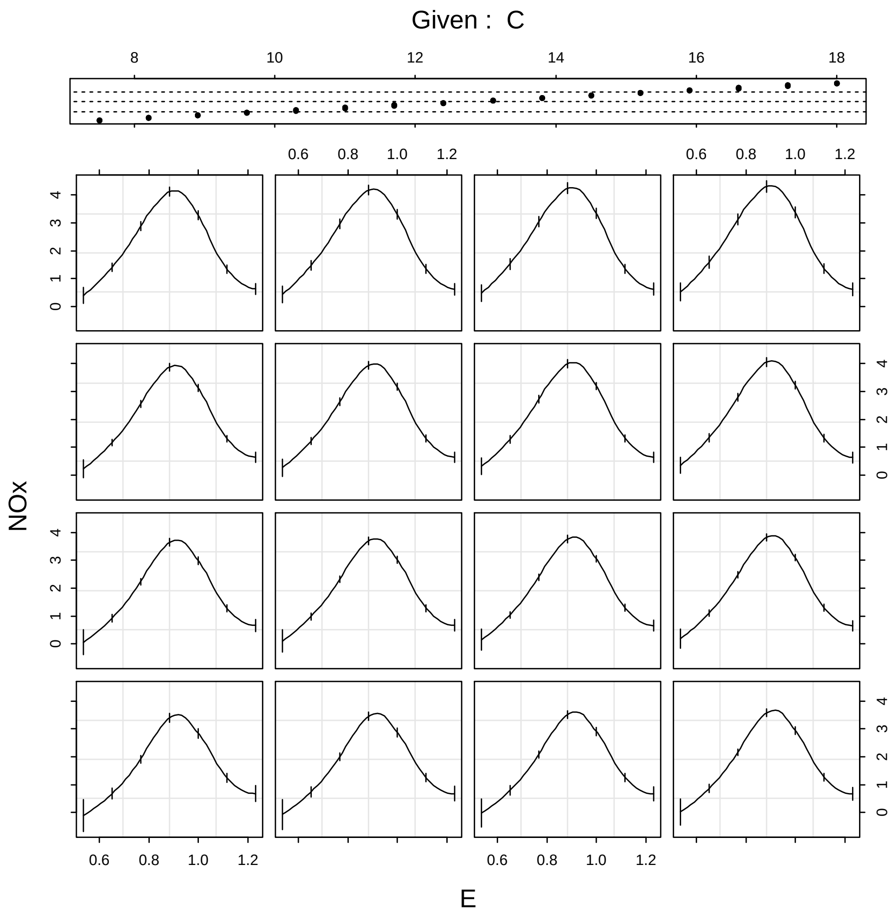

Figure 8: Coplot of conditionally parametric local-regression fit with pointwise 99% confidence intervals.

small amount in an experiment so that the surface is well approximated by a function that is constant in C given E. If the domain of variation of C is enlarged, the next approximation is a linear function of C given E, and then quadratic. This argument carries with it a corollary; we would expect that, overall, the influence of the variables in which the surface is conditionally parametric would be less than the influence of the other variables. The coplots in Figure 7 and Figure 8 show this to be the case for the engine data; the surface changes less as a function of C, the factor in which the surface is conditionally parametric, than as a function of E.

Appendix: Obtaining the Loess Routines Electronically

The C and Fortran routines, all of which are freely available, may be obtained by sending electronic mail to

```
netlib@research.att.com
```

a mailbox at AT&T Bell Laboratories in Murray Hill, NJ. The message

```
send dloess from a
```

should be sent. The routines are double precision.

The file `dloess` is a so-called "shell archive" or "bundle". Moreover, in order to send this 172 kilobyte file to you by email, netlib breaks it into pieces which are themselves shell archives. So you'll need to run `sh` once on each piece of mail to reconstruct the file `dloess`, then run `sh dloess` to finally reconstruct all the source files.

Subroutines from linpack, which are called by the Fortan code, are not included. If they are not already on your system, send the message

```
send d1mach dnrm2 dsvdc dqrdc ddot dqrsl idamax from linpack core
```

to the same address. When installing, don't forget to uncomment the appropriate DATA statements in d1mach, as described by the comments in those functions.

The PostScript file for this user manual is also available by email

```
send cloess.ps from a
```

but since it is over half a megabyte, `ftp` is a better choice

```
ftp research.att.com
login: netlib
password: <your email address>
binary
cd a
get cloess.ps.Z
quit
uncompress cloess.ps
```

Bug reports will receive prompt attention. Send electronic mail to

```
shyu@research.att.com
```

or send paper mail to

```
Ming-Jen Shyu
AT&T Bell Laboratories
600 Mountain Avenue, room 2C-263
Murray Hill NJ 07974
USA
```

References

[1] L. Breiman and J. H. Friedman. Estimating optimal transformations for correlation and regression. *Journal of the American Statistical Association*, 80:580–598, 1985.

[2] D. R. Brillinger. Consistent nonparametric regression. *The Annals of Statistics*, 5:622–623, 1977. Discussion of paper by C. J. Stone.

[3] N. D. Brinkman. Ethanol fuel — a single-cylinder engine study of efficiency and exhaust emissions. *SAE Transactions*, 90(810345):1410–1424, 1981.

[4] W. S. Cleveland. Robust locally-weighted regression and smoothing scatterplots. *J. Am. Statist. Ass.*, 74:829–836, 1979.

[5] W. S. Cleveland. *Visualizing Data.* Hobart Press, Summit, New Jersey, 1993, to appear.

[6] W. S. Cleveland, S. J. Devlin, , and E. Grosse. Regression by local fitting: methods, properties, and computational algorithms. *J. Econometrics*, 37:87–114, 1988.

[7] W. S. Cleveland and S. J. Devlin. Locally-weighted regression: an approach to regression analysis by local fitting. *J. Am. Statist. Ass.*, 83:596–610, 1988.

[8] W. S. Cleveland and E. Grosse. Computational methods for local regression. *Statistics and Computing*, 1:47–62, 1991.

[9] W. S Cleveland, E. Grosse, and W. M. Shyu. Local regression models. In J. M. Chambers and T. Hastie, editors, *Statistical Models in S*, pages 309–376. Wadsworth, Pacific Grove, California, 1991.

[10] J. Fan. Local linear regression smoothers and their minimax efficiency. *Annals of Statistics*, 1991a, (to appear).

[11] J. Fan. Design-adaptive nonparametric regression. *J. Amer. Statist. Assn.*, 87:998–1004, 1992a.

[12] J. Fan. Local linear regression smoothers and their minimax efficiency. *Ann. Statist.*, 1992b.

[13] J. Fan. Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 1992b, (to appear).

[14] J. H. Friedman, E. H. Grosse, and W. Stuetzle. Multidimensional additive spline approximation. *SIAM J. Sci. Stat. Comp.*, 4:291–301, 1983.

[15] J. H. Friedman and W. Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76:817–823, 1981.

[16] T. Gasser and H. G. Müller. Kernel estimation of regression functions. *Lecture Notes in Mathematics*, 757:23–68, 1979. Springer-Verlag, Berlin.

[17] T. Hastie and C. Loader. Local regression: Automatic kernel carpentry. *Statistical Science*, 1993, to appear.

[18] T. Hastie and R. Tibshirani. *Generalized Additive Models.* Chapman and Hall, London, 1990.

[19] T. Hastie and R. Tibshirani. Varying-coefficient models. *Journal of the Royal Statistical Society*, to appear, 1993.

[20] F. R. Macauley. *The Smoothing of Time Series.* National Bureau of Economic Research, New York, 1931.

[21] E. A. Nadaraya. On estimating regression. *Theor. Probab. Appl.*, 9:141–142, 1964.

[22] C. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10:177–183, 1967.

[23] R. N. Rodriguez. A comparison of the ACE and MORALS algorithms in an application to engine exhaust emissions modeling. In *Computer Science and Statistics: Proceedings of the Sixteenth Symposium on the Interface*, pages 159–167, New York, 1985. North-Holland.

[24] B. W. Silverman. Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. R. Statist. Soc.*, 47:1–52, 1985.

[25] C. J. Stone. Consistent nonparametric regression. *Ann. of Stat.*, 5:595–620, 1977.

[26] G. Wahba. Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society B*, 40:364–372, 1978.

[27] G. S. Watson. Smooth regression analysis. *Sankhya*, A, 26:359–372, 1964.