THE VARIAGRAPH CONCEPT AND A LACK-OF-FIT TEST

by

Andrew P. Robinson and Sanford Weisberg

December 1996

Staff Paper Series no. 119

DEPARTMENT OF FOREST RESOURCES

College of Natural Resources
and
Minnesota Agricultural Experiment Station
University of Minnesota
St. Paul, Minnesota

THE VARIAGRAPH CONCEPT AND A LACK-OF-FIT TEST

by

Andrew P. Robinson and Sanford Weisberg[1]

December 1996

Staff Paper Series no. 119

---

[1]Research assistant, Department of Forest Resources, and Professor, Department of Applied Statistics, University of Minnesota, St. Paul. Research supported by the College of Natural Resources and Agricultural Experiment Station, University of Minnesota, St. Paul, Minnesota, and the McIntire-Stennis Cooperative Forestry Research Program. Published as paper no. 964420004 of the Minnesota Agricultural Experiment Station.

ABSTRACT

We present the variagraph, a graphical diagnostic technique for assessing lack of fit and estimating pure error without replication for regression modelling. The behavior of the variagraph is explored by simulation, and its use demonstrated on an example dataset previously applied in lack-of-fit testing.

INTRODUCTION.

In this paper we reconsider the problem of testing for lack of fit in a linear regression model. We introduce a graphical diagnostic procedure called the variagraph, which is simple in concept and in implementation, flexible, and informative in many situations. We present a lack-of-fit test derived from the variagraph along with examples and further possible developments.

THE VARIAGRAPH

The variagraph in its most general manifestation is simply a plot of some measure of deviance against a bandwidth across which that deviance is assessed (Figure 1). The variagraph may be considered as a generalization of earlier attempts to assess the behavior of variance as a function of bandwidth, for example the variogram in spatial statistics (Isaaks and Srivastava 1989), and lacunarity in pattern matching (Plotnick et al. 1993).

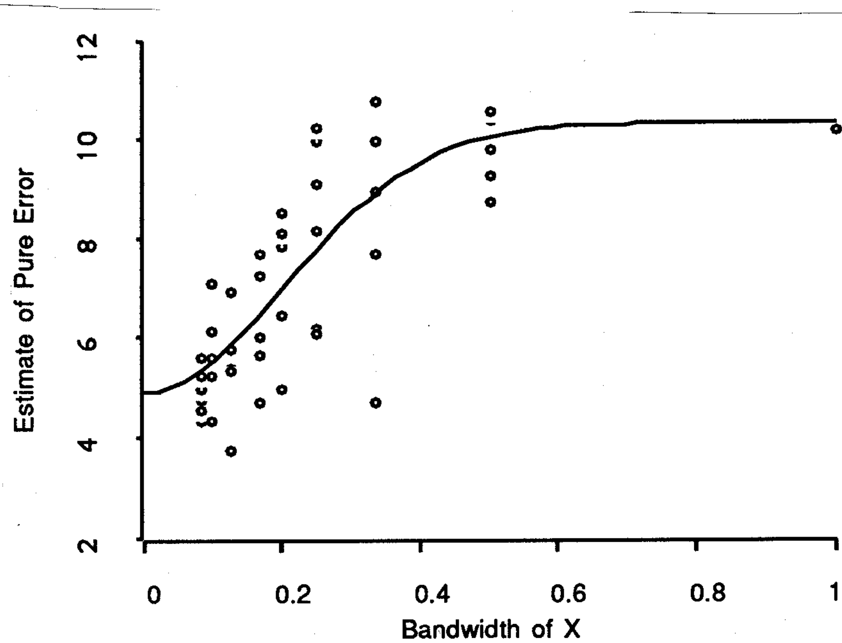

Figure 1. Variagraph of linear regression on mileage data.

The variagraph is viable for linear models with a univariate and continuous response variable. The applications of the variagraph which we will demonstrate are: assessing the lack of fit of a mean function, estimating the underlying pure error, and examining the behavior of the variation across different bandwidths. It will probably be most useful for detecting curvature which the analyst may not be able to discern visually in dense clouds of points.

OUTLINE

Consider the traditional one-dimensional modelling situation where y is a univariate response and x the predictor. We assume:

$E(y \mid x) = g(x)$, an unknown smooth function, and
$\mathrm{Var}(y \mid x) = \sigma^2$.

We wish to estimate $E(y \mid x)$. Given a collection of n data tuples, say $\{x_i, y_i\}$, we start with a model: $E(y \mid x) = M(x,\theta)$ where θ is a vector of unknown parameters of interest. We choose as an estimate of θ:

$$\hat{\theta} = arg\ min \left\{ \overline{D} = \frac{1}{n} \sum_{i=1}^{n} d\big(M(x_i,\theta), y_i\big) \right\}$$

where $d(a,b)$ is a goodness-of-fit measure evaluated at a point; for example

$$d(a,b) = (a - b)^2$$

leads to $\hat{\theta}$ being the least-squares estimate.

We assume that $d(a,b)$ is strictly convex as a function of its first argument. Now let

$$\theta^* = arg\ min \left\{ E_{X,Y} d\big(M(x,\theta), y\big) \right\}$$

Then by definition, $\hat{\theta}$ is Fisher consistent for θ^* (Cox and Hinkley 1974).

We denote the estimate of $M(x,\theta)$ by $M(x,\hat{\theta})$, which is a n x 1 vector of fitted values. We fit the model under scrutiny, $M(x,\hat{\theta})$, and let r_i denote the residual at the i^{th} datum, so $r_i^2 = \left[y_i - M(x_i,\hat{\theta})\right]^2$. Then for large samples,

$$r_i^2 \rightarrow \left[y_i - M(x_i,\theta^*)\right]^2 ,$$

and

$$E(r_i^2|x_i) \to E[y_i - M(x_i,\theta^*)]^2$$

$$= \begin{cases} \sigma^2, \text{if NH is true,} \\ \sigma^2 + [g(x_i) - M(x_i,\theta^*)]^2 \text{ otherwise} \end{cases}$$

since $E(y \mid x) = g(x)$ by Assumption 1. This represents nothing more complicated than the breakdown into model and residual sums of squares arising from analysis of variance. Under the null hypothesis,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} r_i^2$$

is a consistent estimator of σ^2; we usually replace $\frac{1}{n}$ with $\frac{1}{n-2}$. If the null hypothesis is not true then $\hat{\sigma}^2$ is in general too large.

LACK-OF-FIT TESTING

Replication

Considerable attention has been paid to the testing of lack of fit for linear models. It is well documented that such testing is trivial if the data are replicated across the predictor variables; estimating σ^2 based on pure error under these circumstances is straightforward, and a test can be easily constructed (Neill and Johnson 1984, Weisberg 1985). We summarize the key points below.

Suppose we have a certain amount of replication in the predictor variables, such that the n values of x consist of d distinct values. Further, index the response y by y_{ij}, which indicates the j^{th} response in the i^{th} replicate group, $j = 1 \ldots n_i$, $i = 1 \ldots d$.

Let

$$SS_{PE} = \sum_{i=1}^{d}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i+})^2,$$

and

$$df_{PE} = \sum_{i=1}^{d}(n_i - 1) = n - d.$$

Now for any mean function $E(SS_{PE} \mid X) = (n - d)\sigma^2$, and hence SS_{PE}/df_{PE} is unbiased for σ^2 under both hypotheses. Under the assumption of normality, an F test can be constructed in the usual way (Hocking 1996).

Near Replication

The case where there is near replication across the predictor variable or variables has also been extensively examined. In the past 25 years there have been nearly ten different attempts at assessing lack of fit under these circumstances (Neill and Johnson 1984, Christensen 1989, Joglekar et al. 1989). Nearly all have involved subjectively dividing the data across the predictor space and trying to obtain some estimate of pure error to construct an exact or approximate F statistic.

Commonly, an underlying group structure is assumed, defining the predictors as comprising replicated values with some random perturbation, and the groups are subjectively defined by analysis of scatterplots. This subjective clustering of otherwise unclustered data has been a stumbling block for most approaches. It is not clear how to interpret a significance test which is conditioned on the subjective clustering of points into supposedly similar groups.

Informally, the following approach avoids this subjectivity by assessing the pure error and lack of fit at a range of different bandwidths with several different starting points, then modelling the variation as a function of bandwidth.

We start with the data ordered by to their fitted values from the model under scrutiny. Divide the range of the fitted values $\mathrm{M}(x,\hat{\theta})$ into d slices, each of equal relative length b, $b \in [0,1]$, with $n_i \approx \left[\frac{n}{d}\right]$ points in each on average, $i = 1 \ldots d$. Let S_{ib} be the set of x's in slice i at bandwidth b, then fit the model M() within each slice and let $\hat{\theta}_{ib}$ be the vector of parameter estimates coming only from the i^{th} slice. Then calculate the residuals within each slice:

$$r_{ij}(b) = y_{ij} - \mathrm{M}(x_{ij},\hat{\theta}_{ib})$$

where r_{ij} is the j^{th} residual within the i^{th} slice, $\mathrm{M}(x_{ij},\hat{\theta}_{ib})$ is the fit of the model to the data in the slice only and b controls the proportion of the range of $\mathrm{M}(x,\hat{\theta})$ which is included in S_{ib}, (recalling $d * b = 1$). We can now write the expression of deviance as a function of the bandwidth:

$$\hat{\sigma}_b^2 = \frac{1}{n-pd}\sum_{i=1}^{d}\sum_{j=1}^{n_i}\left(r_{ij}(b)\right)^2.$$

If the null hypothesis is false, then as an estimate of σ^2 this will be too large on the average, because

$$E\left[\left(r_{ij}(b)\right)^2\right]=\sigma^2+\left[g\left(x_{ij}\right)-\mathrm{M}\left(x_{ij},\hat{\theta}_{ib}\right)\right]^2,$$

however, if $g(x)$ is continuous $\hat{\sigma}_b^2 \to \sigma^2$ as $n \to \infty$ and $b \to 0$, because as $n \to \infty$ and $b \to 0$, $\mathrm{M}\left(x_{ij},\hat{\theta}_{ib}\right) \to g\left(x_{ij}\right)$ under the conditions noted in the outline above.

The estimation procedure is repeated m times, with different start points for the slices. The different start points could be chosen randomly or systematically from the range of the predictor variable. The empirical variagraph is a plot of all the estimated $\hat{\sigma}_b^2$ for the different start points against b (Figure 1), and we estimate σ^2 by a fitted regression at $b = 0$.

Previously, the disadvantage of clustering or slicing the data is that the results are dependent on the starting points of the slices, much as is the case with creating a histogram (Scott 1985). This technique answers concerns which have been voiced about clustering (e.g. Joglekar et al. 1989). By repeating the slicing procedure with different starting points we guard against subjectivity. This is analogous to the average shifted histograms of Scott (1985), or WARPing the data (Härdle 1991).

FITTING THE VARIAGRAPH

Several different approaches can be taken after plotting the empirical variagraph. We elected to fit a simple variation of the Gaussian semi-variogram (Isaaks and Srivastava 1989), believing it to be flexible enough to represent most of the possible behavior of the data whilst allowing fairly simple estimation of parameters. The semi-variogram is a statistical tool for summarising the variation of a spatial variable, which is one for which the relative values depend at least partially on their relative location in the space which contains them. Informally, it describes the covariance of the responses as a function of their proximity.

The standard Gaussian semi-variogram (Isaaks and Srivastava 1989) is:

$$y=\theta_0+\theta_1\left(1-e^{-3x^2/\theta_2^2}\right). \qquad (1)$$

Each of the parameters has a physical interpretation. θ_0 estimates the *nugget effect*, which is all the variation which occurs at any distances smaller than the proximity of neighboring points. It includes, but is not limited to, measurement error. It is called the nugget effect because of the association of early spatial statistics with mining exploration, and the discontinuity resulting from microvariation, or nuggets (Cressie 1991). θ_1 is the sill, which is the difference between the nugget and the asymptotic limit which the variogram reaches. That limit is considered to be the global variation, and is only meaningful if the generating process is stationary. θ_2 is called the range, and is the distance value at which the variogram reaches the sill.

Our reinterpretation of these parameters is as follows: the nugget is analogous to pure-error sums of squares, and the sill is analogous to the lack-of-fit sums of squares, in lack-of-fit testing. We choose, as a measure of lack of fit, the ratio of the sill over the nugget, which is analogous to the F statistic in the case of true replication.

Fitting even this simple model to the extremely wide variation of forms that the empirical variagraph produced proved extremely difficult. Although the model fits well for classical shapes such as that in Figure 1, if there were no underlying structure then the empirical variagraph could be anything from a straight line to an inverted sigmoid shape. None of the nonlinear fitting mechanisms tried could cope with this range, and all frequently failed to converge. It was necessary to apply some simplification.

For our purposes, θ_2 could be considered a nuisance variable. It added flexibility to the family of curves under consideration but little information to the problem at hand. The point of interest for us was the size of the sill, relative to the nugget, and the distance at which the sill was reached, although of some interest, was not central to the problem. For simplicity, we set this value as $\theta_2 = 1$, which leads to the following form:

$$s = p + d\left(1 - e^{-3b^2}\right) \tag{2}$$

where s represents the error sum of squares, b the bandwidth, p the pure error (nugget effect), and d the difference between the total variance and the pure error (sill), and therefore could be considered as a measure of the lack of fit. This was a more stable manifestation of equation (1), and could be fit using simple linear regression on the appropriate transformation of the predictor variable.

There is no theoretical reason for preferring these to any other models which would fit the points and produce parameter estimates which have interpretive value. Isaaks and Srivastava (1989) note that the Gaussian model is often used to model highly continuous spatial processes. It has an inflection point and will give a higher estimate for the nugget than the spherical or exponential models for the same data, so for our purposes will be more conservative.

We do not know what the distribution of the parameters will be. We suspect that the estimate of pure error should be consistent, and simulation studies (see below) indicate that it is an acceptable estimate. We used $\hat{d}/\hat{p}$, the ratio of the lack of fit estimate to the pure error estimate, as a summary of the variation profile, and tested it for significance via the non-parametric bootstrap (Efron and Tibsharani 1993).

EXAMPLE

To demonstrate the variagraph we examined the mileage vs. engine size data from Velleman and Hoaglin (1981), used to demonstrate lack-of-fit tests by Neill and Johnson (1989), who found lack of fit under linear regression of varying degrees of significance. The dataset includes mileage, measured as miles per gallon and engine size, measured as displacement in cubic inches, for thirty-two 1976 model cars (Appendix 1). There are four values at which the predictor variable is replicated, giving 5 *d.f.* for the standard lack-of-fit test. The data are plotted in Figure 2.

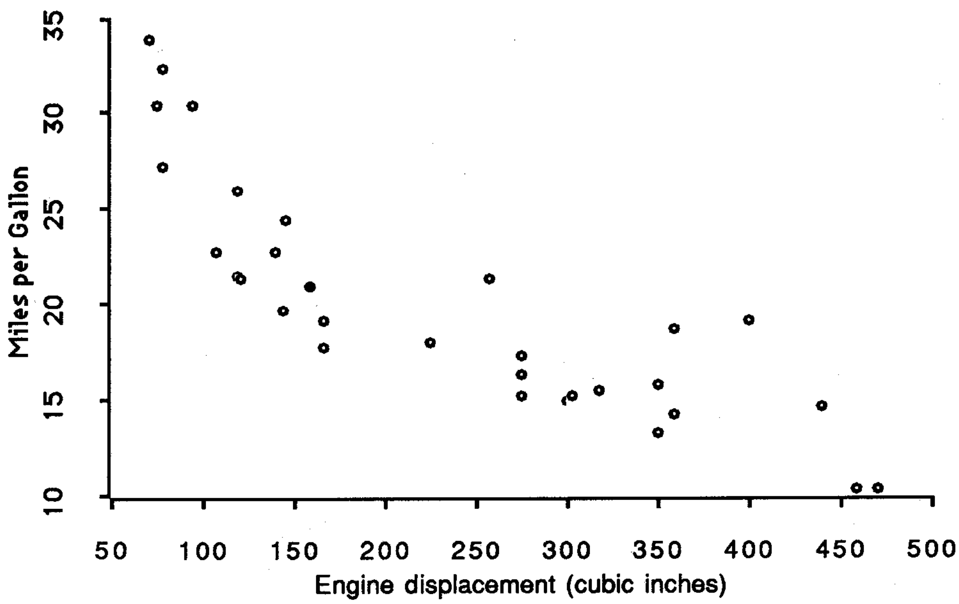

Figure 2. A scatterplot of engine displacement against mileage. The black dot represents two points.

Using the R-code statistical package (Cook and Weisberg 1994) we applied a linear regression, and fit a lowess smooth with weight 0.7 to the residual plot (Figure 3). The diagnostic information from the model provided some evidence of lack of fit ($F_{5,25} = 4.72$; $p = 0.046$)

We applied the variagraph to the model, choosing a minimum bandwidth of 0.1, to allow an average of at least three observations per slice, and (arbitrarily) 6 systematic start points. We fit Equation 2 to the resulting empirical variagraph by least squares (Figure 1), which returned parameter estimates of 5.2 for the pure error, and 1.01 for the ratio, and a significant lack of fit at $\alpha = 0.01$ when compared to 0.33, which was the bootstrap estimate of the 99th percentile of the distribution of the difference under the null hypothesis of no lack of fit.

Figure 1 can be interpreted in the following way: there is a considerable difference between the overall and the local variation, leading us to suspect that there is lack of fit. Furthermore, there is considerable vertical variation within some bandwidths (e.g. 0.33), which means that if a bandwidth of say 0.33 had been chosen and applied as per the subjective techniques discussed earlier, then the resulting estimate of pure error would have varied considerably depending on the starting point.

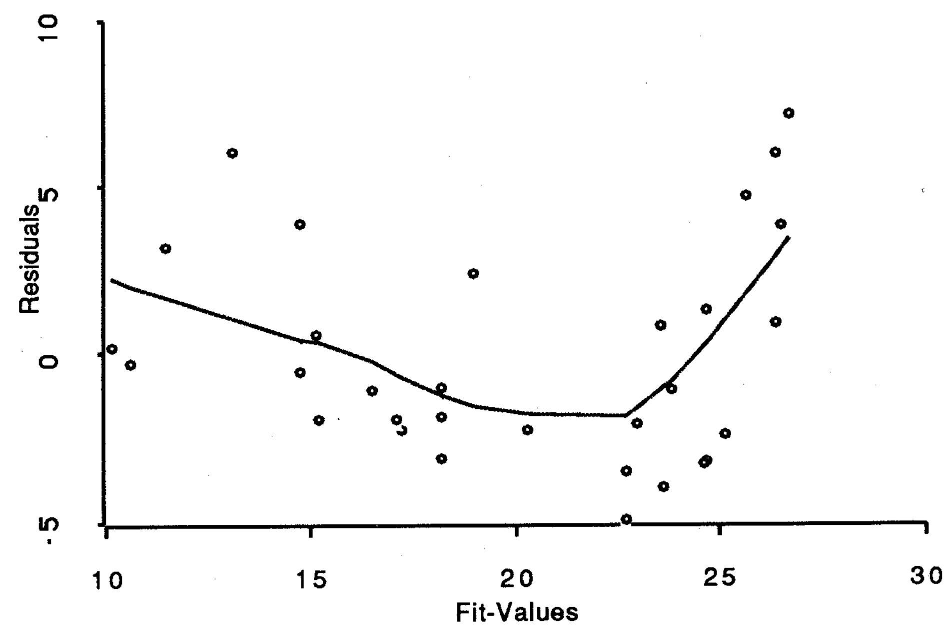

Figure 3. Plot of residuals against fitted values for a linear regression of the mileage data, with a lowess smooth of weight 0.7 superimposed.

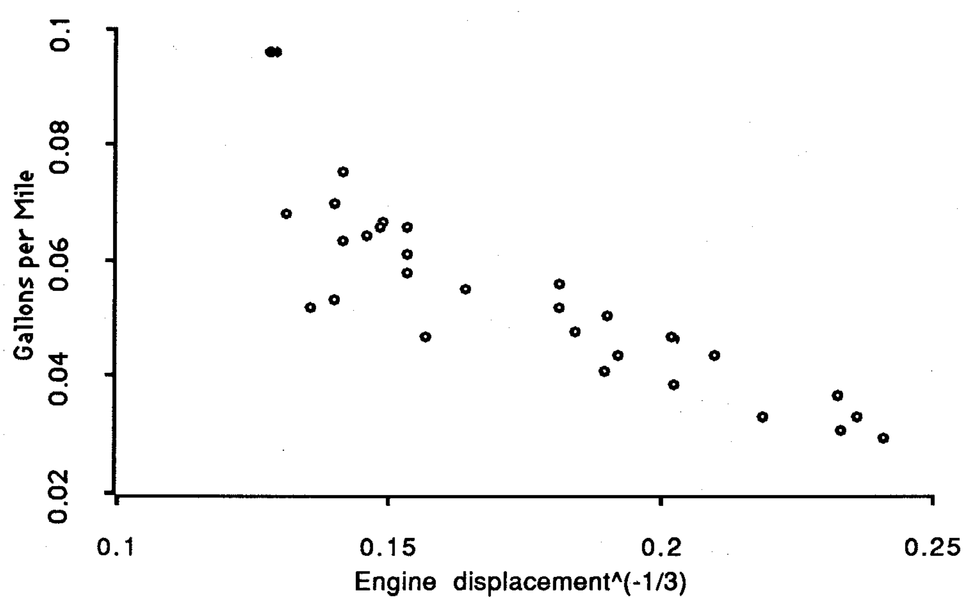

Figure 4. Scatterplot of transformed mileage against engine size - black dots are possible outliers.

After establishing evidence of lack of fit in applying the model to the dataset, Neill and Johnson (1989) noted that Velleman and Hoaglin (1981) transformed the data and refit the model, using an inverse transform on the mileage and an inverse cube root on the engine size (Figure 4).

Neill and Johnson (1989) also performed the lack-of-fit tests on this transformed data, with significant lack of fit indicated for each test if the outliers were included, and not if the outliers were excluded. We applied the variagraph to the model with the outliers included, which gave no evidence of lack of fit (4.57e-06 against estimated 99^{th} percentile of 1.35e-05), unlike all the lack-of-fit tests used by Neill and Johnson (1989). However, we note that examination of Figure 4 indicates that the assumption of homoskedasticity is probably violated whether or not the outliers are included. Furthermore, it is a moot point whether the model is misspecified or the two extreme points are outliers.

We consider this a salutary lesson in the strengths and weaknesses of the variagraph. The variagraph and the associated lack-of-fit test are robust against outliers, since it simply resamples them. It is vital to use it in concert with other diagnostic plots and devices, such as plots of residuals against fitted values, and Cook's Distances.

SIMULATION.

We used a simulation to explore the behavior of the lack-of-fit test associaated with the variagraph under different situations. All the code for the simulation was written under the statistical computing environment XLISP-Stat (Tierney 1990). Nine different configurations, of 200 pairs of points each, were chosen by the combination of three arbitrarily chosen predictor distributions;

1 Uniform (-3,3),
2 Normal (0, 1.2) and
3 Chi-Squared (4),

with three arbitrary mean functions;

Linear: $y = x + 3$, Quadratic: $y = x + 3 + \frac{x^2}{4}$, Cubic: $y = x + 3 - \frac{x^3}{6}$.

The distributions were chosen to have different shapes but similar ranges, as the degree of curvature resulting from the functions depended on the range of the predictors. The linear function was included to examine type 1 error patterns, and the curved functions to examine the power of the test. The coefficients for the functions were chosen by examining panels such as Figure 5, and choosing those which led to somewhat ambiguous shapes.

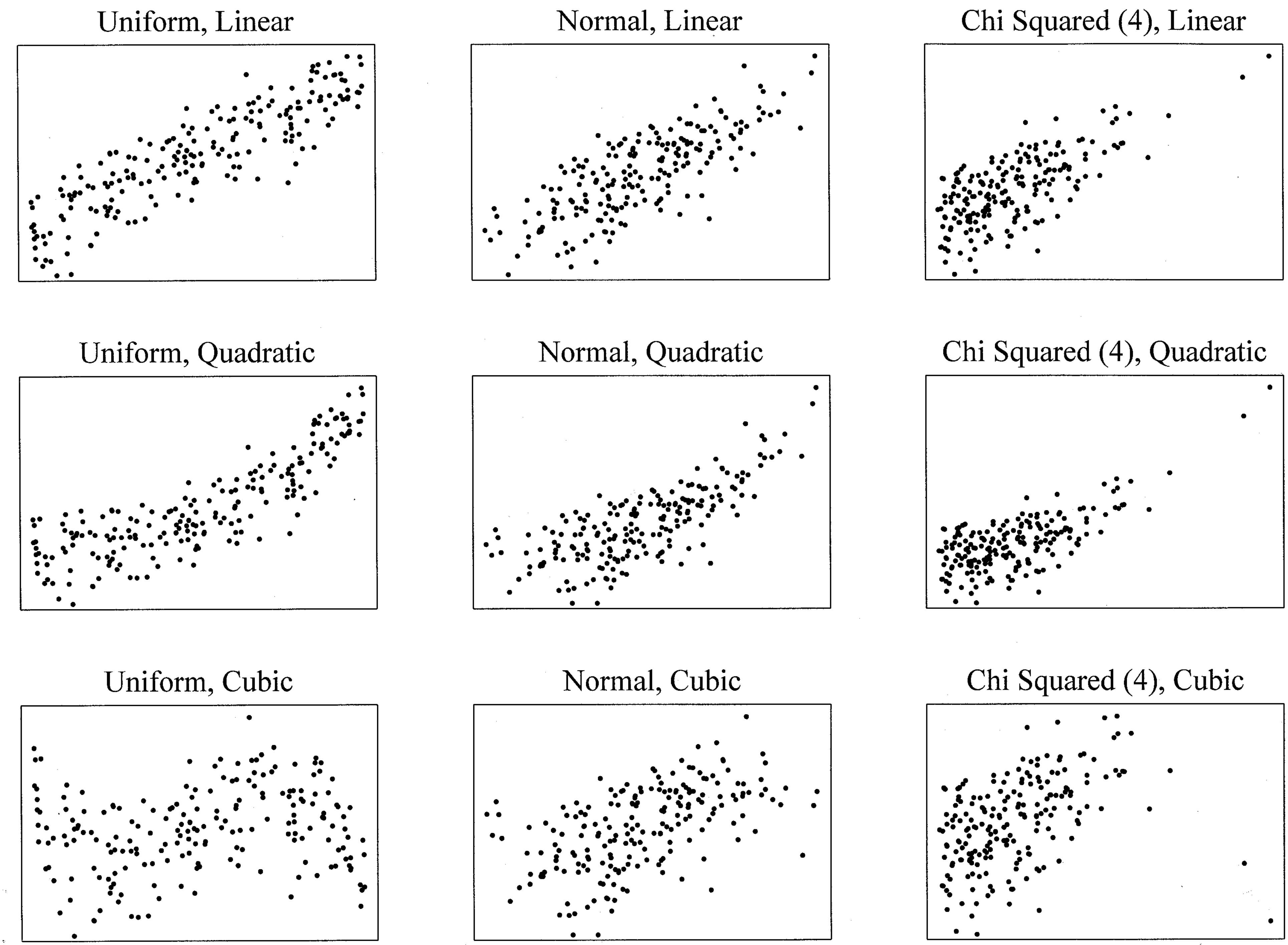

Figure 5. Samples of the functions used for the simulation. Plots are labelled with predictor distribution and function form.

At each replication, of which there were 20, a single set of 200 error terms ~N(0,1) was randomly generated, and added to each of the nine response variables. Therefore the underlying pure error variance was approximately 1. We then fit a linear regression, applied the variagraph to the residuals and fitted values. A boxplot of the estimate of the pure error is presented in Figure 6. Table 2 presents the number of times significant lack of fit was detected ($\alpha = 0.05$). The values for the linear function can be considered an estimate of the true size of the test, and those for the others, an estimate of the power of the lack-of-fit test for the given shapes.

Table 1. Number of times lack of fit was detected at the 0.05 level out of 100 simulation runs after fitting a straight line with linear regression.

Significant Lack of Fit	Linear	Quadratic	Cubic
Uniform	6	100	100
Normal	6	100	99
Chi-Squared	2	82	81

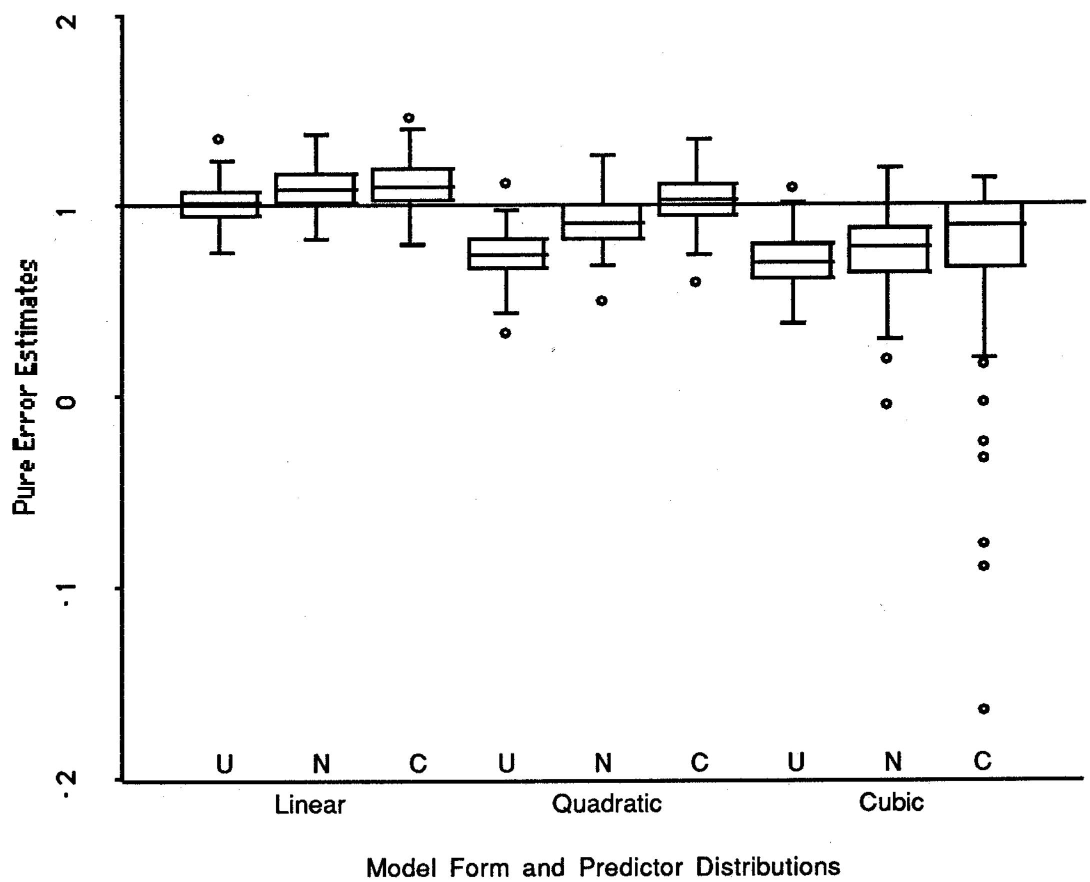

Figure 6. Boxplot of pure error estimates. Predictor distributions are Uniform, Normal, and Chi-squared. Points at -4 and -8 have been deleted from the chi-squared cubic for scaling purposes.

DISCUSSION

Results from the simulation were encouraging. The variagraph detected curvature in shapes which we considered ambiguous, with points of high visual leverage and patterns within clouds of points which were not obvious to the eye. It offered acceptable size 1 error realizations, and displayed encouraging power for the configurations considered (Table 2).

Our approach fits in to the broader sphere of techniques which test lack of fit by locally fitting either the null model (Shillington 1979, Joglekar et al. 1989), a linear approximation (Johnson and McCulloch 1987) or polynomial pieces (Green 1971), as we fit the model to disjoint slices of data. However, unlike any of the preceding, we vary both the slice width and the starting point systematically and observe the effect on the quality of fit.

The basic principles of WARPing and systematically varying the slice size can also be applied to any of the techniques above which rely on clustering the data, although the distributional implications would need to be considered, or short-cut via the bootstrap as we have done here. This would be one way to allay the concerns about artificial clustering expressed in Joglekar et al. (1989).

Contrary to recent developments in smoothing, we compute the deviance based on a fixed, rather than variable, bandwidth procedure. We are interested in exploring the behavior of the variation across different bandwidths, rather than obtaining the best possible fit at any given iteration, and variable bandwidth smoothing would not make sense.

As we fitted it, the variagraph ignores edge effects, which may well explain its relatively poor performance with skewed predictor variables (see simulation section). The edge slices are treated as though they were full size, which will probably be unsatisfactory with small sample sizes.

The variagraph will be invariant under any transformation of the predictors equivalent to the fitted mean function, since it is being performed on the residuals from the model under study. It will not be invariant under different transformations of the predictors since these will alter the mean function and therefore the status of the fit of the model. It will not be invariant under transformation of both the predictors and response as the pattern of the residual variance will change.

A most important point concerning the use of the variagraph is the assumption of homoskedasticity. The technique is therefore capable of detecting lack of fit in the mean function but not the variance function, and misspecification of the variance function will invalidate the results. Visually checking a scatterplot of the points to which the variagraph is being applied should be an adequate check of this assumption if one is willing to make it for the broader modelling process anyway.

It is possible to observe a variagraph with a difference estimate less than zero. There are several possible interpretations, and other diagnostic tools such as residual plots or Cook's Distances will be needed to differentiate between them. For example, it is possible that the assumption of constant variance has been violated. It is also possible that there is significant lack of fit within clusters (Christensen 1989). In either case there still may be lack of fit in the mean function, and the variagraph will not be powerful in detecting it. It is also possible that the minimum slice size being used is too small.

The variagraph can be applied to some problems with more than one predictor, assuming the problem has more structure. Suppose that we have

$$E(y \mid x) = g(\beta^T x)$$

for some unknown vector of parameters β and for some unknown link function g. The variagraph can be used to test the null hypothesis $g = g_0$ against the alternative $g \neq g_0$. The first requirement is to obtain a consistent estimate of β that does not rquire knowledge of g: for this to be done, we need to use a fitting method based on a convex objective function, as observed earlier, and we also need the predictors in x to be linearly related. When these two conditions hold, the Li-Duan lemma (Li and Duan 1989) guarantees that the resulting estimate of β will be consistent, see also Cook and Weisberg (1994).

Given a consistent estimate of β, we can compute fitted values under the null hypothesis, and create a two-dimensional problem with response y and predictor given by the fitted values. The methodology of this paper, including the use of the bootstrap to obtain a test under specified significance levels, can be applied to this two-dimensional problem.

LITERATURE CITED:

Christensen, R. 1989. Lack-of-fit tests based on near or exact replicates. *The Annals of Statistics* 17(2):673-683.

Cook, R. D., and S. Weisberg. 1994. *An introduction to regression graphics.* John Wiley & Sons.

Cox, D. R., and D. V. Hinkley. 1974. *Theoretical statistics.* Chapman and Hall.

Cressie, N. A. C. 1991. *Statistics for spatial data.* John Wiley & Sons.

Efron, B., and R. J. Tibsharani. 1993. *An introduction to the bootstrap.* Chapman and Hall.

Green, J. R. 1971. Testing departure from a regression, without using replication. *Technometrics* 13:609-615.

Härdle, W. 1991. *Smoothing techniques with implementation in S.* Springer-Verlag.

Hocking, R. R. 1996. *Methods and applications of linear models.* John Wiley & Sons.

Isaaks, E. H. and R. M. Srivastava. 1989. *An introduction to applied geostatistics.* Oxford University Press.

Joglekar, J. H., J. H. Schuenemeyer, and V. LaRiccia. 1989. Lack-of-fit testing when replicates are not available. *The American Statistician* 43(3):135-143.

Johnson, B. W., and R. E. McCulloch. 1987. Added-variable plots in linear regression. *Technometrics* 29(4):427-433.

Li, K. C., and N. Duan. 1989. Regression analysis under link violation. *Annals of Statistics* 17:1009-1052.

Neill, J. W., and D. E. Johnson. 1984. Testing for lack of fit in regression - a review. *Communications in Statistics, Part A--Theory and Methods* 13(4):485-511.

Neill, J. W., and D. E. Johnson. 1989. A comparison of some lack-of-fit tests based on near replicates. *Communications in Statistics, Part A--Theory and Methods* 18(10):3533-3570.

Plotnick, R. E., R. H. Gardner, and R. V. O'Neill. 1993. Lacunarity indices as measures of landscape texture. *Landscape Ecology* 8(3):201-211.

Scott, D. W. 1985. Average shifted histograms: Effective nonparametric density estimators in several dimensions. *Annals of Statistics* 13(3):1024-1040.

Shillington, E. R. 1979. Testing lack of fit in regression without replication. *The Canadian Journal of Statistics* 7:137-146.

Tierney, L. 1990. *LISP-STAT. An object-oriented environment for statistical computing and dynamic graphics.* John Wiley & Sons.

Velleman, P. F., and D. C. Hoaglin. 1981. *Applications, basics, and computing of exploratory data analysis.* Duxbury.

Weisberg, S. 1985. *Applied Linear Regression.* John Wiley & Sons.

APPENDIX 1. Mileage data from Velleman and Hoaglin (1981) via Neill and Johnson (1989)

Observation No.	Mileage (mpg)	Engine Size (cubic inches)
1	33.9	71.1
2	30.4	75.7
3	32.4	78.7
4	27.3	79.0
5	30.4	95.1
6	22.8	108.0
7	21.5	120.1
8	26.0	120.3
9	21.4	121.0
10	22.8	140.8
11	19.7	145.0
12	24.4	146.7
13, 14	21.0, 21.0	160.0
15, 16	19.2, 17.8	167.6
17	18.1	225.0
18	21.4	258.0
19, 20, 21	16.4, 17.3, 15.2	275.8
22	15.0	301.0
23	15.2	304.0
24	15.5	318.0
25	13.3	350.0
26	15.8	351.0
27, 28	18.7, 14.3	360.0
29	19.2	400.0
30	14.7	440.0
31	10.4	460.0
32	10.4	472.0