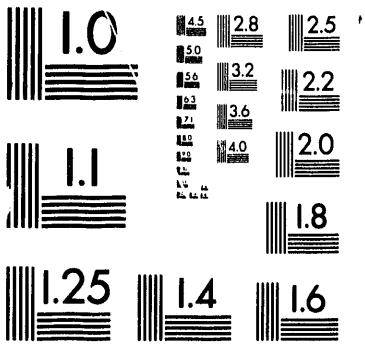
1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6

1 of 1

Conf. 940312--29

SAND93-2419C

MULTIPLE WEIGHT STEPWISE REGRESSION

Joel Atkins[1,2] and James Campbell[2]

[1] Statistics Department
University of California
Berkeley, CA 94720
[2] Manufacturing Systems Modeling
Division 6613
Sandia National Laboratories
P. O. Box 5800
Albuquerque, NM 87185

This work was supported by the United States Department of Energy under Contract DE-AC04-94AL85000.

INTRODUCTION

Background

In many science and engineering applications, there is an interest in predicting the outputs of a process for given levels of inputs. In order to develop a model, one could run the process (or a simulation of the process) at a number of points (a point would be one run at one set of possible input values) and observe the values of the outputs at those points. These observations can be used to predict the values of the outputs for other values of the inputs. Since the outputs are a function of the the inputs, we can generate a surface in the space of possible inputs and outputs. This surface is called a response surface. In some cases, collecting data needed to generate a response surface can e very expensive. Thus, in these cases, there is a powerful incentive to minimize the sample size while building better response surfaces. One such case is the semiconductor equipment manufacturing industry. Semiconductor manufacturing equipment is complex and expensive. Depending upon the type of equipment, the number of control parameters may range from 10 to 30 with perhaps 5 to 10 being important. Since a single run can cost hundreds or thousands of dollars, it is very important to have efficient methods for building response surfaces.

A current approach to this problem is to do the experiment in two stages. First, a traditional design (such as fractional factorial) is used to screen variables. After deciding which variables are significant, additional runs of the experiment are conducted. The original runs and the new runs are used to build a model with the significant variables. However, the orignal (screening) runs are not as helpful for building this model as some other points might have been. This paper presents a point selection scheme that is more efficient than traditional designs.

Approach

Latin Hypercube Sampling (Iman and Conover) is used to select design points. Stratification intervals are chosen in a way that will help us fit a better quadratic screening model. Once the selected input points have been run, a stepwise method is used to add the terms that give the most significant improvement in the model. Each time a term is added, terms that are no longer significant are dropped.

Once we have fit a quadratic model, that model is used to screen input variables and decide which are important. Input variables identified as important are then used in a similar method to fit the final response surface.

Notation

Assume that the space of possible input values is rectangular, or in other words, that the possible values of any given input variable do not depend on the values the other input variables take. This is equivalent to assuming that $\vec{x}_m \in \prod_{i=1}^{n}[\min(x_i), \max(x_i)]$.

Throughout this discussion, we will use the following notation: d for the number of input variables, n for the number of points in the design, y for the output (which we will assume is one-dimensional), $\vec{x}^m$ for the values of the input variables at the mth point in our design, and $x_{i,m}$ for the value of the ith input variable at the mth point.

Initial Screening Model

Our model will include a mean, and a linear and square term for each independent variable. It will also include interactions between pairs of independent variables, i.e. $x_j x_k$. Thus our initial model will be of the form:

$$y = \mu + \sum \beta_i x_i + \sum \alpha_i x_i^2 + \sum_{j<k} \gamma_{j,k} x_j x_k$$

CHOOSING THE DESIGN POINTS

We want our screening model to be "good" with respect to some importance function, $\lambda(\vec{x})$. (Assume that $\lambda(\vec{x}) = \prod_i^n \lambda_i(x_i)$).

For One Input Variable

Here, we choose the values of one input variable, x_i at the design points. We are interested in developing the best quadratic fit for y, i.e. $\alpha_i x_i^2 + \beta_i x_i + c_i$ which has the minimum mean squared error of all quadratic functions under weight $\lambda_i(x)$. For our purposes we can treat $\lambda_i(x_i)$ as a probability density. If we let $d_{i,j} = E(x_i^j|\lambda_i)$, then we will have:

$$\begin{bmatrix} \alpha_i \\ \beta_i \\ c_i \end{bmatrix} = \frac{\begin{bmatrix} d_{i,2} - d_{i,1}^2 & d_{i,2}d_{i,1} - d_{i,3} & d_{i,3}d_{i,1} - d_{i,2}^2 \\ d_{i,2}d_{i,1} - d_{i,3} & d_{i,4} - d_{i,2}^2 & d_{i,3}d_{i,2} - d_{i,4}d_{i,1} \\ d_{i,3}d_{i,1} - d_{i,2}^2 & d_{i,3}d_{i,2} - d_{i,4}d_{i,1} & d_{i,4}d_{i,2} - d_{i,3}^2 \end{bmatrix}}{d_{i,4}d_{i,2} + 2d_{i,3}d_{i,2}d_{i,1} - d_{i,4}d_{i,1}^2 - d_{i,2}^3 - d_{i,3}^2} \begin{bmatrix} E(x_i^2 y|\lambda_i) \\ E(x_i y|\lambda_i) \\ E(y|\lambda_i) \end{bmatrix}$$

The above equation is true because the weight measure is a product of one-dimensional weight measures and thus α_i and β_i are orthogonal to all of the other terms in the model.

Thus, α_i, β_i, and c_i will be linear functions of $E(y|\lambda_i)$, $E(x_i y|\lambda_i)$, and $E(x_i^2 y|\lambda_i)$. We can divide the range of x into n strata of the form $[l_{i,m-1}, l_{i,m}]$ and choose a point from each

strata using the distribution $\lambda_{i,m}(x) = \frac{\lambda_i(x)}{\Lambda_i(l_{i,m})-\Lambda_i(l_{i,m-1})} 1_{(l_{i,m-1} \le x \le l_{i,m})}$. After we choose the points,

$$\sum_{m=1}^{n} [\Lambda_i(l_{i,m}) - \Lambda_i(l_{i,m-1})] y_m,$$
$$\sum_{m=1}^{n} [\Lambda_i(l_{i,m}) - \Lambda_i(l_{i,m-1})] x_{i,m} y_m, \text{ and}$$
$$\sum_{m=1}^{n} [\Lambda_i(l_{i,m}) - \Lambda_i(l_{i,m-1})] x_{i,m}^2 y_m$$

will be unbiased estimates of $E(y)$, $E(x_i y)$, and $E(x_i^2 y)$. If we set $w_{i,m} = \Lambda_i(l_{i,m}) - \Lambda_i(l_{i,m-1})$, then:

$$\begin{aligned} \hat{E(y)} &= \sum_{m=1}^{n} w_{i,m} y_m, \\ \hat{E(x_i y)} &= \sum_{m=1}^{n} w_{i,m} x_{i,m} y_m, \text{ and} \\ \hat{E(x_i^2 y)} &= \sum_{m=1}^{n} w_{i,m} x_{i,m}^2 y_m. \end{aligned}$$

Thus,

$$\begin{aligned} \hat{\alpha}_i &= \sum_{m=1}^{n} w_{i,m} [a_0 y_m + a_1 x_{i,m} y_m + a_2 x_{i,m}^2 y_m], \\ \hat{\beta}_i &= \sum_{i=1}^{n} w_{i,m} [b_0 y_i + b_1 x_i y_i + b_2 x_i^2 y_i], \\ \text{and } \hat{c} &= \sum_{i=1}^{n} w_{i,m} [c_0 y_i + c_1 x_i y_i + c_2 x_i^2 y_i], \end{aligned}$$

are unbiased estimates of a, b, and c. We want to choose our intervals to minimize the expected value of the mean squared errors of our estimates, or:

$$MSE_\lambda = \int_{\hat{a},\hat{b},\hat{c} \in \Psi} \left[\int_{\min(x)}^{\max(x)} E\left((y - \hat{a}x^2 - \hat{b}x - \hat{c})^2 \right) \lambda(x) dx \right] d\psi$$

We can not minimize this function, or even evaluate it, if we do not know y. However, we can approximate this function for a given set of intervals, and then minimize our approximation. Towards this end, we will break y into $f(x) = E(y|x)$ and $g(\vec{x}, \epsilon) = y - f(x)$. Thus, $E(g(\vec{x}, \epsilon)) = 0$, and $\text{var}(g(\vec{x}, \epsilon)|x) = \sigma^2(x)$. We can now argue that within each interval $g(\vec{x}, \epsilon)$ will have more variation than $f(x)$ will have. If we assume that $\sigma^2(x)$ is fairly constant, we can minimize the part of the variance of $\hat{a}$, $\hat{b}$, and $\hat{c}$ that comes from $g(\vec{x}, \epsilon)$ by choosing shorter intervals near the ends of the range and longer intervals near the middle.

Since nothing we do in the one variable case depends on y, we will use the same methodology if there is more than one output variable of interest.

More than one Independent Variable

Once we have chosen $x_i^m : m = 1, \ldots, n$, for each i, these marginal points are combined to create the design points. In creating the design points, we want to minimize the correlations between each pairs of independent variables. One popular method for doing this is using a Cholesky decomposition variation of Latin Hypercube Sampling (Iman).

When the problem involves more than one input variable, we must consider interaction terms. We can show that $\hat{\gamma}_{j,k} = \sum_{m=1}^{n} w_{j,k,m}(d_{j,k,0} + d_{j,k,1}x_{j,m}x_{k,m})y_m$. For $w_{j,k,m}$, we will use the average of the two relevant weights, giving us $w_{j,k,m} = \frac{w_{j,m}+w_{k,m}}{2}$.

To estimate the mean, we will use the average of the weights corresponding to significant terms. Thus, if x_1, x_2, andx_3 are all significant, then we would let $\hat{\mu} = \sum_{m=1}^{n} \frac{w_{1,m}+w_{2,m}+w_{3,m}}{3} y_m$.

FITTING A MODEL AT EACH STEP IN THE STEPWISE ALGORITHM

Stepwise Technique

To build our model, we will cycle through two steps. In the first, we will consider whether adding any term will improve the fit. If any terms do improve the fit, then we add the term which gives the biggest improvement. If no term improves the fit, then we stop and keep the model from the last iteration. Each time that we add a new term, we will go through the second step, dropping any terms which are no longer significant. We will continue going through this cycle until both (a) there are no terms left which would significantly improve the mode, and (b) all of the terms already in the model are significant.

Estimating Individual Terms

As we mentioned in the previous section,

$$\begin{aligned}
\hat{\mu} &= \sum_{m=1}^{n} \frac{\sum_{i \in \text{model}} w_{i,m}}{|\text{model}|} y_m \\
\hat{\alpha}_i &= \sum_{m=1}^{n} w_{i,m}[a_0 y_m + a_1 x_{i,m} y_m + a_2 x_{i,m}^2 y_m], \\
\hat{\beta}_i &= \sum_{i=1}^{n} w_{i,m}[b_0 y_i + b_1 x_i y_i + b_2 x_i^2 y_i], \\
\hat{\gamma}_{j,k} &= \sum_{m=1}^{n} w_{j,k,m}(d_{j,k,0} + d_{j,k,1}x_{j,m}x_{k,m})y_m.
\end{aligned}$$

Unfortunately, this leaves a great deal of confounding between the different effects. In the next subsection, dealing with this confounding will be discussed.

Estimating Terms Simultaneously

Here, we will discuss getting around the confounding and the different sets of weights. To do this, we will show an example where we are interested in fitting the model $\hat{y} = \mu + \beta_1 x_1 + \alpha_2 x_2^2 + \gamma_{3,4} x_3 x_4$. To do this, we can look for values of $\hat{\mu}, \hat{\beta}_1, \hat{\alpha}_2$, and$\hat{\gamma}_{3,4}$, so that evaluating $\hat{y}$ at each of the design points, and then using these values of $\hat{y}$ to estimate the individual terms will give us the same values of $\hat{\mu}, \hat{\beta}_1, \hat{\alpha}_2$, and$\hat{\gamma}_{3,4}$ as we used to construct $\hat{y}$. We can find these four values by solving the following system of linear equations.

$$\sum_{i=1}^{n} \frac{w_{1,m} + w_{2,m} + w_{3,m} + w_{4,m}}{4} y_m$$

$$= \sum_{i=1}^{n} \frac{w_{1,m} + w_{2,m} + w_{3,m} + w_{4,m}}{4} (\hat{\mu} + \hat{\beta}_1 x_{1,i} + \hat{\alpha}_2 x_{2,i}^2 + \hat{\gamma_{3,4}} x_{3,i} x_{4,i})$$

$$\sum_{i=1}^{n} w_{1,m}(b_{1,0} + b_{1,1} x_{1,i} + b_{1,2} x_{1,i}^2) y_i$$

$$= \sum_{i=1}^{n} w_{1,m}(b_{1,0} + b_{1,1} x_{1,i} + b_{1,2} x_{1,i}^2)(\hat{\mu} + \hat{\beta}_1 x_{1,i} + \hat{\alpha}_2 x_{2,i}^2 + \hat{\gamma_{3,4}} x_{3,i} x_{4,i})$$

$$\sum_{i=1}^{n} w_{2,m}(a_{2,0} + a_{2,1} x_{2,i} + a_{2,2} x_{2,i}^2) y_i$$

$$= \sum_{i=1}^{n} w_{2,m}(a_{2,0} + a_{2,1} x_{2,i} + a_{2,2} x_{2,i}^2)(\hat{\mu} + \hat{\beta}_1 x_{1,i} + \hat{\alpha}_2 x_{2,i}^2 + \hat{\gamma_{3,4}} x_{3,i} x_{4,i})$$

$$\sum_{i=1}^{n} \frac{w_{3,m} + w_{4,m}}{2} (d_{3,4,0} + d_{3,4,1} x_{3,i} x_{4,i}) y_i$$

$$= \sum_{i=1}^{n} \frac{w_{3,m} + w_{4,m}}{2} (d_{3,4,0} + d_{3,4,1} x_{3,i} x_{4,i})(\hat{\mu} + \hat{\beta}_1 x_{1,i} + \hat{\alpha}_2 x_{2,i}^2 + \hat{\gamma_{3,4}} x_{3,i} x_{4,i})$$

SIMULATION RESULTS

Here, we ran three simulations do see how our method would compare to stepwise regression. We used both uniform strata intervals and the improved intervals which our method generated. We measured the mean squared error between our fitted models and the true model. We used a uniform measure on the space $[-1, 1]^d$. All of the simulations used experiments of thirty runs for each point in the simulation.

$x_1 + x_2 + x_3 + x_4 + x_5 + \epsilon$ with four additional nonsense input variables.

sd(ϵ)		my method improved intervals	regression improved intervals	my method uniform intervals	regression uniform intervals
.1	ave(MSE)	.0012773	.001694	.003019	.241090
	sd(MSE)	.000613	.001024	.002091	.752062
.4	ave(MSE)	.120925	.210063	.034327	.040045
	sd(MSE)	.329010	.586223	.025154	.023383
.7	ave(MSE)	.519136	.514908	.421598	.528845
	sd(MSE)	.571021	.442535	.394462	.493094

$\sqrt{x_1+1} + sin(\frac{\pi x_2}{2}) + cos(\pi x_3) + x_4 + x_1 x_2 + \epsilon$ with five additional nonsense input variables.

sd(ϵ)		my method improved inte· als	regression improved intervals	my method uniform intervals	regression uniform intervals
0	ave(MSE)	.385236	.371297	.413790	.413790
	sd(MSE)	.136567	.075868	.069418	.069418
.1	ave(MSE)	.476141	.403006	.406254	.406254
	sd(MSE)	.291063	.158053	.053048	.053048
.4	ave(MSE)	.4427	.4738	.4028	.4348
	sd(MSE)	.172750	.153947	.073289	.091597

Thus, we feel that in some cases, when the function is not well approximated by the model or when there is a lot of noise, it will be better to use our method for choosing points, and to then use our analysis as well as regression and use the better of the two models.

REFERENCES

Iman, R.L., and Conover, W.J. (1982), A Distribution Free Approach to Inducing Rank Correlation among Input Variables, *Commun. Statist. - Simula. Computat.* 11, 3, pp. 311-334.

Li, K. C. (1992), On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Applicaton of Stein's Lemma, *J. Amer. Statist. Assc.* 87 1025-1039.

DATE
FILMED
2/7/94
END