Conference on Applied Statistics in Agriculture 1992 - 4th Annual Conference Proceedings

BEYOND LINEARITY AND INDEPENDENCE

J. Stuart Hunter

BEYOND LINEARITY AND INDEPENDENCE

J. Stuart Hunter
503 Lake Drive
Princeton, NJ 08540

This brief lecture discusses statistical problems associated with postulating and fitting models in engineering and the sciences. Particular emphasis is placed on the two-model problem: the employment of both deterministic and stochastic components within a model. Further, the use of empirical versus theoretical models on the part of both statisticians and experimenters is examined.

Key words: linear models, non-linear models, non-independence, empiricism.

1: Introduction, The "Two-Model" Problem

Consider a single recorded observation y measured on a continuous scale. The observation y is commonly viewed as two separate functions added together, $y = \eta + \epsilon$, one deterministic identified as η and the second stochastic identified as ϵ. Parametric models are now postulated for both η and ϵ creating the statistician's "two-model" problem. In its simplest form η is taken to be a constant and ϵ a random independent event with zero mean, $E(\epsilon) = 0$, and constant variance, $V(\epsilon) = \sigma^2$.

More sophisticated models are then postulated for both the deterministic and stochastic components. For example write $y = \eta + \beta + \epsilon$ where η is a constant and β is an additional independent stochastic event with $E(\beta) = 0$ and variance σ^2_β. Extensions of this "random effects" model lead to the statistician's components of variance analyses. Or equally simple, let $\eta = \theta_o + \theta_1\xi_1$ be a known theoretical function relating a forcing factor ξ_1 to the response η, then $y = \theta_o + \theta_1\xi_1 + \epsilon$. Extensions of this "straight line" model lead to regression analyses. In general, deterministic functions $\eta = f(\underline{\xi}, \underline{\theta})$ of some complexity may be selected containing many factors $\underline{\xi}$ and parameters $\underline{\theta}$, and the stochastic event ϵ may be taken to be an occurence arising from a distribution function $h(\underline{\Sigma})$ composed of many parameters $\underline{\Sigma}$.

2: The Linear Model

When the theoretical response η is reasonably 'smooth' over the ranges of the $\underline{\xi}$ employed by an experimenter the deterministic function $f(\underline{\xi}, \underline{\theta})$ is usually assumed to be well *approximated*

locally by:

$$\eta = f(\underline{x},\underline{\theta}) \approx \theta_o + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_p x_p = {}_i\Sigma_p \theta_i x_i ,$$

the consequence of a Taylor's series expansion of $f(\underline{\xi},\underline{\theta})$ about some set of values $\underline{\theta}_o$. This surrogate function is linear in the parameters $\underline{\theta}$. Further, the "regressor" variables $\underline{\mathbf{x}}$ are functions of the forcing factors $\underline{\xi}$ *only*. Usually the x_i are simple surrogates for different versions of $\underline{\xi}$, for example, for the levels of controlled factors such as rpm or concentration, for the recorded levels of uncontrolled factors such as ambient temperature or soil moisture, or used as dummy indicators to identify qualitative versions of factors such as types of machines or plant varieties.

Let $\underline{\mathbf{X}}$ be the nxp matrix whose rows identify the n individual settings of the p regressor variables and at each setting let a single observation y be recorded. We now have:

$$\underline{\mathbf{Y}} = \underline{\mathbf{X}}\underline{\theta} + \underline{\epsilon}$$

where $\underline{\mathbf{Y}}$ is an n element column vector of observations, $\underline{\theta}$ a p element column vector of unknown coefficients and $\underline{\epsilon}$ an n element vector of independent events drawn from $h(\underline{\Sigma})$. If we chose $h(\underline{\Sigma})$ to be a multivariate Normal distribution with $\underline{\Sigma} = \underline{\mathbf{I}}_n\sigma^2$ then the ordinary least squares estimates of $\underline{\theta}$ are given by $\hat{\underline{\theta}} = [\underline{\mathbf{X}}'\underline{\mathbf{X}}]^{-1}\underline{\mathbf{X}}'\underline{\mathbf{Y}}$ Any deleterious effect of colinearity amongst the column vectors of $\underline{\mathbf{X}}$ should be reduced by careful experimental design. The fitted model becomes $\hat{\underline{\mathbf{Y}}} = \underline{\mathbf{X}}\hat{\underline{\theta}}$, the variance of the estimated coefficients $V(\hat{\underline{\theta}}) = [\underline{\mathbf{X}}'\underline{\mathbf{X}}]^{-1}\sigma^2$ and the estimate of the stochastic parameter σ^2 given by $s^2 = [\underline{\mathbf{Y}}'\underline{\mathbf{Y}} - \hat{\underline{\theta}}'\underline{\mathbf{X}}'\underline{\mathbf{Y}}]/(n-p)$. When experiments are repeated, the replicate values of the observations may also be used to obtain a separate estimate of σ^2, and this estimate used in a 'lack-of-fit' test to check the adequancy of the postulated linear model.

The adequacy of the fitted empirical model may also be checked using other lack of fit procedures, many graphical. The influence of particular observations upon the estimated coefficients or predictions may be determined. The appropriateness of the stochastic model assumed for the $\underline{\epsilon}$ is more difficult to appraise, but plots of the residuals $\underline{\mathbf{Y}} - \hat{\underline{\mathbf{Y}}}$ on probability paper, against predicted values, in time sequence, and against the regressor variables are usually informative. Once tests of adequacy of the empirical model to represent the actual functional relationship $f(\underline{\xi},\underline{\theta})$ are passed, a vast panoply of standard hypothesis testing, interval estimation and graphical exposition procedures follow. (Draper & Smith).

Should any of these tests for adequacy fail, the analyst proceeds to change the deterministic and/or stochastic models postulated. Holding to the constraints of independence and a linear model, regressor variables may be omitted, transformed, or

new ones added. Keeping in mind that the average variance of a forecast over the region defined by the factors is given by $V(\hat{y}) = p\sigma^2/n$ where p is the number of parameters employed (Box, Hunter & Hunter, pg 524), parsimony should reign. Transformation of the observed response is often employed. Of course, transforming the y's alters the assumptions about the errors $\underline{\epsilon}$ but usually, and fortunately, when a simplifying response transformation is employed both normality and the assumptions about $\underline{\Sigma}$ become more reasonable.

Experience testifies that many multifactor theoretical functions $f(\underline{\xi},\underline{\theta})$ have been successfully approximated by empirical linear models over the chosen space of the factors $\underline{\xi}$. For these situations standard experimental designs are available for everyday use: the factorials, the 2^{k-p} small fractional factorials coupled to first order models, and response surface designs employing second order models, all available in blocks of varying size.

3: A Linear Model Example

An an example consider the following 3^2 factorial design used to explore the simultaneous role of two factors, ξ_1 (air/fuel ratio) and ξ_2 (ethanol concentration) upon the response η (CO concentration) in the exhaust of a standard automobile engine, (Hunter). The ranges of ξ_1 and ξ_2 were chosen after considerable consultation with the engineers. The resulting design, the proposed second order empirical model and associated data are given in Figure (1).

3^2 FACTORIAL DESIGN

Investigate η = CO concentration as a function of ξ_1 = ethanol concentration and ξ_2 = air/fuel ratio.

x_1	x_2	Obs.			Averages		
-	-	66	62		-	0	+
0	-	78	81	x_1	66.83	75.83	75.83
+	-	90	94	x_2	78.50	75.50	64.50
-	0	72	67				
0	0	80	81				
+	0	75	78				
-	+	68	66		$s^2 = 5.17, \quad \nu = 9$		
0	+	66	69				
+	+	60	58				

Proposed second order model:

$$\eta = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \text{ , and}$$

$$y = \eta + \epsilon \text{ , with } \epsilon \rightarrow \text{Normal iid}(0,\sigma^2) \text{ .}$$

empirical description of events. Of course, additional responses, NO_x or other air toxics, could also be fitted and their contour systems superimposed on those of CO to give multivariate impressions of the joint effects ξ_1 and ξ_2 . A great deal of information has resulted from the application of this simple experimental design and linear model.

4: Coupling

The key element in a multi-factor empirical model is $\beta_{ij}x_ix_j$, the cross product, coupled, or two factor interaction term (to use its several names). If in this example the coefficient β_{12} is zero, the two forcing factors could be separetely investigated and their individually fitted models merely later added together. But the knowledgable engineer knows in advance that ethanol and air-fuel ratio will very likely have a *coupled* influence upon the CO response even though the true functional relationship is unknown. What is known is that the function is not likely to be the simple addition of two separate linear functions. The commonplace graphical display of a two-factor interaction as two non-parallel straight lines superimposed should always be accompanied by descriptions of the concept of coupled effects.

If engineers and scientists are to value the use of the standard statistical models and experimental designs it will be because they are recognized as the tools of an *enlightened empiricism*.

5: Empiricism carried too far

Experiment design is often taught as though the experimenter's mind were a *tabula rasa*. Designs are chosen with almost no concern over the true functional model. Fortunately many standard experimental designs provide an associated linear model so over-parameterized that a parsimonious linear model can almost always be found amply to exposit the experimenter's response function. And if perchance an important factor be omitted and left uncontrolled, its biasing influence upon the estimated model coefficients will of course be reduced through randomization.

However, in selling experimental design strategies to engineers and scientists heavy emphasis is often placed on the ability to investigate many factors simultaneously in very few experimental trials. The approach leads to the common practice of employing many factors ξ in an experimental design in the hope of finding the 'vital few'. This practice is both insidous and dangerous when the design employed is a low resolution fractional factorial. To illustrate, consider the example displayed in Figure 3 which employs a 3^{4-2} fractional factorial design, the L9 hyper-Graeco-Latin square used as a fractional factorial and popularized by the Taguchi school.

The 3^{4-2} Fractional Factorial Design
The 4x4 Hyper-Graeco-Latin Square
The L9 Design

x_1	x_2	x_3	x_4	Obs.		Averages			
							-	0	+
-	-	-	-	66	62				
0	-	0	0	78	81	x_1	66.83	75.83	75.83
+	-	+	+	90	94	x_2	78.50	75.50	64.50
-	0	+	0	72	67	x_3	67.83	74.33	76.33
0	0	-	+	80	81	x_4	69.33	69.33	79.83
+	0	0	-	75	78				
-	+	0	+	68	66	$s^2 = 5.17, \quad \nu = 9$			
0	+	+	-	66	69				
+	+	-	0	60	58				

Alternative view of the 4x4 Hyper-Graeco-Latin Square		1	2	3	4
	I	$A\alpha$	$B\beta$	$C\gamma$	$D\delta$
	II	$B\delta$	$A\gamma$	$D\beta$	$C\alpha$
	III	$C\beta$	$D\alpha$	$A\delta$	$B\gamma$
	IV	$D\gamma$	$C\delta$	$B\alpha$	$A\beta$

Figure 3

Plots of the average response at each of the three levels of the four factors is displayed in Figure 4 and the 95% confidence interval is indicated about each plotted average.

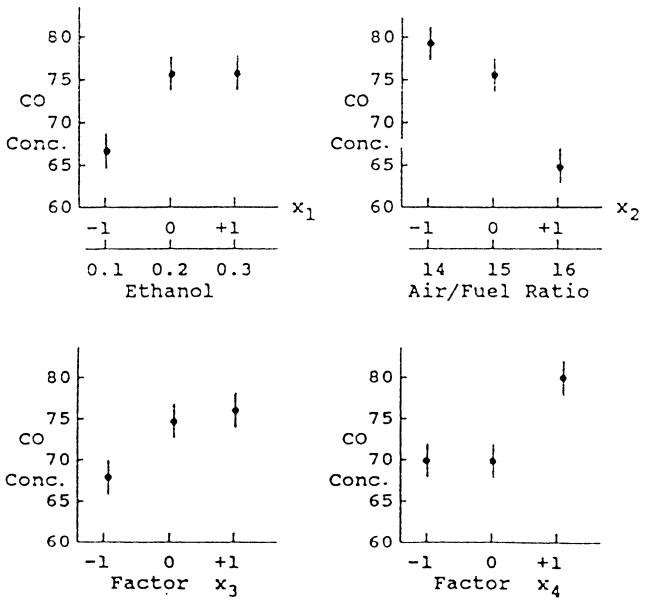

Figure 4: Plots of average response

It is clear that statistically significant effects are present for each factor, and that curvilinearity exists throughout. The trouble with this example is that factors x_3 and x_4 are dummy factors added to the 3^2 design described earlier. The observed influences of x_3 and x_4 are *mirages*.

Crucial here is the role of the crossproduct (interaction) term x_1x_2 . The linear model associated with the analysis of the 3^{4-2} does not contain this important contributing factor and thus the x_3 and x_4 coefficients, instead of approximating zero as they should, are biassed. They are mirages provided by the unestimated contribution of the x_1x_2 term. The 3^{4-2} is a resolution III design and first order estimates are alaised, biassed, (corrupted!) by coupled influences (the two-factor interactions). When a linear model is used as a surrogate for a model likely to be non-linear, special care must be taken to insure that the design-model combination has the ability to estimate at least the coupled influences of the factors. Simplistic empiricism is easily oversold.

6: A Non-linear Model

A recent article in the American Statistician (Kopas & McAllister) describes a series of hands-on exercises for reinforcing concepts taught in introductory statistics and design of experiments courses. One exercise requires the dropping of a pellet into a glass cylinder containing a viscous fluid and measuring the time it takes the pellet to fall to the bottom. The students are asked to plan a sequence of experiments to study the effects of four or five factors in order to design a fluid to meet a specified target drop time with minimun variance. They are told that other customers were interested in their process and that "it is vital to be able to design new fluids", and needed was a "an understanding of the cause and effect mechanisms operative in your process." Team dynamics, brainstorming, and statistical tools such as fishbone diagrams, Pareto charts, control charts, fractional factorial, factorial, mixture designs, components of variance and response surface methods are all encouraged along with heavy emphasis on graphical exposition.

Now it is true that useful approximations of response functions are possible using linear models, and recent work in the applications of splines and non-parametric estimation methods have only added further to the value of empirical approaches. But empiricism, no matter how enlightened, can not replace good theory. It is interesting to contemplate how an engineering student might contemplate this pellet drop problem.

Most engineering students know that force equals mass times acceleration, $F = ma$, that acceleration $a = dv/dt$ is the time rate of change of velocity v , and that velocity $v = d\ell/dt$ where ℓ is the distance travelled in time t. Manipulating these simple dynamic expressions, it is easy to show that the distance

traveled by a freely falling body equals $\ell = v_o t + at^2/2$ where v_o is the initial velocity at time t = 0. But suppose the falling body meets resistance proportional to the square of its velocity. A point in the fall will be reached wherein the acceleration will equal zero and the limiting velocity equal V. This produces the equation

$$\frac{w}{g}\frac{d^2\ell}{dt^2} = w - \lambda\left(\frac{d\ell}{dt}\right)^2$$

where w is the weight of the pellet, g the gravitational constant (w = gm), and λ a constant characterizing the density of the fluid, (Reddick & Miller). Solving this second order differential equation gives:

$$\ell = \frac{1}{g}\left(\frac{w}{\lambda}\right)\log\left[\cosh gt\left(\frac{w}{\lambda}\right)^{-0.5}\right]$$

The engineering student might then conclude that given a set of values of t and ℓ , with w and g known, one might now obtain an estimate of λ. Statisticians recognize this as an example of non-linear estimation.

Models *non-linear in their parameters* are ubiquitous in all the sciences. One popular class is the ratio of polynomials, that is models of the form

$$\eta = \frac{\beta_o + \beta_1 x_1 + \beta_2 x_2}{\alpha_o + \alpha_1 x_2}$$

or equations with mixtures of polynomial and exponential terms

$$\eta = \beta_1 x_1 + \beta_2 x_2 e^{-(\alpha_1 x_1 + \alpha_2 x_2)}$$

or in the natural sciences the popular logistics function

$$\eta = \frac{1}{1 + e^{(\alpha + \beta x)}}$$

The estimation of the parameters in non-linear models

can be quite difficult, most particularly if the data are gathered haphazardly. At the beginning of the estimation procedure, and for construction of a non-linear experimental design, initial guessed values of the unknown parameters $\underline{\theta}$ are required to determine the derivatives $\partial\eta/\partial\theta|_{\theta_0}$ the "sensitivity" elements entering the matrix of derivatives $\underline{\mathbf{X}}$. In the absence of experimental design considerations the matrix $\underline{\mathbf{X}}$ is often poorly conditioned. The Gaussian iterant may be used to find the least squares estimates, speeded perhaps through the use of the Levenberg or Marquardt algorithms (Levenberg, K.), (Marquardt, D. W.). It is not uncommon to obtain different estimates resulting from different starting points, a reflection in part to the influences of numerical rounding errors. Non-linear estimation and non-linear experimental design is a task for the modern high-speed computer. Nor are non-linear models necessarily always best. In the 3^2 factorial example discussed earlier there was good reason to believe that the non-linear model

$$\eta = \alpha_1 e^{\beta_1 x_1} + \alpha_2 e^{\beta_2 x_2}$$

would be far superior to the second order polynomial. The fitted proved a great disappointment.

$$y = 28.47e^{0.15} + 43.81e^{-0.15}$$

7: Employing Prior Knowledge

Of course the purpose of the pellet dropping exercise was to teach students something about the beginning arts of statistics, about *enlightened* empiricism. It was not meant to be a serious effort to determine the characteristics of a viscous fluid. However, in teaching engineers and scientists the statistician must be prepared to draw down on all the information available. A great deal of prior knowledge may exist both with respect to the form of the model and to the magnitudes of important parameters. Knowledge concerning λ is useful in reducing the iterations required in non-linear estimation procedures. Prior knowledge can often be formally employed via Bayesian approaches to lessen both the experimental effort and to increase the precision of the estimate, (Racine, Grieve & Fluhler). Employing the prior knowledge of the subject matter expert can only enhance the role of the statistician.

As further statistical concerns in this pellet-drop example, since the quantities $\ell = \mathcal{L} + \epsilon_{\mathcal{L}}$ and $t = \tau + \epsilon_{\tau}$ are both measured with error shouldn't an *interval* estimate of the parameter λ be obtained? Would anyone want to hazard a guess as to the number of replicate trails needed to get a standard error for λ equal to , say, 0.1λ? Could fewer trials be managed if different levels of t, or ℓ, or combinations of t and ψ were chosen? Might other

factors influence λ: the ambient temperature, the diameter of the cylinder and size of pellet?. The answers to these obviously non-trivial questions are within the realm of the modern statistician. The ability to answer such questions resourcefully will condition the statistician's unfolding future role in the sciences.

8: Dynamic Models and Box-Jenkins

Of course, one doesn't have to go far in the sciences before dynamical considerations become preeminent. Many laws of physics are initially quite simple in their structure, as for example Ohm's Law

$$E = IR$$

where E measures voltage, I measures the current flow in amperes and R measures the resistance to flow in ohms. However, anyamperesinductance in a circuit L resists a *change* in current and thus also influences voltage. The voltage drop induced by the inductance is thus $E_L = L(dI/dt)$. Ohm's Law becomes:

$$E = IR + L\frac{dI}{dt} .$$

solving gives:

$$E = IR\left[\frac{1}{1 - e^{-Rt/L}}\right],$$

an equation clearly no longer linear in its parameters.

But amperes I measures the rate of change of the number of electrons $Q = 6.25x10^{18}$ electrons, that is, $I = dQ/dt$ and $Q = CE$ where C is the capacitance of the circuit. Putting this altogether gives the following ordinary second order differential equation:

$$E = \frac{Q}{C} + R\frac{dQ}{dt} + L\frac{d^2Q}{dt^2} .$$

The solution of this second order equation will take either the form of a sum of two exponentials or a function containing sinusoidal terms, both models non-linear in their parameters and, as indicated earlier, not easily fitted. Similar dynamical equations can be found in the natural sciences showing the growth of populations under varying stresses.

But just as the Taylor's series approximation provides a useful empirical linear model approach to the fitting of non-linear functions, a similar array of easily applied models exists for the dynamic case. The Box-Jenkins ARIMA modeling methods for fitting dynamical models directly to time series data sets has

proved of great practical value, (Box, Jenkins).

Suppose that η is a continuous function and that its rate of change is proportional to the amount of η remaining as it approaches some final asymptotic value η_∞ . The model is then

$$\frac{d\eta}{dt} \propto (\eta_\infty - \eta_t)$$

$$T\frac{d\eta}{dt} + \eta_t = gx$$

where T is the time constant, and η_∞ = gx where here g is the "gain" (units adjuster) and x the excitation. The particular solution to this dynamical equation is:
at equally spaced time intervals Δt. Then the discrete first

$$\eta = \eta_\infty(1 - e^{-t/T}) \quad .$$

Now let η_t, (t = 1, 2, ..., n), denote descrete events occuring at equally spaced time intervals Δt. Then the discrete first order *difference* equation equivalent of the continuous first order *differential* equation given above is:

$$\eta_t - \phi\eta_{t-1} = g(1 - \phi)x_{t-1}$$

where

$$\phi = e^{-\frac{\Delta t}{T}}$$

Using the "backward" operator where

$$B\eta_t = \eta_{t-1}, \; B^2 = \eta_{t-2} \text{ and } (1-B)\eta_t = \eta_t - \eta_{t-1} = \Delta\eta_t$$

the difference equation may be written as:

$$(1 - \phi B)\eta_t = g(1 - \phi)x_{t-1} = a_t$$

where a_t has now replaced $g(1 - \rho)x_{t-1}$. Note well that with η_t thus defined the statistician's two-model problem can now be re-written with a time subscript t assigned to each component:

$$y_t = \eta_t + b_t$$

$$y_t = \frac{1}{(1 - \theta B)}a_t + b_t$$

where b_t is the error or 'noise' due to measurement and/or observation. In the Box-Jenkins series of models the a_t are

taken to be independent, normally distributed model *excitations* with mean zero and fixed variance σ^2_a , and the b_t as *observational errors* similarly independent, normal, mean zero with fixed variance σ^2_b with the a_t and b_t mutually independent.

The difference equation $(1 - \phi B)\eta_t = a_t$ is classified as an AR(1) model, an auto-regressive model of order 1. The AR(2) model

$$(1 - \phi_1 B - \phi_2 B^2)\eta_t = a_t$$

is the discrete equivalent of a second order differential equation where, once again, the excitation is a stochastic shock a_t . Higher order models, AR(p), are also possible. Factoring the quadratic $(1 - \phi_1 B - \phi_2 B^2)$ in B gives $(1 - \phi_1')(1 - \phi_2')\eta_t = a_t$. Suppose $\phi_1' \rightarrow$ zero. The model now becomes a first order AR(1) working upon the $(1 - B)\eta_t$, the lst differences of the η_t. Such models are non-stationary, that is, they produce data traces with continually increasing variance, or viewed another way, data traces without a mean, there is *no* expected value. And if both the ϕ's are close to zero the model considers 2nd differences in the η_t. Models considering differences of order d are possible.

Henceforth, to comply with Box-Jenkins notation let $z_t = (\eta_t - \tau)$, the deviation from some desired constant target value τ, thus forcing z_t to be hopefully zero.

When the successive shocks to a system are independent then $z_t = a_t$. But the stochastic shocks to a system can also be structured as for example:

$$z_t = a_t - \theta a_{t-1} = (1 - \theta B)a_t ,$$

producing the MA(1) model, the moving average model of order one. Higher order moving average models MA(q) are possible. Thus the MA(2) model is

$$z_t = (1 - \theta_1 B - \theta_2 B^2)a_t .$$

Mixed ARMA models are possible, as for example the AR(2) combined with the MA(2) model to give:

$$(1 - \phi_1 B - \phi_2 B^2)z_t = (1 - \theta_1 B - \theta_2 B^2)a_t.$$

and mixed ARMA models involving differences are also possible, the Box-Jenkins ARIMA models of order p,d,q. One favorite is the 0,1,1 model:

$$(1 - B)z_t = (1 - \theta B)a_t$$

which identifies the z_t as an exponentially weighted moving average.

The identification of ARIMA models begins with an

investigation of their associated time series autocorrelation functions: plots of the lagged autocorrelation coefficients $\rho_k = E(z_t z_{t-k})/\sigma^2$ plotted against the time lag k. For AR(p) models, the theoretical autocorrelation structure takes on an appearance analogous to the transient time trace of the corresponding order p continuous ordinary differential equation. Thus given a time series y_t, the plot of the *estimated* autocorrelation coefficients r_k can be employed to identify the appropriate AR(p) model.

Choosing an appropriate ARIMA model can be quite difficult, and alternative models easily postulated based on the information provided by the estimates r_k. Observational errors b_t serve to make variance of the observations y_t larger and thus serve to obscure the pattern presented by the sample autocorrelation function. Also, when the sampling interval Δt is large relative to the dynamics of the system under study, identification of the autocorrelation structure can become almost impossible.

If a simple autoregressive model has been identified, its coefficients can be obtained through ordinary least squares. In general however time series models usually involve both AR and MA components and non-linear estimation procedures must be evoked to obtain estimates of the model parameters. Many software programs exist to aid in both the identification and estimation of ARIMA models, (Pankratz, A).

An interesting variation on the two-model problem occurs when η_t is *not* dynamic. For example, consider the case where $\eta = \beta_o + \beta_1 x_1$ but where the stochastic elements ϵ_t are no longer independent but structured. Let the model for ϵ_t be the ARIMA (1,0,1) model, that is, $(1 - \phi B)\epsilon_t = (1 - \theta B)a_t$. The two-model "problem" for the observations now yields the expression:

$$\eta_t = \beta_o + \beta_1 x_1 + \left(\frac{1 - \theta B}{1 - \phi B}\right) a_t$$

The need for such a model occurs when, after having fitted the model $\hat{y} = b_o + b_1 x_1$ using ordinary least squares the residuals $y_t - \hat{y}_t$ are found to be time auto-correlated, (Durban & Watson). An ARIMA model should then be fit to the residuals, and using the estimates for ϕ and θ as starting points the entire model refitted by non-linear least squares.

9: A Two-model Example

An early example of the successful use of a combined deterministic-stochastic approach to modelling concerned the forecasting of temperature and water flow in the Ohio River, (McMichael & Hunter). Six years of daily data where employed. Initially, to reflect the influence of the seasons, a simple deterministic cyclic model $y_t = \beta_o + \beta_1 \cos(\omega t - \alpha) + \epsilon_t$ was fitted by ordinary least squares. The residuals $y_t - \hat{y}_t$ were found to be highly autocorrelated. Attempts to fit higher order Fourier

series proved futile, many too many coefficients required to get an acceptably adequate fit. A fully stochastic modelling approach was then attempted that recognized both the daily and annual variability. The result was the well fitting five parameter autoregressive-moving average model:

$$(1 - \phi_1 B)(1 - \phi_{365} B^{356})(y_t - y) = (1 - \theta_1 B)(1 - \theta_{365} B^{365}) a_t \quad .$$

However, this fully stochastic model failed as a forecasting instrument since one day's thunderstorm would lead to forecasts of unusual river temperature and flow on the same day in successive years. Finally, and in retrospect one might say obviously, a combined deterministic-stochastic model was postulated:

$$\eta_t = \alpha_0 + \alpha_1 \cos(\omega t + \alpha_2) + \left(\frac{1 - \theta B}{1 - \phi B}\right) a_t$$

to reflect the cyclic contributions of the seasons and the daily stochastic dependence of weather. The residuals $y_t - \hat{y}_t$ remaining after fitting this non-linear five parameter model passed all tests and the model proved a useful forecast instrument.

10: Conclusion

If statisticians are to be succesful in attracting the attention of scientists and engineers they must be prepared to go beyond the commonplace linear model with independent errors, to go beyond enlightened empiricism. Most models in the sciences are non-linear in their parameters, often dynamic and frequently subject to noise regimes that can not be assumed to be independent. Fortunately, most computational obstacles have been essentially resolved. Remaining are the important challenges of creating experimental designs for different classes of non-linear models and under various non-independent error structures. Needed are methods for checking and comparing alternatives for the deteriministic and stochastic components of the two-model problem. Clearly, with so much yet to be accomplished the interface between statistics and the sciences promises to remain vibrant and rewarding to all.

Bibliography

Box, G. E. P. and Jenkins, G. M. (1976) Time Series Analysis: Forecasting and COntrol, Holden Day, San Francisco.

Box, G. E. P., Hunter, W. G. & Hunter, J. S. (1978) Statistics for Experimenters, John Wiley, New York.

Draper, N. & Smith H. (1981) Applied Regression Analysis, 2nd Edition, John Wiley, NY.

Durbin J. and Watson G. S. (1971) "Testing for serial correlation in least squares regression, III", Biometrika 58, 1-19.

Hunter, J. S. (1989) "Let's all beware the Latin Square," Quality Engineering, I, pp453-465.

Kopas, D. A. and McAllister, P. R. (1992) "Process improvement exercises for the chemical industry," The American Statistician, 46 No 1., 34-41.

Levenberg, K. (1944) "A method for the solution of certain non-linear problems in least squares," Quart. pf Appl. Math. 2, 164-168.

Marquardt, D. W. (1963) "An algorithm for least squares estimation of non-linear parameters," J. Soc. Inc. Appl. Math., 11, 431-441.

McMichael, F. C. and Hunter, J. S. (1972) "Stochastic modeling of temperature and flow in rivers," Water Resources Research, 8, No. 1, pp 87-98.

Pankratz, A. (1991) Forecasting with Dynamic Regression Models, John Wiley, NY. pg xiii.

Racine, A., Grieve, A. P. & Fluhler, H. (1986) "Bayesian methods in practice: experiences in the pharmaceutical industry," Applied Statistics, 35, 99 93-150

Reddick, H. W. and Miller, F. H. (1938) Advanced Mathematics for Engineers, John Wiley & Sons, NY.